Proceedings of the

First Workshop on

Psycho-computational Models of Human Language Acquisition

Held in cooperation with COLING-2004

28-29 August 2004
Geneva, Switzerland

INVITED SPEAKERS:

Walter Daelemans	(University of Antwerp, Belgium and Tilburg University, the Netherlands)
B. Elan Dresher	(University of Toronto, Canada)
Charles Yang	(Yale University, USA)

ORGANIZER:

William Gregory Sakas	(City University of New York, USA)

PROGRAM COMMITTEE:

Robert Berwick	(MIT, USA)
Antal van den Bosch	(Tilburg University, the Netherlands)
Ted Briscoe	(University of Cambridge, UK)
Damir Cavar	(Indiana University, USA)
Morten H. Christiansen	(Cornell University, USA)
Stephen Clark	(University of Edinburgh, UK)
James Cussens	(University of York, UK)
Walter Daelemans	(University of Antwerp, Belgium and Tilburg University, the Netherlands)
Jeffrey Elman	(University of California, San Diego, USA)
Gerard Kempen	(Leiden University, the Netherlands and the Max Planck Institute, Nijmegen)
Vincenzo Lombardo	(University of Torino, Italy)
Larry Moss	(University of Indiana, USA)
Miles Osborne	(University of Edinburgh, UK)
Dan Roth	(University of Illinois at Urbana-Champaign, USA)
Ivan Sag	(Stanford University, USA)
Jeffrey Siskind	(Purdue University, USA)
Mark Steedman	(University of Edinburgh, UK)
Menno van Zaanen	(Tilburg University, the Netherlands)
Charles Yang	(Yale University, USA)

WORKSHOP ASSISTANTS:

Xuan Nga Cao (City University of New York, USA)
Mari Fujimoto (City University of New York, USA)
Lydiya Tornyova (City University of New York, USA)

SPONSOR:

The 20th International Conference on Computational Linguistics

FURTHER INFORMATION:

William Gregory Sakas
Ph.D. Programs in Linguistics and Computer Science
Department of Computer Science, North Bldg1008
Hunter College, City University of New York
695 Park Ave
New York, NY 10021
USA

email: sakas@hunter.cuny.edu
psycho.comp@hunter.cuny.edu

WWW: http://www.colag.cs.hunter.cuny.edu/psychocomp

Introduction

Every day, we use language so effortlessly that we often overlook its complexity. The fact that language *is* complex is indisputable. Indeed, even after decades of scrutiny, highly-trained adult scientists cannot agree on a definitive analysis of the underlying mechanism that ultimately determines how our sounds, words, and sentences go together – but such an effortless task for a child! Children as young as one-and-a-half-years-old (and younger) continually exploit much of language's underpinnings while going about the business of making sense of the linguistic environment that surrounds them. By the time a child reaches kindergarten, he or she has almost full mastery of an elaborate structure that eludes adequate scientific description. How children accomplish this – how they come to acquire 'knowledge' of language's essential organization – is one of the most fundamental, beguiling, and surprisingly open questions of modern science.

This workshop brings together researchers whose (at least one) line of investigation is to computationally model the acquisition process and ascertain substantive interrelationships between a model and linguistic and psycholinguistic theory. Progress in this agenda not only directly informs developmental psycholinguistic and linguistic research, but in my opinion, will also have the long term benefit of informing applied computational linguistics in areas that involve the automated acquisition of knowledge from a human or human-computer linguistic environment.

The level of sophistication and breadth of applied computational linguistics techniques has skyrocketed in the past two decades. There is now a battery of computational formalisms and statistical methods to 'choose from,' all which have yielded remarkable success in many applied domains that involve the computer learning of natural language (e.g. speech recognition, web technologies, corpus analysis, etc). These achievements have dramatically spurred even more research and funding to the point where the evolution of the science of computational linguistics can be seen as quickly outpacing that of psycholinguistics.

However, there are signs that the computational linguistics community has been progressively more aware that language technologies might benefit by incorporating learning strategies employed by humans. Although research involving the psycho-computational modeling of human language acquisition has been long active in the areas of psycholinguistics, cognitive science and formal learning theory, it has, arguably, only recently become a growing part of the computational linguistics agenda. This is evidenced by the occasional special session at an ACL meeting (e.g., ACL-1999 – Thematic Session on Computational Psycholinguistics), current workshops at both COLING-2004 (this workshop) and ACL-2004 (Incremental Parsing: Bringing Engineering and Cognition Together), and regular invitations to developmental psycholinguists to deliver plenary addresses at recent ACL meetings. This cross-discipline attentiveness is clearly very healthy and might well help reduce the possibility that applied research will run into a *psycho-computational bottleneck* – when state-of-the-art computational methods cannot be improved further in the development of user-transparent computer-human language applications – by incorporating theoretical advances in computational psycholinguistics into computational language learning technologies.

This workshop brings together a wide range of computational psycholinguistics research that is involved with the study of language acquisition: 34% of author contributions come from researchers holding positions in computer science or related departments, 33% from linguistics departments, 30% from psychology or cognitive science departments, and 3% from other departments.[1] The articles present investigations involving a broad diversity of formalisms, learning strategies, modeling techniques and linguistic phenomena. Linguistic footings range from (variations on): Universal Grammar, constructionist frameworks, and categorial grammar, to novel formulations of structural representation, to 'none.' Learning strategies include: distributional and corpus techniques, connectionist implementations, cue-based learning, and hybrid models that apply several strategies. Phenomena that are modeled include: the acquisition of semantics, linguistic (principles and) parameter setting, lexical subcategorization, child language production, atypical acquisition, phonological acquisition and morphological acquisition. Several papers involve cross-linguistic research and/or use actual child-directed speech (from corpora).

[1] An "author contribution" is calculated as 1 / the number of authors on a paper.

Notably, most papers (not all) address acquisition at the sub-word, word, or multi-word level. Few models assign structure or meaning to an entire utterance (or discourse) although many papers suggest that a presented model could be (easily) scaled-up – a worthwhile direction for future research. It is also worth remarking on the fact that articles addressing formal learning issues (e.g., PAC learning, identification in the limit, grammar induction, etc.) or that incorporate formalisms from mainstream computational linguistics (e.g., any of the many variants of probabilistic grammars) are underrepresented (the workshop contains one such). Future meetings along the lines of this workshop might benefit from attracting research efforts related to these approaches.

I would sincerely like to thank the program committee for above-and-beyond effort given the tight timetable, the diversity of the papers, and the several frustrating problems caused by spam-blockers; the workshop assistants who were a tremendous help with collating the reviews, organizing the articles for the proceedings, dealing with email and designing the conference web site; and, finally, the members of the COLING-2004 Workshop Program Committee, who were extremely helpful (and patient) on more than one occasion.

William Gregory Sakas
New York City
June 2004

Table of Contents

A quantitative evaluation of naturalistic models of language acquisition; the efficiency of the Triggering Learning Algorithm compared to a Categorial Grammar Learner

Paula Buttery
Natural Language and Information Processing Group,
Computer Laboratory, Cambridge University,
15 JJ Thomson Avenue, Cambridge, CB3 0FD, UK
paula.buttery@cl.cam.ac.uk

Abstract

Naturalistic theories of language acquisition assume learners to be endowed with some innate language knowledge. The purpose of this innate knowledge is to facilitate language acquisition by constraining a learner's hypothesis space. This paper discusses a naturalistic learning system (a Categorial Grammar Learner (CGL)) that differs from previous learners (such as the Triggering Learning Algorithm (TLA) (Gibson and Wexler, 1994)) by employing a dynamic definition of the hypothesis-space which is driven by the Bayesian Incremental Parameter Setting algorithm (Briscoe, 1999). We compare the efficiency of the TLA with the CGL when acquiring an independently and identically distributed English-like language in noiseless conditions. We show that when convergence to the target grammar occurs (which is not guaranteed), the expected number of steps to convergence for the TLA is shorter than that for the CGL initialized with uniform priors. However, the CGL converges more reliably than the TLA. We discuss the trade-off of efficiency against more reliable convergence to the target grammar.

1 Introduction

A normal child acquires the language of her environment without any specific training. Chomsky (1965) claims that, given the "relatively slight exposure" to examples and "remarkable complexity" of language, it would be "an extraordinary intellectual achievement" for a child to acquire a language if not specifically designed to do so. His *Argument from the Poverty of the Stimulus* suggests that if we know X, and X is undetermined by learning experience then X must be innate. For an example consider structure dependency in language syntax:

A question in English can be formed by inverting the auxiliary verb and subject noun-phrase: (1a) "*Dinah* **was** drinking a saucer of milk"; (1b) "**was** *Dinah* drinking a saucer of milk?"

Upon exposure to this example, a child could hypothesize infinitely many question-formation rules, such as: (i) *swap the first and second words in the sentence*; (ii) *front the first auxiliary verb*; (iii) *front words beginning with w*.

The first two of these rules are refuted if the child encounters the following: (2a) "*the cat who was grinning at Alice* **was** disappearing"; (2b) "**was** *the cat who was grinning at Alice* disappearing?"

If a child is to converge upon the correct hypothesis unaided she must be exposed to sufficient examples so that all false hypotheses are refuted. Unfortunately such examples are not readily available in child-directed speech; even the constructions in examples **(2a)** and **(2b)** are rare (Legate, 1999). To compensate for this lack of data Chomsky suggests that some principles of language are already available in the child's mind. For example, if the child had innately "known" that all grammar rules are structurally-dependent upon syntax she would never have hypothesized rules **(i)** and **(iii)**. Thus, Chomsky theorizes that a human mind contains a Universal Grammar which defines a hypothesis-space of "legal" grammars.[1] This hypothesis-space must be both large enough to contain grammar's for all of the world's languages and small enough to ensure successful acquisition given the sparsity of data. Language acquisition is the process of searching the hypothesis-space for the grammar that most closely describes the language of the environment. With estimates of the number of living languages being around 6800 (Ethnologue, 2004) it is not sensible to model the hypothesis-space of grammars explicitly, rather it must be modeled parametrically. Language acquisition is then the process of setting these parameters. Chomsky (1981) suggested that parameters should represent points of variation between languages, however the only requirement for parameters is that they define the current hypothesis-space.

[1] Discussion of structural dependence as evidence of the Argument from the Poverty of Stimulus is illustrative, the significance being that innate knowledge in any form will place constraints on the hypothesis-space

The properties of the parameters used by this learner (the CGL) are as follows: (1) Parameters are lexical; (2) Parameters are inheritance based; (3) Parameter setting is statistical.

1 - Lexical Parameters

The CGL employs parameter setting as a means to acquire a lexicon; differing from other parametric learners, (such as the Triggering Learning Algorithm (TLA) (Gibson and Wexler, 1994) and the Structural Triggers Learner (STL) (Fodor, 1998b), (Sakas and Fodor, 2001)) which acquire general syntactic information rather than the syntactic properties associated with individual words.[2]

In particular, a categorial grammar is acquired. The syntactic properties of a word are contained in its lexical entry in the form of a syntactic category. A word that may be used in multiple syntactic situations (or sub-categorization frames) will have multiple entries in the lexicon.

Syntactic categories are constructed from a finite set of primitive categories combined with two operators ($/$ and $\backslash$) and are defined by their members ability to combine with other constituents; thus constituents may be thought of as either functions or arguments.

The arguments of a functional constituent are shown to the right of the operators and the result to the left. The forward slash operator ($/$) indicates that the argument must appear to the right of the function and a backward slash ($\backslash$) indicates that it must appear on the left. Consider the following CFG structure which describes the properties of a transitive verb:

s	→	**np vp**
vp	→	**tv np**
tv	→	gets, finds, ...

Assume that there is a set of primitive categories $\{\mathbf{s}, \mathbf{np}\}$. A $\mathbf{vp}$ must be in the category of functional constituents that takes a $\mathbf{np}$ from the left and returns an $\mathbf{s}$. This can be written $\mathbf{s}\backslash\mathbf{np}$. Likewise a $\mathbf{tv}$ takes an $\mathbf{np}$ from the right and returns a $\mathbf{vp}$ (whose type we already know). A $\mathbf{tv}$ may be written $(\mathbf{s}\backslash\mathbf{np})/\mathbf{np}$.

Rules may be used to combine categories. We assume that our learner is innately endowed with the rules of function application, function composition and generalized weak permutation (Briscoe, 1999) (see figures 1 and 2).

- Forward Application ($>$)
  $X/Y\ Y \rightarrow X$
- Backward Application ($<$)
  $Y\ X\backslash Y \rightarrow X$
- Forward Composition ($> B$)
  $X/Y\ Y/Z \rightarrow X/Z$
- Backward Composition ($< B$)
  $Y\backslash X\ Z\backslash Y \rightarrow X\backslash Z$
- Generalized Weak Permutation (P)
  $((X \mid Y_1)... \mid Y_n) \rightarrow ((X \mid Y_n)... \mid Y_1)$
  where $\mid$ is a variable over $\backslash$ and $/$.

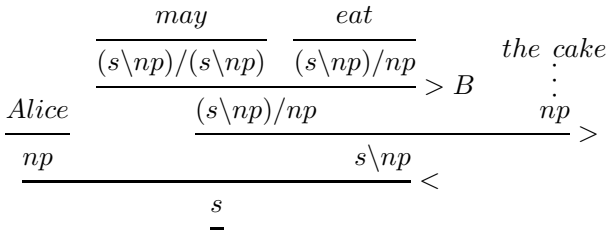


Figure 1: Illustration of forward/backward application ($>$, $<$) and forward composition ($> B$)

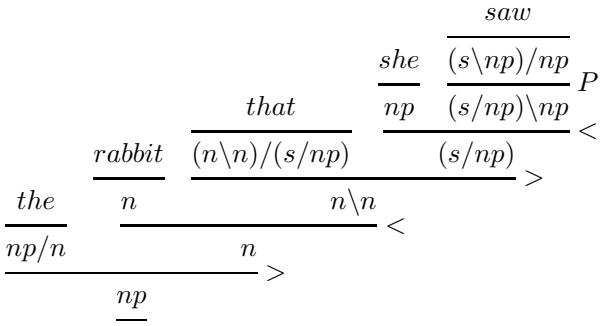


Figure 2: Illustration of generalized weak permutation (P)

The lexicon for a language will contain a finite subset of all possible syntactic categories, the size of which depends on the language. Steedman (2000) suggests that for English the lexical functional categories never need more than five arguments and that these are needed only in a limited number of cases such as for the verb *bet* in the sentence *I bet you five pounds for England to win*.

The categorial grammar parameters of the CGL are concerned with defining the set of syntactic categories present in the language of the environment. Converging on the correct set aids acquisition by constraining the learner's hypothesized syntactic categories for an unknown word. A parameter (with

[2]The concept of lexical parameters and the lexical-linking of parameters is to be attributed to Borer (1984).

value of either ACTIVE or INACTIVE) is associated with every possible syntactic category to indicate whether the learner considers the category to be part of the target grammar.

Some previous parametric learners (TLA and STL) have been primarily concerned with overall syntactic phenomena rather than the syntactic properties of individual words. Movement parameters (such as the $V2$ parameter of the TLA) may be captured by the CGL using innate rules or multiple lexical entries. For instance, Dutch and German word order is captured by assuming that verbs in these languages systematically have two categories, one determining main clause order and the other subordinate clause orders.

2 - Inheritance Based Parameters

The complex syntactic categories of a categorial grammar are a sub-categorization of simpler categories; consequently categories may be arranged in a hierarchy with more complex categories inheriting from simpler ones. Figure 3 shows a fragment of a possible hierarchy. This hierarchical organization of parameters provides the learner with several benefits: (1) The hierarchy can enforce an order on learning; constraints may be imposed such that a parent parameter must be acquired before a child parameter (for example, in Figure 3, the learner must acquire intransitive verbs before transitive verbs may be hypothesized). (2) Parameter values may be inherited as a method of acquisition. (3) The parameters are stored efficiently.

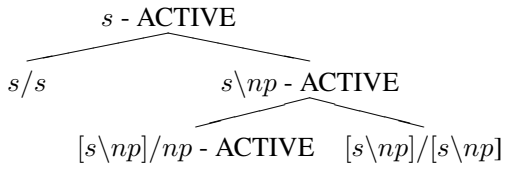


Figure 3: Partial hierarchy of syntactic categories. Each category is associated with a parameter indicating either ACTIVE or INACTIVE status.

3 - Statistical Parameter Setting

The learner uses a statistical method to track relative frequencies of parameter-setting-utterances in the input.[3] We use the Bayesian Incremental Parameter Setting (BIPS) algorithm (Briscoe, 1999) to set the categorial parameters. Such an approach sets the parameters to the values that are most likely given all the accumulated evidence. This represents a compromise between two extremes: implementations of the TLA are memoryless allowing a parameter values to oscillate; some implementations of the STL set a parameter once, for all time.

Using the BIPS algorithm, evidence from an input utterance will either strengthen the current parameter settings or weaken them. Either way, there is re-estimation of the probabilities associated with possible parameter values. Values are only assigned when sufficient evidence has been accumulated, i.e. once the associated probability reaches a threshold value. By employing this method, it becomes unlikely for parameters to switch between settings as the consequence of an erroneous utterance.

Another advantage of using a Bayesian approach is that we may set default parameter values by assigning Bayesian priors; if a parameter's default value is strongly biased against the accumulated evidence then it will be difficult to switch. Also, we no longer need to worry about ambiguity in parameter-setting-utterances (Clark, 1992) (Fodor, 1998b): the Bayesian approach allows us to solve this problem "for free" since indeterminacy just becomes another case of error due to misclassification of input data (Buttery and Briscoe, 2004).

2 Overview of the Categorial Grammar Learner

The learning system is composed of a three modules: a semantics learning module, syntax learning module and memory module. For each utterance heard the learner receives an input stream of word tokens paired with possible semantic hypotheses. For example, on hearing the utterance "Dinah drinks milk" the learner may receive the pairing: ({*dinah*, *drinks*, *milk*}, **drinks(dinah, milk)**).

2.1 The Semantic Module

The semantic module attempts to learn the mapping between word tokens and semantic symbols, building a lexicon containing the meaning associated with each word sense. This is achieved by analyzing each input utterance and its associated semantic hypotheses using cross-situational techniques (following Siskind (1996)).

For a trivial example consider the utterances "Alice laughs" and "Alice eats cookies"; they might have word tokens paired with semantic expressions as follows: ({*alice*, *laughs*}, **laugh(alice)**), ({*alice*, *eats*, *cookies*}, **eat(alice, cookies)**).

From these two utterances it is possible to ascertain that the meaning associated with the word token *alice* must be **alice** since it is the only semantic element that is common to both utterances.

[3] Other statistical parameter setting models include Yang's Variational model (2002) and the Guessing STL (Fodor, 1998a)

2.2 The Syntactic Module

The learning system links the semantic module and syntactic module by using a typing assumption: *the semantic arity of a word is usually the same as its number of syntactic arguments.* For example, if it is known that *likes* maps to $\mathbf{like}(\mathbf{x}, \mathbf{y})$, then the typing assumption suggests that its syntactic category will be in one of the following forms: $a\backslash b\backslash c$, $a/b\backslash c$, $a\backslash b/c$, $a/b/c$ or more concisely $a \mid b \mid c$ (where a, b and c may be basic or complex syntactic categories themselves).

By employing the typing assumption the number of arguments in a word's syntactic category can be hypothesized. Thus, the objective of the syntactic module is to discover the arguments' category types and locations.

The module attempts to create valid parse trees starting from the syntactic information already assumed by the typing assumption (following Buttery (2003)). A valid parse is one that adheres to the rules of the categorial grammar as well as the constraints imposed by the current settings of the parameters. If a valid parse can not be found the learner assumes the typing assumption to have failed and backtracks to allow type raising.

2.3 Memory Module

The memory module records the current state of the hypothesis-space. The syntactic module refers to this information to place constraints upon which syntactic categories may be hypothesized. The module consists of two hierarchies of parameters which may be set using the BIPS algorithm:

Categorial Parameters determine whether a category is in use within the learner's current model of the input language. An inheritance hierarchy of all possible syntactic categories (for up to five arguments) is defined and a parameter associated with each one (Villavicencio, 2002). Every parameter (except those associated with primitive categories such as S) is originally set to INACTIVE, i.e. no categories (except primitives) are known upon the commencement of learning. A categorial parameter may only be set to ACTIVE if its parent category is already active and there has been satisfactory evidence that the associated category is present in the language of the environment.

Word Order Parameters determine the underlying order in which constituents occur. They may be set to either FORWARD or BACKWARD depending on whether the constituents involved are generally located to the right or left. An example is the parameter that specifies the direction of the subject of a verb: if the language of the environment is English this parameter would be set to BACKWARD since subjects generally appear to the left of the verb. Evidence for the setting of word order parameters is collected from word order statistics of the input language.

3 The acquisition of an English-type language

The English-like language of the three-parameter system studied by Gibson and Wexler has the parameter settings and associated unembedded surface-strings as shown in Figure 4. For this task we assume that the surface-strings of the English-like language are independent and identically distributed in the input to the learner.

Specifier	Complement	V2
0 (*Left*)	1 (*Right*)	0 (*off*)

1. Subj Verb
2. Subj Verb Obj
3. Subj Verb Obj Obj
4. Subj Aux Verb
5. Subj Aux Verb Obj
6. Subj Aux Verb Obj Obj
7. Adv Subj Verb
8. Adv Subj Verb Obj
9. Adv Subj Verb Obj Obj
10. Adv Subj Aux Verb
11. Adv Subj Aux Verb Obj
12. Adv Subj Aux Verb Obj Obj

Figure 4: Parameter settings and surface-strings of Gibson and Wexler's English-like Language.

3.1 Efficiency of Trigger Learning Algorithm

For the TLA to be successful it must converge to the correct parameter settings of the English-like language. Berwick and Niyogi (1996) modeled the TLA as a Markov process (see Figure 5).

Using this model it is possible to calculate the probability of converging to the target from each starting grammar and the expected number of steps before convergence.

Probability of Convergence:

Consider starting from Grammar 3, after the process finishes looping it has a $3/5$ probability of moving to Grammar 4 (from which it will never converge) and a $2/5$ probability of moving to Grammar 7 (from which it will definitely converge), therefore there is a 40% probability of converging to the target grammar when starting at Grammar 3.

Expected number of Steps to Convergence:
Let S_n be the expected number of steps from state n to the target state. For starting grammars 6, 7 and 8, which definitely converge, we know:

$$S_6 = 1 + \frac{5}{6}S_6 \quad (1)$$

$$S_7 = 1 + \frac{2}{3}S_7 + \frac{1}{18}S_8 \quad (2)$$

$$S_8 = 1 + \frac{1}{12}S_6 + \frac{1}{36}S_7 + \frac{8}{9}S_8 \quad (3)$$

and for the times when we do converge from grammars 3 and 1 we can expect:

$$S_1 = 1 + \frac{3}{5}S_1 \quad (4)$$

$$S_3 = 1 + \frac{31}{33}S_3 \quad (5)$$

Figure 6 shows the probability of convergence and expected number of steps to convergence for each of the starting grammars. The expected number of steps to convergence ranges from infinity (for starting grammars 2 and 4) down to 2.5 for Grammar 1. If the distribution over the starting grammars is uniform then the overall probability of converging is the sum of the probabilities of converging from each state divided by the total number of states:

$$\frac{1.00 + 1.00 + 1.00 + 1.00 + 0.40 + 0.66}{8} = 0.63 \quad (6)$$

and the expected number of steps given that you converge is the weighted average of the number of steps from each possibly converging state:

$$\frac{5.47 + 14.87 + 6 + 21.98 \times 0.4 + 2.5 \times 0.66}{1.00 + 1.00 + 1.00 + 1.00 + 0.40 + 0.66} = 7.26 \quad (7)$$

3.2 Efficiency of Categorial Grammar Learner

The input data to the CGL would usually be an utterance annotated with a logical form; the only data available here however, is surface-strings consisting of word types. Hence, for the purpose of comparison with the TLA the semantic module of our learner is by-passed; we assume that mappings to semantic forms have previously been acquired and that the subject and objects of surface-strings are known. For example, given surface-string 1 (*Subj Verb*) we assume the mapping *Verb* $\mapsto$ **verb(x)**, which provides *Verb* with a syntactic category of the form $a|b$ by the typing assumption (where a, b are unknown syntactic categories and $|$ is an operator over $\backslash$ and $/$); we also assume *Subj* to map to a primitive syntactic category SB, since it is the subject of *Verb*.

The criteria for success for the CGL when acquiring Gibson and Wexler's English-like language is a lexicon containing the following:[4]

Adv	S/S	**Aux**	$[S\backslash SB]/[S\backslash SB]$
Obj	OB	**Verb**	$S\backslash SB$
Subj	SB		$[S\backslash SB]/OB$
			$[[S\backslash SB]/OB]/OB$

where S (sentence), SB (subject) and OB (object) are primitive categories which are innate to the learner with SB and OB assumed to be derivable from the semantic module.

During the learning process the CGL will have constructed a category hierarchy by setting appropriate categorial parameters to true (see Figure 7). The learner will have also constructed a word-order hierarchy (Figure 8), setting parameters to FORWARD or BACKWARD. These hierarchies are used during the learning process to constrain hypothesized syntactic categories. For this task the setting of the word-order parameters becomes trivial and their role in constraining hypotheses negligible; consequently, the rest of our argument will relate to categorial parameters only. For the purpose of this analysis parameters are initialized with uniform priors and are originally set INACTIVE. Since the input is noiseless, the switching threshold is set such that parameters may be set ACTIVE upon the evidence from one surface-string.

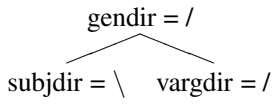


Figure 8: Word-order parameter settings required to parse Gibson and Wexler's English-like language.

It is a requirement of the parameter setting device that the parent-types of hypothesized syntax categories are ACTIVE before new parameters are set. Thus, the learner is not allowed to hypothesize the syntactic category for a transitive verb $[[S\backslash SB]/OB]$ before it has learnt the category for an intransitive verb $[S\backslash SB]$; this behaviour constrains over-generation. Additionally, it is usually not possible to derive a word's full syntactic category (i.e. without any remaining unknowns) unless it is the only new word in the clause.

As a consequence of these issues, the order in which the surface-strings appear to the learner af-

[4]Note that the lexicon would usually contain orthographic entries for the words in the language rather than word type entries.

fects the speed of acquisition. For instance, the learner prefers to see the surface-string *Subj Verb* before *Subj Verb Obj* so that it can acquire the maximum information without wasting any strings. For the English-type language described by Gibson and Wexler the learner can optimally acquire the whole lexicon after seeing only 5 surface-strings (one string needed for each new complex syntactic category to be learnt). However, the strings appear to the learner in a random order so it is necessary to calculate the expected number of strings (or steps) before convergence.

The learner must necessarily see the string *Subj Verb* before it can learn any other information. With 12 surface-strings the probability of seeing *Subj Verb* is $1/12$ and the expected number of strings before it is seen is 12. The learner can now learn from 3 surface-strings: *Subj Verb Obj*, *Subj Aux Verb* and *Adv Subj Verb*. Figure 9 shows a Markov structure of the process. From the model we can calculate the expected number of steps to converge to be 24.53.

4 Conclusions

The TLA and CGL were compared for efficiency (expected number of steps to convergence) when acquiring the English-type grammar of the three-parameter system studied by Gibson and Wexler. The expected number of steps for the TLA was found to be 7.26 but the algorithm only converged 63% of the time. The expected number of steps for the CGL is 24.53 but the learner converges more reliably; a trade off between efficiency and success. With noiseless input the CGL can only fail if there is insufficient input strings or if Bayesian priors are heavily biased against the target. Furthermore, the CGL can be made robust to noise by increasing the probability threshold at which a parameter may be set ACTIVE; the TLA has no mechanism for coping with noisy data.

The CGL learns incrementally; the hypothesis-space from which it can select possible syntactic categories expands dynamically and, as a consequence of the hierarchical structure of parameters, the speed of acquisition increases over time. For instance, in the starting state there is only a $1/12$ probability of learning from surface-strings whereas in state k (when all but one category has been acquired) there is a $1/2$ probability. It is likely that with a more complex learning task the benefits of this incremental approach will outweigh the slow starting costs. Related work on the effects of incremental learning on STL performance (Sakas, 2000) draws similar conclusions. Future work hopes to compare the CGL with other parametric learners (such as the STL) in larger domains.

References

R Berwick and P Niyogi. 1996. Learning from triggers. *Linguistic Inquiry*, 27(4):605–622.

H Borer. 1984. *Parametric Syntax: Case Studies in Semitic and Romance Languages*. Foris, Dordrecht.

E Briscoe. 1999. The acquisition of grammar in an evolving population of language agents. *Machine Intelligence*, 16.

P Buttery and T Briscoe. 2004. The significance of errors to parametric models of language acquisition. Technical Report SS-04-05, American Association of Artificial Intelligence, March.

P Buttery. 2003. A computational model for first language acquisition. In *CLUK-6*, Edinburgh.

N Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press.

N Chomsky. 1981. *Lectures on Government and Binding*. Foris Publications.

R Clark. 1992. The selection of syntactic knowledge. *Language Acquisition*, 2(2):83–149.

Ethnologue. 2004. Languages of the world, 14th edition. SIL International. http://www.ethnologue.com/.

J Fodor. 1998a. Parsing to learn. *Journal of Psycholinguistic Research*, 27(3):339–374.

J Fodor. 1998b. Unambiguous triggers. *Linguistic Inquiry*, 29(1):1–36.

E Gibson and K Wexler. 1994. Triggers. *Linguistic Inquiry*, 25(3):407–454.

J Legate. 1999. Was the argument that was made empirical? Ms, Massachusetts Institute of Technology.

W Sakas and J Fodor. 2001. The structural triggers learner. In S Bertolo, editor, *Language Acquisition and Learnability*, chapter 5. Cambridge University Press, Cambridge, UK.

W Sakas. 2000. *Ambiguity and the Computational Feasibility of Syntax Acquisition*. Ph.D. thesis, City University of New York.

J Siskind. 1996. A computational study of cross situational techniques for learning word-to-meaning mappings. *Cognition*, 61(1-2):39–91, Nov/Oct.

M Steedman. 2000. *The Syntactic Process*. MIT Press/Bradford Books.

A Villavicencio. 2002. *The acquisition of a unification-based generalised categorial grammar*. Ph.D. thesis, University of Cambridge.

C Yang. 2002. *Knowledge and Learning in Natural Language*. Oxford University Press.

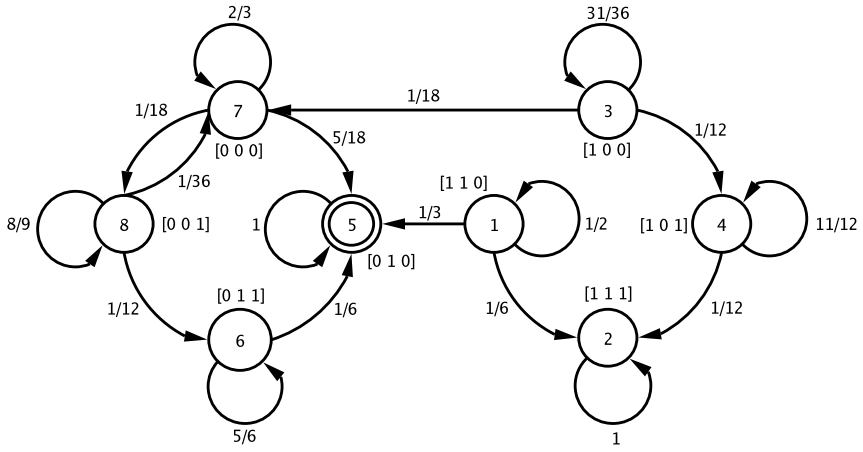


Figure 5: Gibson and Wexler's TLA as a Markov structure. Circles represent possible grammars (a configuration of parameter settings). The target grammar lies at the centre of the structure. Arrows represent the possible transitions between grammars. Note that the TLA is constrained to only allow movement between grammars that differ by one parameter value. The probability of moving between Grammar G_i and Grammar G_j is a measure of the number of target surface-strings that are in G_j but not G_i normalized by the total number of target surface-strings as well as the number of alternate grammars the learner can move to. For example the probability of moving from Grammar 3 to Grammar 7 is $2/12 * 1/3 = 1/18$ since there are 2 target surface-strings allowed by Grammar 7 that are not allowed by Grammar 3 out of a possible of 12 and three grammars that differ from Grammar 3 by one parameter value.

Initial Language	Initial Grammar	Prob. of Converging	Expected no. of Steps
VOS -V2	110	0.66	2.50
VOS +V2	111	0.00	n/a
OVS -V2	100	0.40	21.98
OVS +V2	101	0.00	n/a
SVO -V2	010	1.00	0.00
SVO +V2	011	1.00	6.00
SOV -V2	000	1.00	5.47
SOV +V2	001	1.00	14.87

Figure 6: Probability and expected number of steps to convergence from each starting grammar to an English-like grammar (SVO -V2) when using the TLA.

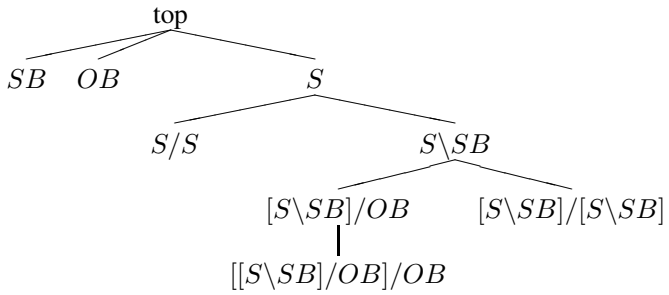


Figure 7: Category hierarchy required to parse Gibson and Wexler's English-like language.

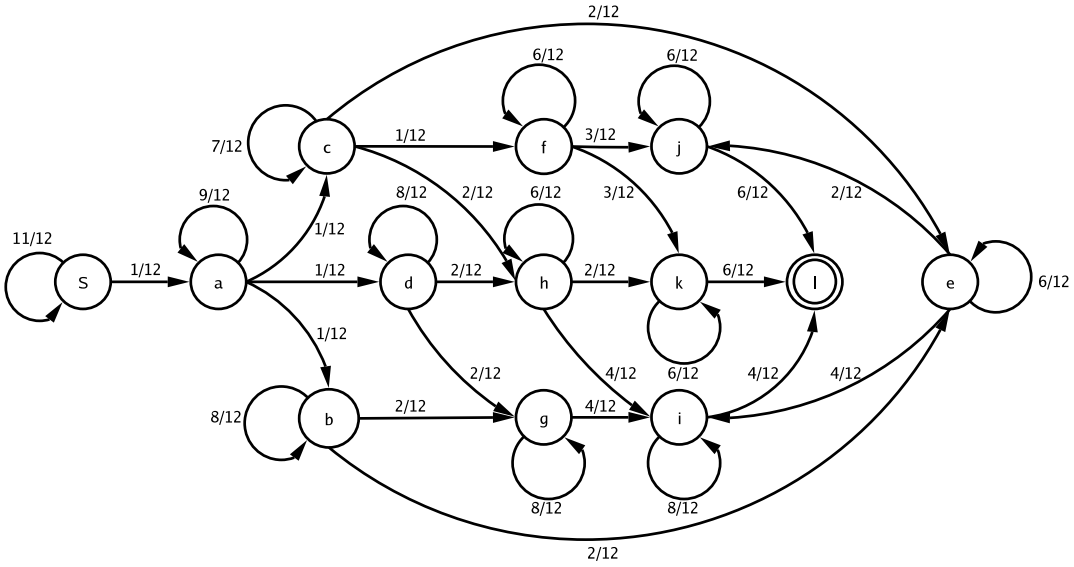


Figure 9: The CGL as a Markov structure. The states represent the set of known syntactic categories: state S - $\{\}$, state a - $\{S\backslash SB\}$, state b - $\{S\backslash SB,\ S/S\}$, state c - $\{S\backslash SB,\ [S\backslash SB]/OB\}$, state d - $\{S\backslash SB,\ [S\backslash SB]/[S\backslash SB]\}$, state e - $\{S\backslash SB,\ S/S,\ [S\backslash SB]/OB\}$, state f - $\{S\backslash SB,\ [S\backslash SB]/OB,\ [[S\backslash SB]/OB]/OB\}$, state g - $\{S\backslash SB,\ [S\backslash SB]/[S\backslash SB],\ S/S\}$ state h - $\{S\backslash SB,\ [S\backslash SB]/[S\backslash SB],\ [S\backslash SB]/OB\}$, state i - $\{S\backslash SB,\ S/S,\ [S\backslash SB]/OB,\ [S\backslash SB]/[S\backslash SB]\}$, state j - $\{S\backslash SB,\ S/S,\ [S\backslash SB]/OB,\ [[S\backslash SB]/OB]/OB\}$, state k - $\{S\backslash SB,\ [S\backslash SB]/OB,\ [[S\backslash SB]/OB]/OB,\ [S\backslash SB]/[S\backslash SB]\}$, state l - $\{S\backslash SB, [S\backslash SB]/OB, [[S\backslash SB]/OB]/OB, [S\backslash SB]/[S\backslash SB], S/S\}$.

On Statistical Parameter Setting

Damir ĆAVAR, Joshua HERRING, Toshikazu IKUTA, Paul RODRIGUES
Linguistics Dept., Indiana University
Bloomington, IN, 46405
dcavar@indiana.edu

Giancarlo SCHREMENTI
Computer Science, Indiana University
Bloomington, IN, 47405
gischrem@indiana.edu

Abstract

We present a model and an experimental platform of a bootstrapping approach to statistical induction of natural language properties that is constraint based with voting components. The system is incremental and unsupervised. In the following discussion we focus on the components for morphological induction. We show that the much harder problem of incremental unsupervised morphological induction can outperform comparable all-at-once algorithms with respect to precision. We discuss how we use such systems to identify cues for induction in a cross-level architecture.

1 Introduction

In recent years there has been a growing amount of work focusing on the computational modeling of language processing and acquisition, implying a cognitive and theoretical relevance both of the models as such, as well as of the language properties extracted from raw linguistic data.[1] In the computational linguistic literature several attempts to induce grammar or linguistic knowledge from such data have shown that at different levels a high amount of information can be extracted, even with no or minimal supervision.

Different approaches tried to show how various puzzles of language induction could be solved. From this perspective, language acquisition is the process of segmentation of non-discrete acoustic input, mapping of segments to symbolic representations, mapping representations on higher-level representations such as phonology, morphology and syntax, and even induction of semantic properties. Due to space restrictions, we cannot discuss all these approaches in detail. We will focus on the close domain of morphology.

Approaches to the induction of morphology as presented in e.g. Schone and Jurafsky (2001) or Goldsmith (2001) show that the morphological properties of a small subset of languages can be induced with high accuracy, most of the existing approaches are motivated by applied or engineering concerns, and thus make assumptions that are less cognitively plausible: a. Large corpora are processed all at once, though unsupervised incremental induction of grammars is rather the approach that would be relevant from a psycholinguistic perspective; b. Arbitrary decisions about selections of sets of elements are made, based on frequency or frequency profile rank,[2] though such decisions should rather be derived or avoided in general.

However, the most important aspects missing in these approaches, however, are the link to different linguistic levels and the support of a general learning model that makes predictions about how knowledge is induced on different linguistic levels and what the dependencies between information at these levels are. Further, there is no study focusing on the type of supervision that might be necessary for the guidance of different algorithm types towards grammars that resemble theoretical and empirical facts about language acquisition, and processing and the final knowledge of language.

While many theoretical models of language acquisition use innateness as a crutch to avoid outstanding difficulties, both on the general and abstract level of I-language as well as the more detailed level of E-language, (see, among others, Lightfoot (1999) and Fodor and Teller (2000), there is also significant research being done which shows that children take advantage of statistical regularities in the input for use in the language-learning task (see Batchelder (1997) and related references within).

In language acquisition theories the dominant view is that knowledge of one linguistic level is bootstrapped from knowledge of one, or even several different levels. Just to mention such approaches: Grimshaw (1981), and Pinker (1984)

[1] See Batchelder (1998) for a discussion of these aspects.

[2] Just to mention some of the arbitrary decisions made in various approaches, e.g. Mintz (1996) selects a small set of all words, the most frequent words, to induce word types via clustering ; Schone and Jurafsky (2001) select words with frequency higher than 5 to induce morphological segmentation.

assume that semantic properties are used to bootstrap syntactic knowledge, and Mazuka (1998) suggested that prosodic properties of language establish a bias for specific syntactic properties, e.g. headedness or branching direction of constituents. However, these approaches are based on conceptual considerations and psycholinguistc empirical grounds, the formal models and computational experiments are missing. It is unclear how the induction processes across linguistic domains might work algorithmically, and the quantitative experiments on large scale data are missing.

As for algorithmic approaches to cross-level induction, the best example of an initial attempt to exploit cues from one level to induce properties of another is presented in Déjean (1998), where morphological cues are identified for induction of syntactic structure. Along these lines, we will argue for a model of statistical cue-based learning, introducing a view on bootstrapping as proposed in Elghamry (2004), and Elghamry and Ćavar (2004), that relies on identification of elementary cues in the language input and incremental induction and further cue identification across all linguistic levels.

1.1 Cue-based learning

Presupposing input driven learning, it has been shown in the literature that initial segmenations into words (or word-like units) is possible with unsupervised methods (e.g. Brent and Cartwright (1996)), that induction of morphology is possible (e.g. Goldsmith (2001), Schone and Jurafsky (2001)) and even the induction of syntactic structures (e.g. Van Zaanen (2001)). As mentioned earlier, the main drawback of these approaches is the lack of incrementality, certain arbitrary decisions about the properties of elements taken into account, and the lack of integration into a general model of bootstrapping across linguistic levels.

As proposed in Elghamry (2004), cues are elementary language units that can be identified at each linguistic level, dependent or independent of prior induction processes. That is, intrinsic properties of elements like segments, syllables, morphemes, words, phrases etc. are the ones available for induction procedures. Intrinsic properties are for example the frequency of these units, their size, and the number of other units they are build of. Extrinsic properties are taken into account as well, where extrinsic stands for distributional properties, the context, relations to other units of the same type on one, as well as across linguistic levels. In this model, extrinsic and intrinsic properties of elementary language units are the cues that are used for grammar induction only.

As shown in Elghamry (2004) and Elghamry and Ćavar (2004), there are efficient ways to identify a kernel set of such units in an unsupervised fashion without any arbitrary decision where to cut the set of elements and on the basis of what kind of features. They present an algorithm that selects the set of kernel cues on the lexical and syntactic level, as the smallest set of words that co-occurs with all other words. Using this set of words it is possible to cluster the lexical inventory into open and closed class words, as well as to identify the subclasses of nouns and verbs in the open class. The direction of the selectional preferences of the language is derived as an average of point-wise Mutual Information on each side of the identified cues and types, which is a self-supervision aspect that biases the search direction for a specific language. This resulting information is understood as derivation of secondary cues, which then can be used to induce selectional properties of verbs (frames), as shown in Elghamry (2004).

The general claim thus is:

- Cues can be identified in an unsupervised fashion in the input.
- These cues can be used to induce properties of the target grammar.
- These properties represent cues that can be used to induce further cues, and so on.

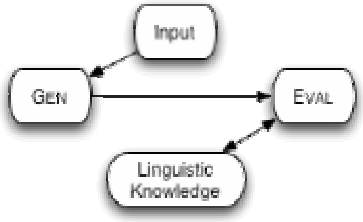
The hypothesis is that this snowball effect can reduce the search space of the target grammar incrementally. The main research questions are now, to what extend do different algorithms provide cues for other linguistic levels and what kind of information do they require as supervision in the system, in order to gain the highest accuracy at each linguistic level, and how does the linguistic information of one level contribute to the information on another.

In the following, the architectural considerations of such a computational model are discussed, resulting in an example implementation that is applied to morphology induction, where morphological properties are understood to represent cues for lexical clustering as well as syntactic structure, and vice versa, similar to the ideas formulated in Déjean (1998), among others.

1.2 Incremental Induction Architecture

The basic architectural principle we presuppose is incrementality, where incrementally utterances are processed. The basic language unit is an utterance, with clear prosodic breaks before and after. The induction algorithm consumes such utterances and breaks them into basic linguistic units, generating for each step hypotheses about

the linguistic structure of each utterance, based on the grammar built so far and statistical properties of the single linguistic units. Here we presuppose a successful segmentation into words, i.e. feeding the system utterances with unambiguous word boundaries. We implemented the following pipeline architecture:



The GEN module consumes input and generates hypotheses about its structural descriptions (SD). EVAL consumes a set of SDs and selects the set of best SDs to be added to the knowledge base. The knowledge base is a component that not only stores SDs but also organizes them into optimal representations, here morphology grammars.

All three modules are modular, containing a set of algorithms that are organized in a specific fashion. Our intention is to provide a general platform that can serve for the evaluation and comparison of different approaches at every level of the induction process. Thus, the system is designed to be more general, applicable to the problem of segmentation, as well as type and grammar induction.

We assume for the input to consist of an alphabet: a non-empty set A of n symbols $\{s_1, s_2,... s_n\}$. A word w is a non-empty list of symbols $w = [s_1,s_2,... s_n]$, with $s \in A$. The corpus is a non-empty list C of words $C = [w_1,w_2,... w_n]$.

In the following, the individual modules for the morphology induction task are described in detail.

1.2.1 GEN

For the morphology task GEN is compiled from a set of basically two algorithms. One algorithm is a variant of Alignment Based Learning (ABL), as described in Van Zaanen (2001).

The basic ideas in ABL go back to concepts of *substitutability* and/or *complementarity*, as discussed in Harris (1961). The concept of *substitutability* generally applies to central part of the induction procedure itself, i.e. substitutable elements (e.g. substrings, words, structures) are assumed to be of the same type (represented e.g. with the same symbol).

The advantage of ABL for grammar induction is its constraining characteristics with respect to the set of hypotheses about potential structural properties of a given input. While a brute-force method would generate all possible structural representations for the input in a first order explosion and subsequently filter out irrelevant hypotheses, ABL reduces the set of possible SDs from the outset to the ones that are motivated by previous experience/input or a pre-existing grammar.

Such constraining characteristics make ABL attractive from a cognitive point of view, both because hopefully the computational complexity is reduced on account of the smaller set of potential hypotheses, and also because learning of new items, rules, or structural properties is related to a general learning strategy and previous experience only. The approaches that are based on a brute-force first order explosion of all possible hypotheses with subsequent filtering of relevant or irrelevant structures are both memory-intensive and require more computational effort.

The algorithm is not supposed to make any assumptions about types of morphemes. There is no expectation, including use of notions like *stem*, *prefix*, or *suffix*. We assume only linear sequences. The properties of single morphemes, being stems or suffixes, should be a side effect of their statistical properties (including their frequency and co-occurrence patterns, as will be explained in the following), and their alignment in the corpus, or rather within words.

There are no rules about language built-in, such as what a morpheme must contain or how frequent it should be. All of this knowledge is induced statistically.

In the ABL Hypotheses Generation, a given word in the utterance is checked against morphemes in the grammar. If an existing morpheme LEX aligns with the input word INP, a hypothesis is generated suggesting a morphological boundary at the alignment positions:

INP (*speaks*) + LEX (*speak*) = HYP [*speak*, *s*]

Another design criterion for the algorithm is complete language independence. It should be able to identify morphological structures of Indo-European type of languages, as well as agglutinative languages (e.g. Japanese and Turkish) and polysynthetic languages like some Bantu dialects or American Indian languages. In order to guarantee this behavior, we extended the Alignment Based hypothesis generation with a pattern identifier that extracts patterns of character sequences of the types:

1. A — B — A
2. A — B — A — B
3. A — B — A — C

This component is realized with cascaded regular expressions that are able to identify and

return the substrings that correspond to the repeating sequences.[3]

All possible alignments for the existing grammar at the current state, are collected in a hypothesis list and sent to the EVAL component, described in the following. A hypothesis is defined as a tuple:

$H = <w, f, g>$, with w the input word, f its frequency in C, and g a list of substrings that represent a linear list of morphemes in w, $g = [m_1, m_2, ... m_n]$.

1.2.2 EVAL

EVAL is a voting based algorithm that subsumes a set of independent algorithms that judge the list of SDs from the GEN component, using statistical and information theoretic criteria. The specific algorithms are grouped into memory and usability oriented constraints.

Taken as a whole, the system assumes two (often competing) cognitive considerations. The first of these forms a class of what we term "time-based" constraints on learning. These constraints are concerned with the processing time required of a system to make sense of items in an input stream, whereby "time" is understood to mean the number of steps required to generate or parse SDs rather than the actual temporal duration of the process. To that end, they seek to minimize the amount of structure assigned to an utterance, which is to say they prefer to deal with as few rules as possible. The second of these cognitive considerations forms a class of "memory-based" constraints. Here, we are talking about constraints that seek to minimize the amount of memory space required to store an utterance by maximizing the efficiency of the storage process. In the specific case of our model, which deals with morphological structure, this means that the memory-based constraints search the input string for regularities (in the form of repeated substrings) that then need only be stored once (as a pointer) rather than each time they are found. In the extreme case, the time-based constraints prefer storing the input "as is", without any processing at all, where the memory-based constraints prefer a rule for every character, as this would assign maximum structure to the input. Parsable information falls out of the tension between these two conflicting constraints, which can then be applied to organize the input into potential syntactic categories. These can then be used to set the parameters for the internal adult parsing system.

Each algorithm is weighted. In the current implementation these weights are set manually. In future studies we hope to use the weighting for self-supervision.[4] Each algorithm assigns a numerical rank to each hypothesis multiplied with the corresponding weight, a real number between 0 and 1.

On the one hand, our main interest lies in the comparison of the different algorithms and a possible interaction or dependency between them. Also, we expect the different algorithms to be of varying importance for different types of languages.

Mutual Information (MI)

For the purpose of this experiment we use a variant of standard Mutual Information (MI), see e.g. MacKay (2003). Information theory tells us that the presence of a given morpheme restricts the possibilities of the occurrence of morphemes to the left and right, thus lowering the amount of bits needed to store its neighbors. Thus we should be able to calculate the amount of bits needed by a morpheme to predict its right and left neighbors respectively. To calculate this, we have designed a variant of mutual information that is concerned with a single direction of information.

This is calculated in the following way. For every morpheme y that occurs to the right of x we sum the point-wise MI between x and y, but we relativize the point-wise MI by the probability that y follows x, given that x occurs. This then gives us the expectation of the amount of information that x tells us about which morpheme will be to its right. Note that $p(<xy>)$ is the probability of the bigram $<xy>$ occurring and is not equal to $p(<yx>)$ which is the probability of the bigram $<yx>$ occurring.

We calculate the MI on the right side of $x \in G$ by:

$$\sum_{y \in \{<xY>\}} p(< xy > | x) \lg \frac{p(< xy >)}{p(x)p(y)}$$

and the MI on the left of $x \in G$ respectively by:

$$\sum_{y \in \{<Yx>)} p(< yx > | x) \lg \frac{p(< yx >)}{p(y)p(x)}$$

One way we use this as a metric, is by summing up the left and right MI for each morpheme in a

[3] This addition might be understood to be a sort of *supervision* in the system. However, as shown in recent research on human cognitive abilities, and especially on the ability to identify patterns in the speech signal by very young infants (Marcus et al, 1999) shows that we can assume such an ability to be part of the cognitive abilities, maybe not even language specific

[4] One possible way to self-supervise the weights in this architecture is by taking into account the revisions subsequent components make when they optimize the grammar. If rules or hypotheses have to be removed from the grammar due to general optimization constraints on the grammars as such, the weight of the responsible algorithm can be lowered, decreasing its general value in the system on the long run. The relevant evaluations with this approach are not yet finished.

hypothesis. We then look for the hypothesis that results in the maximal value of this sum. The tendency for this to favor hypotheses with many morphemes is countered by our criterion of favoring hypotheses that have fewer morphemes, discussed later.

Another way to use the left and right MI is in judging the quality of morpheme boundaries. In a good boundary, the morpheme on the left side should have high right MI and the morpheme on the right should have high left MI. Unfortunately, MI is not reliable in the beginning because of the low frequency of morphemes. However, as the lexicon is extended during the induction procedure, reliable frequencies are bootstrapping this segmentation evaluation.

Minimum Description Length (DL)

The principle of Minimum Description Length (MDL), as used in recent work on grammar induction and unsupervised language acquisition, e.g. Goldsmith (2001) and De Marcken (1996), explains the grammar induction process as an iterative minimization procedure of the grammar size, where the smaller grammar corresponds to the *best* grammar for the given data/corpus.

The description length metric, as we use it here, tells us how many bits of information would be required to store a word given a hypothesis of the morpheme boundaries, using the so far generated grammar. For each morpheme in the hypothesis that doesn't occur in the grammar we need to store the string representing the morpheme. For morphemes that do occur in our grammar we just need to store a pointer to that morphemes entry in the grammar. We use a simplified calculation, taken from Goldsmith (2001), of the cost of storing a string that takes the number of bits of information required to store a letter of the alphabet and multiply it by the length of the string.

$$\lg(len(alphabet)) * len(morpheme)$$

We have two different methods of calculating the cost of the pointer. The first assigns a variable the cost based on the frequency of the morpheme that it is pointing to. So first we calculate the frequency rank of the morpheme being pointed to, (e.g. the most frequent has rank 1, the second rank 2, etc.). We then calculate:

$$floor(\lg(freq\_rank) - 1)$$

to get a number of bits similar to the way Morse code assigns lengths to various letters.

The second is simpler and only calculates the entropy of the grammar of morphemes and uses this as the cost of all pointers to the grammar. The entropy equation is as follows:

$$\sum_{x \in G} p(x) \lg \frac{1}{p(x)}$$

The second equation doesn't give variable pointer lengths, but it is preferred since it doesn't carry the heavy computational burden of calculating the frequency rank.

We calculate the description length for each GEN hypothesis only,[5] by summing up the cost of each morpheme in the hypothesis. Those with low description lengths are favored.

Relative Entropy (RE)

We are using RE as a measure for the cost of adding a hypothesis to the existing grammar. We look for hypotheses that when added to the grammar will result in a low divergence from the original grammar.

We calculate RE as a variant of the Kullback-Leibler Divergence, see MacKay (2003). Given grammar G_1, the grammar generated so far, and G_2 the grammar with the extension generated for the new input increment, $P(X)$ is the probability mass function (*pmf*) for grammar G_2, and $Q(X)$ the *pmf* for grammar G_1:

$$\sum_{x \in X} P(x) \lg \frac{P(x)}{Q(x)}$$

Note that with every new iteration a new element can appear, that is not part of G_1. Our variant of RE takes this into account by calculating the costs for such a new element x to be the point-wise entropy of this element in $P(X)$, summing up over all new elements:

$$\sum_{x \in X} P(x) \lg \frac{1}{P(x)}$$

These two sums then form the RE between the original grammar and the new grammar with the addition of the hypothesis. Hypotheses with low RE are favored.

This metric behaves similarly to description length, that is discussed above, in that both are calculating the distance between our original grammar and the grammar with the inclusion of the new hypothesis. The primary difference is RE also takes into account how the *pmf* differs in the two grammars and that our variation punishes new morphemes based upon their frequency relative to the frequency of other morphemes. Our implementation of MDL does not consider frequency in this way, which is why we are including RE as an independent metric.

Further Metrics

In addition to the mentioned metric, we take into account the following criteria: a. Frequency of

[5] We do not calculate the sizes of the grammars with and without the given hypothesis, just the amount each given hypothesis would add to the grammar, favoring the least increase of total grammar size.

morpheme boundaries; b. Number of morpheme boundaries; c. Length of morphemes.

The frequency of morpheme boundaries is given by the number of hypotheses that contain this boundary. The basic intuition is that the higher this number is, i.e. the more alignments are found at a certain position within a word, the more likely this position represents a morpheme boundary. We favor hypotheses with high values for this criterion.

The number of morpheme boundaries indicates how many morphemes the word was split into. To prevent the algorithm from degenerating into the state where each letter is identified as a morpheme, we favor hypotheses with low number of morpheme boundaries.

The length of the morphemes is also taken into account. We favor hypotheses with long morphemes to prevent the same degenerate state as the above criterion.

1.2.3 Linguistic Knowledge

The acquired lexicon is stored in a hypothesis space which keeps track of the words from the input and the corresponding hypotheses. The hypothesis space is defined as a list of hypotheses:

Hypotheses space: $S = [H_1, H_2, ... H_n]$

Further, each morpheme that occurred in the SDs of words in the hypothesis space is kept with its frequency information, as well as bigrams that consist of morpheme pairs in the SDs and their frequency.[6]

Similar to the specification of signatures in Goldsmith (2001), we list every morpheme with the set of morphemes it co-occurs. Signatures are lists of morphemes. Grammar construction is performed by replacement of morphemes with a symbol, if they have equal signatures.

The hypothesis space is virtually divided into two sections, long term and short term storage. Long term storage is not revised further, in the current version of the algorithm. The short term storage is cyclically cleaned up by eliminating the signatures with a low likelihood, given the long term storage.

2 The experimental setting

In the following we discuss the experimental setting. We used the Brown corpus,[7] the child-oriented speech portion of the CHILDES Peter corpus,[8] and Caesar's "De Bello Gallico" in Latin.[9]

From the Brown corpus we used the files ck01 – ck09, with an average number of 2000 words per chapter. The total number of words in these files is 18071. The randomly selected portion of "De Bello Gallico" contained 8300 words. The randomly selected portion of the Peter corpus contains 58057 words.

The system reads in each file and dumps log information during runtime that contains the information for online and offline evaluation, as described below in detail.

The gold standard for evaluation is based on human segmentation of the words in the respective corpora. We create for every word a manual segmentation for the given corpora, used for online evaluation of the system for accuracy of hypothesis generation during runtime. Due to complicated cases, where linguist are undecided about the accurate morphological segmentation, a team of 5 linguists was cooperating with this task.

The offline evaluation is based on the grammar that is generated and dumped during runtime after each input file is processed. The grammar is manually annotated by a team of linguists, indicating for each construction whether it was segmented correctly and exhaustively. An additional evaluation criterion was to mark undecided cases, where even linguists do not agree. This information was however not used in the final evaluation.

2.1 Evaluation

We used two methods to evaluate the performance of the algorithm. The first analyzes the accuracy of the morphological rules produced by the algorithm after an increment of *n* words. The second looks at how accurately the algorithm parsed each word that it encountered as it progressed through the corpus.

The morphological rule analysis looks at each grammar rule generated by the algorithm and judges it on the correctness of the rule and the resulting parse. A grammar rule consists of a stem and the suffixes and prefixes that can be attached to it, similar to the signatures used in Goldsmith (2001). The grammar rule was then marked as to whether it consisted of legitimate suffixes and prefixes for that stem, and also as to whether the

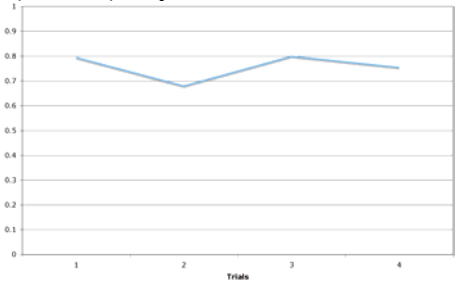
[6] Due to space restrictions we do not formalize this further. A complete documentation and the source code is available at: http://jones.ling.indiana.edu/~abugi/.

[7] The Brown Corpus of Standard American English, consisting of 1,156,329 words from American texts printed in 1961 organized into 59,503 utterances and compiled by W.N. Francis and H. Kucera at Brown University.

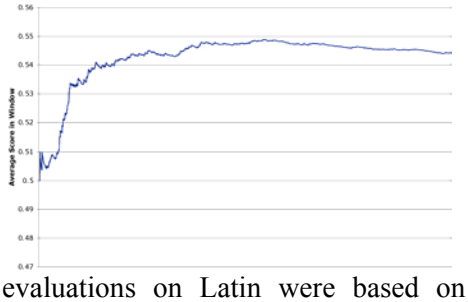
[8] Documented in L. Bloom (1970) and available at http://xml.talkbank.org:8888/talkbank/file/CHILDES/Eng-USA/Bloom70/Peter/.

[9] This was taken from the Gutenberg archive at: http://www.gutenberg.net/etext/10657. The Gutenberg header and footer were removed for the experimental run.

stem of the rule was a true stem, as opposed to a stem plus another morpheme that wasn't identified by the algorithm. The number of rules that were correct in these two categories were then summed, and precision and recall figures were calculated for the trial. The trials described in the graph below were run on three increasingly large portions of the general fiction section of the Brown Corpus. The first trial was run on one randomly chosen chapter, the second trial on two chapters, and the third on three chapters. The graph shows the harmonic average (F-score) of precision and recall.



The second analysis is conducted as the algorithm is running and examines each parse the system produces. The algorithm's parses are compared with the "correct" morphological parse of the word using the following method to derive a numerical score for a particular parse. The first part of the score is the distance in characters between each morphological boundary in the two parses, with a score of one point for each character space. The second part is a penalty of two points for each morphological boundary that occurs in one parse and not the other. These scores were examined within a moving window of words that progressed through the corpus as the algorithm ran. The average scores of words in each such window were calculated as the window advanced. The purpose of this method was to allow the performance of the algorithm to be judged at a given point without prior performance in the corpus affecting the analysis of the current window. The following graph shows how the average performance of the windows of analyzed words as the algorithm progresses through five randomly chosen chapters of general fiction in the Brown Corpus amounting to around 10,000 words. The window size for the following graph was set to 40 words.

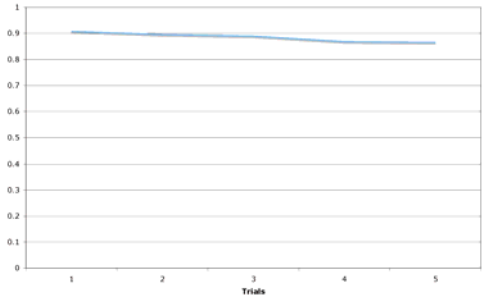


The evaluations on Latin were based on the initial 4000 words of "De Bello Gallico" in a pretest. In the very initial phase we reached a precision of 99.5% and a recall of 13.2%. This is however the preliminary result for the initial phase only. We expect that for a larger corpus the recall will increase much higher, given the rich morphology of Latin, potentially with negative consequences for precision.

The results on the Peter corpus are shown in the following table:

After file	precision	recall
01	.9957	.8326
01-03	.9968	.8121
01-05	.9972	.8019
01-07	.9911	.7710
01-09	.9912	.7666

We notice a more or less stable precision value with decreasing recall, due to a higher number of words. The Peter corpus contains also many very specific transcriptions and tokens that are indeed unique, thus it is rather surprising to get such results at all. The following graphics shows the F-score for the Peter corpus:



3 Conclusion

The evaluations on two related morphology systems show that with a restrictive setting of the parameters in the described algorithm, approx 99% precision can be reached, with a recall higher than 60% for the portion of the Brown corpus, and even higher for the Peter corpus.

We are able to identify phases in the generation of rules that turn out to be for English: a. initially inflectional morphology on verbs, with the plural "s" on nouns, and b. subsequently other types of morphemes. We believe that this phenomenon is purely driven by the frequency of these morphemes in the corpora. In the manually segmented portion of the Brown corpus we identified on the token level 11.3% inflectional morphemes, 6.4% derivational morphemes, and 82.1% stems. In average there are twice as many inflectional morphemes in the corpus, than derivational.

Given a very strict parameters, focusing on the description length of the grammar, our system would need long time till it would discover prefixes, not to mention infixes. By relaxing the weight of description length we can inhibit the

generation and identification of prefixing rules, however, to the cost of precision.

Given these results, the inflectional paradigms can be claimed to be extractable even with an incremental approach. As such, this means that central parts of the lexicon can be induced very early along the time line.

The existing signatures for each morpheme can be used as simple clustering criteria.[10] Clustering will separate dependent (affixes) from independent morphemes (stems). Their basic distinction is that affixes will usually have a long signature, i.e. many elements they co-occur with, as well as a high frequency, while for stems the opposite is true.[11] Along these lines, morphemes with a similar signature can be replaced by symbols, expressing the same type information and compressing the grammar further. This type information, especially for rare morphemes is essential in subsequent induction of syntactic structure. Due to space limitations, we cannot discuss in detail subsequent steps in the cross-level induction procedures. Nevertheless, the model presented here provides an important pointer to the mechanics of how grammatical parameters might come to be set.

Additionally, we provide a method by which to test the roles different statistical algorithms play in this process. By adjusting the weights of the contributions made by various constraints, we can approach an understanding of the optimal ordering of algorithms that play a role in the computational framework of language acquisition.

This is but a first step to what we hope will eventually finish a platform for a detailed study of various induction algorithms and evaluation metrics.

References

E. O. Batchelder. 1997. *Computational evidence for the use of frequency information in discovery of the infant's first lexicon*. PhD dissertation, CUNY.

E. O. Batchelder. 1998. Can a computer really model cognition? A case study of six computational models of infant word discovery. In M. A. Gernsbacher and S. J. Derry, editors, *Proceedings of the 20th Annual Conference of the Cognitive Science Society*, pages 120–125. Lawrence Erlbaum, University of Wisconsin-Madison.

L. Bloom, L. Hood, and P. Lightbown. 1974. Imitation in language development: If, when and why. *Cognitive Psychology,* 6, 380–420.

M.R. Brent and T.A. Cartwright. 1996. Distributional regularity and phonotactic constraints are useful for segmentation. Cognition 61: 93-125.

H. Déjean. 1998. *Concepts et algorithmes pour la découverte des structures formelles des langues*. Doctoral dissertation, Université de Caen Basse Normandie.

K. Elghamry. 2004. *A generalized cue-based approach to the automatic acquisition of subcategorization frames*. Doctoral dissertation, Indiana University.

K. Elghamry and D. Ćavar. 2004. *Bootstrapping cues for cue-based bootstrapping*. Mscr. Indiana University.

J. Fodor and V. Teller. 2000. Decoding syntactic parameters: The superparser as oracle. Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society, 136-141.

J. Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics* 27(2): 153-198.

Z.S. Harris. 1961. *Structural linguistics*. University of Chicago Press. Chicago.

J. Grimshaw. 1981. Form, function, and the language acquisition device. In C.L. Baker and J.J. McCarthy (eds.), *The Logical Problem of Language Acquisition*. Cambridge, MA: MIT Press.

D.J.C. MacKay. 2003. *Information Theory, Inference, and Learning Algorithms*. Cambridge: Cambridge University Press.

C.G. de Marcken. 1996. *Unsupervised Language Acquisition*. Phd dissertation, MIT.

G.F. Marcus, S. Vijayan, S. Bandi Rao, and P.M. Vishton. 1999. Rule-learning in seven-month-old infants. *Science* 283:77-80.

R. Mazuka. 1998. The Development of Language Processing Strategies: A cross-linguistic study between Japanese and English. Lawrence Erlbaum.

T.H. Mintz. 1996. *The roles of linguistic input and innate mechanisms in children's acquisition of grammatical categories.* Unpublished doctoral dissertation, University of Rochester.

S. Pinker. 1984. Language Learnability and Language Development, Harvard University Press, Cambridge, MA.

S. Pinker. 1994. *The language instinct*. New York, NY: W. Morrow and Co.

P. Schone and D. Jurafsky. 2001. *Knowledge-Free Induction of Inflectional Morphologies*. In Proceedings of NAACL-2001. Pittsburgh, PA, June 2001.

M.M. Van Zaanen and Pieter Adriaans. 2001. Comparing two unsupervised grammar induction systems: Alignment-based learning vs. EMILE. Tech. Rep. TR2001.05, University of Leeds.

M.M. Van Zaanen. 2001. Bootstrapping Structure into Language: Alignment-Based Learning. Doctoral dissertation, The University of Leeds.

[10] Length of the signature and frequency of each morpheme are mapped on a feature vector.

[11] This way, similar to the clustering of words into open and closed class on the basis of feature vectors, as described in Elghamry and Ćavar (2004), the morphemes can be separated into open and closed class.

Putting Meaning into Grammar Learning

Nancy Chang
UC Berkeley, Dept. of Computer Science and
International Computer Science Institute
1947 Center St., Suite 600
Berkeley, CA 94704 USA
nchang@icsi.berkeley.edu

Abstract

This paper proposes a formulation of grammar learning in which meaning plays a fundamental role. We present a computational model that aims to satisfy convergent constraints from cognitive linguistics and crosslinguistic developmental evidence within a statistically driven framework. The target grammar, input data and goal of learning are all designed to allow a tight coupling between language learning and comprehension that drives the acquisition of new constructions. The model is applied to learn lexically specific multi-word constructions from annotated child-directed transcript data.

1 Introduction

What role does meaning play in the acquisition of grammar? Computational approaches to grammar learning have tended to exclude semantic information entirely, or else relegate it to lexical representations. Starting with Gold's (1967) influential early work on language identifiability in the limit and continuing with work in the formalist learnability paradigm, grammar learning has been equated with syntax learning, with the target of learning consisting of relatively abstract structures that govern the combination of symbolic linguistic units. Statistical, corpus-based efforts have likewise restricted their attention to inducing syntactic patterns, though in part due to more practical considerations, such as the lack of large-scale semantically tagged corpora.

But a variety of cognitive, linguistic and developmental considerations suggest that meaning plays a central role in the acquisition of linguistic units at all levels. We start with the proposition that language *use* should drive language learning — that is, the learner's goal is to improve its ability to communicate, via comprehension and production. Cognitive and constructional approaches to grammar assume that the basic unit of linguistic knowledge needed to support language use consists of pairings of form and meaning, or **constructions** (Langacker, 1987; Goldberg, 1995; Fillmore and Kay, 1999). Moreover, by the time children make the leap from single words to complex combinations, they have amassed considerable conceptual knowledge, including familiarity with a wide variety of entities and events and sophisticated pragmatic skills (such as using joint attention to infer communicative intentions (Tomasello, 1995) and subtle lexical distinctions (Bloom, 2000)). The developmental evidence thus suggests that the input to grammar learning may in principle include not just surface strings but also meaningful situation descriptions with rich semantic and pragmatic information.

This paper formalizes the grammar learning problem in line with the observations above, taking seriously the ideas that the target of learning, for both lexical items and larger phrasal and clausal units, is a bipolar structure in which meaning is on par with form, and that meaningful language use drives language learning. The resulting core computational problem can be seen as a restricted type of relational learning. In particular, a key step of the learning task can be cast as learning *relational correspondences*, that is, associations between form relations (typically word order) and meaning relations (typically role-filler bindings). Such correlations are essential for capturing complex multi-unit constructions, both lexically specific constructions and more general grammatical constructions.

The remainder of the paper is structured as follows. Section 2 states the learning task and provides an overview of the model and its assumptions. We then present algorithms for inducing structured mappings, based on either specific input examples or the current set of constructions (Section 3), and describe how these are evaluated using criteria based on minimum description length (Rissanen, 1978). Initial results from applying the learning algorithms to a small corpus of child-directed utterances demonstrate the viability of the approach (Section 4). We conclude with a discussion of the broader implications of this approach for language learning and use.

2 Overview of the learning problem

We begin with an informal description of our learning task, to be formalized below. At all stages of language learning, children are assumed to exploit general cognitive abilities to make sense of the flow of objects and events they experience. To make sense of linguistic events — sounds and gestures used in their environments for communicative purposes — they also draw on specifically linguistic knowledge of how forms map to meanings, i.e., constructions. Comprehension consists of two stages: identifying the constructions involved and how their meanings are related (**analysis**), and matching these constructionally sanctioned meanings to the actual participants and relations present in context (**resolution**). The set of linguistic constructions will typically provide only a *partial* analysis of the utterance in the given context; when this happens, the agent may still draw on general inference to match even a partial analysis to the context.

The goal of construction learning is to acquire a *useful* set of constructions, or **grammar**. This grammar should allow constructional analysis to produce increasingly complete interpretations of utterances in context, thus requiring minimal recourse to general resolution and inference procedures. In the limit the grammar should stabilize, while still being useful for comprehending novel input. A useful grammar should also reflect the statistical properties of the input data, in that more frequent or specific constructions should be learned before more infrequent and more general constructions.

Formally, we define our learning task as follows: Given an initial grammar G and a sequence of training examples consisting of an utterance paired with its context, find the best grammar G' to fit seen data and generalize to new data. The remainder of this section describes the hypothesis space, prior knowledge and input data relevant to the task.

2.1 Hypothesis space: embodied constructions

The space of possible grammars (or sets of constructions) is defined by Embodied Construction Grammar (ECG), a computationally explicit unification-based formalism for capturing insights from the construction grammar and cognitive linguistics literature (Bergen and Chang, in press; Chang et al., 2002). ECG is designed to support the analysis process mentioned above, which determines what constructions and schematic meanings are present in an utterance, resulting in a *semantic specification* (or *semspec*).[1]

We highlight a few relevant aspects of the formalism, exemplified in Figure 1. Each construction has sections labeled **form** and **meaning** listing the entities (or roles) and constraints (type constraints marked with :, filler constraints marked with $\longleftarrow$, and identification (or coindexation) constraints marked with $\longleftrightarrow$) of the respective domains. These two sections, also called the form and meaning **poles**, capture the basic intuition that constructions are form-meaning pairs. A subscripted f or m allows reference to the form or meaning pole of any construction, and the keyword **self** allows self-reference. Thus, the THROW construction simply links a form whose orthography role (or feature) is bound to the string "throw" to a meaning that is constrained to be of type Throw, a separately defined conceptual schema corresponding to throwing events (including roles for a thrower and throwee). (Although not shown in the examples, the formalism also includes a **subcase of** notation for expressing constructional inheritance.)

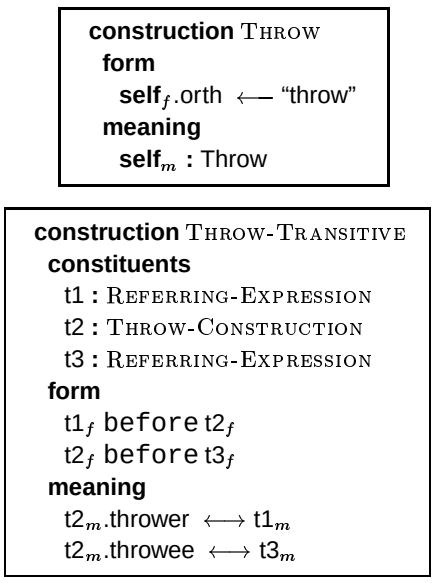


Figure 1: Embodied Construction Grammar representation of the lexical THROW and lexically specific THROW-TRANSITIVE construction (licensing expressions like *You throw the ball*).

Multi-unit constructions such as the THROW-TRANSITIVE construction also list their **constituents**, each of which is itself a form-meaning construction. These multi-unit constructions serve as the target representation for the specific learning task at hand. The key representational insight here is that the form and meaning constraints typi-

[1] ECG is intended to support a simulation-based model of language understanding, with the semspec parameterizing a simulation using active representations (or *embodied schemas*) to produce context-sensitive inferences. See Bergen and Chang (in press) for details.

cally involve *relations* among the form and meaning poles of the constructional constituents. For current purposes we limit the potential form relations to word order, although many other form relations are in principle allowed. In the meaning domain, the primary relation is *identification*, or unification, between two meaning entities. In particular, we will focus on role-filler bindings, in which a role of one constituent is identified with another constituent or with one of its roles. The example construction pairs two word order constraints over its constituents' form poles with two identification constraints over its constituents' meaning poles (these specify the fillers of the thrower and throwee roles of a Throw event, respectively).

Note that both lexical constructions and the multi-unit constructions needed to express grammatical patterns can be seen as graphs of varying complexity. Each domain (form or meaning) can be represented as a subgraph of elements and relations among them. Lexical constructions involve a simple mapping between these two subgraphs, whereas complex constructions with constituents require *structured relational mappings* over the two domains, that is, mappings between form and meaning relations whose arguments are themselves linked by known constructions.

2.2 Prior knowledge

The model makes a number of assumptions based on the child language literature about prior knowledge brought to the task, including conceptual knowledge, lexical knowledge and the language comprehension process described earlier. Figure 2 depicts how these are related in a simple example; each is described in more detail below.

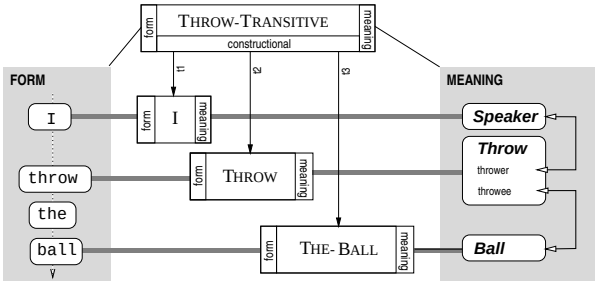


Figure 2: A constructional analysis of *I throw the ball*, with form elements on the left, meaning elements (conceptual schemas) on the right and constructions linking the two domains in the center.

2.2.1 Conceptual knowledge

Conceptual knowledge is represented using an ontology of typed feature structures, or **schemas**. These include schemas for people, objects (e.g. Ball in the figure), locations, and actions familiar to children by the time they enter the two-word stage (typically toward the end of the second year). Actions like the Throw schema referred to in the example THROW construction and in the figure have roles whose fillers are subject to type constraints, reflecting children's knowledge of what kinds of entities can take place in different events.

2.2.2 Lexical constructions

The input to learning includes a set of lexical constructions, represented using the ECG formalism, linking simple forms (i.e. words) to specific conceptual items. Examples of these include the I and BALL constructions in the figure, as well as the THROW construction formally defined in Figure 1. Lexical learning is not the focus of the current work, but a number of previous computational approaches have shown how simple mappings may be acquired from experience (Regier, 1996; Bailey, 1997; Roy and Pentland, 1998).

2.2.3 Construction analyzer

As mentioned earlier, the ECG construction formalism is designed to support processes of language use. In particular, the model makes use of a construction analyzer that identifies the constructions responsible for a given utterance, much like a syntactic parser in a traditional language understanding system identifies which parse rules are responsible. In this case, however, the basic representational unit is a form-meaning pair. The analyzer must therefore also supply a semantic interpretation, called the semspec, indicating which conceptual schemas are involved and how they are related. The analyzer is also required to be robust to input that is not covered by its current grammar, since that situation is the norm during language learning.

Bryant (2003) describes an implemented construction analyzer program that meets these needs. The construction analyzer takes as input a set of ECG constructions (linguistic knowledge), a set of ECG schemas (conceptual knowledge) and an utterance. The analyzer draws on partial parsing techniques previously applied to syntactic parsing (Abney, 1996): utterances not covered by known constructions yield partially filled semspecs, and unknown forms in the input are skipped. As a result, even a small set of simple constructions can provide skeletal interpretations of complex utterances.

Figure 2 gives an iconic representation of the result of analyzing the utterance *I throw the ball* using the THROW-TRANSITIVE and THROW constructions shown earlier, along with some additional

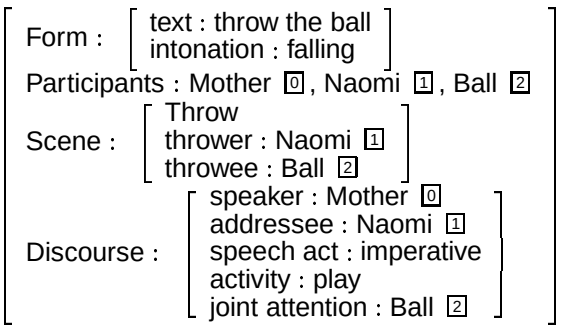
lexical constructions (not shown). The analyzer matches each input form with its lexical construction (if available) and corresponding meaning, and then matches the clausal construction by checking the relevant word order relations (implicitly represented by the dotted arrow in the figure) and role bindings (denoted by the double-headed arrows within the meaning domain) asserted on its candidate constituents. Note that at the stage shown, no construction for *the* has yet been learned, resulting in a partial analysis. At an even earlier stage of learning, before the THROW-TRANSITIVE construction is learned, the lexical constructions are matched without resulting in the role-filler bindings on the Throw action schema.

Finally, note that the semspec produced by constructional analysis (right-hand side of the figure) must be matched to the current situational context using a contextual interpretation, or resolution, process. Like other resolution (e.g. reference resolution) procedures, this process relies on category/type constraints and (provisional) identification bindings. The resolution procedure attempts to unify each schema and constraint appearing in the semspec with a type-compatible entity or relation in the context. In the example, the schemas on the right-hand side of the figure should be identified during resolution with particular schema instances available in context (e.g., the Speaker schema should be linked to the specific contextually available discourse speaker, the Ball schema to a particular ball instance, etc.).

2.3 Input data

The input is characterized as a set of **input tokens**, each consisting of an utterance form (a string of known and novel word-forms) paired with a specific communicative context (a set of linked conceptual schemas corresponding to the participants, salient scene and discourse information available in the situation). The learning model receives only positive examples, as in the child learning case. Note, however, that the interpretation a given utterance has in context depends on the current state of linguistic knowledge. Thus the same utterance at different stages may lead to different learning behavior.

The specific training corpus used in learning experiments is a subset of the Sachs corpus of the CHILDES database of parent-child transcripts(Sachs, 1983; MacWhinney, 1991), with additional annotations made by developmental psychologists as part of a study of motion utterances (Dan I. Slobin, p.c.). These annotations indicate semantic and pragmatic features available in the scene. A simple feature structure representation of a sample input token is shown here; boxed numbers indicate that the relevant entities are identified:



Many details have been omitted, and a number of simplifying assumptions have been made. But the rough outline given here nevertheless captures the core computational problem faced by the child learner in acquiring multi-word constructions in a framework putting meaning on par with form.

3 Learning algorithms

We model the learning task as a search through the space of possible grammars, with new constructions incrementally added based on encountered data. As in the child learning situation, the goal of learning is to converge on an optimal set of constructions, i.e., a grammar that is both general enough to encompass significant novel data and specific enough to accurately predict previously seen data.

A suitable overarching computational framework for guiding the search is provided by the minimum description length (MDL) heuristic (Rissanen, 1978), which is used to find the optimal analysis of data in terms of (a) a compact representation of the data (i.e., a grammar); and (b) a compact means of describing the original data in terms of the compressed representation (i.e., constructional analyses using the grammar). The MDL heuristic exploits a tradeoff between competing preferences for smaller grammars (encouraging generalization) and for simpler analyses of the data (encouraging the retention of specific/frequent constructions).

The rest of this section makes the learning framework concrete. Section 3.1 describes several heuristics for moving through the space of grammars (i.e., how to update a grammar with new constructions based on input data), and Section 3.2 describes how to chose among these candidate moves to find optimal points in the search space (i.e., specific MDL criteria for evaluating new grammars). These specifications extend previous methods to accommodate the relational structures of the ECG formalism and the process-based assumptions of the model.

3.1 Updating the grammar

The grammar may be updated in three ways:

hypothesis forming new structured maps to account for mappings present in the input but unexplained by the current grammar;

reorganization exploiting regularities in the set of known constructions (merge two similar constructions into a more general construction, or compose two constructions that cooccur into a larger construction); and

reinforcement incrementing the weight associated with constructions that are successfully used during comprehension.

Hypothesis. The first operation addresses the core computational challenge of learning new structured maps. The key idea here is that the learner is assumed to have access to a partial analysis based on linguistic knowledge, as well as a fuller situation interpretation it can infer from context. Any difference between the two can directly prompt the formation of new constructions that will improve the agent's ability to handle subsequent instances of similar utterances in similar contexts. In particular, certain form and meaning relations that are unmatched by the analysis but present in context may be mapped using the procedure in Figure 3.

Hypothesize construction. Given utterance U in situational context S and current grammar G:

1. Call the construction analysis/resolution processes on (U,S,G) to produce a semspec consisting of form and meaning graphs F and M. Nodes and edges of F and M are marked as matched or unmatched by the analysis.
2. Find $\text{rel}_f(\text{A}_f,\text{B}_f)$, an unmatched edge in F corresponding to an unused form relation over the matched form poles of two constructs A and B.
3. Find $\text{rel}_m(\text{A}_m, \text{B}_m)$, an unmatched edge (or subgraph) in M corresponding to an unused meaning relation (or set of bindings) over the corresponding matched meaning poles A_m and B_m. $\text{rel}_m(\text{A}_m,\text{B}_m)$ is required to be *pseudo-isomorphic* to $\text{rel}_f(\text{A}_f,\text{B}_f)$.
4. Create a new construction C with constituents A and B and form and meaning constraints corresponding to $\text{rel}_f(\text{A}_f,\text{B}_f)$ and $\text{rel}_m(\text{A}_m,\text{B}_m)$, respectively.

Figure 3: Construction hypothesis.

The algorithm creates new constructions mapping form and meaning relations whose arguments are already constructionally mapped. It is best illustrated by example, based on the sample input token shown in Section 2.3 and depicted schematically in Figure 4. Given the utterance "throw the ball" and a grammar including constructions for *throw* and *ball* (but not *the*), the analyzer produces a semspec including a Ball schema and a Throw schema, without indicating any relations between them. The resolution process matches these schemas to the actual context, which includes a particular throwing event in which the addressee (Naomi) is the thrower of a particular ball. The resulting resolved analysis looks like Figure 4 but without the new construction (marked with dashed lines): the two lexical constructions are shown mapping to particular utterance forms and contextual items.

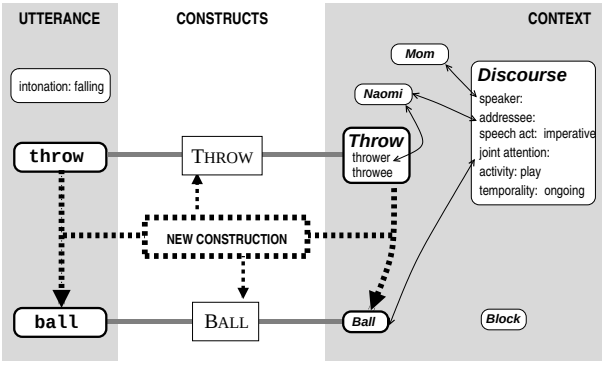


Figure 4: Hypothesizing a relational mapping for the utterance *throw ball*. Heavy solid lines indicate structures matched during analysis; heavy dotted lines indicate the newly hypothesized mapping.

Next, an unmatched form relation (the word order edge between *throw* and *ball*) is found, followed by a corresponding unmatched meaning relation (the binding between the Throw.throwee role and the specific Ball in context); these are shown in the figure using heavy dashed lines. Crucially, these relations meet the condition in step 3 that the relations be *pseudo-isomorphic*. This condition captures three common patterns of relational form-meaning mappings, i.e., ways in which a meaning relation rel_m over A_m and B_m can be correlated with a form relation rel_f over A_f and B_f (e.g., word order); these are illustrated in Figure 5, where we assume a simple form relation:

(a) strictly isomorphic: B_m is a role-filler of A_m (or vice versa) ($\text{A}_m.\text{r1} \longleftrightarrow \text{B}_m$)

(b) shared role-filler: A_m and B_m each have a role filled by the same entity ($\text{A}_m.\text{r1} \longleftrightarrow \text{B}_m.\text{r2}$)

(c) sibling role-fillers: A_m and B_m fill roles of the same schema ($\text{Y.r1} \longleftrightarrow \text{A}_m$, $\text{Y.r2} \longleftrightarrow \text{B}_m$)

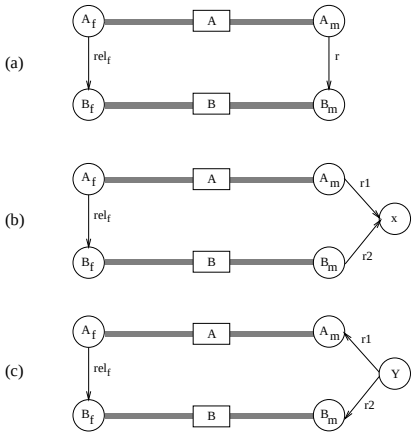


Figure 5: Pseudo-isomorphic relational mappings over constructs A and B: (a) strictly isomorphic; (b) shared role-filler; and (c) sibling role-fillers.

This condition enforces structural similarity between the two relations while recognizing that constructions may involve relations that are not strictly isomorphic. (The example mapping shown in the figure is strictly isomorphic.) The resulting construction is shown formally in Figure 6.

construction THROW-BALL
constituents
t1 : THROW
t2 : BALL
form
$t1_f$ `before` $t2_f$
meaning
$t1_m$.throwee $\longleftrightarrow$ $t2_m$

Figure 6: Example learned construction.

Reorganization. Besides hypothesizing constructions based on new data, the model also allows new constructions to be formed via constructional reorganization, essentially by applying general categorization principles to the current grammar, as described in Figure 7.

For example, the THROW-BALL construction and a similar THROW-BLOCK construction can be merged into a general THROW-OBJECT construction; the resulting subcase constructions each retain the appropriate type constraint. Similarly, a general HUMAN-THROW and THROW-OBJECT construction may occur in many analyses in which they compete for the THROW constituent. Since they have compatible constraints in both form and meaning (in the latter case based on the same conceptual Throw schema), repeated co-occurrence may lead to the formation of a larger construction that includes all three constituents.

Reorganize constructions. Reorganize G to consolidate similar and co-occurring constructions:

- **Merge:** Pairs of constructions with significant shared structure (same number of constituents, minimal ontological distance (i.e., distance in the type ontology) between corresponding constituents, maximal overlap in constraints) may be merged into a new construction containing the shared structure; the original constructions are rewritten as subcases of the new construction along with the non-overlapping information.
- **Compose:** Pairs of constructions that co-occur frequently with compatible constraints (are part of competing analyses using the same constituent, or appear in a constituency relationship) may be composed into one construction.

Figure 7: Construction reorganization.

Reinforcement. Each construction is associated with a weight, which is incremented each time it is used in an analysis that is successfully matched to the context. A successful match covers a majority of the contextually available bindings.

Both hypothesis and reorganization provide means of proposing new constructions; we now specify how proposed constructions are evaluated.

3.2 Evaluating grammar cost

The MDL criteria used in the model is based on the *cost* of the grammar G given the data D:

$$\begin{aligned}
\text{cost}(G|D) &= m \cdot \text{size}(G) + n \cdot \text{cost}(D|G) \\
\text{size}(G) &= \sum_{c \in G} \text{size}(c) \\
\text{size}(c) &= n_c + r_c + \sum_{e \in c} \text{length}(e) \\
\text{cost}(D|G) &= \sum_{d \in D} \text{score}(d) \\
\text{score}(d) &= \sum_{x \in d} (\text{weight}_x + p \cdot \sum_{r \in x} |\text{type}_r|) \\
&\quad + \text{height}_d + \text{semfit}_d
\end{aligned}$$

where m and n are learning parameters that control the relative bias toward model simplicity and data compactness. The size(G) is the sum over the size of each construction c in the grammar (n_c is the number of constituents in c, r_c is the number of constraints in c, and each element reference e in c has a length, measured as slot chain length). The cost (complexity) of the data D given G is the sum of the analysis scores of each input token d using G. This score sums over the constructions

x in the analysis of d, where weight_x reflects relative (in)frequency, $|\text{type}_r|$ denotes the number of ontology items of type r, summed over all the constituents in the analysis and discounted by parameter p. The score also includes terms for the height of the derivation graph and the semantic fit provided by the analyzer as a measure of semantic coherence.

In sum, these criteria favor constructions that are simply described (relative to the available meaning representations and the current set of constructions), frequently useful in analysis, and specific to the data encountered. The MDL criteria thus approximate Bayesian learning, where the minimizing of cost corresponds to maximizing the posterior probability, the structural prior corresponds to the grammar size, and likelihood corresponds to the complexity of the data relative to the grammar.

4 Learning verb islands

The model was applied to the data set described in Section 2.3 to determine whether lexically specific multi-word constructions could be learned using the MDL learning framework described. This task represents an important first step toward general grammatical constructions, and is of cognitive interest, since item-based patterns appear to be learned on independent trajectories (i.e., each verb forms its own "island" of organization (Tomasello, 2003)). We give results for *drop* (n=10 examples), *throw* (n=25), and *fall* (n=50).

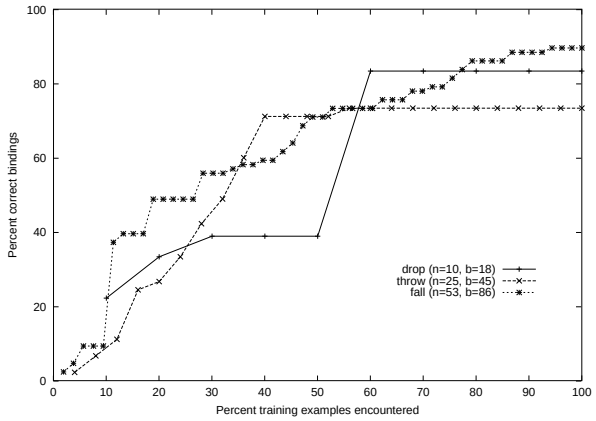


Figure 8: Incrementally improving comprehension for three verb islands.

Given the small corpus sizes, the focus for this experiment is not on the details of the statistical learning framework but instead on a qualitative evaluation of whether learned constructions improve the model's comprehension over time, and how verbs may differ in their learning trajectories. Qualitatively, the model first learned item-specific constructions as expected (e.g. *throw bear*, *throw books*, *you throw*), later in learning generalizing over different event participants (*throw OBJECT*, *PERSON throw*, etc.).

A quantitative measure of comprehension over time, **coverage**, was defined as the percentage of total bindings b in the data accounted for at each learning step. This metric indicates how new constructions incrementally improve the model's comprehensive capacity, shown in Figure 8. The *throw* subset, for example, contains 45 bindings to the roles of the Throw schema (thrower, throwee, and goal location). At the start of learning, the model has no combinatorial constructions and can account for none of these. But the model gradually amasses constructions with greater coverage, and by the tenth input token, the model learns new constructions that account for the majority of the bindings in the data.

The learning trajectories do appear distinct: *throw* constructions show a gradual build-up before plateauing, while *fall* has a more fitful climb converging at a higher coverage rate than *throw*. It is interesting to note that the *throw* subset has a much higher percentage of imperative utterances than *fall* (since throwing is pragmatically more likely to be done on command); the learning strategy used in the current model focuses on relational mappings and misses the association of an imperative speech-act with the lack of an expressed agent, providing a possible explanation for the different trajectories.

While further experimentation with larger training sets is needed, the results indicate that the model is able to acquire useful item-based constructions like those learned by children from a small number examples. More importantly, the learned constructions permit a limited degree of generalization that allows for increasingly complete coverage (or comprehension) of new utterances, fulfilling the goal of the learning model. Differences in verb learning lend support to the verb island hypothesis and illustrate how the particular semantic, pragmatic and statistical properties of different verbs can affect their learning course.

5 Discussion and future directions

The work presented here is intended to offer an alternative formulation of the grammar learning problem in which meaning in context plays a pivotal role in the acquisition of grammar. Specifically, meaning is incorporated directly into the target grammar (via the construction representation), the input data (via the context representation) and the evaluation criteria (which is usage-based, i.e. to improve comprehension). To the extent possible, the assump-

tions made with respect to structures and processes available to a human language learner in this stage are consistent with evidence from across the cognitive spectrum. Though only preliminary conclusions can be made, the model is a concrete computational step toward validating a meaning-oriented approach to grammar learning.

The model draws from a number of computational forerunners from both logical and probabilistic traditions, including Bayesian models of word learning (Bailey, 1997; Stolcke, 1994) for the overall optimization model, and work by Wolff (1982) modeling language acquisition (primarily production rules) using data compression techniques similar to the MDL approach taken here. The use of the results of analysis to hypothesize new mappings can be seen as related to both explanation-based learning (DeJong and Mooney, 1986) and inductive logic programming (Muggleton and Raedt, 1994). The model also has some precedents in the work of Siskind (1997) and Thompson (1998), both of which based learning on the discovery of isomorphic structures in syntactic and semantic representations, though in less linguistically rich formalisms.

In current work we are applying the model to the full corpus of English verbs, as well as crosslinguistic data including Russian case markers and Mandarin directional particles and aspect markers. These experiments will further test the robustness of the model's theoretical assumptions and protect against model overfitting and typological bias. We are also developing alternative means of evaluating the system's progress based on a rudimentary model of production, which would enable it to label scene descriptions using its current grammar and thus facilitate detailed studies of how the system generalizes (and overgeneralizes) to unseen data.

Acknowledgments

We thank members of the ICSI Neural Theory of Language group and two anonymous reviewers.

References

Steven Abney. 1996. Partial parsing via finite-state cascades. In *Workshop on Robust Parsing, 8th European Summer School in Logic, Language and Information*, pages 8–15, Prague, Czech Republic.

David R. Bailey. 1997. *When Push Comes to Shove: A Computational Model of the Role of Motor Control in the Acquisition of Action Verbs*. Ph.D. thesis, University of California at Berkeley.

Benjamin K. Bergen and Nancy Chang. in press. Simulation-based language understanding in Embodied Construction Grammar. In *Construction Grammar(s): Cognitive and Cross-language dimensions*. John Benjamins.

Paul Bloom. 2000. *How Children Learn the Meanings of Words*. MIT Press, Cambridge, MA.

John Bryant. 2003. Constructional analysis. Master's thesis, University of California at Berkeley.

Nancy Chang, Jerome Feldman, Robert Porzel, and Keith Sanders. 2002. Scaling cognitive linguistics: Formalisms for language understanding. In *Proc. 1st International Workshop on Scalable Natural Language Understanding*, Heidelberg, Germany.

G.F. DeJong and R. Mooney. 1986. Explanation-based learning: An alternative view. *Machine Learning*, 1(2):145–176.

Charles Fillmore and Paul Kay. 1999. *Construction grammar*. CSLI, Stanford, CA. To appear.

E.M. Gold. 1967. Language identification in the limit. *Information and Control*, 16:447–474.

Adele E. Goldberg. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press.

Ronald W. Langacker. 1987. *Foundations of Cognitive Grammar, Vol. 1*. Stanford University Press.

Brian MacWhinney. 1991. *The CHILDES project: Tools for analyzing talk*. Erlbaum, Hillsdale, NJ.

Stephen Muggleton and Luc De Raedt. 1994. Inductive logic programming: Theory and methods. *Journal of Logic Programming*, 19/20:629–679.

Terry Regier. 1996. *The Human Semantic Potential*. MIT Press, Cambridge, MA.

Jorma Rissanen. 1978. Modeling by shortest data description. *Automatica*, 14:465–471.

Deb Roy and Alex Pentland. 1998. Learning audio-visually grounded words from natural input. In *Proc. AAAI workshop, Grounding Word Meaning*.

J. Sachs. 1983. Talking about the there and then: the emergence of displaced reference in parent-child discourse. In K.E. Nelson, editor, *Children's language*, volume 4, pages 1–28. Lawrence Erlbaum Associates.

Jeffrey Mark Siskind, 1997. *A computational study of cross-situational techniques for learning word-to-meaning mappings*, chapter 2. MIT Press.

Andreas Stolcke. 1994. *Bayesian Learning of Probabilistic Language Models*. Ph.D. thesis, Computer Science Division, University of California at Berkeley.

Cynthia A. Thompson. 1998. *Semantic Lexicon Acquisition for Learning Natural Language Interfaces*. Ph.D. thesis, Department of Computer Sciences, University of Texas at Austin, December.

Michael Tomasello. 1995. Joint attention as social cognition. In C.D. Moore P., editor, *Joint attention: Its origins and role in development*. Lawrence Erlbaum Associates, Hillsdale, NJ. Educ/Psych BF720.A85.J65 1995.

Michael Tomasello. 2003. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press, Cambridge, MA.

J. Gerard Wolff. 1982. Language acquisition, data compression and generalization. *Language & Communication*, 2(1):57–89.

Grammatical Inference and First Language Acquisition

Alexander Clark (asc@aclark.demon.co.uk)
ISSCO / TIM, University of Geneva
UNI-MAIL, Boulevard du Pont-d'Arve,
CH-1211 Genève 4, Switzerland

Abstract

One argument for parametric models of language has been learnability in the context of first language acquisition. The claim is made that "logical" arguments from learnability theory require non-trivial constraints on the class of languages. Initial formalisations of the problem (Gold, 1967) are however inapplicable to this particular situation. In this paper we construct an appropriate formalisation of the problem using a modern vocabulary drawn from statistical learning theory and grammatical inference and looking in detail at the relevant empirical facts. We claim that a variant of the Probably Approximately Correct (PAC) learning framework (Valiant, 1984) with positive samples only, modified so it is not completely distribution free is the appropriate choice. Some negative results derived from cryptographic problems (Kearns et al., 1994) appear to apply in this situation but the existence of algorithms with provably good performance (Ron et al., 1995) and subsequent work, shows how these negative results are not as strong as they initially appear, and that recent algorithms for learning regular languages partially satisfy our criteria. We then discuss the applicability of these results to parametric and non-parametric models.

1 Introduction

For some years, the relevance of formal results in grammatical inference to the empirical question of first language acquisition by infant children has been recognised (Wexler and Culicover, 1980). Unfortunately, for many researchers, with a few notable exceptions (Abe, 1988), this begins and ends with Gold's famous negative results in the identification in the limit paradigm. This paradigm, though still widely used in the grammatical inference community, is clearly of limited relevance to the issue at hand, since it requires the model to be able to exactly identify the target language even when an adversary can pick arbitrarily misleading sequences of examples to provide. Moreover, the paradigm as stated has no bounds on the amount of data or computation required for the learner. In spite of the inapplicability of this particular paradigm, in a suitable analysis there are quite strong arguments that bear directly on this problem.

Grammatical inference is the study of machine learning of formal languages. It has a vast formal vocabulary and has been applied to a wide selection of different problems, where the "languages" under study can be (representations of) parts of natural languages, sequences of nucleotides, moves of a robot, or some other sequence data. For any conclusions that we draw from formal discussions to have any applicability to the real world, we must be sure to select, or construct, from the rich set of formal devices available an appropriate formalisation. Even then, we should be very cautious about making inferences about how the infant child must or cannot learn language: subsequent developments in GI might allow a more nuanced description in which these conclusions are not valid. The situation is complicated by the fact that the field of grammtical inference, much like the wider field of machine learning in general, is in a state of rapid change.

In this paper we hope to address this problem by justifying the selection of the appropriate learning framework starting by looking at the actual situation the child is in, rather than from an *a priori* decision about the right framework. We will not attempt a survey of grammatical inference techniques; nor shall we provide proofs of the theorems we use here. Arguments based on formal learnability have been used to support the idea of parameter based theories of language (Chomsky, 1986). As we shall see below, under our analysis of the problem these arguments are weak. Indeed, they are more pertinent to questions about the autonomy and modularity of language learning: the question whether learning of some level of linguistic knowledge – morphology or syntax, for example – can take place in isolation from other forms of learning, such as the acquisition of word meaning, and without interaction, grounding and so on.

Positive results can help us to understand how humans might learn languages by outlining the class of algorithms that *might* be used by humans, considered as computational systems at a suitable abstract level. Conversely, negative results might be helpful if they could demonstrate that no algorithms of a certain class could perform the task – in this case we could know that the human child learns his language in some other way.

We shall proceed as follows: after briefly describing FLA, we describe the various elements of a model of learning, or framework. We then make a series of decisions based on the empirical facts about FLA, to construct an appropriate model or models, avoiding unnecessary idealisation wherever possible. We proceed to some strong negative results, well-known in the GI community that bear on the questions at hand. The most powerful of these (Kearns et al., 1994) appears to apply quite directly to our chosen model. We then discuss an interesting algorithm (Ron et al., 1995) which shows that this can be circumvented, at least for a subclass of regular languages. Finally, after discussing the possibilities for extending this result to all regular languages, and beyond, we conclude with a discussion of the implications of the results presented for the distinction between parametric and non-parametric models.

2 First Language Acquisition

Let us first examine the phenomenon we are concerned with: first language acquisition. In the space of a few years, children almost invariably acquire, in the absence of explicit instruction, one or more of the languages that they are exposed to. A multitude of subsidiary debates have sprung up around this central issue covering questions about critical periods – the ages at which this can take place, the exact nature of the evidence available to the child, and the various phases of linguistic use through which the infant child passes. In the opinion of many researchers, explaining this ability is one of the most important challenges facing linguists and cognitive scientists today.

A difficulty for us in this paper is that many of the idealisations made in the study of this field are in fact demonstrably false. Classical assumptions, such as the existence of uniform communities of language users, are well-motivated in the study of the "steady state" of a system, but less so when studying acquisition and change. There is a regrettable tendency to slip from viewing these idealisations correctly – as counter-factual idealizations – to viewing them as empirical facts that need to be explained. Thus, when looking for an appropriate formulation of the problem, we should recall for example the fact that different children do not converge to exactly the same knowledge of language as is sometimes claimed, nor do all of them acquire a language competently at all, since there is a small proportion of children who though apparently neurologically normal fail to acquire language. In the context of our discussion later on, these observations lead us to accept slightly less stringent criteria where we allow a small probability of failure and do not demand perfect equality of hypothesis and target.

3 Grammatical Inference

The general field of machine learning has a specialised subfield that deals with the learning of formal languages. This field, Grammatical Inference (GI), is characterised above all by an interest in formal results, both in terms of formal characterisations of the target languages, and in terms of formal proofs either that particular algorithms can learn according to particular definitions, or that sets of language cannot be learnt. In spite of its theoretical bent, GI algorithms have also been applied with some success. Natural language, however is not the only source of real-world applications for GI. Other domains include biological sequence data, artificial languages, such as discovering XML schemas, or sequences of moves of a robot. The field is also driven by technical motives and the intrinsic elegance and interest of the mathematical ideas employed. In summary it is not just about language, and accordingly it has developed a rich vocabulary to deal with the wide range of its subject matter.

In particular, researchers are often concerned with formal results – that is we want algorithms where we can *prove* that they will perform in a certain way. Often, we may be able to empirically establish that a particular algorithm performs well, in the sense of reliably producing an accurate model, while we may be unable to prove formally that the algorithm will always perform in this way. This can be for a number of reasons: the mathematics required in the derivation of the bounds on the errors may be difficult or obscure, or the algorithm may behave strangely when dealing with sets of data which are ill-behaved in some way.

The basic framework can be considered as a game played between two players. One player, the teacher, provides information to another, the learner, and from that information the learner must identify the underlying language. We can break down this situation further into a number of elements. We assume that the languages to be learned are drawn

in some way from a possibly infinite class of languages, $\mathcal{L}$, which is a set of formal mathematical objects. The teacher selects one of these languages, which we call the *target*, and then gives the learner a certain amount of information of various types about the target. After a while, the learner then returns its guess, the hypothesis, which in general will be a language drawn from the same class $\mathcal{L}$. Ideally the learner has been able to deduce or induce or abduce something about the target from the information we have given it, and in this case the hypothesis it returns will be identical to, or close in some technical sense, to the target. If the learner can conistently do this, under whatever constraints we choose, then we say it can learn that class of languages. To turn this vague description into something more concrete requires us to specify a number of things.

- What sort of mathematical object should we use to represent a language?
- What is the target class of languages?
- What information is the learner given?
- What computational constraints does the learner operate under?
- How close must the target be to the hypothesis, and how do we measure it?

This paper addresses the extent to which negative results in GI could be relevant to this real world situation. As always, when negative results from theory are being applied, a certain amount of caution is appropriate in examining the underlying assumptions of the theory and the extent to which these are applicable. As we shall see, in our opinion, none of the current negative results, though powerful, are applicable to the empirical situation. We shall accordingly, at various points, make strong pessimistic assumptions about the learning environment of the child, and show that even under these unrealistically stringent stipulations, the negative results are still inapplicable. This will make the conclusions we come to a little sharper. Conversely, if we wanted to show that the negative results did apply, to be convincing we would have to make rather optimistic assumptions about the learning environment.

4 Applying GI to FLA

We now have the delicate task of selecting, or rather constructing, a formal model by identifying the various components we have identified above. We want to choose the model that is the best representation of the learning task or tasks that the infant child must perform. We consider that some of the empirical questions do not yet have clear answers. In those cases, we shall make the choice that makes the learning task more difficult. In other cases, we may not have a clear idea of how to formalise some information source. We shall start by making a significant idealisation: we consider language acquisition as being a single task. Natural languages as traditionally describe have different levels. At the very least we have morphology and syntax; one might also consider inter-sentential or discourse as an additional level. We conflate all of these into a single task: learning a formal language; in the discussion below, for the sake of concreteness and clarity, we shall talk in terms of learning syntax.

4.1 The Language

The first question we must answer concerns the language itself. A formal language is normally defined as follows. Given a finite alphabet Σ, we define the set of all strings (the free monoid) over Σ as $\Sigma*$. We want to learn a language $L \subset \Sigma^*$. The alphabet Σ could be a set of phonemes, or characters, or a set of words, or a set of lexical categories (part of speech tags). The language could be the set of well-formed sentences, or the set of words that obey the phonotactics of the language, and so on. We reduce all of the different learning tasks in language to a single abstract task – identifying a possibly infinite set of strings. This is overly simplistic since transductions, i.e. mappings from one string to another, are probably also necessary. We are using here a standard definition of a language where every string is unambiguously either in or not in the language.. This may appear unrealistic – if the formal language is meant to represent the set of grammatical sentences, there are well-known methodological problems with deciding where exactly to draw the line between grammatical and ungrammatical sentences. An alternative might be to consider acceptability rather than grammaticality as the defining criterion for inclusion in the set. Moreover, there is a certain amount of noise in the input – There are other possibilities. We could for example use a fuzzy set – i.e. a function from $\Sigma^* \rightarrow [0, 1]$ where each string has a degree of membership between 0 and 1. This would seem to create more problems than it solves. A more appealing option is to learn distributions, again functions f from $\Sigma^* \rightarrow [0, 1]$ but where $\sum_{s \in L} f(s) = 1$. This is of course the classic problem of language modelling, and is compelling for two reasons. First, it is empirically well grounded – the probability of a string is related to its frequency of occurrence, and secondly, we can de-

duce from the speech recognition capability of humans that they must have some similar capability.

Both possibilities – crisp languages, and distributions – are reasonable. The choice depends on what one considers the key phenomena to be explained are – grammaticality judgments by native speakers, or natural use and comprehension of the language. We favour the latter, and accordingly think that learning distributions is a more accurate and more difficult choice.

4.2 The class of languages

A common confusion in some discussions of this topic is between languages and classes of languages. Learnability is a property of *classes* of languages. If there is only one language in the class of languages to be learned then the learner can just guess that language and succeed. A class with two languages is again trivially learnable if you have an efficient algorithm for testing membership. It is only when the set of languages is exponentially large or infinite, that the problem becomes non-trivial, from a theoretical point of view. The class of languages we need is a class of languages that includes all attested human languages and additionally all "possible" human languages. Natural languages are thought to fall into the class of mildly context-sensitive languages, (Vijay-Shanker and Weir, 1994), so clearly this class is large enough. It is, however, not necessary that our class be this large. Indeed it is essential for learnability that it is not. As we shall see below, even the class of regular languages contains some subclasses that are computationally hard to learn. Indeed, we claim it is reasonable to define our class so it does *not* contain languages that are clearly not possible human languages.

4.3 Information sources

Next we must specify the information that our learning algorithm has access to. Clearly the primary source of data is the *primary linguistic data* (PLD), namely the utterances that occur in the child's environment. These will consist of both child-directed speech and adult-to-adult speech. These are generally acceptable sentences that is to say sentences that are in the language to be learned. These are called *positive* samples. One of the most long-running debates in this field is over whether the child has access to negative data – unacceptable sentences that are marked in some way as such. The consensus (Marcus, 1993) appears to be that they do not. In middle-class Western families, children are provided with some sort of feedback about the well-formedness of their utterances, but this is unreliable and erratic, not a universal of global child-raising. Furthermore this appears to have no effect on the child. Children do also get indirect pragmatic feedback if their utterances are incomprehensible. In our opinion, both of these would be better modelled by what is called a membership query: the algorithm may generate a string and be informed whether that string is in the language or not. However, we feel that this is too erratic to be considered an essential part of the process. Another question is whether the input data is presented as a flat string or annotated with some sort of structural evidence, which might be derived from prosodic or semantic information. Unfortunately there is little agreement on what the constituent structure should be – indeed many linguistic theories do not have a level of constituent structure at all, but just dependency structure.

Semantic information is also claimed as an important source. The hypothesis is that children can use lexical semantics, coupled with rich sources of real-world knowlege to infer the meaning of utterances from the situational context. That would be an extremely powerful piece of information, but it is clearly absurd to claim that the meaning of an utterance is uniquely specified by the situational context. If true, there would be no need for communication or information transfer at all. Of course the context puts some constraints on the sentences that will be uttered, but it is not clear how to incorporate this fact without being far too generous. In summary it appears that only positive evidence can be unequivocally relied upon though this may seem a harsh and unrealistic environment.

4.4 Presentation

We have now decided that the only evidence available to the learner will be unadorned positive samples drawn from the target language. There are various possibilities for how the samples are selected. The choice that is most favourable for the learner is where they are slected by a helpful teacher to make the learning process as easy as possible (Goldman and Mathias, 1996). While it is certainly true that carers speak to small children in sentences of simple structure (Motherese), this is not true for all of the data that the child has access to, nor is it universally valid. Moreover, there are serious technical problems with formalising this, namely what is called 'collusion' where the teacher provides examples that encode the grammar itself, thus trivialising the learning process. Though attempts have been made to limit this problem, they are not yet completely satisfactory. The next alternative is that the examples are selected randomly from some fixed

distribution. This appears to us to be the appropriate choice, subject to some limitations on the distributions that we discuss below. The final option, the most difficult for the learner, is where the sequence of samples can be selected by an intelligent adversary, in an attempt to make the learner fail, subject only to the weak requirement that each string in the language appears at least once. This is the approach taken in the identification in the limit paradigm (Gold, 1967), and is clearly too stringent. The remaining question then regards the distribution from which the samples are drawn: whether the learner has to be able to learn for every possible distribution, or only for distributions from a particular class, or only for one particular distribution.

4.5 Resources

Beyond the requirement of computability we will wish to place additional limitations on the computational resources that the learner can use. Since children learn the language in a limited period of time, which limits both the amount of data they have access to and the amount of computation they can use, it seems appropriate to disallow algorithms that use unbounded or very large amounts of data or time. As normal, we shall formalise this by putting polynomial bounds on the *sample complexity* and *computational complexity*. Since the individual samples are of varying length, we need to allow the computational complexity to depend on the total length of the sample. A key question is what the parameters of the sample complexity polynomial should be. We shall discuss this further below.

4.6 Convergence Criteria

Next we address the issue of reliability: the extent to which all children acquire language. First, variability in achievement of particular linguistic milestones is high. There are numerous causes including deafness, mental retardation, cerebral palsy, specific language impairment and autism. Generally, autistic children appear neurologically and physically normal, but about half may never speak. Autism, on some accounts, has an incidence of about 0.2%. Therefore we can require learning to happen with arbitrarily high probability, but requiring it to happen with probability one is unreasonable. A related question concerns convergence: the extent to which children exposed to a linguistic environment end up with the same language as others. Clearly they are very close since otherwise communication could not happen, but there is ample evidence from studies of variation (Labov, 1975), that there are nontrivial differences between adults, who have grown up with near-identical linguistic experiences, about the interpretation and syntactic acceptability of simple sentences, quite apart from the wide purely lexical variation that is easily detected. A famous example in English is “Each of the boys didn’t come”.

Moreover, language change *requires* some children to end up with slightly different grammars from the older generation. At the very most, we should require that the hypothesis should be close to the target. The function we use to measure the ’distance’ between hypothesis and target depends on whether we are learnng crisp languages or distributions. If we are learning distributions then the obvious choice is the Kullback-Leibler divergence – a very strict measure. For crisp languages, the probability of the symmetric difference with respect to some distribution is natural.

4.7 PAC-learning

These considerations lead us to some variant of the Probably Approximately Correct (PAC) model of learning (Valiant, 1984). We require the algorithm to produce with arbitrarily high probability a good hypothesis. We formalise this by saying that for any $\delta > 0$ it must produce a good hypothesis with probability more than $1 - \delta$. Next we require a good hypothesis to be arbitrarily close to the target, so we have a precision ϵ and we say that for any $\epsilon > 0$, the hypothesis must be less than ϵ away from the target. We allow the amount of data it can use to increase as the confidence and precision get smaller. We define PAC-learning in the following way: given a finite alphabet Σ, and a class of languages $\mathcal{L}$ over Σ, an algorithm PAC-learns the class $\mathcal{L}$, if there is a polynomial q, such that for every confidence $\delta > 0$ and precision $\epsilon > 0$, for every distribution D over Σ^*, for every language L in $\mathcal{L}$, whenever the number of samples exceeds $q(1/\epsilon, 1/\delta, |\Sigma|, |L|)$, the algorithm must produce a hypothesis H such that with probability greater than $1 - \delta$, $Pr_D(H \Delta L > \epsilon)$. Here we use $A \Delta B$ to mean the symmetric difference between two sets. The polynomial q is called the sample complexity polynomial. We also limit the amount of computation to some polynomial in the total length of the data it has seen. Note first of all that this is a worst case bound – we are not requiring merely that on average it comes close. Additionally this model is what is called ’distribution-free’. This means that the algorithm must work for every combination of distribution and language. This is a very stringent requirement, only mitigated by the fact that the error is calculated with respect to the same distribution that the samples are drawn from. Thus, if there is a subset of Σ^* with low aggregate probability under D, the algorithm will not get many sam-

ples from this region but will not be penalised very much for errors in that region. From our point of view, there are two problems with this framework: first, we only want to draw positive samples, but the distributions are over all strings in Σ^*, and include some that give a zero probability to all strings in the language concerned. Secondly, this is too pessimistic because the distribution has no relation to the language: intuitively it's reasonable to expect the distribution to be derived in some way from the language, or the structure of a grammar generating the language. Indeed there is a causal connection in reality since the sample of the language the child is exposed to is generated by people who do in fact know the language.

One alternative that has been suggested is the PAC learning with simple distributions model introduced by (Denis, 2001). This is based on ideas from complexity theory where the samples are drawn according to a universal distribution defined by the conditional Kolmogorov complexity. While mathematically correct this is inappropriate as a model of FLA for a number of reasons. First, learnability is proven only on a single very unusual distribution, and relies on particular properties of this distribution, and secondly there are some very large constants in the sample complexity polynomial.

The solution we favour is to define some natural class of distributions based on a grammar or automaton generating the language. Given a class of languages defined by some generative device, there is normally a natural stochastic variant of the device which defines a distribution over that language. Thus regular languages can be defined by a finite-state automaton, and these can be naturally extended to Probabilistic finite state automaton. Similarly context free languages are normally defined by context-free grammmars which can be extended again to to Probabilistic or stochastic CFG. We therefore propose a slight modification of the PAC-framework. For every class of languages $\mathcal{L}$, defined by some formal device define a class of distributions defined by a stochastic variant of that device. $\mathcal{D}$. Then for each language L, we select the set of distributions whose support is equal to the language and subject to a polynomial bound (q)on the complexity of the distribution in terms of the complexity of the target language: $D_L^+ = \{D \in \mathcal{D} : L = supp(D) \wedge |D| < q(|L|)\}$. Samples are drawn from one of these distributions.

There are two technical problems here: first, this doesn't penalise over-generalisation. Since the distribution is over positive examples, negative examples have zero weight, so we need some penalty function over negative examples or alternatively require the hypothesis to be a subset of the target. Secondly, this definition is too vague. The exact way in which you extend the "crisp" language to a stochastic one can have serious consequences. When dealing with regular languages, for example, though the class of languages defined by deterministic automata is the same as that defined by non-deterministic languages, the same is not true for their stochastic variants. Additionally, one can have exponential blow-ups in the number of states when determinising automata. Similarly, with CFGs, (Abney et al., 1999) showed that converting between two parametrisations of stochastic Context Free languages are equivalent but that there are blow-ups in both directions. We do not have a completely satisfactory solution to this problem at the moment; an alternative is to consider learning the distributions rather than the languages.

In the case of learning distributions, we have the same framework, but the samples are drawn according to the distribution being learned T, and we require that the hypothesis H has small divergence from the target: $D(T||H) < \epsilon$. Since the divergence is infinite if the hypothesis gives probability zero to a string in the target, this will have the consequence that the target must assign a non-zero probability to every string.

5 Negative Results

Now that we have a fairly clear idea of various ways of formalising the situation we can consider the extent to which formal results apply. We start by considering negative results, which in Machine Learning come in two types. First, there are information-theoretic bounds on sample complexity, derived from the Vapnik-Chervonenkis (VC) dimension of the space of languages, a measure of the complexity of the set of hypotheses. If we add a parameter to the sample complexity polynomial that represents the complexity of the concept to be learned then this will remove these problems. This can be the size of a representation of the target which will be a polynomial in the number of states, or simply the number of non-terminals or states. This is very standard in most fields of machine learning.

The second problem relates not to the amount of information but to the computation involved. Results derived from cryptographic limitations on computational complexity, can be proved based on widely held and well supported assumptions that certain hard cryptographic problems are insoluble. In what follows we assume that there are no efficient algorithms for common cryptographic prob-

lems such as factoring Blum integers, inverting RSA function, recognizing quadratic residues or learning noisy parity functions.

There may be algorithms that will learn with reasonable amounts of data but that require unfeasibly large amounts of computation to find. There are a number of powerful negative results on learning in the purely distribution-free situation we considered and rejected above. (Kearns and Valiant, 1989) showed that acyclic deterministic automata are not learnable even with positive and negative examples. Similarly, (Abe and Warmuth, 1992) showed a slightly weaker representation dependent result on learning with a large alphabet for non-deterministic automata, by showing that there are strings such that maximising the likelihood of the string is NP-hard. Again this does not strictly apply to the partially distribution free situation we have chosen.

However there is one very strong result that appears to apply. A straightforward consequence of (Kearns et al., 1994) shows that Acyclic Deterministic Probabilistic FSA over a two letter alphabet cannot be learned under another cryptographic assumption (the noisy parity assumption). Therefore any class of languages that includes this comparatively weak family will not be learnable in out framework.

But this rests upon the assumption that the class of possible human languages must include some cryptographically hard functions. It appears that our formal apparatus does not distinguish between these cryptographic functions which hav been consciously designed to be hard to learn, and natural languages which presumably have evolved to be easy to learn since there is no evolutionary pressure to make them hard to decrypt – no intelligent predators eavesdropping for example. Clearly this is a flaw in our analysis: we need to find some more nuanced description for the class of possible human languages that excludes these hard languages or distributions.

6 Positive results

There is a positive result that shows a way forward. A PDFA is μ-distinguishable the distributions generated from any two states differ by at least μ in the L_∞-norm, i.e. there is a string with a difference in probability of at least μ. (Ron et al., 1995) showed that μ-distinguishable acyclic PDFAs can be PAC-learned using the KLD as error function in time polynomial in $n, 1/\epsilon, 1/\delta, 1/\mu, |\Sigma|$. They use a variant of a standard state-merging algorithm. Since these are acyclic the languages they define are always finite. This additional criterion of distinguishability suffices to guarantee learnability. This work can be extended to cyclic automata (Clark and Thollard, 2004a; Clark and Thollard, 2004b), and thus the class of all regular languages, with the addition of a further parameter which bounds the expected length of a string generated from any state. The use of distinguishability seems innocuous; in syntactic terms it is a consequence of the plausible condition that for any pair of distinct non-terminals there is some fairly likely string generated by one and not the other. Similarly strings of symbols in natural language tend to have limited length. An alternate way of formalising this is to define a class of distinguishable automata, where the distinguishability of the automata is lower bounded by an inverse polynomial in the number of states. This is formally equivalent, but avoids adding terms to the sample complexity polynomial. In summary this would be a valid solution if all human languages actually lay within the class of regular languages. Note also the general properties of this kind of algorithm: provably learning an infinite class of languages with infinite support using only polynomial amounts of data and computation.

It is worth pointing out that the algorithm does not need to "know" the values of the parameters. Define a new parameter t, and set, for example $n = t, L = t, \delta = e^{-t}, \epsilon = t^{-1}$ and $\mu = t^{-1}$. This gives a sample complexity polynomial in one parameter $q(t)$. Given a certain amount of data N we can just choose the largest value of t such that $q(t) < N$, and set the parameters accordingly.

7 Parametric models

We can now examine the relevance of these results to the distinction between parametric and non-parametric languages. Parametric models are those where the class of languages is parametrised by a small set of finite-valued (binary) parameters, where the number of paameters is small compared to the $\log_2$ of the complexity of the languages. Without this latter constraint the notion is mathematically vacuous, since, for example, any context free grammar in Chomsky normal form can be parametrised with $N^3 + NM + 1$ binary parameters where N is the number of non-terminals and M the number of terminals. This constraint is also necessary for parametric models to make testable empirical predictions both about language universals, developmental evidence and relationships between the two (Hyams, 1986). We neglect here the important issue of lexical learning: we assume, implausibly, that lexical learning can take place completely before syntax learning commences. It has in the past been stated that the finiteness of a language class

suffices to guarantee learnability even under a PAC-learning criterion (Bertolo, 2001). This is, in general, false, and arises from neglecting constraints on the sample complexity and the computational complexities both of learning and of parsing. The negative result of (Kearns et al., 1994) discussed above applies also to parametric models. The specific class of noisy parity functions that they prove are unlearnable, are parametrised by a number of binary parameters in a way very reminiscent of a parametric model of language. The mere fact that there are a finite number of parameters does not suffice to guarantee learnability, if the resulting class of languages is exponentially large, or if there is no polynomial algorithm for parsing. This does not imply that *all* parametrised classes of languages will be unlearnable, only that having a small number of parameters is neither necessary nor sufficient to guarantee efficient learnability. If the parameters are shallow and relate to easily detectable properties of the languages and are independent then learning can occur efficiently (Yang, 2002). If they are "deep" and inter-related, learning may be impossible. Learnability depends more on simple statistical properties of the distributions of the samples than on the structure of the class of languages.

Our conclusion then is ultimately that the theory of learnability will not be able to resolve disputes about the nature of first language acquisition: these problems will have to be answered by empirical research, rather than by mathematical analysis.

Acknowledgements

This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778, funded in part by the Swiss Federal Office for Education and Science (OFES). This publication only reflects the authors' views.

References

N. Abe and M. K. Warmuth. 1992. On the computational complexity of approximating distributions by probabilistic automata. *Machine Learning*, 9:205–260.

N. Abe. 1988. Feasible learnability of formal grammars and the theory of natural language acquisition. In *Proceedings of COLING 1988*, pages 1–6.

S. Abney, D. McAllester, and F. Pereira. 1999. Relating probabilistic grammars and automata. In *Proceedings of ACL '99*.

Stefano Bertolo. 2001. A brief overview of learnability. In Stefano Bertolo, editor, *Language Acquisition and Learnability*. Cambridge University Press.

Noam Chomsky. 1986. *Knowledge of Language : Its Nature, Origin, and Use*. Praeger.

Alexander Clark and Franck Thollard. 2004a. PAC-learnability of probabilistic deterministic finite state automata. *Journal of Machine Learning Research*, 5:473–497, May.

Alexander Clark and Franck Thollard. 2004b. Partially distribution-free learning of regular languages from positive samples. In *Proceedings of COLING*, Geneva, Switzerland.

F. Denis. 2001. Learning regular languages from simple positive examples. *Machine Learning*, 44(1/2):37–66.

E. M. Gold. 1967. Language indentification in the limit. *Information and control*, 10(5):447 – 474.

S. A. Goldman and H. D. Mathias. 1996. Teaching a smarter learner. *Journal of Computer and System Sciences*, 52(2):255–267.

N. Hyams. 1986. *Language Acquisition and the Theory of Parameters*. D. Reidel.

M. Kearns and G. Valiant. 1989. Cryptographic limitations on learning boolean formulae and finite automata. In *21st annual ACM symposium on Theory of computation*, pages 433–444, New York. ACM, ACM.

M.J. Kearns, Y. Mansour, D. Ron, R. Rubinfeld, R.E. Schapire, and L. Sellie. 1994. On the learnability of discrete distributions. In *Proc. of the 25th Annual ACM Symposium on Theory of Computing*, pages 273–282.

W. Labov. 1975. Empirical foundations of linguistic theory. In R. Austerlitz, editor, *The Scope of American Linguistics*. Peter de Ridder Press.

G. F. Marcus. 1993. Negative evidence in language acquisition. *Cognition*, 46:53–85.

D. Ron, Y. Singer, and N. Tishby. 1995. On the learnability and usage of acyclic probabilistic finite automata. In *COLT 1995*, pages 31–40, Santa Cruz CA USA. ACM.

L. Valiant. 1984. A theory of the learnable. *Communications of the ACM*, 27(11):1134 – 1142.

K. Vijay-Shanker and David J. Weir. 1994. The equivalence of four extensions of context-free grammars. *Mathematical Systems Theory*, 27(6):511–546.

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

C. Yang. 2002. *Knowledge and Learning in Natural Language*. Oxford.

A Developmental Model of Syntax Acquisition in the Construction Grammar Framework with Cross-Linguistic Validation in English and Japanese

Peter Ford Dominey
Sequential Cognition and Language Group
Institut des Sciences Cognitives, CNRS
69675 Bron CEDES, France
dominey@isc.cnrs.fr

Toshio Inui
Graduate School of Informatics,
Kyoto University,
Yoshida-honmachi, Sakyo-ku, 606-8501,
Kyoto, Japan
inui@kyoto-u.ac.jp

Abstract

The current research demonstrates a system inspired by cognitive neuroscience and developmental psychology that learns to construct mappings between the grammatical structure of sentences and the structure of their meaning representations. Sentence to meaning mappings are learned and stored as grammatical constructions. These are stored and retrieved from a construction inventory based on the constellation of closed class items uniquely identifying each construction. These learned mappings allow the system to processes natural language sentences in order to reconstruct complex internal representations of the meanings these sentences describe. The system demonstrates error free performance and systematic generalization for a rich subset of English constructions that includes complex hierarchical grammatical structure, and generalizes systematically to new sentences of the learned construction categories. Further testing demonstrates (1) the capability to accommodate a significantly extended set of constructions, and (2) extension to Japanese, a free word order language that is structurally quite different from English, thus demonstrating the extensibility of the structure mapping model.

1 Introduction

The nativist perspective on the problem of language acquisition holds that the <sentence, meaning> data to which the child is exposed is highly indeterminate, and underspecifies the mapping to be learned. This "poverty of the stimulus" is a central argument for the existence of a genetically specified universal grammar, such that language acquisition consists of configuring the UG for the appropriate target language (Chomsky 1995). In this framework, once a given parameter is set, its use should apply to new constructions in a generalized, generative manner.

An alternative functionalist perspective holds that learning plays a much more central role in language acquisition. The infant develops an inventory of grammatical constructions as mappings from form to meaning (Goldberg 1995). These constructions are initially rather fixed and specific, and later become generalized into a more abstract compositional form employed by the adult (Tomasello 1999, 2003). In this context, construction of the relation between perceptual and cognitive representations and grammatical form plays a central role in learning language (e.g. Feldman et al. 1990, 1996; Langacker 1991; Mandler 1999; Talmy 1998).

These issues of learnability and innateness have provided a rich motivation for simulation studies that have taken a number of different forms. Elman (1990) demonstrated that recurrent networks are sensitive to predictable structure in grammatical sequences. Subsequent studies of grammar induction demonstrate how syntactic structure can be recovered from sentences (e.g. Stolcke & Omohundro 1994). From the "grounding of language in meaning" perspective (e.g. Feldman et al. 1990, 1996; Langacker 1991; Goldberg 1995) Chang & Maia (2001) exploited the relations between action representation and simple verb frames in a construction grammar approach. In effort to consider more complex grammatical forms, Miikkulainen (1996) demonstrated a system that learned the mapping between relative phrase constructions and multiple event representations, based on the use of a stack for maintaining state information during the processing of the next embedded clause in a recursive manner.

In a more generalized approach, Dominey (2000) exploited the regularity that sentence to meaning mapping is encoded in all languages by word order and grammatical marking (bound or free) (Bates et al. 1982). That model was based on

the functional neurophysiology of cognitive sequence and language processing and an associated neural network model that has been demonstrated to simulate interesting aspects of infant (Dominey & Ramus 2000) and adult language processing (Dominey et al. 2003).

2 Structure mapping for language learning

The mapping of sentence form onto meaning (Goldberg 1995) takes place at two distinct levels in the current model: Words are associated with individual components of event descriptions, and grammatical structure is associated with functional roles within scene events. The first level has been addressed by Siskind (1996), Roy & Pentland (2002) and Steels (2001) and we treat it here in a relatively simple but effective manner. Our principle interest lies more in the second level of mapping between scene and sentence structure.

Equations 1-7 implement the model depicted in Figure 1, and are derived from a neurophysiologically motivated model of sensorimotor sequence learning (Dominey et al. 2003).

2.1 Word Meaning

Equation (1) describes the associative memory, WordToReferent, that links word vectors in the OpenClassArray (OCA) with their referent vectors in the SceneEventArray (SEA)[1]. In the initial learning phases there is no influence of syntactic knowledge and the word-referent associations are stored in the WordToReferent matrix (Eqn 1) by associating every word with every referent in the current scene (α = 1), exploiting the cross-situational regularity (Siskind 1996) that a given word will have a higher coincidence with referent to which it refers than with other referents. This initial word learning contributes to learning the mapping between sentence and scene structure (Eqn. 4, 5 & 6 below). Then, knowledge of the syntactic structure, encoded in SentenceToScene can be used to identify the appropriate referent (in the SEA) for a given word (in the OCA), corresponding to a zero value of α in Eqn. 1. In this "syntactic bootstrapping" for the new word "gugle," for example, syntactic knowledge of Agent-Event-Object structure of the sentence "John pushed the gugle" can be used to assign "gugle" to the object of push.

$$\text{WordToReferent}(i,j) = \text{WordToReferent}(i,j) + \text{OCA}(k,i) * \text{SEA}(m,j) * \max(\alpha, \text{SentenceToScene}(m,k)) \quad (1)$$

2.2 Open vs Closed Class Word Categories

Our approach is based on the cross-linguistic observation that open class words (e.g. nouns, verbs, adjectives and adverbs) are assigned to their thematic roles based on word order and/or grammatical function words or morphemes (Bates et al. 1982). Newborn infants are sensitive to the perceptual properties that distinguish these two categories (Shi et al. 1999), and in adults, these categories are processed by dissociable neurophysiological systems (Brown et al. 1999). Similarly, artificial neural networks can also learn to make this function/content distinction (Morgan et al. 1996). Thus, for the speech input that is provided to the learning model open and closed class words are directed to separate processing streams that preserve their order and identity, as indicated in Figure 2.

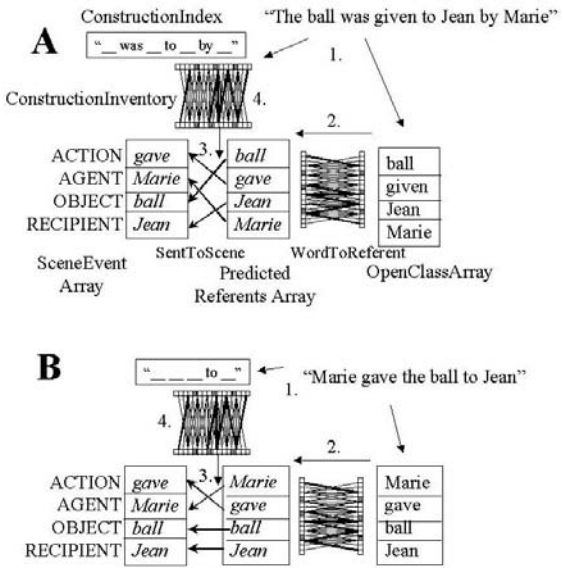


Figure 1. Structure-Mapping Architecture. 1. Lexical categorization. 2. Open class words in Open Class Array are translated to Predicted Referents in the PRA via the WordtoReferent mapping. 3. PRA elements are mapped onto their roles in the SceneEventArray by the SentenceToScene mapping, specific to each sentence type. 4. This mapping is retrieved from Construction Inventory, via the ConstructionIndex that encodes the closed class words that characterize each grammatical construction type.

2.3 Mapping Sentence to Meaning

Meanings are encoded in an event predicate, argument representation corresponding to the SceneEventArray in Figure 1 (e.g. push(Block, triangle) for "The triangle pushed the block"). There, the sentence to meaning mapping can be

[1] In Eqn 1, the index k = 1 to 6, corresponding to the maximum number of words in the open class array (OCA). Index m = 1 to 6, corresponding to the maximum number of elements in the scene event array (SEA). Indices i and j = 1 to 25, corresponding to the word and scene item vector sizes, respectively.

characterized in the following successive steps. First, words in the Open Class Array are decoded into their corresponding scene referents (via the WordToReferent mapping) to yield the Predicted Referents Array that contains the translated words while preserving their original order from the OCA (Eqn 2) [2].

$$PRA(k,j) = \sum_{i=1}^{n} OCA(k,i) * WordToReferent(i,j) \quad (2)$$

Next, each sentence type will correspond to a specific *form to meaning* mapping between the PRA and the SEA. encoded in the SentenceToScene array. The problem will be to retrieve for each sentence type, the appropriate corresponding SentenceToScene mapping. To solve this problem, we recall that each sentence type will have a unique constellation of closed class words and/or bound morphemes (Bates et al. 1982) that can be coded in a ConstructionIndex (Eqn.3) that forms a unique identifier for each sentence type.

The ConstructionIndex is a 25 element vector. Each function word is encoded as a single bit in a 25 element FunctionWord vector. When a function word is encountered during sentence processing, the current contents of ConstructionIndex are shifted (with wrap-around) by n + m bits where n corresponds to the bit that is on in the FunctionWord, and m corresponds to the number of open class words that have been encountered since the previous function word (or the beginning of the sentence). Finally, a vector addition is performed on this result and the FunctionWord vector. Thus, the appropriate SentenceToScene mapping for each sentence type can be indexed in ConstructionInventory by its corresponding ConstructionIndex.

$$ConstructionIndex = f_{circularShift}(ConstructionIndex, FunctionWord) \quad (3)$$

The link between the ConstructionIndex and the corresponding SentenceToScene mapping is established as follows. As each new sentence is processed, we first reconstruct the specific SentenceToScene mapping for that sentence (Eqn 4)[3], by mapping words to referents (in PRA) and referents to scene elements (in SEA). The resulting, SentenceToSceneCurrent encodes the correspondence between word order (that is preserved in the PRA Eqn 2) and thematic roles in the SEA. Note that the quality of SentenceToSceneCurrent will depend on the quality of acquired word meanings in WordToReferent. Thus, syntactic learning requires a minimum baseline of semantic knowledge.

$$SentenceToSceneCurrent(m,k) = \sum_{i=1}^{n} PRA(k,i)*SEA(m,i) \quad (4)$$

Given the SentenceToSceneCurrent mapping for the current sentence, we can now associate it in the ConstructionInventory with the corresponding function word configuration or ConstructionIndex for that sentence, expressed in (Eqn 5)[4].

$$ConstructionInventory(i,j) = ConstructionInventory(i,j) + ConstructionIndex(i) * SentenceToScene\text{-}Current(j) \quad (5)$$

Finally, once this learning has occurred, for new sentences we can now extract the SentenceToScene mapping from the learned ConstructionInventory by using the ConstructionIndex as an index into this associative memory, illustrated in Eqn. 6[5].

$$SentenceToScene(i) = \sum_{i=1}^{n} ConstructionInventory(i,j) * ConstructinIndex(j) \quad (6)$$

To accommodate the dual scenes for complex events Eqns. 4-7 are instantiated twice each, to represent the two components of the dual scene. In the case of simple scenes, the second component of the dual scene representation is null.

We evaluate performance by using the WordToReferent and SentenceToScene knowledge to construct for a given input sentence the "predicted scene". That is, the model will

[2] Index k = 1 to 6, corresponding to the maximum number of scene items in the predicted references array (PRA). Indices i and j = 1 to 25, corresponding to the word and scene item vector sizes, respectively.

[3] Index m = 1 to 6, corresponding to the maximum number of elements in the scene event array (SEA). Index k = 1 to 6, corresponding to the maximum number of words in the predicted references array (PRA). Index i = 1 to 25, corresponding to the word and scene item vector sizes.

[4] Note that we have linearized SentenceToSceneCurrent from 2 to 1 dimensions to make the matrix multiplication more transparent. Thus index j varies from 1 to 36 corresponding to the 6x6 dimensions of SentenceToSceneCurrent.

[5] Again to simplify the matrix multiplication, SentenceToScene has been linearized to one dimension, based on the original 6x6 matrix. Thus, index i = 1 to 36, and index j = 1 to 25 corresponding to the dimension of the ConstructionIndex.

construct an internal representation of the scene that should correspond to the input sentence. This is achieved by first converting the Open-Class-Array into its corresponding scene items in the Predicted-Referents-Array as specified in Eqn. 2. The referents are then re-ordered into the proper scene representation via application of the SentenceToScene transformation as described in Eqn. 7[6].

$$PSA(m,i) = PRA(k,i) * SentenceToScene(m,k) \quad (7)$$

When learning has proceeded correctly, the predicted scene array (PSA) contents should match those of the scene event array (SEA) that is directly derived from input to the model. We then quantify performance error in terms of the number of mismatches between PSA and SEA.

3 Learning Experiments

Three sets of results will be presented. First the demonstration of the model sentence to meaning mapping for a reduced set of constructions is presented as a proof of concept. This will be followed by a test of generalization to a new extended set of grammatical constructions. Finally, in order to validate the cross-linguistic validity of the underlying principals, the model is tested with Japanese, a free word-order language that is qualitatively quite distinct from English.

3.1 Proof of Concept with Two Constructions

3.1.1 Initial Learning of Active Forms for Simple Event Meanings

The first experiment examined learning with sentence, meaning pairs with sentences only in the active voice, corresponding to the grammatical forms 1 and 2.

1. Active: The block pushed the triangle.
2. Dative: The block gave the triangle to the moon.

For this experiment, the model was trained on 544 <sentence, meaning> pairs. Again, meaning is coded in a predicate-argument format, e.g. push(block, triangle) for sentence 1. During the first 200 trials (scene/sentence pairs), value α in Eqn. 1 was 1 and thereafter it was 0. This was necessary in order to avoid the effect of erroneous (random) syntactic knowledge on semantic learning in the initial learning stages. Evaluation of the performance of the model after this training indicated that for all sentences, there was error-free performance. That is, the PredictedScene generated from each sentence corresponded to the actual scene paired with that sentence. An important test of language learning is the ability to generalize to new sentences that have not previously been tested. Generalization in this form also yielded error free performance. In this experiment, only 2 grammatical constructions were learned, and the lexical mapping of words to their scene referents was learned. Word meaning provides the basis for extracting more complex syntactic structure. Thus, these word meanings are fixed and used for the subsequent experiments.

3.1.2 Passive forms

The second experiment examined learning with the introduction of passive grammatical forms, thus employing grammatical forms 1-4.

3. Passive: The triangle was pushed by the block.
4. Dative Passive: The moon was given to the triangle by the block.

A new set of <sentence, scene> pairs was generated that employed grammatical constructions, with two- and three- arguments, and active and passive grammatical forms for the narration. Word meanings learned in Experiment 1 were used, so only the structural mapping from grammatical to scene structure was learned. With exposure to less than 100 <sentence, scene>, error free performance was achieved. Note that only the WordToReferent mappings were retained from Experiment 1. Thus, the 4 grammatical forms were learned from the initial naive state. This means that the ConstructionIndex and ConstructionInventory mechanism correctly discriminates and learns the mappings for the different grammatical constructions. In the generalization test, the learned values were fixed, and the model demonstrated error-free performance on new sentences for all four grammatical forms that had not been used during the training.

3.1.3 Relative forms for Complex Events

The complexity of the scenes/meanings and corresponding grammatical forms in the previous experiments were quite limited. Here we consider complex <sentence, scene> mappings that involve relativised sentences and dual event scenes. A

[6] In Eqn 7, index i = 1 to 25 corresponding to the size of the scene and word vectors. Indices m and k = 1 to 6, corresponding to the dimension of the predicted scene array, and the predicted references array, respectively.

small corpus of complex <sentence, scene> pairs were generated corresponding to the grammatical construction types 5-10

5. The block that pushed the triangle touched the moon.
6. The block pushed the triangle that touched the moon.
7. The block that pushed the triangle was touched by the moon.
8. The block pushed the triangle that was touched the moon.
9. The block that was pushed by the triangle touched the moon.
10. The block was pushed by the triangle that touched the moon.

After exposure to less than 100 sentences generated from these relativised constructions, the model performed without error for these 6 construction types. In the generalization test, the learned values were fixed, and the model demonstrated error-free performance on new sentences for all six grammatical forms that had not been used during the training.

3.1.4 Combined Test

The objective of the final experiment was to verify that the model was capable of learning the 10 grammatical forms together in a single learning session. Training material from the previous experiments were employed that exercised the ensemble of 10 grammatical forms. After exposure to less than 150 <sentence, scene> pairs, the model performed without error. Likewise, in the generalization test the learned values were fixed, and the model demonstrated error-free performance on new sentences for all ten grammatical forms that had not been used during the training.

This set of experiments in ideal conditions demonstrates a proof of concept for the system, though several open questions can be posed based on these results. First, while the demonstration with 10 grammatical constructions is interesting, we can ask if the model will generalize to an extended set of constructions. Second, we know that the English language is quite restricted with respect to its word order, and thus we can ask whether the theoretical framework of the model will generalize to free word order languages such as Japanese. These questions are addressed in the following three sections.

3.2 Generalization to Extended Construction Set

As illustrated above the model can accommodate 10 distinct form-meaning mappings or grammatical constructions, including constructions involving "dual" events in the meaning representation that correspond to relative clauses. Still, this is a relatively limited size for the construction inventory. The current experiment demonstrates how the model generalizes to a number of new and different relative phrases, as well as additional sentence types including: conjoined (John took the key and opened the door), reflexive (The boy said that the dog was chased by the cat), and reflexive pronoun (The block said that it pushed the cylinder) sentence types, for a total of 38 distinct abstract grammatical constructions. The consideration of these sentence types requires us to address how their meanings are represented. Conjoined sentences are represented by the two corresponding events, e.g. *took(John, key), open(John, door)* for the conjoined example above. Reflexives are represented, for example, as *said(boy), chased(cat, dog)*. This assumes indeed, for reflexive verbs (e.g. said, saw), that the meaning representation includes the second event as an argument to the first. Finally, for the reflexive pronoun types, in the meaning representation the pronoun's referent is explicit, as in *said(block), push(block, cylinder)* for "The block said that it pushed the cylinder."

For this testing, the ConstructionInventory is implemented as a lookup table in which the ConstructionIndex is paired with the corresponding SentenceToScene mapping during a single learning trial. Based on the tenets of the construction grammar framework (Goldberg 1995), if a sentence is encountered that has a form (i.e. ConstructionIndex) that does not have a corresponding entry in the ConstructionInventory, then a new construction is defined. Thus, one exposure to a sentence of a new construction type allows the model to generalize to any new sentence of that type. In this sense, developing the capacity to handle a simple initial set of constructions leads to a highly extensible system. Using the training procedures as described above, with a pre-learned lexicon (WordToReferent), the model successfully learned all of the constructions, and demonstrated generalization to new sentences that it was not trained on.

That the model can accommodate these 38 different grammatical constructions with no modifications indicates its capability to generalize. This translates to a (partial) validation of the hypothesis that across languages, thematic role assignment is encoded by a limited set of

parameters including word order and grammatical marking, and that distinct grammatical constructions will have distinct and identifying ensembles of these parameters. However, these results have been obtained with English that is a relatively fixed word-order language, and a more rigorous test of this hypothesis would involve testing with a free word-order language such as Japanese.

3.3 Generalization to Japanese

The current experiment will test the model with sentences in Japanese. Unlike English, Japanese allows extensive liberty in the ordering of words, with grammatical roles explicitly marked by postpositional function words -ga, -ni, -wo, -yotte. This word-order flexibility of Japanese with respect to English is illustrated here with the English active and passive di-transitive forms that each can be expressed in 4 different common manners in Japanese:

1. The block gave the circle to the triangle.
1.1 Block-ga triangle-ni circle-wo watashita .
1.2 Block-ga circle-wo triangle-ni watashita .
1.3 Triangle-ni block-ga circle-wo watashita .
1.4 Circle-wo block-ga triangle-ni watashita .
2. The circle was given to the triangle by the block.
2.1 Circle-ga block-ni-yotte triangle-ni watasareta.
2.2 Block-ni-yotte circle-ga triangle-ni watasareta .
2.3 Block-ni-yotte triangle-ni circle-ga watasareta .
2.4 Triangle-ni circle-ga block-ni-yotte watasareta .

In the "active" Japanese sentences, the postpositional function words -ga, -ni and –wo explicitly mark agent, recipient and, object whereas in the passive, these are marked respectively by –ni-yotte, -ga, and –ni. For both the active and passive forms, there are four different legal word-order permutations that preserve and rely on this marking. Japanese thus provides an interesting test of the model's ability to accommodate such freedom in word order.

Employing the same method as described in the previous experiment, we thus expose the model to <sentence, meaning> pairs generated from 26 Japanese constructions that employ the equivalent of active, passive, relative forms and their permutations. We predicted that by processing the -ga, -ni, -yotte and –wo markers as closed class elements, the model would be able to discriminate and identify the distinct grammatical constructions and learn the corresponding mappings. Indeed, the model successfully discriminates between all of the construction types based on the ConstructionIndex unique to each construction type, and associates the correct SentenceToScene mapping with each of them. As for the English constructions, once learned, a given construction could generalize to new untrained sentences.

This demonstration with Japanese is an important validation that at least for this subset of constructions, the construction-based model is applicable both to fixed word order languages such as English, as well as free word order languages such as Japanese. This also provides further validation for the proposal of Bates and MacWhinney (et al. 1982) that thematic roles are indicated by a constellation of cues including grammatical markers and word order.

3.4 Effects of Noise

The model relies on lexical categorization of open vs. closed class words both for learning lexical semantics, and for building the ConstructionIndex for phrasal semantics. While we can cite strong evidence that this capability is expressed early in development (Shi et al. 1999) it is still likely that there will be errors in lexical categorization. The performance of the model for learning lexical and phrasal semantics for active transitive and ditransitive structures is thus examined under different conditions of lexical categorization errors. A lexical categorization error consists of a given word being assigned to the wrong category and processed as such (e.g. an open class word being processed as a closed class word, or vice-versa). Figure 2 illustrates the performance of the model with random errors of this type introduced at levels of 0 to 20 percent errors.

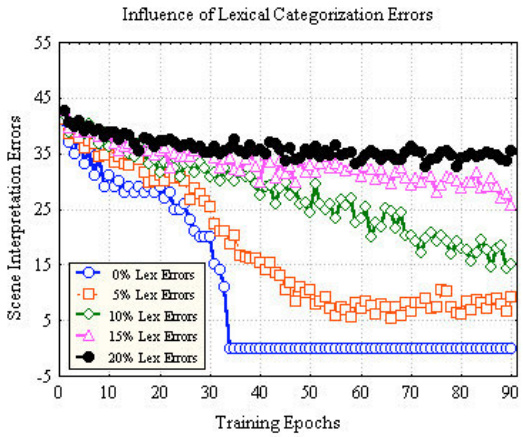


Figure 2. The effects of Lexical Categorization Errors (mis-categorization of an open-class word as a closed-class word or vice-versa) on performance (Scene Interpretation Errors) over Training Epochs. The 0% trace indicates performance in the absences of noise, with a rapid elimination of errors . The successive introduction of categorization errors yields a corresponding progressive impairment in learning. While sensitive to the errors, the system demonstrates a desired graceful degradation

We can observe that there is a graceful degradation, with interpretation errors progressively increasing as categorization errors rise to 20 percent. In order to further asses the learning that was able to occur in the presence of noise, after training with noise, we then tested performance on noise-free input. The interpretation error values in these conditions were 0.0, 0.4, 2.3, 20.7 and 33.6 out of a maximum of 44 for training with 0, 5, 10, 15 and 20 percent lexical categorization errors, respectively. This indicates that up to 10 percent input lexical categorization errors allows almost error free learning. At 15 percent input errors the model has still significantly improved with respect to the random behavior (~45 interpretation errors per epoch). Other than reducing the lexical and phrasal learning rates, no efforts were made to optimize the performance for these degraded conditions, thus there remains a certain degree of freedom for improvement. The main point is that the model does not demonstrate a catastrophic failure in the presence of lexical categorization errors.

4 Discussion

The research demonstrates an implementation of a model of sentence-to-meaning mapping in the developmental and neuropsychologically inspired construction grammar framework. The strength of the model is that with relatively simple "innate" learning mechanisms, it can acquire a variety of grammatical constructions in English and Japanese based on exposure to <sentence, meaning> pairs, with only the lexical categories of open vs. closed class being prespecified. This lexical categorization can be provided by frequency analysis, and/or acoustic properties specific to the two classes (Blanc et al. 2003; Shi et al. 1999). The model learns grammatical constructions, and generalizes in a systematic manner to new sentences within the class of learned constructions. This demonstrates the cross-linguistic validity of our implementation of the construction grammar approach (Goldberg 1995, Tomasello 2003) and of the "cue competition" model for coding of grammatical structure (Bates et al. 1982). The point of the Japanese study was to demonstrate this cross-linguistic validity – i.e. that nothing extra was needed, just the identification of constructions based on lexical category information. Of course a better model for Japanese and Hungarian etc. that exploits the explicit marking of grammatical roles of NPs would have been interesting – but it wouldn't have worked for English!

The obvious weakness is that it does not go further. That is, it cannot accommodate new construction types without first being exposed to a training example of a well formed <sentence, meaning> pair. Interestingly, however, this appears to reflect a characteristic stage of human development, in which the infant relies on the use of constructions that she has previously heard (see Tomasello 2003). Further on in development, however, as pattern finding mechanisms operate on statistically relevant samples of this data, the child begins to recognize structural patterns, corresponding for example to noun phrases (rather than solitary nouns) in relative clauses. When this is achieved, these phrasal units can then be inserted into existing constructions, thus providing the basis for "on the fly" processing of novel relativised constructions. This suggests how the abstract construction model can be extended to a more generalized compositional capability. We are currently addressing this issue in an extension of the proposed model, in which recognition of linguistic markers (e.g. "that", and directly successive NPs) are learned to signal embedded relative phrases (see Miikkulainen 1996).

Future work will address the impact of ambiguous input. The classical example "John saw the girl with the telescope" implies that a given grammatical form can map onto multiple meaning structures. In order to avoid this violation of the one to one mapping, we must concede that form is influenced by context. Thus, the model will fail in the same way that humans do, and should be able to succeed in the same way that humans do. That is, when context is available to disambiguate then ambiguity can be resolved. This will require maintenance of the recent discourse context, and the influence of this on grammatical construction selection to reduce ambiguity.

5 Acknowledgements

This work was supported by the ACI Computational Neuroscience Project, The Eurocores OMLL project and the HFSP Organization.

References

Bates E, McNew S, MacWhinney B, Devescovi A, Smith S (1982) Functional constraints on sentence processing: A cross linguistic study, *Cognition* (11) 245-299.

Blanc JM, Dodane C, Dominey P (2003) Temporal processing for syntax acquisition. Proc. 25th Ann. Mtg. Cog. Science Soc. Boston

Brown CM, Hagoort P, ter Keurs M (1999) Electrophysiological signatures of visual lexical

processing : Open- and closed-class words. *Journal of Cognitive Neuroscience*. 11 :3, 261-281

Chang NC, Maia TV (2001) Grounded learning of grammatical constructions, *AAAI Spring Symp. On Learning Grounded Representations*, Stanford CA.

Chomsky N. (1995) The Minimalist Program. MIT

Crangle C. & Suppes P. (1994) Language and Learning for Robots, CSLI lecture notes: no. 41, Stanford.

Dominey PF, Ramus F (2000) Neural network processing of natural language: I. Sensitivity to serial, temporal and abstract structure of language in the infant. *Lang. and Cognitive Processes*, 15(1) 87-127

Dominey PF (2000) Conceptual Grounding in Simulation Studies of Language Acquisition, *Evolution of Communication*, 4(1), 57-85.

Dominey PF, Hoen M, Lelekov T, Blanc JM (2003) Neurological basis of language and sequential cognition: Evidence from simulation, aphasia and ERP studies, *Brain and Language, 86, 207-225*

Elman J (1990) Finding structure in time. *Cognitive Science*, 14:179-211.

Feldman JA, Lakoff G, Stolcke A, Weber SH (1990) Miniature language acquisition: A touchstone for cognitive science. In *Proceedings of the 12th Ann Conf. Cog. Sci. Soc.* 686-693, MIT, Cambridge MA

Feldman J., G. Lakoff, D. Bailey, S. Narayanan, T. Regier, A. Stolcke (1996). L0: The First Five Years. *Artificial Intelligence Review*, v10 103-129.

Goldberg A (1995) *Constructions*. U Chicago Press, Chicago and London.

Hirsh-Pasek K, Golinkof RM (1996) *The origins of grammar: evidence from early language comprehension.* MIT Press, Boston.

Kotovsky L, Baillargeon R, The development of calibration-based reasoning about collision events in young infants. 1998, *Cognition*, 67, 311-351

Langacker, R. (1991). *Foundations of Cognitive Grammar. Practical Applications, Volume 2*. Stanford University Press, Stanford.

Mandler J (1999) Preverbal representations and language, in P. Bloom, MA Peterson, L Nadel and MF Garrett (Eds) *Language and Space*, MIT Press, 365-384

Miikkulainen R (1996) Subsymbolic case-role analysis of sentences with embedded clauses. *Cognitive Science*, 20:47-73.

Morgan JL, Shi R, Allopenna P (1996) Perceptual bases of rudimentary grammatical categories: Toward a broader conceptualization of bootstrapping, pp 263-286, in Morgan JL, Demuth K (Eds) *Signal to syntax*, Lawrence Erlbaum, Mahwah NJ, USA.

Pollack JB (1990) Recursive distributed representations. Artificial Intelligence, 46:77-105.

Roy D, Pentland A (2002). Learning Words from Sights and Sounds: A Computational Model. *Cognitive Science*, 26(1), 113-146.

Shi R., Werker J.F., Morgan J.L. (1999) Newborn infants' sensitivity to perceptual cues to lexical and grammatical words, *Cognition*, Volume 72, Issue 2, B11-B21.

Siskind JM (1996) A computational study of cross-situational techniques for learning word-to-meaning mappings, *Cognition* (61) 39-91.

Siskind JM (2001) Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. *Journal of AI Research* (15) 31-90

Steels, L. (2001) Language Games for Autonomous Robots. *IEEE Intelligent Systems*, vol. 16, nr. 5, pp. 16-22, New York: IEEE Press.

Stolcke A, Omohundro SM (1994) Inducing probabilistic grammars by Bayesian model merging/ In *Grammatical Inference and Applications: Proc. 2nd Intl. Colloq. On Grammatical Inference*, Springer Verlag.

Talmy L (1988) Force dynamics in language and cognition. *Cognitive Science*, 10(2) 117-149.

Tomasello M (1999) The item-based nature of children's early syntactic development, *Trends in Cognitive Science*, 4(4):156-163

Tomasello, M. (2003) Constructing a language: A usage-based theory of language acquisition. Harvard University Press, Cambridge.

On the Acquisition of Phonological Representations

B. Elan DRESHER
Department of Linguistics
University of Toronto
Toronto, Ontario
Canada M5S 3H1
dresher@chass.utoronto.ca

Abstract

Language learners must acquire the grammar (rules, constraints, principles) of their language as well as representations at various levels. I will argue that representations are part of the grammar and must be acquired together with other aspects of grammar; thus, grammar acquisition may not presuppose knowledge of representations. Further, I will argue that the goal of a learning model should not be to try to match or approximate target forms directly, because strategies to do so are defeated by the disconnect between principles of grammar and the effects they produce. Rather, learners should use target forms as evidence bearing on the selection of the correct grammar. I will draw on two areas of phonology to illustrate these arguments. The first is the grammar of stress, or metrical phonology, which has received much attention in the learning model literature. The second concerns the acquisition of phonological features and contrasts. This aspect of acquisition turns out, contrary to first appearances, to pose challenging problems for learning models.

1 Introduction

I will discuss the extent to which representations are intertwined with the grammar, and consequences of this fact for acquisition models. I will focus on phonological representations, but the argument extends to other components of the grammar.

One might suppose that phonological representations can be acquired directly from the acoustic signal. If, for example, children are equipped with innate phonetic feature detectors, one might suppose that they can use these to extract phonetic features from the signal. These extracted phonetic features would then constitute phonological representations (*surface*, or *phonetic, representations*). Once these are acquired, they can serve as a basis from which learners can acquire the rest of the grammar, namely, the phonological rules (and/or constraints) and the *lexical*, or *underlying, representations*.

This idea of acquisition by stages, with representations preceding rules, has enduring appeal, though details vary with the prevailing theory of grammar; versions of this theory can be found in (Bloch, 1941) and (Pinker, 1994:264–5). The idea could not be implemented in American Structuralist phonology, however (Chomsky, 1964), and I will argue that it remains untenable today. I will discuss two areas of phonology in which representations must be acquired together with the grammar, rather than prior to it. The first concerns the grammar of stress, or metrical phonology. The second concerns the acquisition of phonological features. These pose different sorts of problems for learning models. The first has been the subject of considerable discussion. The second, to my knowledge, has not been discussed in the context of formal learning models. Though it has often been assumed, as mentioned above, that acquisition of features might be the most straightforward aspect of phonological acquisition, I will argue that it presents challenging problems for learning models.

2 Representations of stress

Phonetic representations are not simply bundles of features. Consider stress, for example. Depending on the language, stress may be indicated phonetically by pitch, duration, loudness, or by some combination of these dimensions. So even language learners gifted with phonetic feature detectors will have to sort out what the specific correlates of stress are in their language. For purposes of the ensuing discussion, I will assume that this much can be acquired prior to further acquisition of the phonology.

But simply deciding which syllables have stress does not yield a surface representation of the stress contour of a word. According to metrical theory (Liberman and Prince 1977, Halle and Idsardi 1995, Hayes 1995), stress results from grouping syllables into feet; the strongest foot is assigned the main stress, the other feet are associated with secondary stress. Moreover, some syllables at the edges of the stress domain may be designated as extrametrical, and not included in feet.

For example, I assume that learners who have sorted out which acoustic cues signal stress can at some point assign the stress contours depicted in (1) to English words. The height of the column over each syllable, *S*, indicates how much relative stress it has. However, these are not the surface representations. They indicate levels of stress, but no metrical organization.

(1) Representations of stress contours before setting metrical parameters

a. *América*

```
   x
   x
 x x x x
 S S S S
 America
```

b. *Mànitóba*

```
     x
   x   x
 x x x   x
 S S S   S
 Manito:ba
```

According to conventional accounts of English stress, the metrical structures assigned to these words are as in (2).

(2) Acquired representations

a. *América*

```
   x                Line 2
  (x)               Line 1
 x(x x)<x>          Line 0
 L L L  L
 Ameri ca
```

b. *Mànitóba*

```
     x              Line 2
   (x   x)          Line 1
 (x x)(x)<x>        Line 0
   L L  H  L
 Mani to:ba
```

Looking at the word *America*, these representations indicate that the first syllable *A* is unfooted, that the next two syllables *meri* constitute a trochaic foot, and that the final syllable *ca* is extrametrical. *Manitoba* has two feet, hence two stresses, of which the second is stronger than the first. The *L*s and *H*s under the first line of the metrical grid designate light and heavy syllables, respectively. The distinction is important in English: The syllable *to:* in *Manitoba* is heavy, hence capable of making up a foot by itself, and it receives the stress. If it were light, then *Manitoba* would have stress on the antepenultimate syllable, as in *America*.

How does a learner know to assign these surface structures? Not just from the acoustic signal, or from the schematic stress contours in (1). Observe that an unstressed syllable can have several metrical representations: it can be footed, like the first syllable in *America*; it can be the weak position of a foot, like the second syllable of *Manitoba*; or it can be extrametrical, like the final syllables in both words. One cannot tell from the sound which of these representations to assign. The only way to know this is to acquire the grammar of stress, based on evidence drawn from the observed contours in (1).

Similar remarks hold for determining syllable quantity. English divides syllables into light and heavy: a light syllable ends in a short vowel, and a heavy syllable contains either a long vowel or is closed by a consonant. In many other languages, though, a closed syllable containing a short vowel is considered to be light, contrary to the English categorization. Learners must decide how to

classify such syllables, and the decision cannot be made on phonetic grounds alone.

3 Acquisition of metrical structure

How, then, are these aspects of phonological structure acquired? Following Chomsky (1981), I will suppose that metrical structures are governed by a finite number of parameters, whose value is to be set on the basis of experience. The possible values of a parameter are limited and given in advance.[1]

Parameter setting models must overcome a basic problem: the relation between a parameter and what it does is indirect, due to the fact that there are many parameters, and they interact in complex ways (Dresher and Kaye, 1990). For example, in English main stress is tied to the right edge of the word. But that does not mean that stress is always on the last syllable: it could be on the penultimate syllable, as in *Manitoba*, or on the antepenultimate, as in *America*. What is consistent in these examples is that main stress devolves onto the strong syllable of the rightmost foot. Where this syllable and foot is in any given word depends on how a variety of parameters are set. Some surprising consequences follow from the nontransparent relationship between a parameter and its effects.

The first one is that a learner who has some incorrectly set parameters might know that something is wrong, but might not know which parameter is the source of the problem. This is known as the *Credit Problem* (cf. Clark 1989, 1992, who calls this the *Selection Problem*): a learner cannot reliably assign credit or blame to individual parameters when something is wrong.

There is a second way in which parameters can pose problems to a learner. Some parameters are stated in terms of abstract entities and theory-internal concepts that the learner may not initially be able to identify. For example, the theory of stress is couched in terms of concepts such as heavy syllables, heads, feet, and so on. In syntax, various parameters have been posited that refer specifically to anaphors, or to functional projections of various types. These entities do not come labelled as such in the input, but must themselves be constructed by the learner. So, to echo the title character in Plato's dialogue *The Meno*, how can learners determine if main stress falls on the first or last foot if they do not know what a foot is, or how to identify one? This can be called the *Epistemological Problem*: in this case we know about something in the abstract, but we do not recognize that thing when it is front of us.

Because of the Credit Problem and the Epistemological Problem, parameter setting is not like learning to hit a target, where one can correct one's aim by observing where previous shots land. The relation between number of parameters correct and apparent closeness to the target is not *smooth* (Turkel, 1996): one parameter wrong may result in forms that appear to be way off the target, whereas many parameters wrong may produce results that appear to be better (Dresher, 1999). This discrepancy between grammar and outputs defeats learning models that blindly try to match output forms (Gibson and Wexler, 1994), or that are based on a notion of goodness-of-fit (Clark and Roberts, 1993). In terms of Fodor (1998), there are no unambiguous triggers: thus, learning models that seek them in individual target forms are unlikely to be successful.

I have argued (Dresher, 1999) that Plato's solution – a series of questions posed in a specified order – is the best approach we have. One version of this approach is the *cue-based* learner of (Dresher and Kaye, 1990). In this model, not only are the principles and parameters of Universal Grammar innate, but learners must be born with some kind of a road map that guides them in setting the parameters. Some ingredients of this road map are the following:

First, Universal Grammar associates every parameter with a *cue*, something in the data

[1] For some other approaches to the acquisition of stress see (Daelemans Gillis and Durieux, 1994), (Gupta and Touretzky, 1994), (Tesar, 1998, 2004), and (Tesar and Smolensky, 1998).

that signals the learner how that parameter is to be set. The cue might be a pattern that the learner must look for, or simply the presence of some element in a particular context.

Second, parameter setting proceeds in a (partial) order set by Universal Grammar: this ordering specifies a *learning path* (Lightfoot 1989). The setting of a parameter later on the learning path depends on the results of earlier ones.

Hence, cues can become increasingly abstract and grammar-internal the further along the learning path they are. As learners acquire more of the system, their representations become more sophisticated, and they are able to build on what they have already learned to set more parameters.[2]

If this approach is correct, there is no parameter-independent learning algorithm. This is because the learning path is dependent on the particular parameters. Also, the cues must be discovered for each parameter. Thus, a learning algorithm for one part of the grammar cannot be applied to another part of the grammar in an automatic way.[3]

4. Segmental representations

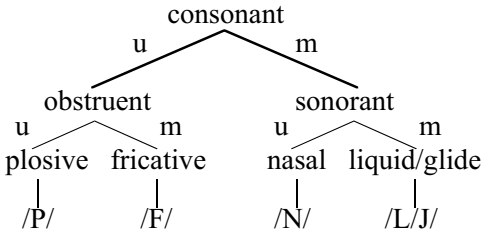
Up to now we have been looking at an aspect of phonological representation above the level of the segment. I have argued that acquisition of this aspect of surface phonological representation cannot simply be based on attending to the acoustic signal, but requires a more elaborate learning model. But what about acquisition of the phonemic inventory of a language? One might suppose that this be achieved prior to the acquisition of the phonology itself.

Since the pioneering work of Trubetzkoy and Jakobson, phonological theory has posited that phonemes are characterized in terms of a limited set of *distinctive features*. Therefore, to identify a phoneme one must be able to assign to it a representation in terms of feature specifications. What are these representations? Since Saussure, it has been a central assumption of much linguistic theory that a unit is defined not only in terms of its substance, but also in negative terms, with respect to the units it contrasts with. On this way of thinking, an /i/ that is part of a three-vowel system /i a u/ is not necessarily the same thing as an /i/ that is part of a seven-vowel system /i ɪ e a o ʊ u/. In a three-vowel system, no more than two features are required to distinguish each vowel from all the others; in a seven-vowel system, at least one more feature is required.

Jakobson and Halle (1956) suggested that distinctive features are necessarily binary because of how they are *acquired*, through a series of 'binary fissions'. They propose that the order of these contrastive splits, which form what I will call a *contrastive hierarchy* (Dresher 2003a, b) is partially fixed, thereby allowing for certain developmental sequences and ruling out others. This idea has been fruitfully applied in acquisition studies, where it is a natural way of describing developing phonological inventories (Pye Ingram and List, 1987), (Ingram, 1989), (Levelt, 1989), (Dinnsen et al., 1990), (Dinnsen, 1992), and (Rice and Avery, 1995).

Consider, for example, the development of segment types in onset position in Dutch (Fikkert, 1994):

(3) Development of Dutch onset consonants (Fikkert 1994)



At first there are no contrasts. The value of the consonant defaults to the least marked (*u*) onset, namely an obstruent plosive, desig-

[2]For details of parameter ordering, defaults, and cues in the acquisition of stress, see (Dresher and Kaye, 1990) and (Dresher, 1999).

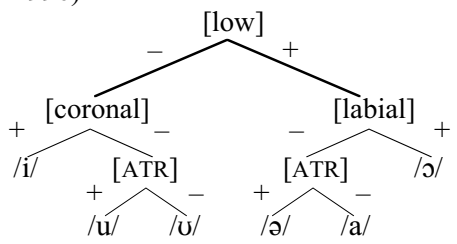
[3] For further discussion and critiques of cue-based models see (Nyberg, 1991), (Gillis Durieux and Daelemans, 1995), (Bertolo et al. 1997), and (Tesar, 2004).

nated here as /P/. The first contrast is between obstruent and sonorant. The former remains the unmarked (*u*), or default, option; the marked (*m*) sonorant defaults to nasal, /N/. At this point children differ. Some expand the obstruent branch first, bringing in marked fricatives, /F/, in contrast with plosives. Others expand the sonorant branch, introducing marked sonorants, which may be either liquids, /L/, or glides, /J/. Continuing in this way we will eventually have a tree that gives all and only the contrasting features in the language.

5. Acquiring segmental representations

Let us consider how such representations might be acquired. To illustrate, we will look at the vowel system of Classical Manchu (Zhang, 1996), which nicely illustrates the types of problems a learning model will have to overcome. Zhang (1996) proposes the contrastive hierarchy in (4) for Classical Manchu, where the order of the features is [low]> [coronal]>[labial]>[ATR].

(4) Classical Manchu vowel system (Zhang 1996)[4]



Part of the evidence for these specifications comes from the following observations:

(5) Evidence for the specifications in (4)
   a. /u/ and /ə/ trigger ATR harmony, but /i/ does not, though /i/ is phonetically [+ATR], suggesting that /i/ lacks a phonological specification for [ATR].
   b. /ɔ/ triggers labial harmony, but /u/ and /ʊ/ do not. Though phonetically [+labial], there is no evidence that /u/ and /ʊ/ are specified for this feature.

Acquiring phonological specifications is not the same as identifying phonetic features. Surface phonetics do not determine the phonological specifications of a segment. Manchu /i/ is phonetically [+ATR], but does not bear the feature phonologically; /u/ and /ʊ/ are phonetically [+labial], but are not specified for that feature. How does a learner deduce phonological (contrastive) specifications from surface phonetics?[5]

It must be the case that phoneme acquisition requires learners to take into account phonological processes, and not just the local phonetics of individual segments (Dresher and van der Hulst, 1995). Thus, the phonological status of Manchu vowels is demonstrated most clearly by attending to the effects of the vowel on neighbouring segments.

This conclusion is strengthened when we consider that the distinction between /u/ and /ʊ/ in Classical Manchu is phonetically evident only after back consonants; elsewhere, they merge to [u]. To determine the underlying identity of a surface [u], therefore, a language learner must observe its patterning with other vowels: if it co-occurs with [+ATR] vowels, it is /u/; otherwise, it is /ʊ/. The nonlocal and diverse character of the evidence bearing on the feature specifications of segments poses a challenge to learning models.

Finally, let us consider the acquisition of the hierarchy of contrastive features in each language. Examples such as the acquisition of Dutch onsets given above appear to accord well with the notion of a learning path, whereby learners proceed to master individual feature contrasts in order. If this order were the same for all languages, then this

[4]Zhang (1996) assumes privative features: [F] vs. the absence of [F], rather than [+F] vs. [–F]. The distinction between privative and binary features is not crucial to the matters under discussion here.

[5]Phonological contrasts that play a role in phonological representations are thus different from their phonetic manifestations, the subject of studies such as (Flemming, 1995).

much would not have to be acquired. However, it appears that the feature hierarchies vary somewhat across languages (Dresher, 2003a, b). The existence of variation raises the question of how learners determine the order for their language. The problem is difficult, because establishing the correct ordering, as shown by the active contrasts in a language, appears to involve different kinds of potentially conflicting evidence. In the case of metrical parameters, the relevant evidence could be reduced to particular cues, or so it appears. Whether the setting of feature hierarchies can be parameterized in a similar way remains to be demonstrated.

6 Conclusion

I will conclude by raising one further problem for learning models that is suggested by the Manchu vowel system. We have observed that in Classical Manchu, /ə/ is the [+ATR] counterpart of /a/. Both vowels are [+low]. Since [low] is ordered first among the vowel features in the Manchu hierarchy, we might suppose that learners determine which vowels are [+low] and which are not at an early stage in the process, before assigning the other features. However, a vowel that is phonetically [ə] is ambiguous as to its featural classification. In many languages, including descendants of Classical Manchu (Zhang, 1996, Dresher & Zhang, 2003) such vowels are classified as [–low]. What helps to place /ə/ as a [+low] vowel in Classical Manchu is the knowledge that it is the [+ATR] counterpart of /a/. That is, in order to assign the feature [+low] to /ə/, it helps to know that it is [+ATR]. But, by hypothesis, [low] is assigned before [ATR]. Similarly, the determination that /i/ is contrastively [+coronal] is tied in with its not being contrastively [–labial]; but [coronal] is assigned prior to [labial].

It appears, then, that whatever order we choose to assign features, it is necessary to have some advance knowledge about classification with respect to features ordered later. Perhaps this paradox is only apparent. However it is resolved, the issue raises an interesting problem for models of acquisition.

7 Acknowledgements

This research was supported in part by grant 410-2003-0913 from the Social Sciences and Humanities Research Council of Canada. I would like to thank the members of the project on Contrast in Phonology at the University of Toronto (http://www.chass.utoronto.ca/~contrast/) for discussion.

References

Stefano Bertolo Kevin Broihir Edward Gibson and Kenneth Wexler. 1997. Characterizing learnability conditions for cue-based learners in parametric language systems. In Tilman Becker and Hans-Ulrich Krieger, editors, *Proceedings of the Fifth Meeting on the Mathematics of Language*. http://www.dfki.de/events/ mol/.

Bernard Bloch. 1941. Phonemic overlapping. *American Speech* 16:278–284. Reprinted in Martin Joos, editor, *Readings in Linguistics I*, Second edition, 93–96. New York: American Council of Learned Societies, 1958.

Noam Chomsky. 1964. Current issues in linguistic theory. In Jerry A. Fodor and Jerrold J. Katz, editors, *The Structure of Language*, 50–118. Englewood Cliffs, NJ: Prentice-Hall.

Noam Chomsky. 1981. Principles and parameters in syntactic theory. In Norbert Hornstein and David Lightfoot, editors, *Explanation In Linguistics: The Logical Problem of Language Acquisition*, 32–75. London: Longman.

Robin Clark. 1989. On the relationship between the input data and parameter setting. In *Proceedings of NELS 19*, 48–62. GLSA, University of Massachusetts, Amherst.

Robin Clark. 1992. The selection of syntactic knowledge. *Language Acquisition* 2:83–149.

Robin Clark and Ian Roberts. 1993. A computational model of language learnability and language change. *Linguistic Inquiry* 24:299–345.

Walter Daelemans Steven Gillis and Gert Durieux. 1994. The acquisition of stress: A data-oriented approach. *Computational Linguistics* 20:421–451.

Daniel A. Dinnsen. 1992. Variation in developing and fully developed phonetic inventories. In Charles Ferguson Lise Menn and Carol Stoel-Gammon, editors, *Phonological Development: Models, Research, Implications*, 191–210. Timonium, MD: York Press,.

Daniel A. Dinnsen Steven B. Chin Mary Elbert and Thomas W. Powell. 1990. Some constraints on functionally disordered phonologies: Phonetic inventories and phonotactics. *Journal of Speech and Hearing Research* 33:28–37.

B. Elan Dresher. 1999. Charting the learning path: Cues to parameter setting. *Linguistic Inquiry* 30:27–67.

B. Elan Dresher. 2003a. Contrast and asymmetries in inventories. In Anna-Maria di Sciullo, editor, *Asymmetry in Grammar, Volume 2: Morphology, Phonology, Acquisition*, 239–257. Amsterdam: John Benjamins.

B. Elan Dresher. 2003b. The contrastive hierarchy in phonology. In Daniel Currie Hall, editor, *Toronto Working Papers in Linguistics (Special Issue on Contrast in Phonology) 20*, 47–62. Toronto: Department of Linguistics, University of Toronto.

B. Elan Dresher and Harry van der Hulst. 1995. Global determinacy and learnability in phonology. In John Archibald, editor, *Phonological Acquisition and Phonological Theory,* 1–21. Hillsdale, NJ: Lawrence Erlbaum.

B. Elan Dresher and Jonathan Kaye. 1990. A computational learning model for metrical phonology. *Cognition* 34:137–195.

B. Elan Dresher and Xi Zhang. 2003. Phonological contrast and phonetics in Manchu vowel systems. Paper presented at the Twenty-Ninth Annual Meeting of the Berkeley Linguistics Society, February 2003. To appear in the Proceedings.

Paula Fikkert. 1994. *On the Acquisition of Prosodic Structure (HIL Dissertations 6)*. Dordrecht: ICG Printing.

Edward Flemming. 1995. Auditory representations in phonology. Doctoral dissertation, UCLA.

Janet Dean Fodor. 1998. Unambiguous triggers. *Linguistic Inquiry* 29:1–36.

Edward Gibson and Kenneth Wexler. 1994. Triggers. *Linguistic Inquiry* 25:407–454.

Steven Gillis Gert Durieux and Walter Daelemans. 1996. A computational model of P&P: Dresher & Kaye (1990) revisited. In Frank Wijnen and Maaike Verrips, editors, *Approaches to Parameter Setting*, 135–173. Vakgroep Algemene Taalwetenschap, Universiteit van Amsterdam.

Prahlad Gupta and David Touretzky. 1994. Connectionist models and linguistic theory: Investigations of stress systems in language. *Cognitive Science* 18:1–50.

Morris Halle and William J. Idsardi. 1995. General properties of stress and metrical structure. In John Goldsmith, editor, *The Handbook of Phonological Theory*, 403–443. Cambridge, MA: Blackwell.

Bruce Hayes. 1995. *Metrical Stress Theory: Principles and Case Studies*. Chicago: University of Chicago Press.

David Ingram. 1989. *First Language Acquisition: Method, Description and Explanation*. Cambridge: Cambridge University Press.

Roman Jakobson and Morris Halle. 1956. *Fundamentals of Language.* The Hague: Mouton.

Clara C. Levelt. 1989. An essay on child phonology. M.A. thesis, Leiden University.

Mark Liberman and Alan Prince. 1977. On stress and linguistic rhythm. *Linguistic Inquiry* 8:249–336.

David Lightfoot. 1989. The child's trigger experience: Degree-0 learnability (with

commentaries). *Behavioral and Brain Sciences* 12:321–375.

Eric H. Nyberg 3rd. 1991. A non-deterministic, success-driven model of parametric setting in language acquisition. Doctoral dissertation, Carnegie Mellon University, Pittsburgh, PA.

Steven Pinker. 1994. *The Language Instinct*. New York: William Morrow.

Plato. *Meno*. Various editions.

Clifton Pye David Ingram and Helen List. 1987. A comparison of initial consonant acquisition in English and Quiché. In Keith E. Nelson and Ann Van Kleeck, editors, *Children's Language (Vol. 6)*, 175–190. Hillsdale, NJ: Erlbaum.

Keren Rice and Peter Avery. 1995. Variability in a deterministic model of language acquisition: A theory of segmental elaboration. In John Archibald editor, *Phonological Acquisition and Phonological Theory*, 23–42. Hillsdale, NJ: Lawrence Erlbaum.

Bruce Tesar. 1998. An iterative strategy for language learning. *Lingua* 104:131–145.

Bruce Tesar. 2004. Using inconsistency detection to overcome structural ambiguity. *Linguistic Inquiry* 35:219–253.

Bruce Tesar and Paul Smolensky. 1998. Learnability in Optimality Theory. *Linguistic Inquiry* 29:229–268.

William J. Turkel. 1996. Smoothness in a parametric subspace. Ms., University of British Columbia, Vancouver.

Xi Zhang. 1996. Vowel systems of the Manchu-Tungus languages of China. Doctoral dissertation, University of Toronto.

Statistics Learning and Universal Grammar: Modeling Word Segmentation

Timothy Gambell
59 Bishop Street
New Haven, CT 06511
USA
timothy.gambell@aya.yale.edu

Charles Yang
Department of Linguistics, Yale University
New Haven, CT 06511
USA
charles.yale.edu@yale.edu

Abstract

This paper describes a computational model of word segmentation and presents simulation results on realistic acquisition In particular, we explore the capacity and limitations of statistical learning mechanisms that have recently gained prominence in cognitive psychology and linguistics.

1 Introduction

Two facts about language learning are indisputable. First, only a human baby, but not her pet kitten, can learn a language. It is clear, then, that there must be some element in our biology that accounts for this unique ability. Chomsky's Universal Grammar (UG), an innate form of knowledge specific to language, is an account of what this ability is. This position gains support from formal learning theory [1-3], which sharpens the logical conclusion [4,5] that no (realistically efficient) learning is possible without priori restrictions on the learning space. Second, it is also clear that no matter how much of a head start the child has through UG, language *is* learned. Phonology, lexicon, and grammar, while governed by universal principles and constraints, do vary from language to language, and they must be learned on the basis of linguistic experience. In other words–indeed a truism–both endowment and learning contribute to language acquisition, the result of which is extremely sophisticated body of linguistic knowledge. Consequently, both must be taken in account, explicitly, in a theory of language acquisition [6,7].

Controversies arise when it comes to the relative contributions by innate knowledge and experience-based learning. Some researchers, in particular linguists, approach language acquisition by characterizing the scope and limits of innate principles of Universal Grammar that govern the world's language. Others, in particular psychologists, tend to emphasize the role of experience and the child's domain-general learning ability. Such division of research agenda understandably stems from the division of labor between endowment and learning–plainly, things that are built in needn't be learned, and things that can be garnered from experience needn't be built in.

The important paper of Saffran, Aslin, & Newport [8] on statistical learning (SL), suggests that children may be powerful learners after all. Very young infants can exploit transitional probabilities between syllables for the task of word segmentation, with only minimum exposure to an artificial language. Subsequent work has demonstrated SL in other domains including artificial grammar learning [9], music [10], vision [11], as well as in other species [12]. This then raises the possibility of learning as an alternative to the innate endowment of linguistic knowledge [13].

We believe that the computational modeling of psychological processes, with special attention to concrete mechanisms and quantitative evaluations, can play an important role in the endowment vs. learning debate. Linguists' investigations of UG are rarely developmental, even less so corpus-oriented. Developmental psychologists, by contrast, often stop at identifying *components* in a cognitive task [14], without an account of how such components work together in an algorithmic manner. On the other hand, if computation is to be of relevance to linguistics, psychology, and cognitive science in general, being merely computational will not suffice. A model must be psychological plausible, and ready to face its implications in the broad empirical contexts [7]. For example, how does it generalize to typologically different languages? How does the model's behavior compare with that of human language learners and processors?

In this article, we will present a simple computational model of word segmentation and some of its formal and developmental issues in child language acquisition. Specifically we show that SL using transitional probabilities cannot reliably segment words when scaled to a realistic setting (e.g., child-directed English). To be successful, it must be constrained by the knowledge of phonological

structure. Indeed, the model reveals that SL may well be an artifact–an impressive one, nonetheless–that plays no role in actual word segmentation in human children.

2 Statistics does not Refute UG

It has been suggested [15, 8] that word segmentation from continuous speech may be achieved by using transitional probabilities (TP) between adjacent syllables A and B, where , TP(A→B) = P(AB)/P(A), with P(AB) being the frequency of B following A, and P(A) the total frequency of A. Word boundaries are postulated at local minima, where the TP is lower than its neighbors. For example, given sufficient amount of exposure to English, the learner may establish that, in the four-syllable sequence "prettybaby", TP(pre→tty) and TP(ba→by) are both higher than TP(tty→ba): a word boundary can be (correctly) postulated. It is remarkable that 8-month-old infants can extract three-syllable words in the continuous speech of an artificial language from only two minutes of exposure [8].

To be effective, a learning algorithm–indeed any algorithm–must have an appropriate representation of the relevant learning data. We thus need to be cautious about the interpretation of the success of SL, as the authors themselves note [16]. If anything, it seems that the findings strengthen, rather than weaken, the case for (innate) linguistic knowledge. A classic argument for innateness [4, 5, 17] comes from the fact that syntactic operations are defined over specific types of data structures–constituents and phrases–but not over, say, linear strings of words, or numerous other logical possibilities. While infants seem to keep track of statistical information, any conclusion drawn from such findings must presuppose children knowing what kind of statistical information to keep track of. After all, an infinite range of statistical correlations exists in the acoustic input: e.g., What is the probability of a syllable rhyming with the next? What is the probability of two adjacent vowels being both nasal? The fact that infants can use SL to segment syllable sequences at all entails that, at the minimum, they know the relevant unit of information over which correlative statistics is gathered: in this case, it is the syllables, rather than segments, or front vowels.

A host of questions then arises. First, How do they know so? It is quite possible that the primacy of syllables as the basic unit of speech is innately available, as suggested in neonate speech perception studies [18]? Second, where do the syllables come from? While the experiments in [8] used uniformly CV syllables, many languages, including English, make use of a far more diverse range of syllabic types. And then, syllabification of speech is far from trivial, which (most likely) involve both innate knowledge of phonological structures as well as discovering language-specific instantiations [14]. All these problems have to be solved before SL for word segmentation can take place.

3 The Model

To give a precise evaluation of SL in a realistic setting, we constructed a series of (embarrassingly simple) computational models tested on child-directed English.

The learning data consists of a random sample of child-directed English sentences from the CHILDES database [19] The words were then phonetically transcribed using the Carnegie Mellon Pronunciation Dictionary, and were then grouped into syllables. Spaces between words are removed; however, utterance breaks are available to the modeled learner. Altogether, there are 226,178 words, consisting of 263,660 syllables.

Implementing SL-based segmentation is straightforward. One first gathers pair-wise TPs from the training data, which are used to identify local minima and postulate word boundaries in the on-line processing of syllable sequences. Scoring is done for each utterance and then averaged. Viewed as an information retrieval problem, it is customary [20] to report both precision and recall of the performance.

The segmentation results using TP local minima are remarkably poor, even under the assumption that the learner has already syllabified the input perfectly. Precision is 41.6%, and recall is 23.3%; over half of the words extracted by the model are not actual English words, while close to 80% of actual words fail to be extracted. And it is straightforward why this is the case. In order for SL to be effective, a TP at an actual word boundary must be lower than its neighbors. Obviously, this condition cannot be met if the input is a sequence of monosyllabic words, for which a space must be postulated for every syllable; there are no local minima to speak of. While the pseudowords in [8] are uniformly three-syllables long, much of child-directed English consists of sequences of monosyllabic words: corpus statistics reveals that on average, a monosyllabic word is followed by another monosyllabic word 85% of time. As long as this is the case, SL cannot, in principle, work.

4 Statistics Needs UG

This is not to say that SL cannot be effective for word segmentation. Its application, must be constrained–like that of any learning algorithm however powerful–as suggested by formal learning theories [1-3]. The performance improves dramatically, in fact, if the learner is equipped with even a small amount of prior knowledge about phonological structures. Specifically, we assume, uncontroversially, that each word can have only one primary stress. (This would not work for functional words, however.) If the learner knows this, then it may limit the search for local minima only in the window between two syllables that both bear primary stress, e.g., between the two **a**'s in the sequence "l**a**nguage**a**cquisition". This assumption is plausible given that 7.5-month-old infants are sensitive to strong/weak prosodic distinctions [14]. When stress information suffices, no SL is employed, so "b**i**gb**a**dw**o**lf" breaks into three words for free. Once this simple principle is built in, the stress-delimited SL algorithm can achieve the precision of 73.5% and 71.2%, which compare favorably to the best performance reported in the literature [20]. (That work, however, uses an computationally prohibitive and psychological implausible algorithm that iteratively optimizes the entire lexicon.)

The computational models complement the experimental study that prosodic information takes priority over statistical information when both are available [21]. Yet again one needs to be cautious about the improved performance, and a number of unresolved issues need to be addressed by future work. It remains possible that SL is not used at all in actual word segmentation. Once the one-word-one-stress principle is built in, we may consider a model that does not use any statistics, hence avoiding the computational cost that is likely to be considerable. (While we don't know how infants keep track of TPs, there are clearly quite some work to do. Syllables in English number in the thousands; now take the quadratic for the potential number of pair-wise TPs.) It simply stores previously extracted words in the memory to bootstrap new words. Young children's familiar segmentation errors–"I was have" from be-have, "hiccing up" from hicc-up, "two dults", from a-dult–suggest that this process does take place. Moreover, there is evidence that 8-month-old infants can store familiar sounds in the memory [22]. And finally, there are plenty of single-word utterances–up to 10% [23]–that give many words for free. The implementation of a purely symbolic learner that recycles known words yields even better performance: a precision of 81.5% and recall of 90.1%.

5 Conclusion

Further work, both experimental and computational, will need to address a few pressing questions, in order to gain a better assessment of the relative contribution of SL and UG to language acquisition. These include, more pertinent to the problem of word segmentation:

- Can statistical learning be used in the acquisition of language-specific phonotactics, a prerequisite to syllabification and a prelude to word segmentation?
- Given that prosodic constraints are critical for the success of SL in word segmentation, future work needs to quantify the availability of stress information in spoken corpora.
- Can further experiments, carried over realistic linguistic input, further tease apart the multiple strategies used in word segmentation [14]? What are the psychological mechanisms (algorithms) that integrate these strategies?
- How does word segmentation, statistical or otherwise, work for agglutinative (e.g., Turkish) and polysynthetic languages (e.g. Mohawk), where the division between words, morphology, and syntax is quite different from more clear-cut cases like English?

Computational modeling can make explicit the balance between statistics and UG, and are in the same vein as the recent findings [24] on when/where SL is effective/possible. UG can help SL by providing specific constraints on its application, and modeling may raise new questions for further experimental studies. In related work [6,7], we have augmented traditional theories of UG–derivational phonology, and the Principles and Parameters framework–with a component of statistical learning, with novel and desirable consequences. Yet in all cases, statistical learning, while perhaps domain-general, is constrained by what appears to be innate and domain-specific knowledge of linguistic structures, such that learning can operate on specific aspects of the input evidence

References

1. Gold, E. M. (1967). Language identification in the limit. *Information and Control*, 10:447-74.
2. Valiant, L. (1984). A theory of the learnable. *Communication of the ACM*. 1134-1142.

3. Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Berlin: Springer.

4. Chomsky, N. (1959). Review of Verbal Behavior by B.F. Skinner. *Language*, 35, 26-57.

5. Chomsky, N. (1975). *Reflections on Language*. New York: Pantheon.

6. Yang, C. D. (1999). A selectionist theory of language development. In *Proceedings of 37th Meeting of the Association for Computational Linguistics*. Maryland, MD. 431-5.

7. Yang, C. D. (2002). *Knowledge and Learning in Natural Language*. Oxford: Oxford University Press.

8. Saffran, J.R., Aslin, R.N., & Newport, E.L. (1996). Statistical learning by 8-month old infants. *Science*, 274, 1926-1928.

9. Gomez, R.L., & Gerken, L.A. (1999). Artificial grammar learning by one-year-olds leads to specific and abstract knowledge. *Cognition*, 70, 109-135.

10. Saffran, J.R., Johnson, E.K., Aslin R.N. & Newport, E.L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, 70, 27-52.

11. Fiser, J., & Aslin, R.N. (2002). Statistical learning of new visual feature combinations by infants. *PNAS*, 99, 15822-6.

12. Hauser, M., Newport, E.L., & Aslin, R.N. (2001). Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins. *Cognition*, 78, B41-B52.

13. Bates, E., & Elman, J. (1996). Learning rediscovered. *Science*, 274, 1849-50.

14. Jusczyk, P.W. (1999). How infants begin to extract words from speech. *Trends in Cognitive Sciences*, 3, 323-8.

15. Chomsky, N. (1955/1975). *The Logical Structure of Linguistic Theory*. Manuscript, Harvard University and Massachusetts Institute of Technology. Published in 1975 by New York: Plenum.

16. Saffran, J.R., Aslin, R.N., & Newport, E.L. (1997). Letters. *Science*, 276, 1177-1181

17. Crain, S., & Nakayama, M. (1987). Structure dependency in grammar formation. *Language*, 63:522-543.

18. Bijeljiac-Babic, R., Bertoncini, J., & Mehler, J. (1993). How do four-day-old infants categorize multisyllabic utterances. *Developmental psychology*, 29, 711-21.

19. MacWhinney, B. (1995). *The CHILDES Project: Tools for Analyzing Talk*. Hillsdale: Lawrence Erlbaum.

20. Brent, M. (1999). Speech segmentation and word discovery: a computational perspective. *Trends in Cognitive Science*, 3, 294-301.

21. Johnson, E.K. & Jusczyk, P.W. (2001) Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language*, 44, 1-20.

22. Jusczyk, P. W., & Hohne, E. A. (1997). Infants' memory for spoken words. *Science*, 277, 1984-6.

23. Brent, M.R., & Siskind, J.M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, 81, B33-44.

24. Newport, E.L., & Aslin, R.N. (2004). Learning at a distance: I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, 48, 127-62.

Modelling syntactic development in a cross-linguistic context

Fernand GOBET
Centre for Cognition and Neuroimaging
Department of Human Sciences
Brunel University
Uxbridge UB8 3PH, UK
Fernand.Gobet@Brunel.ac.uk

Daniel FREUDENTHAL
Julian M. PINE
School of Psychology
University of Nottingham
Nottingham NG7 2RD, UK
Daniel.Freudenthal@psyc.nott.ac.uk
Julian.Pine@psyc.nott.ac.uk

Abstract

Mainstream linguistic theory has traditionally assumed that children come into the world with rich innate knowledge about language and grammar. More recently, computational work using distributional algorithms has shown that the information contained in the input is much richer than proposed by the nativist approach. However, neither of these approaches has been developed to the point of providing detailed and quantitative predictions about the developmental data. In this paper, we champion a third approach, in which computational models learn from naturalistic input and produce utterances that can be directly compared with the utterances of language-learning children. We demonstrate the feasibility of this approach by showing how MOSAIC, a simple distributional analyser, simulates the optional-infinitive phenomenon in English, Dutch, and Spanish. The model accounts for young children's tendency to use both correct finites and incorrect (optional) infinitives in finite contexts, for the generality of this phenomenon across languages, and for the sparseness of other types of errors (e.g., word order errors). It thus shows how these phenomena, which have traditionally been taken as evidence for innate knowledge of Universal Grammar, can be explained in terms of a simple distributional analysis of the language to which children are exposed.

1 Introduction

Children acquiring the syntax of their native language are faced with a task of considerable complexity, which they must solve using only noisy and potentially inconsistent input. Mainstream linguistic theory has addressed this 'learnability problem' by proposing the nativist hypothesis that children come into the world with rich innate knowledge about language and grammar (Chomsky, 1981; Piattelli-Palmarini, 2002; Pinker, 1984). However, there is also strong empirical evidence that the amount of information present in the input is considerably greater than has traditionally been assumed by the nativist approach. In particular, computer simulations have shown that a distributional analysis of the statistics of the input can provide a significant amount of syntactic information (Redington & Chater, 1997).

One limitation of the distributional approach is that analyses have rarely been done with naturalistic input (e.g. mothers' child-directed speech) and have so far not been linked to the detailed analysis of a linguistic phenomenon found in human data, (e.g., Christiansen & Chater, 2001). Indeed, neither the nativist nor the distributional approach has been developed to the point of providing detailed and quantitative predictions about the developmental dynamics of the acquisition of language. In order to remedy this weakness, our group has recently been exploring a different approach. This approach, which we think is a more powerful way of understanding how children acquire their native language, has involved developing a computational model (MOSAIC; Model Of Syntax Acquisition In Children) that learns from naturalistic input, and produces utterances that can be directly compared with the utterances of language-learning children. This makes it possible to derive quantitative predictions about empirical phenomena observed in children learning different languages and about the developmental dynamics of these phenomena.

MOSAIC, which is based upon a simple distributional analyser, has been used to simulate a number of phenomena in language acquisition. These include: the verb-island phenomenon (Gobet & Pine, 1997; Jones, Gobet, & Pine, 2000); negation errors in English (Croker, Pine, & Gobet, 2003); patterns of pronoun case marking error in English (Croker, Pine, & Gobet, 2001); patterns of subject omission error in English (Freudenthal, Pine, & Gobet, 2002b); and the optional-infinitive phenomenon (Freudenthal, Pine, & Gobet, 2001, 2002a, 2003). MOSAIC has also been used to simulate data from three different languages (English, Dutch, and Spanish), which has helped us to

understand how these phenomena are affected by differences in the structure of the language that the child is learning.

In this paper, we illustrate our approach by showing how MOSAIC can account in detail for the 'optional-infinitive phenomenon' in two languages (English and Dutch) and its quasi-absence in a third language (Spanish). This phenomenon is of particular interest as it has generally been taken to reflect innate grammatical knowledge on the part of the child (Wexler, 1994, 1998).

We begin by highlighting the theoretical challenges faced in applying our model to data from three different languages. Then, after describing the optional-infinitive phenomenon, we describe MOSAIC, with an emphasis on the mechanisms that will be crucial in explaining the empirical data. We then consider the data from the three languages, and show to what extent the same model can simulate these data. When dealing with English, we describe the methods used to collect and analyse children's data in some detail. While these details may seem out of place in a conference on computational linguistics, we emphasise that they are critical to our approach: first, our approach requires fine-grained empirical data, and, second, the analysis of the data produced by the model is as close as possible to that used with children's data. We conclude by discussing the implications of our approach for developmental psycholinguistics.

2 Three languages: three challenges

The attempt to use MOSAIC to model data in three different languages involves facing up to a number of challenges, each of which is instructive for different reasons. An obvious problem when modelling English data is that English has an impoverished system of verb morphology that makes it difficult to determine which form of the verb a child is producing in any given utterance. This problem militates against conducting objective quantitative analyses of children's early verb use and has resulted in there being no detailed quantitative description of the developmental patterning of the optional infinitive phenomenon in English (in contrast to other languages like Dutch). We have addressed this problem by using exactly the same (automated) methods to classify the utterances produced by the child and by the model. These methods, which do not rely on the subjective judgment of the coder (e.g. on Bloom's, 1970, method of rich interpretation) proved to be sufficiently powerful to capture the development of the optional infinitive in English, and to do so at a relatively fine level of detail.

One potential criticism of these simulations of English is that we may have tuned the model's parameters in order to optimise the goodness of fit to the human data. An obvious consequence of overfitting the data in this way would be that MOSAIC's ability to simulate the phenomenon would break down when the model was applied to a new language. The simulations of Dutch show that this is not the case: with this language, which has a richer morphology than English, the model was still able to reproduce the key characteristics of the optional-infinitive stage.

Spanish, the syntax of which is quite different to English and Dutch, offered an even more sensitive test of the model's mechanisms. The Dutch simulations relied heavily on the presence of compound finites in the child-directed speech used as input. However, although Spanish child-directed speech has a *higher* proportion of compound finites than Dutch, children learning Spanish produce optional-infinitive errors *less* often than children learning Dutch. Somewhat counter-intuitively, the simulations correctly reproduce the relative scarcity of optional-infinitive errors in Spanish, showing that the model is sensitive to subtle regularities in the way compound finites are used in Dutch and Spanish.

3 The optional-infinitive phenomenon

Between two and three years of age, children learning English often produce utterances that appear to lack inflections, such as past tense markers or third person singular agreement markers. For example, children may produce utterances as:

(1a) *That go there**
(2a) *He walk home**

instead of:

(1b) *That goes there*
(2b) *He walked home*

Traditionally, such utterances have been interpreted in terms of absence of knowledge of the appropriate inflections (Brown, 1973) or the dropping of inflections as a result of performance limitations in production (L. Bloom, 1970; P. Bloom, 1990; Pinker, 1984; Valian, 1991). More recently, however, it has been argued that they reflect the child's optional use of (root) infinitives (e.g. *go*) in contexts where a finite form (e.g. *went*, *goes*) is obligatory in the adult language (Wexler, 1994, 1998).

This interpretation reflects the fact that children produce (root) infinitives not only in English, where the infinitive is a zero-marked form, but also in languages such as Dutch where the infinitive carries its own infinitival marker. For instance,

children learning Dutch may produce utterances such as:

(3a) *Pappa eten** *(Daddy to eat)*
(4a) *Mamma drinken** *(Mummy to drink)*

instead of:

(3b) *Pappa eet* *(Daddy eats)*
(4b) *Mamma drinkt* *(Mummy drinks)*

The optional infinitive phenomenon is particularly interesting as it occurs in languages that differ considerably in their underlying grammar, and is subject to considerable developmental and cross-linguistic variation. It is also intriguing because children in the optional infinitive stage typically make few other grammatical errors. For example, they make few errors in their use of the basic word order of their language: English-speaking children may say *he go*, but not *go he*.

Technically, the optional infinitive phenomenon revolves around the notion of 'finiteness'. Finite forms are forms that are marked for Tense and/or Agreement (e.g. *went, goes*). Non-finite forms are forms that are not marked for Tense or Agreement. This includes the infinitive form (*go*), the past participle (*gone*), and the progressive participle (*going*). In English, finiteness marking increases with development: as they grow older, children produce an increasing proportion of unambiguous finite forms.

4 Description of the model

MOSAIC is a computational model that analyses the distributional characteristics present in the input. It learns to produce increasingly long utterances from naturalistic (child-directed) input, and produces output consisting of actual utterances, which can be directly compared to children's speech. This allows for a direct comparison of the output of the model at different stages with the children's developmental data.

The model learns from text-based input (i.e., it is assumed that the phonological stream has been segmented into words). Utterances are processed in a left to right fashion. MOSAIC uses two learning mechanisms, based on discrimination and generalisation, respectively. The first mechanism grows an n-ary discrimination network (Feigenbaum & Simon, 1984; Gobet et al., 2001) consisting of nodes connected by test links. Nodes encode single words or phrases. Test links encode the difference between the contents of consecutive nodes. (Figure 1 illustrates the structure of the type of discrimination net used.) As the model sees more and more input, the number of nodes and links increases, and so does the amount of information held in the nodes, and, as a consequence, the average length of the phrases it can output. The node creation probability (NCP) is computed as follows:

$$NCP = (N / M)^L$$

where M is a parameter arbitrarily set to 70,000 in the English and Spanish simulations, N = number of nodes in the net ($N \leq M$), and L = length of the phrase being encoded. Node creation probability is thus dependent both on the length of the utterance (longer utterances are less likely to yield learning) and on the amount of knowledge already acquired. In a small net, learning is slow. When the number of nodes in the net increases, the node creation probability increases and, as a result, the learning rate also increases. This is consistent with data showing that children learn new words more easily as they get older (Bates & Carnavale, 1993).

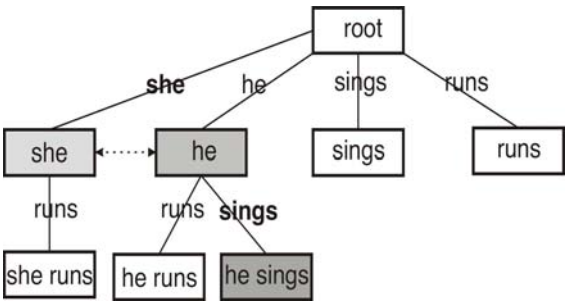


Figure 1: Illustration of a MOSAIC discrimination net. The Figure also illustrates how an utterance can be generated. Because *she* and *h*e have a generative link, the model can output the novel utterance *she sings*. (For simplicity, preceding context is ignored in this Figure.)

While the first learning mechanism is based on discrimination, the second is based on generalisation. When two nodes share a certain percentage (set to 10% for these simulations) of nodes (phrases) following and preceding them, a new type of link, a *generative link* is created between them (see Figure 1 for an example). Generative links connect words that have occurred in similar contexts in the input, and thus are likely to be of the same word class. As no linguistic constructs are given to the model, the development of approximate linguistic classes, such as those of noun or verb, is an emergent property of the distributional analysis of the input. An important feature of MOSAIC is that the creation and removal of generative links is dynamic. Since new nodes are constantly being created in the network, the percentage overlap between two nodes varies over time; as a

consequence, a generative link may drop below the threshold and so be removed.

The model generates output by traversing the network and outputting the contents of the visited links. When the model traverses test links only, the utterances it produces must have been present in the input. Where the model traverses generative links during output, novel utterances can be generated. An utterance is generated only if its final word was the final word in the utterance when it was encoded (this is accomplished by the use of an *end marker*). Thus, the model is biased towards generating utterances from sentence final position, which is consistent with empirical data from language-learning children (Naigles & Hoff-Ginsberg, 1998; Shady & Gerken, 1999; Wijnen, Kempen, & Gillis, 2001).

5 Modelling the optional-infinitive phenomenon in English

Despite the theoretical interest of the optional-infinitive phenomenon, there is, to our knowledge, no quantitative description of the developmental dynamics of the use of optional infinitives in English, with detail comparable to that provided in other languages, such as Dutch (Wijnen et al., 2001). The following analyses fill this gap.

5.1 Children's data: Methods

We selected the speech of two children (Anne, from 1 year 10 months to 2 years 9 months; and Becky, from 2 years to 2 years 11 months). These data were taken from the Manchester corpus (Theakston, Lieven, Pine, & Rowland, 2001), which is available in the CHILDES data base (MacWhinney, 2000). Recordings were made twice every three weeks over a period of one year and lasted for approximately one hour per session.

Given that optional-infinitive phenomena are harder to identify in English than in languages such as Dutch or German (due to the relatively low number of unambiguous finite forms), the analysis focused on the subset of utterances that contain a verb with *he*, *she*, *it*, *this* (*one*), or *that* (*one*) as its subject. Restricting the analysis in this way avoids utterances such as *I go*, which could be classified both as non-finite and finite, and therefore makes it possible to more clearly separate non-finites, simple finites, compound finites, and ambiguous utterances.

Identical (automatic) analyses of the data and model were carried out in a way consistent with previous work on Dutch (Wijnen et al., 2001). Utterances that had the copula (i.e., forms of the verb *to be*) as a main verb were removed. Utterances that contained a non-finite form as the only verb were classified as non-finites. Utterances with an unambiguous finite form (*walks*, *went*) were counted as finite, while those containing a finite verb form plus a non-finite form (*has gone*) were classified as compound finites. The remaining utterances were classified as ambiguous and counted separately; they contained an ambiguous form (such as *bought* in *he bought*) as the main verb, which can be classified either as a finite past tense form or as a (non-finite) perfect participle (in the phrase *he bought*, the word *has* may have been omitted).

5.2 Children's data: Results

The children's speech was partitioned into three developmental stages, defined by mean length of utterance (MLU). The resulting distributions, portrayed in Figure 2, show that the proportion of non-finites decreases as a function of MLU, while the proportion of compound finites increases. There is also a slight increase in the proportion of simple finites, although this is much less pronounced than the increase in the proportion of compound finites.

5.3 Simulations

The model received as input speech from the children's respective mothers. The size of the input was 33,000 utterances for Anne's model, and 27,000 for Becky's model. Note that, while the analyses are restricted to a subset of the children's corpora, the entire mothers' corpora were used as input during learning. The input was fed through the model several times, and output was generated after every run of the model, until the MLU of the output was comparable to that of the end stage in the two children. The output files were then compared to the children's data on the basis of MLU.

The model shows a steady decline in the proportion of non-finites as a function of MLU coupled with a steady increase in the proportion of compound finites (Figure 3). On average, the model's production of optional infinitives in third person singular contexts drops from an average of 31.5% to 16% compared with 47% to 12.5% in children. MOSAIC thus provides a good fit to the developmental pattern in the children's data (not including the 'ambiguous' category: $r^2 = .65$, $p < .01$, RMSD = 0.096 for Anne and her model; $r^2 = .88$, $p < .001$, RMSD = 0.104 for Becky and her model). One obvious discrepancy between the model's and the children's output is that both models at MLU 2.1 produce too many simple finite utterances. Further inspection of these utterances reveals that they contain a relatively high proportion of finite modals such as *can* and *will* and finite forms of the dummy modal *do* such as *does* and *did*. These forms are unlikely to be used as the only verb in children's early utterances as their function is to

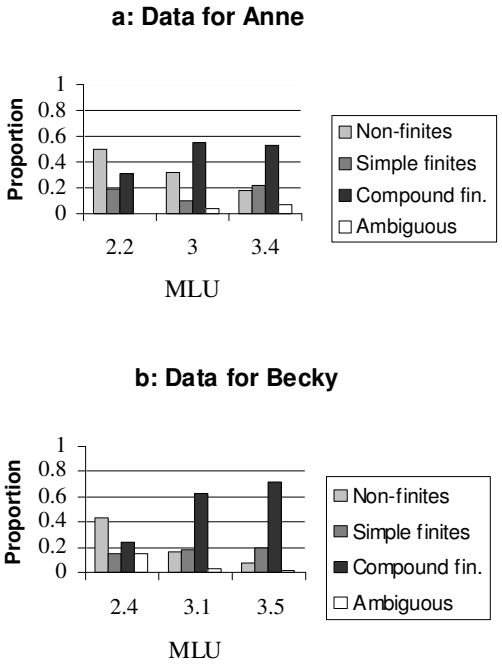


Figure 2: Distribution of non-finites, simple finites, compound finites, and ambiguous utterances for Anne and Becky as a function of developmental phase. Only utterances with *he*, *she*, *it*, *that* (*one*), or *this* (*one*) as a subject are included.

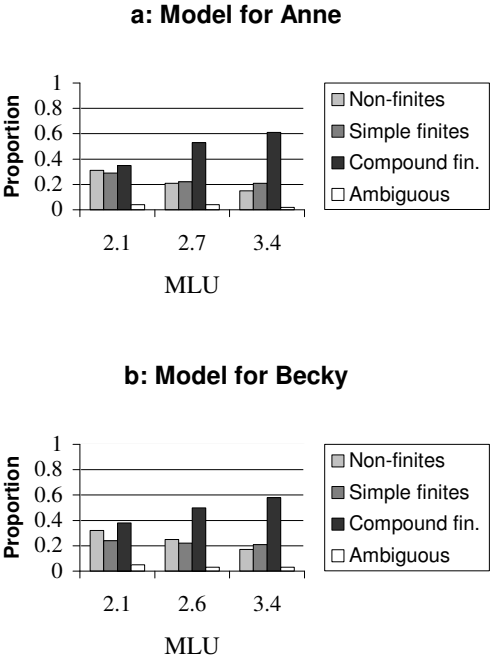


Figure 3: Distribution of non-finites, simple finites, compound finites, and ambiguous utterances for the models of Anne and Becky as a function of developmental phase. Only utterances with *he*, *she*, *it*, *that* (*one*), or *this* (*one*) as a subject are included.

modulate the meaning of the main verb rather than to encode the central relational meaning of the sentence.

An important reason why MOSAIC accounts for the data is that it is biased towards producing sentence final utterances. In English, non-finite utterances can be learned from compound finite questions in which finiteness is marked on the auxiliary rather than the lexical verb. A phrase like *He walk home* can be learned from *Did he walk home?*, and a phrase like *That go there* can be learned from *Does that go there?* As MLU increases, the relative frequency of non-finite utterances in the output decreases, because the model learns to produce more and more of the compound finite utterances from which these utterances have been learned. MOSAIC therefore predicts that as the proportion of non-finite utterances decreases, there will be a complementary increase in the proportion of compound finites.

6 Modelling optional infinitives in Dutch

Children acquiring Dutch seem to use a larger proportion of non-finite verbs in finite contexts (e.g., *hij lopen*, *bal trappen*) than children learning English. Thus, in Dutch, a very high percentage of children's early utterances with verbs (about 80%) are optional-infinitive errors. This percentage decreases to around 20% by MLU 3.5 (Wijnen, Kempen & Gillis, 2001).

As in English, optional infinitives in Dutch can be learned from compound finites (auxiliary/modal + infinitive). However, an important difference between English and Dutch is that in Dutch verb position is dependent on finiteness. Thus, in the simple finite utterance *Hij drinkt koffie (He drinks coffee)* the finite verb form *drinkt* precedes its object argument *koffie* whereas in the compound finite utterance *Hij wil koffie drinken (He wants coffee drink),* the non-finite verb form *drinken* is restricted to utterance final position and is hence preceded by its object argument: *koffie*. Interestingly, children appear to be sensitive to this feature of Dutch from very early in development and MOSAIC is able to simulate this sensitivity. However, the fact that verb position is dependent on finiteness in Dutch also means that whereas non-finite verb forms are restricted to sentence final position, finite verb forms tend to occur earlier in the utterance. MOSAIC therefore simulates the very high proportion of optional infinitives in early child Dutch as a function of the interaction between its utterance final bias and increasing MLU. That is, the high proportion of non-finites early on is explained by the fact that the model mostly produces sentence-final phrases, which, as a result of

Dutch grammar, have a large proportion of non-finites.

As shown in Figure 4, the model's production of optional infinitives drops from 69% to 28% compared with 77% to 18% in the data of the child on whose input data the model had been trained. In these simulations, the input data consisted of a sample of approximately 13,000 utterances of child-directed speech. Because of the lower input size, the M used in the NCP formula was set to 50,000.

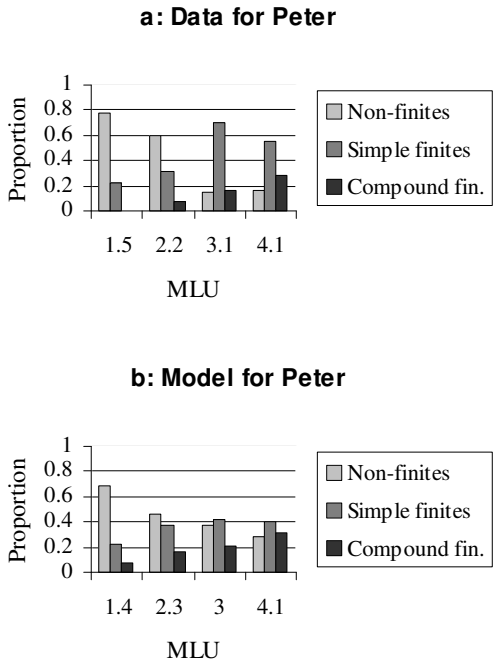


Figure 4: Distribution of non-finites, simple finites and compound finites for Peter and his model, as a function of developmental phase.

7 Modelling optional infinitives in Spanish

Wexler (1994, 1998) argues that the optional-infinitive stage does not occur in pro-drop languages, that is, languages like Spanish in which verbs do not require an overt subject. Whether MOSAIC can simulate the low frequency of optional-infinitive errors in early child Spanish is therefore of considerable theoretical interest, since the ability of Wexler's theory to explain cross-linguistic data is presented as one of its main strengths. Note that simulating the pattern of finiteness marking in early child Spanish is not a trivial task. This is because although optional-infinitive errors are much less common in Spanish than they are in Dutch, compound finites are actually more common in Spanish child-directed speech than they are in Dutch child-directed speech (in the corpora we have used, they make up 36% and 30% of all parents' utterances including verbs, respectively).

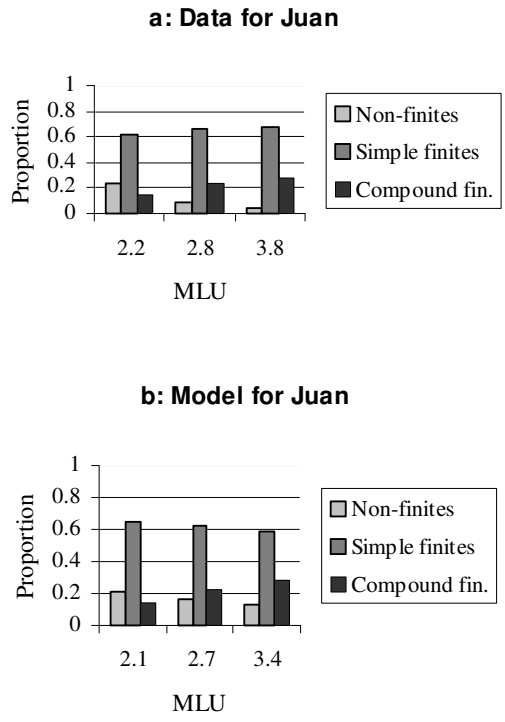


Figure 5: Distribution of non-finites, simple finites and compound finites for Juan and his model, as a function of developmental phase.

Figure 5a shows the data for a Spanish child, Juan (Aguado Orea & Pine, 2002), and Figure 5b the outcome of the simulations run using MOSAIC. The parental corpus used as input consisted of about 27,000 utterances. The model's production of optional infinitives drops from 21% to 13% compared with 23% to 4% in the child. Both the child and the model show a lower proportion of optional-infinitive errors than in Dutch. The presence of (some rare) optional-infinitive errors in the model's output is explained by the same mechanism as in English and Dutch: a bias towards learning the end of utterances. For example, the input *¿Quieres beber café? (Do you want to drink coffee?)* may later lead to the production of *beber café.* But why does the model produce so few optional-infinitive errors in Spanish when the Spanish input data contain so many compound finites? The answer is that finite verb forms are much more likely to occur in utterance final position in Spanish than they are in Dutch, which makes them much easier to learn.

8 Conclusion

In this paper, we have shown that the same simple model accounts for the data in three languages that differ substantially in their underlying structure. To our knowledge, this is the only model of language acquisition which simultaneously (1) learns from naturalistic input (actual child-directed utterances), where the statistics and frequency distribution of the input are similar to that experienced by children; (2) produces actual utterances, which can be directly compared to those of children; (3) has a developmental component; (4) accounts for speech generativity and increasing MLU; (5) makes quantitative predictions; and (6) has simulated phenomena from more than one language.

An essential feature of our approach is to limit the number of degrees of freedom in the simulations. We have used an identical model for simulating the same class of phenomena in three languages. The method of data analysis was also the same, and, in all cases, the model's and child's output were coded automatically and identically. The use of realistic input was also crucial in that it guaranteed that cross-linguistic differences were reflected in the input.

The simulations showed that simple mechanisms were sufficient for obtaining a good fit to the data in three different languages, in spite of obvious syntactic differences and very different proportions of optional-infinitive errors. The interaction between a sentence final processing bias and increasing MLU enabled us to capture the reason why English, Dutch and Spanish offer different patterns of optional-infinitive errors: the difference in the relative position of finites and non-finites is larger in Dutch than in English, and Spanish verbs are predominantly finite. We suggest that any model that learns to produce progressively longer utterances from realistic input, and in which learning is biased towards the end of utterances, will simulate these results.

The production of actual utterances (as opposed to abstract output) by the model makes it possible to analyse the output with respect to several (seemingly) unrelated phenomena, so that the nontrivial predictions of the learning mechanisms can be assessed. Thus, the same output can be utilized to study phenomena such as optional-infinitive errors (as in this paper), evidence for verb-islands (Jones et al., 2000), negation errors (Croker et al., 2003), and subject omission (Freudenthal et al., 2002b). It also makes it possible to assess the relative importance of factors such as increasing MLU that are implicitly assumed by many current theorists but not explicitly factored into their models.

An important contribution of Wexler's (1994, 1998) nativist theory of the optional-infinitive stage has been to provide an integrated account of the different patterns of results observed across languages, of the fact that children use both correct finite forms and incorrect (optional) infinitives, and of the scarcity of other types of errors (e.g. verb placement errors). His approach, however, requires a complex theoretical apparatus to explain the data, and does not provide any quantitative predictions. Here, we have shown how a simple model with few mechanisms and no free parameters can account for the same phenomena not only qualitatively, but also quantitatively.

The simplicity of the model inevitably means that some aspects of the data are ignored. Children learning a language have access to a range of sources of information (e.g. phonology, semantics), which the model does not take into consideration. Also, generating output from the model means producing everything the model can output. Clearly, children produce only a subset of what they can say. Furthermore, any rote-learned utterance that the model produces early on in its development will continue to be produced during the later stages. This inability to unlearn is clearly a weakness of the model, but one that we hope to correct in subsequent research.

The results clearly show that the interaction between a simple distributional analyser and the statistical properties of naturalistic child-directed speech can explain a considerable amount of the developmental data, without the need to appeal to innate linguistic knowledge. The fact that such a relatively simple model provides such a good fit to the developmental data in three languages suggests that (1) aspects of children's multi-word speech data such as the optional-infinitive phenomenon do not necessarily require a nativist interpretation, and (2) nativist theories of syntax acquisition need to pay more attention to the role of input statistics and increasing MLU as determinants of the shape of the developmental data.

9 Acknowledgements

This research was supported by the Leverhulme Trust and the Economic and Social Research Council. We thank Javier Aguado-Orea for sharing his corpora on early language acquisition in Spanish children and for discussions on the Spanish simulations.

References

Aguado-Orea, J., & Pine, J. M. (2002). Assessing the productivity of verb morphology in early child Spanish. Paper presented at *the IX International Congress for the Study of Child Language*, Madison, Wisconsin.

Bates, E., & Carnavale, G. F. (1993). New directions in research on child development. *Developmental Review, 13*, 436-470.

Bloom, L. (1970). *Language development: Form and function in emerging grammars*. Cambridge, MA: MIT Press.

Bloom, P. (1990). Subjectless sentences in child language. *Linguistic Inquiry, 21*, 491-504.

Brown, R. (1973). *A first language*. Boston, MA: Harvard University Press.

Chomsky, N. (1981). *Lectures on government and binding*. Dordrecht, NL: Foris.

Christiansen, M. H., & Chater, N. (2001). Connectionist psycholinguistics: Capturing the empirical data. *Trends in Cognitive Sciences, 5*, 82-88.

Croker, S., Pine, J. M., & Gobet, F. (2001). Modelling children's case-marking errors with MOSAIC. In E. M. Altmann, A. Cleeremans, C. D. Schunn & W. D. Gray (Eds.), *Proceedings of the 4th International Conference on Cognitive Modeling* (pp. 55-60). Mahwah, NJ: Erlbaum.

Croker, S., Pine, J. M., & Gobet, F. (2003). Modelling children's negation errors using probabilistic learning in MOSAIC. In F. Detje, D. Dörner & H. Schaub (Eds.), *Proceedings of the Fifth International Conference on Cognitive Modeling* (pp. 69-74). Bamberg:Universitäts-Verlag.

Feigenbaum, E. A., & Simon, H. A. (1984). EPAM-like models of recognition and learning. *Cognitive Science, 8*, 305-336.

Freudenthal, D., Pine, J. M., & Gobet, F. (2001). Modelling the optional infinitive stage in MOSAIC: A generalisation to Dutch. In E. M. Altmann, A. Cleeremans, C. D. Schunn & W. D. Gray (Eds.), *Proceedings of the 4th International Conference on Cognitive Modeling* (pp. 79-84). Mahwah, NJ: Erlbaum.

Freudenthal, D., Pine, J. M., & Gobet, F. (2002a). Modelling the development of Dutch optional infinitives in MOSAIC. In W. Gray & C. D. Schunn (Eds.), *Proceedings of the 24th Annual Meeting of the Cognitive Science Society* (pp. 322-327). Mahwah, NJ: Erlbaum.

Freudenthal, D., Pine, J. M., & Gobet, F. (2002b). Subject omission in children's language: The case for performance limitations in learning. In W. Gray & C. D. Schunn (Eds.), *Proceedings of the 24th Annual Meeting of the Cognitive Science Society* (pp. 328-333). Mahwah, NJ: Erlbaum.

Freudenthal, D., Pine, J. M., & Gobet, F. (2003). The role of input size and generativity in simulating language acquisition. In F. Schmalhofer, R. M. Young & G. Katz (Eds.), *Proceedings of EuroCogSci 03: The European Cognitive Science Conference 2003* (pp. 121-126). Mahwah NJ: Erlbaum.

Gobet, F., Lane, P. C. R., Croker, S., Cheng, P. C. H., Jones, G., Oliver, I., & Pine, J. M. (2001). Chunking mechanisms in human learning. *Trends in Cognitive Sciences, 5*, 236-243.

Gobet, F., & Pine, J. M. (1997). Modelling the acquisition of syntactic categories. In *Proceedings of the 19th Annual Meeting of the Cognitive Science Society* (pp. 265-270). Hillsdale, NJ: Erlbaum.

Jones, G., Gobet, F., & Pine, J. M. (2000). A process model of children's early verb use. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the Twenty Second Annual Meeting of the Cognitive Science Society* (pp. 723-728). Mahwah, N.J.: Erlbaum.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Mahwah, NJ: Erlbaum.

Naigles, L., & Hoff-Ginsberg, E. (1998). Why are some verbs learned before other verbs: Effects of input frequency and structure on children's early verb use. *Journal of Child Language, 25*, 95-120.

Piattelli-Palmarini, M. (2002). The barest essentials. *Nature, 416*, 129.

Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.

Redington, M., & Chater, N. (1997). Probabilistic and distributional approaches to language acquisition. *Trends in Cognitive Sciences, 1*, 273-279.

Shady, M., & Gerken, L. (1999). Grammatical and caregiver cue in early sentence comprehension. *Journal of Child Language, 26*, 163-176.

Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language, 28*, 127-152.

Valian, V. (1991). Syntactic subjects in the early speech of American and Italian children. *Cognition, 40*, 21-81.

Wexler, K. (1994). Optional infinitives, head movement and the economy of derivation in child grammar. In N. Hornstein & D. Lightfoot (Eds.), *Verb movement*. Cambridge, MA: Cambridge University Press.

Wexler, K. (1998). Very early parameter setting and the unique checking constraint: A new explanation of the optional infinitive stage. *Lingua, 106*, 23-79.

Wijnen, F., Kempen, M., & Gillis, S. (2001). Root infinitives in Dutch early child language. *Journal of Child Language, 28*, 629-660.

A Computational Model of Emergent Simple Syntax: Supporting the Natural Transition from the One-Word Stage to the Two-Word Stage.

Kris Jack, Chris Reed, and Annalu Waller
Division of Applied Computing,
University of Dundee,
Dundee, Scotland, DD1 4HN
[kjack | creed | awaller]@computing.dundee.ac.uk

Abstract

This paper introduces a system that simulates the transition from the one-word stage to the two-word stage in child language production. Two-word descriptions are syntactically generated and compete against one-word descriptions from the outset. Two-word descriptions become dominant as word combinations are repeatedly recognised, forming syntactic categories; resulting in an emergent simple syntax. The system demonstrates a similar maturation as children as evidenced by phenomena such as overextensions and mismatching, and the use of one-word descriptions being replaced by two-word descriptions over time.

1 Introduction

Studies of first language acquisition in children have documented general stages in linguistic development. Neither the trigger nor the mechanism that takes a child from one stage to the next are known. Stages arise gradually with no precise start or end points, overlapping one another (Ingram, 1989).

The aim of this research is to develop a system that autonomously acquires conceptual representations of individual words (the "one-word stage") and also, simultaneously, is capable of developing representations of valid multi-word structures i.e. simple syntax (the "two-word stage"). Two-word descriptions are expected to emerge as a result of the system state and not be artificially triggered.

The system accepts sentences containing a maximum of two words. It is designed to be scalable, allowing larger, more natural sentence sizes also. System input is therefore a mixture of both one-word and two-word sentences. The system is required to produce valid descriptions, particularly in the two-word stage. Rules that enforce syntactic order, and allow for the production of semantically correct descriptions from novel concepts, are desirable.

This paper is sectioned as follows; pre-one-word stage linguistic abilities in children are briefly discussed to explain why initial system functionality assumptions are made; the defining characteristics of both the one-word stage and two-word stage in children are introduced as possible benchmarks for the system; a detailed description of system design and implementation with examples of the learning process and games played by the system are presented; a discussion of current results along with their possible implications follows; a brief review of related works that have influenced this research, citing major influences; the direction and aims of future research is described briefly; and finally, conclusions are drawn.

2 Pre-One-Word Stage Children

Linguistic abilities can be found in children prior to word production. In terms of comprehension, children can distinguish between their mother's voice and a stranger's voice, male and female voices, and sentences spoken in their mother's native language and sentences spoken in a different language. They also show categorical perception to voice, can use formant transition information to mark articulation, and show intonation sensitivity (Pinker, 1994, Jusczyk, 1999).

In terms of production, children produce noises, such as discomfort noises (0-2 months), comfort noises (2-4 months), and "play" vocally with pitch and loudness variations (4-7 months) (Pinker, 1994). The babbling stage (6-8 months) is characterised with the production of recognisable syllables. The syllables are often repeated, such as [mamama] and [papapa], with the easiest to produce sounds often being associated with members of the family (Jakobson, 1971).

From this evidence it is reasonable to draw conclusions about linguistic abilities in the young child that can be used to frame assumptions for use in the system. It is assumed that the system can receive and produce strings that can be broken down into their component words. These words can be compared and equalities can be detected.

3 One-Word Stage and Two-Word Stages

The system is required to produce one-word descriptions in early stages that develop into two-word descriptions, where appropriate, in latter stages.. The recognition of each stage is based on the number of words that the system uses at a particular point. In children, the one and two-word stages have notable features.

The one-word, or holophrastic, stage (9-18 months), is characterised by one-word vocalisations that are consistently associated with concepts. These concepts can be either concrete or abstract, such as "mama", referring to the concrete concept of the child's mother, and "more", an abstract concept which can be applied in a variety of situations (Piaget, 1960).

Two phenomena that occur during this stage are underextensions and overextensions. An underextension is the formation of a word to concept association that is too narrow, such as "dog" referring to only the family dog. Overextension, similarly, is an association that is too broad, such as "dog" referring to all four legged animals. Mismatches, or idiosyncratic referencing also occur, resulting in a word being associated with an unrelated concept, such as "dog" referring to a table (Pinker, 1994). These associations change over time.

The two-word stage (18-24 months) introduces simple syntax into the child's language faculty. Children appear to determine the most important words in a sentence and, almost all of the time, use them in the same order as an adult would (Gleitman and Newport, 1995). Brown (1973) defines a typology to express semantic relations in the two-word stage. It contains ten sets of relations, but only one will be considered in this paper; attribute + entity ("red circle"). During this stage, children already demonstrate a three word comprehension level (Tomasello and Kruger, 1992). The concepts relating to their sentences may therefore be more detailed than the phrases themselves.

The system is expected to make the transition from the one-word stage to the two-word stage without changes to the functionality of the system. Once the system begins to run, input is restricted to that of sensory (concept based) and vocal (string representation) data.

4 System Design and Implementation

4.1 Introduction

The system is designed to learn phrase-to-concept associations and demonstrate it through playing games: a guessing game and a naming game. Games are often used to test, and encourage system learning (Steels and Kaplan, 2001). The learning process involves a user selecting an object in a scene and naming it. The guessing game involves a user saying a phrase, and the system pointing to the object that the phrase refers to. The naming game involves a user pointing to an object and the system naming it The system is not physically grounded, so all games are simulated.

The learning process allows the system to acquire associations between phrases and concepts while the games test system comprehension and system production respectively. The learning process takes a **string** and **concept** as input, and produces no output. Comprehension takes a string as input, and produces a concept as output, whereas production takes a concept as input, and produces a string as output.

4.2 Strings and Concepts

A string is a list of characters with a fixed order. A blank space is used to separate **words** within the string, of which there can be either one or two. The system can break strings down into their component words.

A concept is a list of **feature values**. The system recognises six feature values; red, blue, green, white, circle, and square. There are no in-built associations between any of the feature values. This form of learning is supported by the imageability theory (Paiviom 1971). No claims concerning concept acquisition and formation are made in this paper. All concepts are hard coded from the outset.

The full list of objects used in the games are derived from shape and colour combinations; *red square*, *red circle*, *blue square*, *blue circle*, *green square*, *green circle*, *white square*, and *white circle*. Individual feature values can also act as concepts, therefore the full list is concepts is the list of object plus the list of feature values.

4.3 Groups

To associate a string with a concept, the system stores a list of **groups**. Each group contains an **ID**, one or more **description pairs**, an **observed frequency**, and zero or more **occurrence supporter links**.

The ID acts as a unique identifier, allowing the group to be found. A description pair is a string and a concept. Groups must have at least one description pair since their primary function is to relate a string to a concept. The observed frequency represents the number of times that the description pair's components have been associated through system input.

The occurrence supporter links are a set of group IDs. Each ID in the set refers to a group that

contains a superset of either the description pair, or the same value for one component of the description pair and a superset of the other e.g. The description pair ["red"; *red*] [1] would be supported by the description pair ["red square"; *red square*]. A worked example is provided in the next section. The links therefore record the number of occurrences of the group's description pair. The occurrence supporter link reinforces the description pair's association and increases the **total frequency** of the group. The total frequency is the group's observed frequency plus the observed frequency of all of its supporters, never including a supporter more than once.

Finally, group equality is defined by groups sharing the same description pair.

4.4 The Learning Process

At each stage in the learning process, a description pair is entered into the system. The system does not attempt to parse the correctness of the description. All data is considered to be positive. The general learning process algorithm is detailed in the rest of this section. Specific examples are also provided in Table 1, showing the groups' values; ID, description pair, occurrence frequency (OF), occurrence supporter links (OSLs), and total frequency (TF). Five steps are followed to incorporate the new data:

1. Identify the description pair.
2. Find equal and unequal parts.
3. Update system based on equal parts..
4. Update system based on unequal parts.
5. Re-enter new groups into the system.

4.4.1 Identify the description pair

If the description pair exists in a group that is already in the system, then that group's observed frequency is incremented. Otherwise, the system creates a new group containing the new description. It is given a unique ID and an observed frequency of one. Assume that the system already contains a group based on the description pair ["red circle"; *red circle*]. This has an ID of one. Assume also that the new description pair entered is ["red square"; *red square*]. Its group has an ID of two (group #2).

All description pairs entered into the system are called **concrete description pairs**, this is, the system has encountered them directly as input. The new group is referred to as a **concrete group**, since it contains a concrete description pair.

ID	Description Pair	OF	OSLs	TF
#1	["red circle"; *red circle*]	1	[]	1
#2	["red square"; *red square*]	1	[]	1
#3	["red"; *red*]	0	[#1,# 2]	2
#4	["#3 circle"; *#3 circle*]	0	[#1]	1
#5	["#3 square"; *#3 square*]	0	[#2]	1
#6	["circle"; *circle*], ["square"; *square*]	0	[]	0
#7	["#3 #6"; *#3 #6*]	0	[#2]	1

Table 1: Sample data

4.4.2 Find equal and unequal parts

The new group is compared to all of the groups in the system. Comparisons are based on the groups' description pairs alone. Strings are compared separately from concepts. A string match is found if one of the strings is a subset, or exact match, of the other. Subsets of strings must contain complete words. Words are regarded as atomic units. Concepts are compared in the same fashion as strings, where feature values are the atomic units. Successful comparisons create a set of equal parts and unequal parts. Comparison results are only used when equal parts exist. This approach is similar to alignment based learning, but with the additional component of concepts (van Zaanen, 2000).

In comparing the new group, group #2, to the existing group, group #1, the equal part ["red"; *red*] and the unequal part ["circle"; *circle*], ["square"; *square*] are found. The comparison algorithm is essential to the operation of the system. It is used in the learning process and in the games. Without it, no string or concept relations could be drawn[2].

4.4.3 Update system based on equal parts

When an equal part is found, a new group is created. In the example, an equal part is found between group #1 and group #2. Group #3 is created as a result. The new group is given an observed frequency of zero. The IDs of the groups that were compared (group #1 and group #2) are added to the new group's (group #3) occurrence supporter links. If the group already exists, then as well as the existing group's observed frequency being incremented, the IDs of the groups that were compared are added to the occurrence supporter links. IDs can only appear once in the set of occurrence supporters links, so if an ID is already in it, then it is not added.

[1] The convention of strings appearing in quotes ("red"), and concepts appearing in italics (*red*) is adopted throughout this paper.

[2] The system assumes full compositionality. Idioms and metaphors are not considered at this stage.

Up until this point, all groups' description pairs have contained a string and concept. Description pairs can also contain links to other groups' strings and groups' concepts. These description pairs are referred to as **abstract description pairs**. If all elements of the abstract description pair are links to other groups then it is **fully abstract**, else it is **partially abstract**. A group that contains an abstract description pair is called an **abstract group**. The group is fully abstract if its abstract description pair is fully abstract, else it is a **partially abstract group**. Once a group has been created (as group #3 was), based on a description comparison, the system attempts to make two abstract groups.

The new abstract groups (group #4 and group #5) are based on substitutions of the new group's ID (group #3) into each of the groups that were originally compared. Group #4 is therefore created by substituting group #3 into group #1. Similarly, group #5 is created by substituting group #3 into group #2.

The new abstract groups are given an observed frequency of zero (ID's equal four and five). Note that abstract groups always have an observed frequency of zero as they can never been directly observed. The ID of the appropriate group used in comparison and later creation is added to the occurrence supporters links. Each abstract group therefore has a total frequency equal to that of the group of which it is an abstract form.

4.4.4 Update system based on unequal parts

Unequal parts are only considered if equal parts are found in the comparison. Otherwise, the unequal parts would be the complete set of data from both groups, which does not provide useful information for comparisson. For every set of unequal parts that is found, a new group is created. If there is more than one unequal part then the group will contain more than one description pair. Such a group is referred to as a **multi-group**. Two unequal parts were found earlier in comparing group #1 and group #2. They are ["circle"; *circle*] and ["square"; *square*]. Group #6 is therefore created using these two description pairs.

The creation of a multi-group allows for a fully abstract group to be created. The system uses the data from the new multi-group (group #6) and the group created through equal parts (group #3). Both groups are substituted back into the group that was originally being compared (group #1). The resulting group (group #7) is fully abstract as both equal parts and unequal parts have been used to reconstruct the original group (group #1).

4.4.5 Re-enter new groups into the system

All groups that have been created through steps 3 and 4 are compared to all other groups in the system. Results of comparisons are dealt with by repeating steps 3-5 with the new results. By use a recursive step like this, all groups are compared to one another in the system. All group equalities are therefore created when the round is complete. The amount of information available from every new group entered into the system is therefore maximised.

4.5 The Significance of Groups Types

Four different types of group have been identified in the previous section. Although all groups share the same properties, they can be seen to represent difference aspects of language. It is the combination and interaction of these groups that gives rise to emergent simple syntax. This syntax is bi-gram collocations, but since the system is scalable, it is referred to as simple syntax.

4.5.1 Concrete Groups

Concrete groups acquire the meaning of individual lexemes (associate concepts with strings). They are verifiable in the real world through the use of scene based games.

4.5.2 Multi-Groups

Multi-groups form syntactic categories based on similarities between description pair usage. Under the current system, groups can only have a maximum of two description pairs. If this were to be expanded, it is clear that large syntactic categories such as noun and verb equivalents would arise.

4.5.3 Partially and Fully Abstract Groups

Partially and fully abstract groups act as phrasal rules in the system. Abstract values contained within the group's description pairs can relate to both concrete groups and multi-groups. Abstract groups that relate to multi-groups offer a choice of substitutions.

For example, group #7 (Table 1) relates a single group to a multi-group. By substitution of groups #3 and #6 into group #7, the concrete pairings of ["red circle"; *red circle*] and ["red square"; *red square*] are produced. The string data are directly equivalent to:

S -> Adj. N,
where Adj. = {"red"}
and N = {"circle", "square"}

When a description pair is entered into the system, the process of semantic bootstrapping takes place. Lexical items (strings) are associated with their meanings (concepts). When group

comparisons are made, syntactic bootstrapping begins. Associations are made between all combinations of lexical items throughout the system, and all combinations of meanings throughout the system.

The system stores lexical item-meaning associations, lexical item-lexical item associations and meaning-meaning associations. This basic framework allows for the production of complex phrasal rules.

4.6 Comprehension and Production Through Games

The guessing game tests comprehension while the naming game test production. Comprehension takes a string as input, and produces a concept as output, whereas production takes a concept as input, and produces a string as output. The comprehension and the production algorithms are the same, except the first is string based, and the second is concept based.

The algorithm performs two tasks: finding concrete groups with exact matches to the input, and finding abstract groups with possible matches to the input. Holophrastic matching uses only concrete groups. Syntactic matching performs holophrastic matching, followed by further matches using abstract groups. Note that the system only performs syntactic matching, which includes holophrastic matching. Holophrastic matching is never performed alone, unless in testing stages.

For holophrastic matches, the system searches through its list of groups. Their description pairs are compared to the input being searched for. There is therefore re-use of the comparison algorithm introduced in the learning process. When a match is found, the group is added to a list of possible results.

If holophrastic matching is being performed alone, then this list of possible results is sorted by total frequency. The group with the highest total frequency is output by the system.

Syntactic matching begins by performing holophrastic matching, but does not output a result until all abstract groups have been matched too. It is therefore an extension of holophrastic matching. Once a first fun of holophrastic matching is performed, the input is converted into abstract form. This is performed at the word/feature value level. The most likely element is found by searching through the groups, comparing it to the description pair, and selecting the group with the highest total frequency from those found.

The group IDs replace the appropriate element in the input (just as substitutions were made during the learning process). All multi-groups that contain any of the abstract forms are found. Each multi-group's description pair becomes a replacement for the appropriate input's abstract value.

The new input, which is still in abstract form, is searched for, using holophrastic matching again. Since the groups found are not exact matches of the original input, their total frequency is multiplied by an **abstract factor**. The abstract factor is a value between zero and one inclusive. The higher the factor, the greater the effect that abstract groups have on the results. Syntactic matches can therefore produce different results based on the value of abstract factor. The abstract factor is not changed from the initiation to termination of the system.

Groups found during the search are added to a new list of possible results. The appropriate elements are substituted into the groups abstract values to make them concrete. If an abstract value is acting as a substitute (by being found originally in a multi-group) then the original input value is used, not the replacement element. This allows the abstract group to act as a syntactic rule, but it is penalised by the abstract factor so it does not have as much influence as concrete groups, that have been found to occur through direct input associations.

The groups found throughout the entire syntactic search are now contained in a second list of possible results. This list is reduced by removing duplicate groups. For each group that is removed, its observed frequency and occurrence supporter links are added to the duplicate that is kept in the list.

The two lists from each matching routine are merged and sorted by total frequency. The string\concept of the group with the highest total frequency is outputted by the system.

5 Testing and Results

The system is tested within the following areas:

1. Comprehension and production of all fourteen concepts. The rate at which full comprehension and full production are achieved is compared.
2. Correctness of production matches for compound concepts. The correctness of production matches are studied over a number of rounds.
3. Type of production matches for compound concepts. The type of production matches favoured, holophrastic or syntactic, are compared over a number of rounds

A match of concept to word or word to concept is considered correct if the string describes the concept fully. For example, ["red"; *red*] and ["red

square"; *red square*] are correct, but ["red"; *red square*] and ["red square"; *red*] are incorrect. One point is given for each correct match, zero for each incorrect match.

Note that all test results are based on the average of ten different system trials. Each result shows a broad tendency that will likely be smoothed if more trials are run. All input is randomly generated. The abstract factor is set to 0.4 for all tests.

5.1 Comprehension Vs. Production

Full comprehension occurs much sooner (see Figure 1), on average, than full production. This result is found in children also. Although production and comprehension compete quite steadily in early stages of the system, comprehension reaches its maximum, on average, in 20% of the time that production takes to reach its maximum.

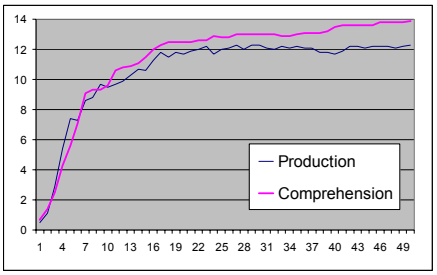


Figure 1: Shows number of correct comprehension and production matches

Full comprehension (fourteen points) is achieved, on average, by round 50, while full production comes at round 250. Both holophrastic data and syntactic data contribute to the successes. Underextensions are found during comprehension. For example, in early rounds, "green" is used to describe only *green squares*. This phenomena is quickly eliminated in the trials but with a larger set of concepts and vocabulary, it is likely to persist for more than a few rounds.

5.2 Correctness of Holophrastic Vs Syntactic Matches

At the end of each round, production is tested using the eight compound concepts alone. These are based on the eight observable objects in the simulated scene. Only compound concepts can demonstrate simple syntax in this system, as singular concepts have associations to single word strings.

The system uses syntactic matching alone, but syntactic matching includes holophrastic matching, as discussed earlier. To determine whether holophrastic data is being used, or syntactic data when a syntactic match is run, the matching algorithm has been split. The number of correct strings produced using holophrastic data and the number of correct strings produced using syntactic data alone are compared (see Figure 2).

The data demonstrate that the system uses mostly holophrastic matches in early rounds (comparable to the one-word stage). This is eliminated in further rounds, in favour or syntactic matches alone (the two-word stage). Note that although the holophrastic stage may appear to be producing two-words, these words are considered to be one-word. For example, "allgone" is considered to be one-word in early stages of linguistic development, as opposed to "all gone" (Ingram, 1989).

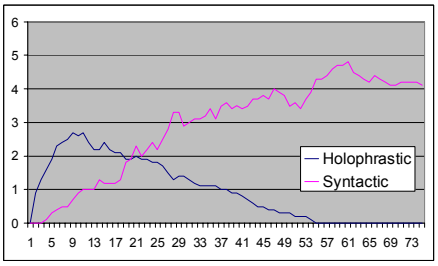


Figure 2: Shows number of correct holophrastic and syntactic matches.

The syntactic data continues to rise, until it achieves full production. The holophrastic stage never achieves full production, but peaks, then reduces to zero. This trend occurs as holophrastic underextensions such as "red" representing *red square* become more likely than "red square" representing *red square*.

Early syntactic matches are based on novel string productions for novel string concepts. Holophrastic matching is incapable of producing novel strings from novel concepts, as it deals with concrete concepts. Abstract concepts however, allow new string combinations to be produced, such as "blue square", from *blue square* even though neither then string nor concept have been encountered before. Such an abstraction may come from a multi-group that associates "blue" with "red", while containing a group that contains "red square" also. The novel string "blue square" is therefore abstracted.

5.3 Use of Holophrastic Vs Syntactic Matches

The system does not always produce the correct strings when a concept is entered. The strings that are produced are a result of either holophrastic or syntactic matching. Regardless of correctness, the amount of times that holophrastic matches are made over syntactic matches can be compared (see Figure 3).

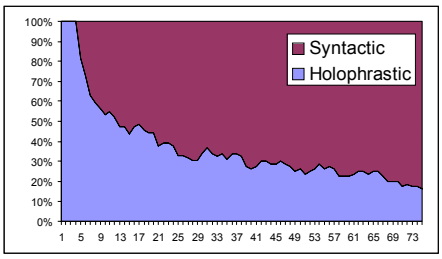


Figure 3: Shows distribution of holophrastic and syntactic matches.

The system relies completely on one-word descriptions at the outset, but soon syntactically derived two-word descriptions become prevalent. It is likely that the one-word stage will last longer if larger concept and vocabulary sets are in use.

The system shows the same form of transition as can be seen in children from the one-word stage to the two-word stage, without the use of an artificial trigger. The shift is gradual although the use of larger concept and vocabulary sets, plus different abstract factor values will affect the transition. The greater the number of words in multi-groups (the greater the size of syntactic categories), the lower the abstract factor is required to encourage the emergence of simple syntax.

6 Related Works

Supporters of computational modelling in language acquisition, often promote the practical importance of running simulations, where evolutionary effects can be recreated in short time periods (Zuidema, 2001).

Although this paper is focussed on an individual system, or agent, acquiring language, it is been influenced by research into social learning (Oliphant and Batali, 1997; Kirby, 1999; Steels and Kaplan, 2002). Social learning demonstrates the convergence upon a common language, or set of languages, from an uncoordinated proto-language, within a population of agents. Social learning allows for the playing of games between agents, similar to those in this paper, with the results being used as further system input, to support, or deny associations. This research can be viewed as a form of social learning with one agent (string and concept generator) performing the teacher role, and the other agent (the system) performing the learner role.

Simulations of both the babbling stage and the one-word stage have been developed (Scheler, 1997; Abidi, 1999). ACCLAIM, a one-word stage simulator, demonstrates that systems can react appropriately to changes in situations. For example, when a cessation event is triggered, it produces “Stop”, and when an object is requested, it produces “More”. Both examples are typical of children during the one-word stage (Bloom, 1973).

Several systems exist that use perceptions to encourage language acquisition (Howell, Becker, and Jankowicz,, 2001; Roy, 2001). ELBA learns both nouns and verbs from video scenes, starting with a blank lexicon. Such systems have helped in the selection of both appropriate input sources and feature values to use in this research. This system will also be physically grounded in future.

The research presented in this paper describes a system that drives linguistic development. Other systems have used similar techniques, based on syntactic and semantic bootstrapping (Howell and Becker, 2001), but have not explained how multiple word acquisition is achieved from a single word basis.

Steels (1998) introduces frames that group lexical elements together by the roles that they play, very similar to groups in this paper. Frames are more dynamic than groups however, structurally adapting when words reoccur. Groups do not adapt in this way. New groups are created to describe similarities rather than adapting existing ones. Steels also introduces multiple word sentences, but it is unclear as to why agents invent a multiple word description over creating a new single word description. The invention is triggered and does not emerge. This research is based on real multiple word inputs, so the reason for invention is not necessary, unlike the reason for adoption i.e. why the system adopts two-word descriptions.

The comparison algorithm, as previously noted, is similar to alignment based learning (van Zaanen, 2000). The system in this research performs perfect alignment requiring exact word matches when finding equal parts and unequal parts. This system also uses concepts, reducing the number of incorrect groupings, or constituents, when there is ambiguity in text. Unsupervised grammar induction can also be found in EMILE (van Zaanen and Adriaans, 2001). EMILE identifies substitution classes by means of clustering. These classes are comparable to this system’s groups although no concepts are used.

7 Future Research

As the system stands, it uses a small input set. Further developments are focussed on expanding the system. All ten of Brown’s relations should be implemented. Larger concept and vocabulary sets are therefore required. Extensions to these sets are likely to affect underextensions, mismatches, the length of pre-syntactic usage time, and the overall growth pattern of simple syntax.

8 Conclusion

This paper offers a potential explanation of the mechanism by which the two-word stage emerges from the one-word stage. It suggests that syntactic data is sought out from the beginning of language acquisition. This syntactic data is always competing with the associations of holophrastic data. Syntax is strengthened when patterns are consistently found between strings and concepts, and is used in favour of holophrastic data when it is sufficiently frequent. The simple syntax continues to grow in strength, ultimately being used in favour of holophrastic data in all production and comprehension tasks.

This system provides the foundation for more complex, hierarchical, syntax to emerge. The type and volume of input is the only constraint upon the system. The entry into post two-word stages is predicted from the system's robust architecture.

9 Acknowledgements

The first author is sponsored by a studentship from the EPSRC.

Thanks to the workshop reviewers for their helpful and much appreciated advice.

References

S. Abidi, 1996. A Neural Network Simulation of Child Language Development at the One-word Stage. *In proceedings of IASTED Int. Conf. on Modelling,* Simulation and Optimization, Gold Coast, Australia.

L. Bloom, 1973. One Word at a Time. The use of single-word utterances before syntax The Hague, Mouton.

R.W. Brown, 1986. Language and categories. In "*A Study of Thinking*", ed. J.S. Bruner, J.J. Goodnow, and G.A. Austin, pages 247-312. New York: John Wiley, 1956. Reprint, New Brunswick: Transaction.

L.R.. Gleitman and Elissa L. Newport, 1995. *The Invention of Language by Children: Environmental and Biological Influences on the Acquisition Language*. In "An Invitation to Cognitive Science", L.R. Gleitman and M. Liberman, 2nd ed., Vol.1, Cambridge, Mass., London, MIT Press.

S.R. Howell and S. Becker, 2001. Modelling language acquisition: Grammar from the Lexicon? In *Proceedings of the Cognitive Science Society*..

S.R. Howell, S. Becker, and D. Jankowicz, 2001. *Modelling Language Acquisition: Lexical Grounding Through Perceptual Features*. In Proceedings of the 2001 Workshop on Developmental Embodied Cognition

J.R. Hurford, M. Studdert-Kennedey, and C. Knight, 1998. The Emergence of Syntax. In "*Approaches to the evolution of language: social and cognitive bases*", Cambridge, Cambridge University Press.

D. Ingram, 1989. *First Language Acquisition. Method, Description and Explanation.* Cambridge: Cambridge University Press.

R. Jakobson, 1971. *Why "mama" and "papa"?* In "Child Language: A Book of Readings", by A. Bar-Adon and W. F. Leopold, ed., pages 213-217. Englewood Cliffs, NJ:Prentice-Hall.

P.W. Jusczyk, 1999 How infants begin to extract words from speech. *Trends in Cognitive Science*, 3 (9, September):323-328.

S. Kirby, 1999. *Syntax out of learning: The cultural evolution of structured communication in a population of induction algorithms*. In Proceedings of ECAL99 European Conference on Artificial Life, D. Floreano et al. ed. pages 694-703, Berlin: Springer-Verlag,

M. Oliphant and J. Batali 1997. Learning and the emergence of coordinated communication. *Centre for Research in Language Newsletter*, 11(1).

A. Paivio, 1971, *Imagery and Verbal Processes*. New York: Holt, Rinehart & Winston.

J. Piaget, 1960. *The Language and Thought of the Child.* Routledge and K. Paul, 3rd ed.,. Routledge Paperbacks.

S. Pinker, 1994. *The Language Instinct. The New Science of Language and Mind*. Allen Lane, Penguin Press.

D. Roy, 2001. Grounded spoken language acquisition: Experiments in word learning. *IEEE Transactions on Multimedia.*

G. Scheler, 1997d. The transition from babbling to the one-word stage: A computational model. In *Proceedings of GALA '97.*

L. Steels and F. Kaplan, 2001. AIBO's first words: The social learning of language and meaning. *Evolution of Communication*, vol. 4(1):3-32. John Benjamin's Publishing Company, Amsterdam, Holland.

L. Steels, 1998. The Origins of Syntax in visually grounded robotic agents. *AI* 103, 1-24.

M. Tomasello, and A.C. Kruger, 1992. Joint attention in action: Acquiring verbs in ostensive and non-ostensive contexts. *Journal of Child Language* 19:311-333.

M. van Zaanen, 2000. Learning structure using alignment based learning. In *Proceedings of the Third Annual Doctoral Research Colloquium (CLUK)*, pages 75-82.

M. van Zaanen and P. Adriaans, 2001. Alignment-based learning versus EMILE: A comparison. In *Proceedings of the Belgian-Dutch Conference on AI (BNAIC)*.

W.H. Zuidema, 2001. Emergent syntax: the unremitting value of computational modelling for understanding the origins of complex language. *ECAL01*, 641-644. Springer, Prague, Sept. 10-14, 2001.

On a possible role for pronouns in the acquisition of verbs

Aarre Laakso and Linda Smith
Department of Psychology
1101 E. 10th Street
Bloomington, IN 47408
{alaakso,smith4}@indiana.edu

Abstract

Given the restrictions on the subjects and objects that any given verb may take, it seems likely that children might learn verbs partly by exploiting statistical regularities in co-occurrences between verbs and noun phrases. Pronouns are the most common NPs in the speech that children hear. We demonstrate that pronouns systematically partition several important classes of verbs, and that a simple statistical learner can exploit these regularities to narrow the range of possible verbs that are consistent with an incomplete utterance. Taken together, these results suggest that children *might* use regularities in pronoun/verb co-occurrences to help learn verbs, though whether this is *actually* so remains a topic for further research.

1 Introduction

Pronouns stand for central elements of adult conceptual schemes—as Quine pointed out, pronouns "are the basic media of reference" (Quine, 1980, p. 13). In fact, most syntactic subjects in spontaneous spoken adult discourse are pronouns (Chafe, 1994), and English-speaking mothers often begin with a high-frequency pronoun when speaking to their children, with *you* and *I* occurring most frequently (e.g., Valian, 1991). Parents use the inanimate pronoun *it* far more frequently as the subject of an intransitive sentence than of an transitive one (Cameron-Faulkner et al., 2003, p. 860). As Cameron-Faulkner et al. note, this suggests that intransitive sentences are used more often than transitives for talking about inanimate objects. It also suggests, we would note, that the use of the inanimate pronoun might be a cue for the child as to whether the verb is transitive or intransitive. Similarly, Lieven and Pine (Lieven et al., 1997; Pine and Lieven, 1993) have suggested that pronouns may form the fixed element in lexically-specific frames acquired by early language learners—so-to-speak "pronoun islands" something like Tomasello's (1992) "verb islands."

Many researchers have suggested that word-word relations in general, and syntactic frames specifically, are particularly important for learning verbs (e.g., Gleitman, 1990; Gleitman and Gillette, 1995). What has not been studied, to our knowledge, is how *pronouns* specifically may help children learn verbs by virtue of systematic co-occurrences. We have begun to address this issue in two steps. First, we measured the statistical regularities among the uses of pronouns and verbs in a large corpus of parent and child speech. We found strong regularities in the use of pronouns with several broad classes of verbs. Second, using the corpus data, we trained a connectionist network to guess which verb belongs in a sentence given only the subject and object, demonstrating that it is possible in principle for a statistical learner to use the regularities in parental speech to deduce information about an unknown verb.

2 Experiment 1

The first experiment consisted of a corpus analysis to identify patterns of co-occurrence between pronouns and verbs in the child's input.

2.1 Method

Parental utterances from the CHILDES database (MacWhinney, 2000) were coded for syntactic categories, then subjected to cluster analysis. The mean age of target children represented in the transcripts that were coded for this experiment was 3;0 (SD≈1;2).

2.1.1 Materials

The following corpora were used: Bates, Bliss, Bloom 1970, Brown, Clark, Cornell, Demetras Working, Gleason, Hall, Higginson, Kuczaj, MacWhinney, Morisset, New England, Post, Sachs, Suppes, Tardiff, Valian, Van Houten, Van Kleeck and Warren-Leubecker. Coding was performed using a custom web application that randomly selected transcripts, assigned them to coders as they became available, collected coding

input, and stored it in a MySQL database. The application occasionally assigned the same transcript to all coders, in order to measure reliability. Five undergraduate coders were trained on the coding task and the use of the system.

2.1.2 Procedure

Each coder was presented, in sequence, with each main tier line of each transcript she was assigned, together with several lines of context; the entire transcript was also available by clicking a link on the coding page. For each line, she indicated (a) whether the speaker was a parent, target child, or other; (b) whether the addressee was a parent, target child, or other; (c) the syntactic frames of up to 3 clauses in the utterance; (d) for each clause, up to 3 subjects, auxiliaries, verbs, direct objects, indirect objects and obliques. Because many utterances were multi-clausal, the unit of analysis for assessing pronoun-verb co-occurrences was the clause rather than the utterance.

The syntactic frames were: no verb, question, passive, copula, intransitive, transitive and ditransitive. These were considered to be mutually exclusive, i.e., each clause was tagged as belonging to one and only one frame, according to which of the following frames it matched first: (1) The *no verb* frame included clauses – such as "Yes" or "OK" – with no main verb. (2) The *question* frame included any clause using a question word – such as "Where did you go?" – or having inverted word order – such as "Did you go to the bank?" – but not merely a question mark – such as "You went to the bank?" (3) The *passive* frame included clauses in the passive voice, such as "John was hit by the ball." (4) The *copula* frame included clauses with the copula as the main verb, such as "John is angry." (5) The *intransitive* frame included clauses with no direct object, such as "John ran." The *transitive* frame included clauses with a direct object but no indirect object, such as "John hit the ball." (6) The *ditransitive* frame included clauses with an indirect object, such as "John gave Mary a kiss."

All nouns were coded in their singular forms, whether they were singular or plural (e.g., "boys" was coded as "boy"), and all verbs were coded in their infinitive forms, whatever tense they were in (e.g., "ran" was coded as "run").

In total, 59,977 utterances were coded from 123 transcripts. *All* of the coders coded 7 of those transcripts for the purpose of measuring reliability. Average inter-coder reliability (measured for each coder as the percentage of items coded exactly the same way they were coded by each other coder) was 86.1%. Given the number of variables, the number of levels of each variable (3 speakers, 3 addressees, 7 frames, and 6 syntactic relations), and the number of coders (5), the probability of chance agreement is very low. Although there are some substantive errors (usually with complex embedded clauses or other unusual constructions), many of the discrepancies are simple spelling mistakes or failures to trim words to their roots.

We only considered parental child-directed speech (PCDS), defined as utterances where the speaker was a parent and the addressee was a target child. A total of 24,286 PCDS utterances were coded, including a total of 28,733 clauses. More than a quarter (28.36%) of the PCDS clauses contained no verb at all; these were excluded from further analysis. Clauses that were questions (16.86%), passives (0.02%), and copulas (11.86%) were also excluded from further analysis. The analysis was conducted using only clauses that were intransitives (17.24% of total PCDS clauses), transitives (24.36%) or ditransitives (1.48%), a total of 12,377 clauses.

2.2 Results

The most frequent nouns in the corpus—both subjects and objects—are pronouns, as shown in Figures 1 and 2. The objects divided the most common verbs into three main classes: verbs that take the pronoun *it* and concrete nouns as objects, verbs that take complement clauses, and verbs that take specific concrete nouns as objects. The subjects divided the most common verbs into four main classes: verbs whose subject is almost always *I*, verbs whose subject is almost always *you*, verbs that take *I* or *you* almost equally as subject, and other verbs. The verbs divided the most common object nouns into a number of classes, including objects of telling and looking verbs, objects of having and wanting verbs, and objects of putting and getting verbs. The verbs also divided the most common subject nouns into a number of classes, including subjects of having and wanting verbs, and subjects of thinking and knowing verbs.

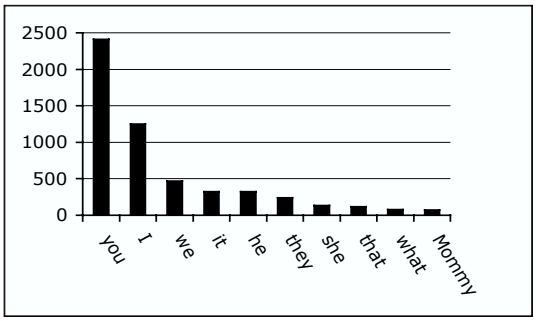


Figure 1: The 10 most frequent subjects in PCDS by their number of occurrences

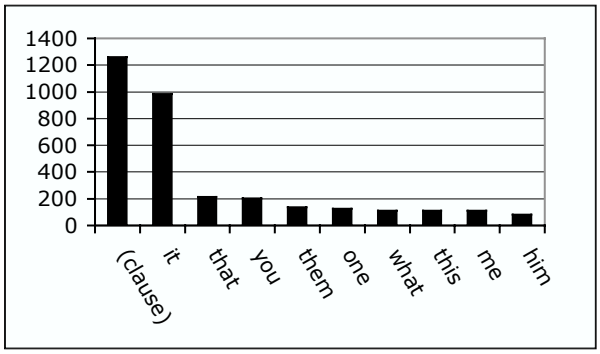


Figure 2: The 10 most frequent objects in PCDS by their number of occurrences.

2.2.1 Verbs that take *it* as an object

The verbs that take *it* as their most common object include verbs of motion and transfer, as shown in Table 1.

2.2.2 Verbs that take complement clauses

Most verbs that did not take *it* as their most common object instead took complement clauses. These are primarily psychological verbs, as shown in Table 2.

2.2.3 Verbs that take concrete nouns as objects

Most remaining verbs in the corpus took unique sets of objects. For example, the most common object used with *read* was *book*, followed by *it* and *story*; the most common object used with *play* was *game*, followed by *it*, *block*, and *house*.

2.2.4 Verbs that take *I* as a subject

Verbs whose most common subject is *I* include *bet* (23 out of 23 uses with a subject, or 100%), *guess* (21/22, 95.4%), *think* (212/263, 80.6%), and *see* (95/207, 45.9%). Parents were not discussing their gambling habits with their children – *bet* was being used to indicate the epistemic status of a subsequent clause, as were the other verbs.

2.2.5 Verbs that take *you* as a subject

Verbs whose most common subject is *you* include *like* (86 out of its 134 total uses with a subject, or 64.2%), *want* (192/270, 71.1%), and *need* (33/65, 50.8%). These verbs are being used to indicate the deontic status of a subsequent clause, including disposition or inclination, volition, and compulsion.

2.2.6 Verbs that take you or I as a subject

Verbs that take *I* and *you* more or less equally as subject include *mean* (15 out of 32 uses, or 46.9%, with *I* and 12 of 32 uses, or 37.5%, with *you*), *know* (*I*: 159/360, 44.2%; *you*: 189/360, 52.5%), and *remember* (*I*: 9/23, 39.1%; *you*: 12/23, 52.2%).

Verb	Total	it (#)	it (%)
turn	56	33	58.9
throw	36	20	55.5
push	25	13	52.0
hold	42	19	45.2
break	36	16	44.4
leave	27	12	44.4
open	36	15	41.7
do	256	105	41.0
wear	25	10	40.0
take off	24	9	37.5
put	276	93	33.7
get	348	74	21.3
take	106	22	20.8
put on	42	8	19.0
buy	50	9	18.0
give	85	14	16.5
have	340	26	7.6

Table 1: Verbs most commonly used with object *it.*

Verb	Total	<clause> (#)	<clause> (%)
think	187	179	95.7
remember	31	23	74.2
let	78	57	73.1
know	207	141	68.1
ask	29	17	58.6
go	55	32	58.2
want	317	183	57.7
mean	25	14	56.0
tell	115	45	39.1
try	51	18	35.3
say	175	53	30.3
look	48	14	29.2
need	64	18	28.1
see	266	73	27.4
like	123	32	26.0
show	36	9	25.0
make	155	23	14.8

Table 2: Verbs most commonly used with complement clauses.

Verb	Total	I (#)	I (%)	you (#)	you (%)
bet	23	23	100	0	0
guess	22	21	95.4	0	0
think	263	212	80.6	38	14.4
see	207	95	45.9	50	24.1
mean	32	15	46.9	12	37.5
know	360	159	44.2	189	52.5
remember	23	9	39.1	12	52.2
like	134	20	14.9	86	64.2
want	270	34	12.6	192	71.1
need	65	5	7.7	33	50.8

Table 3: Some verbs commonly used with subject *I* or *you.*

2.2.7 Objects of *tell* and *look at*

The objects *me*, *us*, *Daddy* and *Mommy* formed a cluster in verb space, appearing frequently with the verbs *tell* and *look at*.

2.2.8 Objects of *put* and *get*

The objects *one*, *stuff*, *box*, and *toy* occurred most frequently with *get*, and frequently with *put*. The objects *them*, *him*, *her*, *bed*, and *mouth* occurred most frequently with *put* and, in some cases, also frequently with *get*.

2.2.9 Objects of *have* and *want*

The objects *cookie*, *some*, *money*, *coffee*, *milk*, and *juice* formed a cluster in verb space, appearing frequently with verbs such as *have* and *want*, as well as, in some cases, *give*, *take*, *pour*, *drink*, and *eat*.

2.2.10 Subjects of *think* and *know*

The subject *I* appeared most frequently with the verbs *think* and *know*.

2.3 Discussion

Although pronouns are semantically "light," their particular referents determinable only from context, they may nonetheless be potent forces on early lexical learning by statistically pointing to some classes of verbs as being more likely than others. The results of Experiment 1 clearly show that there are statistical regularities in the co-occurrences of pronouns and verbs that the child could use to discriminate classes of verbs. Specifically, when followed by *it*, the verb is likely to describe physical motion, transfer, or possession. When followed a relatively complex complement clause, by contrast, the verb is likely to attribute a psychological state. Finer distinctions may also be made with other objects, including proper names and nouns. Verbs followed by *me*, *us*, *Daddy*, and *Mommy* are likely to have to do with telling or looking. Verbs followed by *one*, *stuff*, *them*, *him*, or *her* are likely to have to do with getting or putting. Verbs followed by certain concrete objects such as *cookie*, *milk*, or *juice* are likely to have to do with having or wanting. Fine distinctions may also be made according to subject. If the subject is *I*, the verb is likely to have to do with thinking or knowing, whereas if the subject is *you*, *she*, *we*, *he*, or *they*, the verb is likely to have to do with having or wanting. This regularity most likely reflects the ecology of parents and children—parents "know" and children "want" — but it could nonetheless be useful in distinguishing these two classes of verbs.

The results thus far show that there are potentially usable regularities in the statistical relations between pronouns and verbs. However, they do not show that these regularities can be used to cue the associated words.

3 Experiment 2

To demonstrate that the regularities in pronoun-verb co-occurrences in parental speech to children can actually be exploited by a statistical learner, we trained an autoassociator on the corpus data, then tested it on incomplete utterances to see how well it would "fill in the blanks" when given only a pronoun, or only a verb. An autoassociator is a connectionist network that is trained to take each input pattern and reproduce it at the output. In the process, it compresses the pattern through a small set of hidden units in the middle, forcing the network to find the statistical regularities among the elements in the input data. The network is trained by backpropagation, which iteratively reduces the discrepancies between the network's actual outputs and the target outputs (the same as the inputs for an autoassociator).

In our case, the inputs (and thus the outputs) are subject-verb-object "sentences." Once the network has learned the regularities inherent in a corpus of complete SVO sentences, testing it on incomplete sentences (e.g., "I ___ him") allows us to see what it has gleaned about the relationship between the given parts (subject "I" and object "him" in our example) and the missing parts (the verb in our example).

3.1 Method

3.1.1 Data

The network training data consisted of the subject, verb, and object of all coded utterances that contained the 50 most common subjects, verbs and objects. There were 5,835 such utterances. The inputs used a localist coding wherein there was one and only one input unit out of 50 activated for each subject, and likewise for each verb and each object. Absent and omitted arguments were counted among the 50, so, for example, the utterance "John runs" would have 3 units activated even though it only has 2 words—the third unit being the "no object" unit. With 50 units each for subject, verb and object, there were a total of 150 input units to the network. Active input units had a value of 1, and inactive input units had a value of 0.

3.1.2 Network Architecture

The network consisted of a two-layer 150-8-150 unit autoassociator with a logistic activation function at the hidden layer and a three separate softmax activation functions (one each for the subject, verb and object) at the output layer—see

Figure 3. Using the softmax activation function, which ensures that all the outputs in the bank sum to 1, together with the cross-entropy error measure, allows us to interpret the network outputs as probabilities (Bishop, 1995). The network was trained by the resilient backpropagation algorithm (Riedmiller and Braun, 1993) to map its inputs back onto its outputs. We chose to use eight units in the hidden layer on the basis of some pilot experiments that varied the number of hidden units. Networks with fewer hidden units either did not learn the problem sufficiently well or took a long time to converge, whereas networks with more than about 8 hidden units learned quickly but tended to overfit the data.

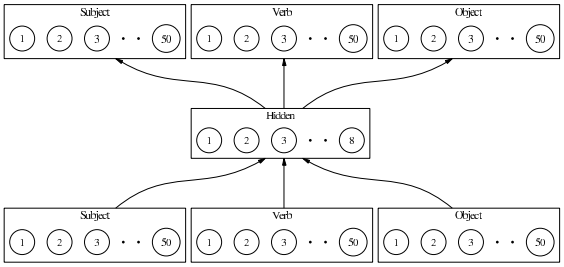


Figure 3: Network architecture

3.1.3 Training

The data was randomly assigned to two groups: 90% of the data was used for training the network, while 10% was reserved for validating the network's performance. Starting from different random initial weights, five networks were trained until the cross-entropy on the validation set reached a minimum for each of them. Training stopped after approximately 150 epochs of training, on average. At that point, the networks were achieving about 81% accuracy on correctly identifying subjects, verbs and objects from the training set. Near perfect accuracy on the training set could have been achieved by further training, with some loss of generalization, but we wanted to avoid overfitting.

3.1.4 Testing

After training, the networks were tested with incomplete inputs corresponding to isolated verbs and pronouns. For example, to see what a network had learned about *it* as a subject, it was tested with a single input unit activated—the one corresponding to *it* as subject. The other input units were set to 0. Activations at the output units were recorded. The results presented below report average activations over all five networks.

3.2 Results

The networks learn many of the co-occurrence regularities observed in the data. For example, when tested on the object *it* (see Figure 4 on page 7 below), the most activated verbs are *get*, *hold*, *take* and *have*, which are among the most common verbs associated with *it* in the input (see Table 1). Similarly, *tell*, *make* and *say* are the most activated verbs when networks are tested with the *clause* unit activated in the object position (figure not shown), and they are also among the verbs most commonly associated with a *clause* in the input (see Table 2).

However, the network does not merely learn the relative frequencies of pronouns with verbs. For example, the verbs most activated by the subject *you* are *have* and *get* (see Figure 5 on page 8 below), neither of which appears in Table 3. The reason for this, we believe, is that the subject *you* is strongly associated with the object *it* (note the strong activation of *it* in the right column of Figure 5), and the object *it*, as mentioned in the previous paragraph, is strongly associated with the verbs *have* and *get*. The difference may be observed most clearly when the network is prompted simultaneously with *you* as the subject and *clause* as the object (see Figure 6 on page 8 below). In that case, the verb *want* is strongly preferred and, though *get* still takes second place, *tell* and *know* rank third and fourth, respectively—consistent with the results in Table 1. This demonstrates that the network model is sensitive to high-order correlations among words in the input, not merely the first-order correlations between pronoun and verb occurrences.

These results do not depend on using an autoassociation network, and we do not claim that children in fact use an autoassociation architecture to learn language. Any statistical learner that is able to discover higher-order correlations will produce results similar to the ones shown here. An autoassociator was chosen only as a simple means of demonstrating in principle that a statistical learner can extract the statistical regularities from the data.

4 Conclusion

We have shown that there are statistical regularities in co-occurrences between pronouns and verbs in the speech that children hear from their parents. We have also shown that a simple statistical learner can learn these regularities, including subtle higher-order regularities that are not obvious in a casual glance at the input data, and use them to predict the verb in an incomplete sentence. How might this help children learn

verbs? In the first place, hearing a verb framed by pronouns may help the child isolate the verb itself—having simple, short consistent, and high-frequency slot fillers could make it that much easier to segment the relevant word in frames like "He ___ it." Second, the information provided by the particular pronouns that are used in a given utterance might help the child isolate the relevant event or action from the blooming, buzzing confusion around it—in English, pronouns can indicate animacy, gender and number, and their order can indicate temporal or causal direction or sequence (e.g., "You ___ it" versus "It ___ you"). Finally, if we suppose that the child has already learned one verb and its pattern of correlations with pronouns, and then hears another verb being used with the same or a similar pattern of correlations, the child may hypothesize that the unknown verb is similar to the known verb. For example, a child who understood "want" but not "need" might observe that "you" is usually the subject of both and conclude that "want," like "need," has to do with his desires and not, for example, a physical motion or someone else's state of mind. The pronoun/verb co-occurrences in the input may thus help the child narrow down the class to which an unknown verb belongs, allowing the learner to focus on further refining her grasp of the verb through subsequent exposures.

Whether children are actually sensitive to these regularities remains an open question. To the extent that children have actually picked up on the regularities, two predictions should follow. The first is that children's utterances should exhibit roughly the same co-occurrence patterns as we found in their parents' speech to them. Therefore, the next step in our research is to determine whether children are using pronouns and verbs together with roughly the same frequencies that they hear in their parents' speech. This is the subject of research in progress using the coded corpus data from Experiment 1. Because our hypothesis concerns broad-class verb acquisition, we are focusing on children younger than the age of 3, by which time most children can produce the most common verbs (Dale and Fenson, 1996).

The second prediction that follows from the hypothesis that children might be sensitive to the regularities demonstrated in this paper is that children's comprehension of ordinary verbs should be better when they are used in frames that are consistent with the regularities in the input than when they are used in frames that are inconsistent with those regularities. Assessing whether this is true requires an experiment testing children's comprehension of real but relatively infrequent verbs in two conditions: a "consistent" condition (in which the verb is used with nouns or pronouns that are consistent with the regularities in the corpus) and an "inconsistent" condition (in which the verb is used with nouns or pronouns that are inconsistent with the regularities in the corpus). This experiment is in the planning stages.

Even if children are sensitive to the regularities, this knowledge might not help them learn new verbs. That is, whether these regularities actually play a role in language acquisition also remains an open question. To the extent that they do, a third prediction follows: children should be better able to generalize comprehension of *novel* verbs when they are presented in frames consistent with these regularities. We are designing an experiment to test this hypothesis.

The argument that the frequency of pronouns and their co-occurrences with verb classes play a role in the acquisition of verbs could be strengthened by showing that it is true in many languages. The present study considered only English, which is a relatively noun-heavy language in which argument ellipsis is rare. Some other languages, by contrast, tend to emphasize verbs and frequently drop nominal arguments. We are especially keen to find out what sorts of cues children might be using to identify verb classes in such languages. Hence, work is underway to collect comparable data from Japanese and Tamil, verb-heavy languages with frequent argument dropping and case-marked pronouns reflecting various degrees of social status.

5 Acknowledgements

This research was supported by NIMH grant number ROI MH 60200. Additional thanks go to our coders, to members of the Cognitive Development Laboratory at IU for useful discussions of these results, and to several anonymous reviewers for helpful comments.

References

Christopher M. Bishop. 1995. *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press.

Thea Cameron-Faulkner, Elena V. M. Lieven, and Michael Tomasello. 2003. A construction-based analysis of child directed speech. *Cognitive Science* 27:843-873.

Wallace L. Chafe. 1994. *Discourse, Consciousness and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. Chicago: University of Chicago Press.

P. S. Dale, and L. Fenson. 1996. Lexical development norms for young children.

Behavioral Research Methods, Instruments & Computers 28:125-127.

Lila R. Gleitman. 1990. The structural sources of word meaning. *Language Acquisition* 1:3-55.

Lila R. Gleitman, and Jane Gillette. 1995. The role of syntax in verb learning. In *The Handbook of Child Language*, eds. Paul Fletcher and Brian MacWhinney, 413-427. Cambridge, MA: Blackwell.

Elena V. M. Lieven, Julian M. Pine, and Gillian Baldwin. 1997. Lexically-based learning and early grammatical development. *Journal of Child Language* 24:187-219.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*.vol. 2: The Database. Mahwah, NJ: Lawrence Erlbaum Associates.

Julian M. Pine, and Elena V. M. Lieven. 1993. Reanalysing rote-learned phrases: Individual differences in the transition to multi-word speech. *Journal of Child Language* 20:551-571.

Willard Van Orman Quine. 1980. On what there is. In *From a Logical Point of View*, ed. Willard Van Orman Quine. Cambridge, MA: Harvard University Press.

Martin Riedmiller, and H. Braun. 1993. A direct adaptive method for faster backpropagation learning: The Rprop algorithm. Paper presented at *IEEE International Conference on Neural Networks 1993 (ICNN 93)*, San Francisco, CA.

Michael Tomasello. 1992. *First Verbs: A Case Study of Early Grammatical Development*. Cambridge: Cambridge University Press.

Virginia Valian. 1991. Syntactic subjects in the early speech of American and Italian children. *Cognition* 40:21-81.

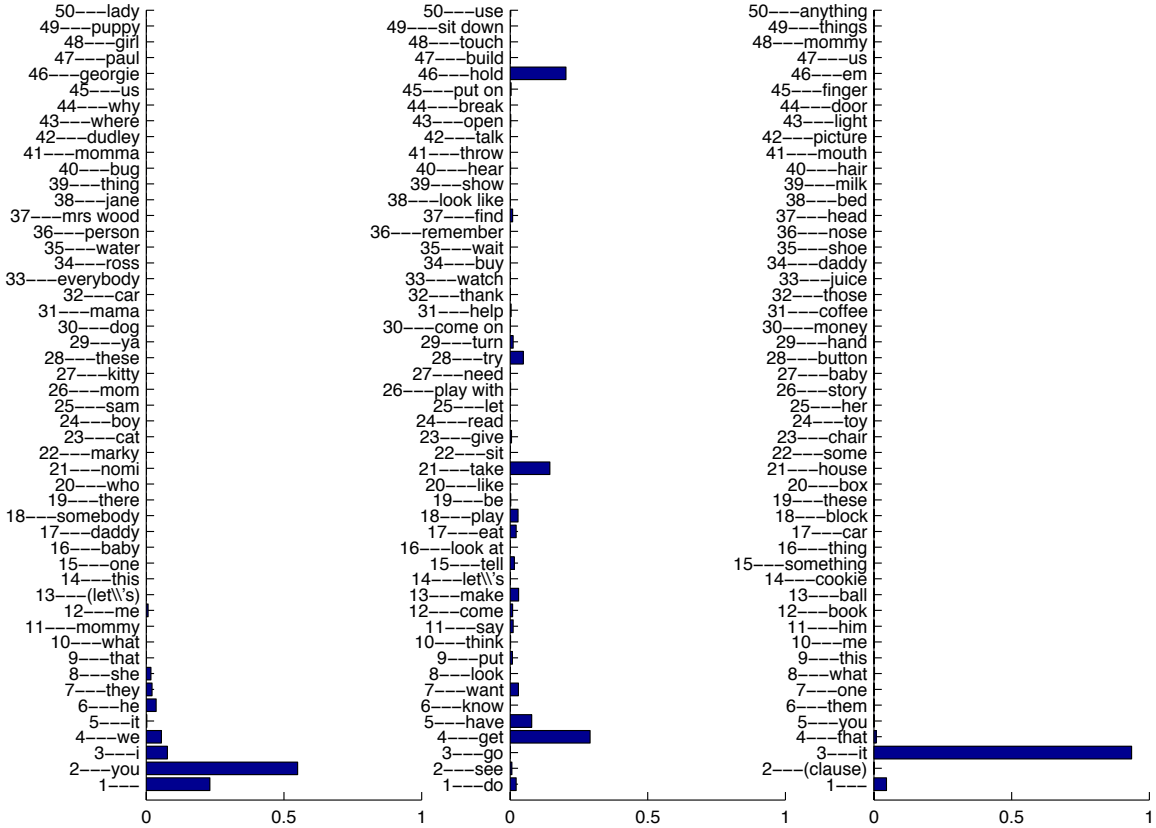


Figure 4: Average network output response to the object *it*. Subjects are shown in the left column, verbs in the middle, and objects on the right. Within each syntactic category, output units are ordered according to the frequency of the corresponding words in the input (lower numbers are higher frequency). The width of each bar reflects the average activation of the corresponding unit in our networks.

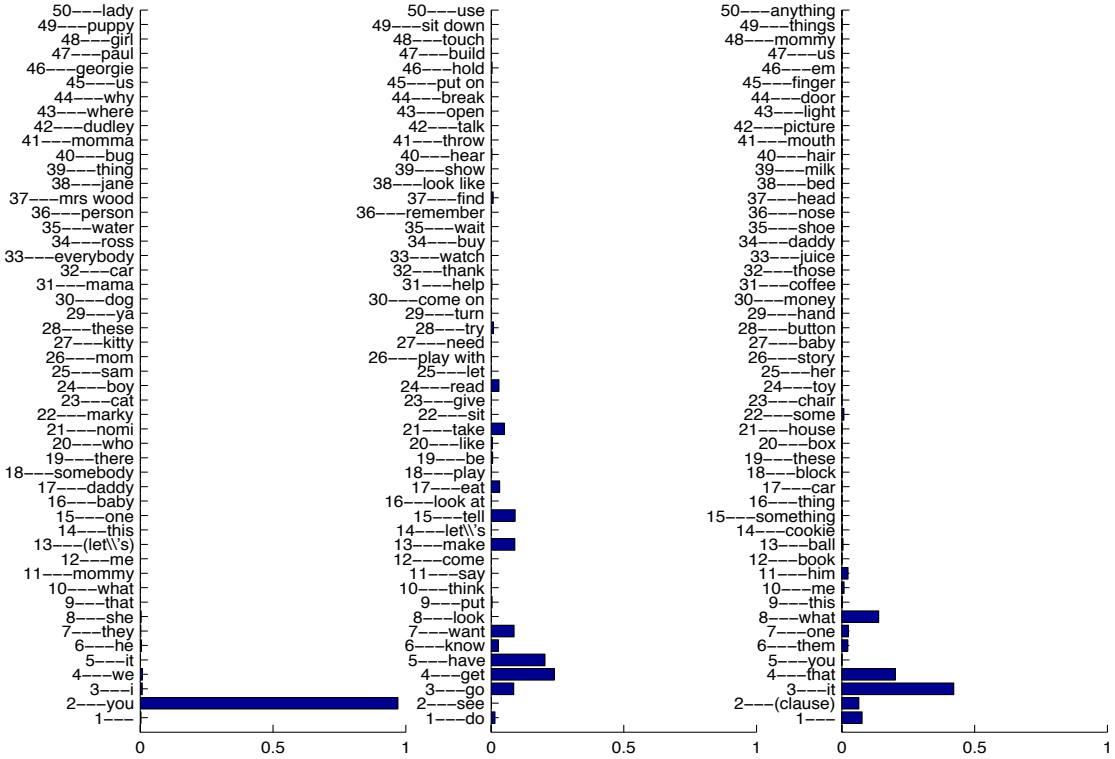


Figure 5: Average network output response to the subject *you*. Same conventions as previous figure.

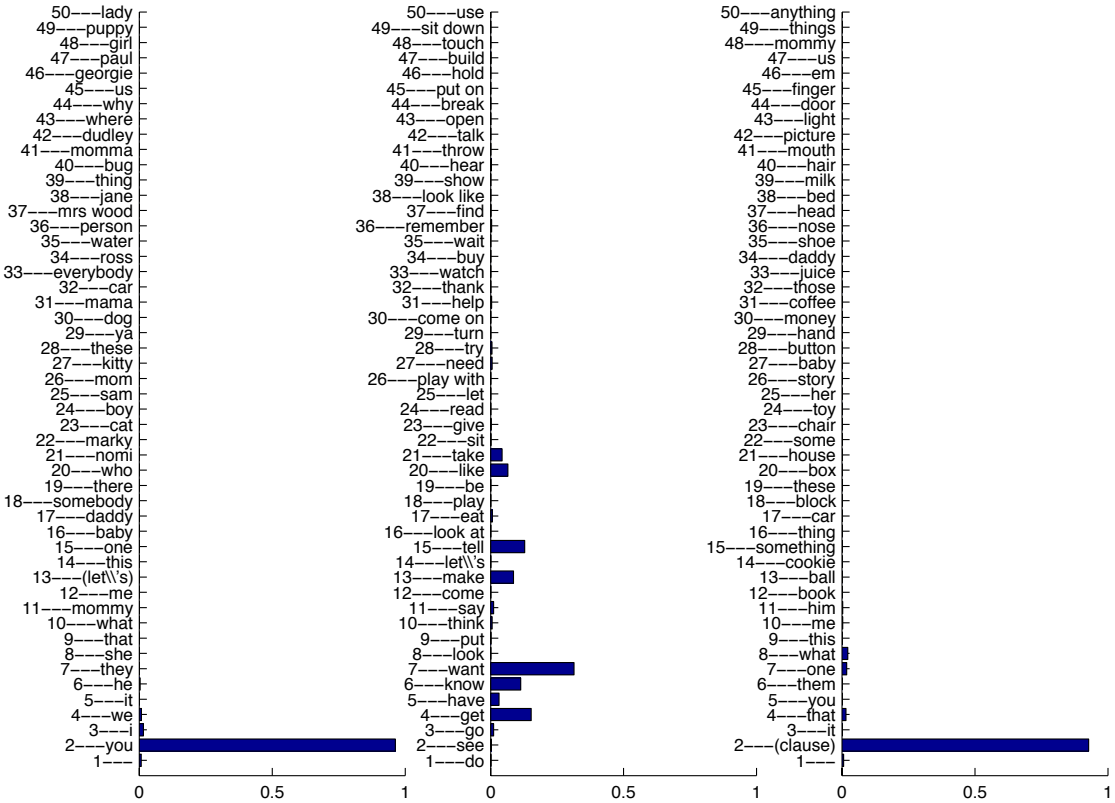


Figure 6: Average network output response to the subject *you* and the object *clause* simultaneously. Same conventions as Figures 4 and 5.

Some Tests of an Unsupervised Model of Language Acquisition

Bo Pedersen and Shimon Edelman
Department of Psychology
Cornell University
Ithaca, NY 14853, USA
{*bp64,se37*}*@cornell.edu*

Zach Solan, David Horn, Eytan Ruppin
Faculty of Exact Sciences
Tel Aviv University
Tel Aviv, Israel 69978
{*zsolan,horn,ruppin*}*@post.tau.ac.il*

Abstract

We outline an unsupervised language acquisition algorithm and offer some psycholinguistic support for a model based on it. Our approach resembles the Construction Grammar in its general philosophy, and the Tree Adjoining Grammar in its computational characteristics. The model is trained on a corpus of transcribed child-directed speech (CHILDES). The model's ability to process novel inputs makes it capable of taking various standard tests of English that rely on forced-choice judgment and on magnitude estimation of linguistic acceptability. We report encouraging results from several such tests, and discuss the limitations revealed by other tests in our present method of dealing with novel stimuli.

1 The empirical problem of language acquisition

The largely unsupervised, amazingly fast and almost invariably successful learning stint that is language acquisition by children has long been the envy of computer scientists (Bod, 1998; Clark, 2001; Roberts and Atwell, 2002) and a daunting enigma for linguists (Chomsky, 1986; Elman et al., 1996). Computational models of language acquisition or "grammar induction" are usually divided into two categories, depending on whether they subscribe to the classical generative theory of syntax, or invoke "general-purpose" statistical learning mechanisms. We believe that polarization between classical and statistical approaches to syntax hampers the integration of the stronger aspects of each method into a common powerful framework. On the one hand, the statistical approach is geared to take advantage of the considerable progress made to date in the areas of distributed representation and probabilistic learning, yet generic "connectionist" architectures are ill-suited to the abstraction and processing of symbolic information. On the other hand, classical rule-based systems excel in just those tasks, yet are brittle and difficult to train.

We are developing an approach to the acquisition of distributional information from raw input (e.g., transcribed speech corpora) that also supports the distillation of structural regularities comparable to those captured by Context Sensitive Grammars out of the accrued statistical knowledge. In thinking about such regularities, we adopt Langacker's notion of grammar as "simply an inventory of linguistic units" ((Langacker, 1987), p.63). To detect potentially useful units, we identify and process partially redundant sentences that share the same word sequences. We note that the detection of paradigmatic variation within a slot in a set of otherwise identical aligned sequences (syntagms) is the basis for the classical distributional theory of language (Harris, 1954), as well as for some modern work (van Zaanen, 2000). Likewise, the *pattern* — the syntagm and the *equivalence class* of complementary-distribution symbols that may appear in its open slot — is the main representational building block of our system, ADIOS (for Automatic DIstillation Of Structure).

Our goal in the present short paper is to illustrate some of the capabilities of the representations learned by our method vis a vis standard tests used by developmental psychologists, by second-language instructors, and by linguists. Thus, the main computational principles behind the ADIOS model are outlined here only briefly. The algorithmic details of our approach and accounts of its learning from CHILDES corpora appear elsewhere (Solan et al., 2003a; Solan et al., 2003b; Solan et al., 2004; Edelman et al., 2004).

2 The principles behind the ADIOS algorithm

The representational power of ADIOS and its capacity for unsupervised learning rest on three principles: (1) probabilistic inference of pattern significance, (2) context-sensitive generalization, and (3) recursive construction of complex patterns. Each of these is described briefly below.

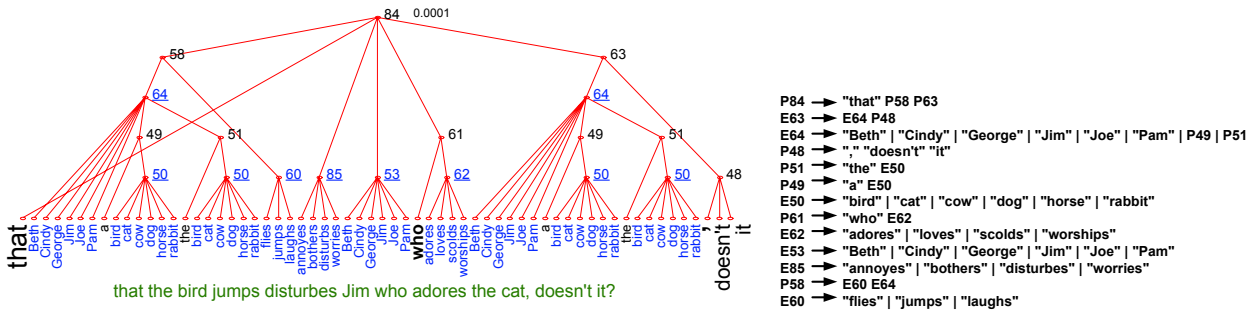


Figure 1: *Left:* a pattern (presented in a tree form), capturing a long range dependency (equivalence class labels are underscored). This and other examples here were distilled from a 400-sentence corpus generated by a 40-rule Context Free Grammar. *Right:* the same pattern recast as a set of rewriting rules that can be seen as a Context Free Grammar fragment.

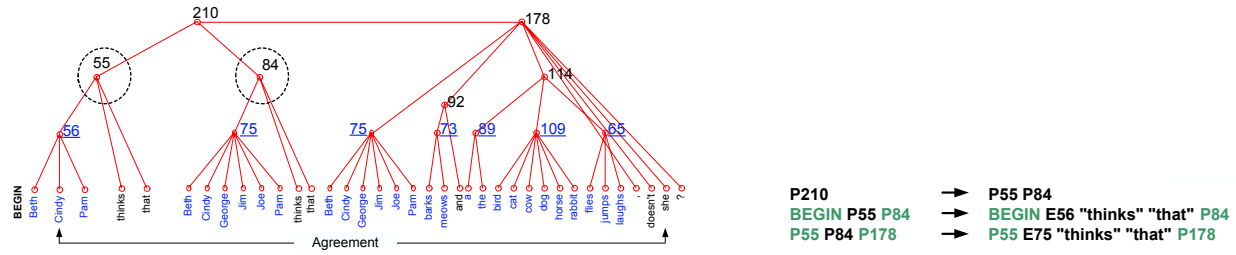


Figure 2: *Left:* because ADIOS does not rewire all the occurrences of a specific pattern, but only those that share the same context, its power is comparable to that of Context Sensitive Grammars. In this example, equivalence class #75 is not extended to subsume the subject position, because that position appears in a different context (e.g., immediately to the right of the symbol BEGIN). Thus, long-range agreement is enforced and over-generalization prevented. *Right:* the context-sensitive "rules" corresponding to pattern #210.

Probabilistic inference of pattern significance. ADIOS represents a corpus of sentences as an initially highly redundant directed graph, which can be informally visualized as a tangle of strands that are partially segregated into *bundles*. Each of these consists of some strands clumped together; a bundle is formed when two or more strands join together and run in parallel and is dissolved when more strands leave the bundle than stay in. In a given corpus, there will be many bundles, with each strand (sentence) possibly participating in several. Our algorithm, described in detail in (Solan et al., 2004), identifies significant bundles that balance high compression (small size of the bundle "lexicon") against good generalization (the ability to generate new grammatical sentences by splicing together various strand fragments each of which belongs to a different bundle).

Context sensitivity of patterns. A pattern is an abstraction of a bundle of sentences that are identical up to variation in one place, where one of several symbols — the members of the equivalence class associated with the pattern — may appear (Figure 1). Because this variation is only allowed in the context specified by the pattern, the generalization afforded by a set of patterns is inherently safer than in approaches that posit globally valid categories ("parts of speech") and rules ("grammar"). The reliance of ADIOS on many context-sensitive patterns rather than on traditional rules can be compared both to the Construction Grammar (discussed later) and to Langacker's concept of the grammar as a collection of "patterns of all intermediate degrees of generality" ((Langacker, 1987), p.46).

Hierarchical structure of patterns. The ADIOS graph is rewired every time a new pattern is detected, so that a bundle of strings subsumed by it is represented by a single new edge. Following the rewiring, which is context-specific, potentially far-apart symbols that used to straddle the newly abstracted pattern become close neighbors. Patterns thus become hierarchically structured in that their elements may be either terminals (i.e., fully specified strings) or other patterns. Moreover, patterns may refer to themselves, which opens the door for recursion.

3 Related computational and linguistic formalisms and psycholinguistic findings

Unlike ADIOS, very few existing algorithms for unsupervised language acquisition use raw, unannotated corpus data (as opposed, say, to sentences converted into sequences of POS tags). The only work described in a recent review (Roberts and Atwell, 2002) as completely unsupervised — the GraSp model (Henrichsen, 2002) — does attempt to induce syntax from raw transcribed speech, yet it is not completely data-driven in that it makes a prior commitment to a particular theory of syntax (Categorial Grammar, complete with a pre-specified set of allowed categories). Because of the unique nature of our chosen challenge — finding structure in language rather than imposing it — the following brief survey of grammar induction focuses on contrasts and comparisons to approaches that generally stop short of attempting to do what our algorithm does. We distinguish between approaches that are motivated computationally (Local Grammar and Variable Order Markov models, and Tree Adjoining Grammar, discussed elsewhere (Edelman et al., 2004), and those whose main motivation is linguistic and cognitive psychological (Cognitive and Construction grammars, discussed below).

Local Grammar and Markov models. In capturing the regularities inherent in multiple crisscrossing paths through a corpus, ADIOS superficially resembles finite-state Local Grammars (Gross, 1997) and Variable Order Markov (VOM) models (Guyon and Pereira, 1995). The VOM approach starts by postulating a maximum-n structure, which is then fitted to the data by maximizing the likelihood of the training corpus. The ADIOS philosophy differs from the VOM approach in several key respects. *First*, rather than fitting a model to the data, we use the data to construct a (recursively structured) graph. Thus, our algorithm naturally addresses the inference of the graph's structure, a task that is more difficult than the estimation of parameters for a given configuration. *Second*, because ADIOS works from the bottom up in a recursive, data-driven fashion, it is less susceptible to complexity issues. It can be used on huge graphs, and may yield very large patterns, which in a VOM model would correspond to an unmanageably high order n. *Third*, ADIOS transcends the idea of VOM structure, in the following sense. Consider a set of patterns of the form $b_1[c_1]b_2[c_2]b_3$, etc. The equivalence classes $[\cdot]$ may include vertices of the graph (both words and word patterns turned into nodes), wild cards (i.e., any node), as well as ambivalent cards (any node or no node). This means that the terminal-level length of the string represented by a pattern does not have to be of a fixed length. This goes conceptually beyond the variable order Markov structure: $b_2[c_2]b_3$ do not have to appear in a Markov chain of a finite order $||b_2||+||c_2||+||b_3||$ because the size of $[c_2]$ is ill-defined, as explained above. *Fourth*, as we showed earlier (Figure 2), ADIOS incorporates both context-sensitive substitution and recursion.

Tree Adjoining Grammar. The proper place in the Chomsky hierarchy for the class of strings accepted by our model is between Context Free and Context Sensitive Languages. The pattern-based representations employed by ADIOS have counterparts for each of the two composition operations, substitution and adjoining, that characterize a Tree Adjoining Grammar, or TAG, developed by Joshi and others (Joshi and Schabes, 1997). Specifically, both substitution and adjoining are subsumed in the relationships that hold among ADIOS patterns, such as the membership of one pattern in another. Consider a pattern $\mathcal{P}_i$ and its equivalence class $\mathcal{E}(\mathcal{P}_i)$; any other pattern $\mathcal{P}_j \in \mathcal{E}(\mathcal{P}_i)$ can be seen as substitutable in $\mathcal{P}_i$. Likewise, if $\mathcal{P}_j \in \mathcal{E}(\mathcal{P}_i)$, $\mathcal{P}_k \in \mathcal{E}(\mathcal{P}_i)$ and $\mathcal{P}_k \in \mathcal{E}(\mathcal{P}_j)$, then the pattern $\mathcal{P}_j$ can be seen as adjoinable to $\mathcal{P}_i$. Because of this correspondence between the TAG operations and the ADIOS patterns, we believe that the latter represent regularities that are best described by Mildly Context-Sensitive Language formalism (Joshi and Schabes, 1997). Importantly, because the ADIOS patterns are learned from data, they already incorporate the constraints on substitution and adjoining that in the original TAG framework must be specified manually.

Psychological and linguistic evidence for pattern-based representations. Recent advances in understanding the psychological role of representations based on what we call patterns, or *constructions* (Goldberg, 2003), focus on the use of statistical cues such as conditional probabilities in pattern learning (Saffran et al., 1996; Gómez, 2002), and on the importance of exemplars and constructions in children's language acquisition (Cameron-Faulkner et al., 2003). Converging evidence for the centrality of pattern-like structures is provided by corpus-based studies of *prefabs* — sequences, continuous or discontinuous, of words that appear to be prefabricated, that is, stored and retrieved as a whole, rather than being subject to syntactic processing (Wray, 2002). Similar ideas concerning the ubiquity in syntax of structural peculiarities hitherto marginalized as "exceptions" are now being voiced by linguists (Culicover, 1999; Croft, 2001).

Cognitive Grammar; Construction Grammar. The main methodological tenets of ADIOS — populating the lexicon with "units" of varying complexity and degree of entrenchment, and using cognition-general mechanisms for learning and representation — fit the spirit of the foundations of Cognitive Grammar (Langacker, 1987). At the same time, whereas the cognitive grammarians typically face the chore of hand-crafting structures that would reflect the logic of language as they perceive it, ADIOS discovers the primitives of grammar empirically and autonomously. The same is true also for the comparison between ADIOS and the various Construction Grammars (Goldberg, 2003; Croft, 2001), which are all hand-crafted. A construction grammar consists of elements that differ in their complexity and in the degree to which they are specified: an idiom such as "big deal" is a fully specified, immutable construction, whereas the expression "the X, the Y" – as in "the more, the better" (Kay and Fillmore, 1999) – is a partially specified template. The patterns learned by ADIOS likewise vary along the dimensions of complexity and specificity (e.g., not every pattern has an equivalence class).

4 ADIOS: a psycholinguistic evaluation

To illustrate the applicability of our method to real data, we first describe briefly the outcome of running it on a subset of the CHILDES collection (MacWhinney and Snow, 1985), consisting of transcribed speech directed at children. The corpus we selected contained $300,000$ sentences (1.3 million tokens) produced by parents. After 14 real-time days, the algorithm (version 7.3) identified 3400 patterns and 3200 equivalence classes. The outcome was encouraging: the algorithm found intuitively significant patterns and produced semantically adequate corresponding equivalence sets. The algorithm's ability to recombine and reuse the acquired patterns is exemplified in the legend of Figure 3, which lists some of the novel sentences it generated.

The input module. The ADIOS system's *input module* allows it to process a novel sentence by forming its distributed representation in terms of activities of existing patterns. *We stress that this module plays a crucial role in the tests described below, all of which require dealing with novel inputs.* Figure 4 shows the activation of two patterns (#141 and #120) by a phrase that contains a word in a novel context (stay), as well as another word never before encountered in any context (5pm).

Acceptability of correct and perturbed novel sentences. To test the quality of the representations

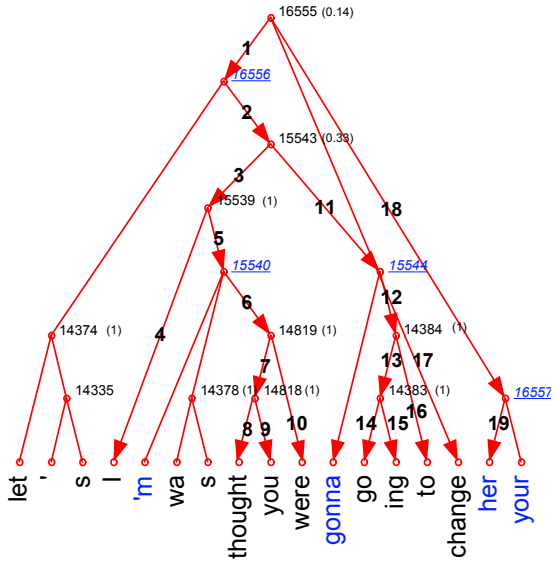


Figure 3: a typical pattern extracted from the CHILDES collection (MacWhinney and Snow, 1985). Hundreds of such patterns and equivalence classes (underscored) together constitute a concise representation of the raw data. Some of the phrases that can be described/generated by these patterns are: let's change her...; I thought you were gonna change her...; I was going to change your...; none of these appear in the training data, illustrating the ability of ADIOS to generalize. The generation process operates as a depth-first search of the tree corresponding to a pattern. For details see (Solan et al., 2003a; Solan et al., 2004).

(patterns and their associated equivalence classes) acquired by ADIOS, we have examined their ability to support various kinds of grammaticality judgments. The first experiment we report sought to make a distinction between a set of (presumably grammatical) CHILDES sentences not seen by the algorithm during training, and the same sentences in which the word order has been perturbed. We first trained the model on $10,000$ sentences from CHILDES, then compared its performance on (1) 1000 previously unseen sentences and (2) the same sentences in each of which a single random word order switch has been carried out. The results, shown in Figure 5, indicate a substantial sensitivity of the ADIOS input module to simple deviations from grammaticality in novel data, even after a very brief training.

Learnability of nonadjacent dependencies Within the ADIOS framework, the "nonadjacent dependencies" that characterize the artificial languages used by (Gómez, 2002) translate, simply, into patterns with embedded equivalence classes.

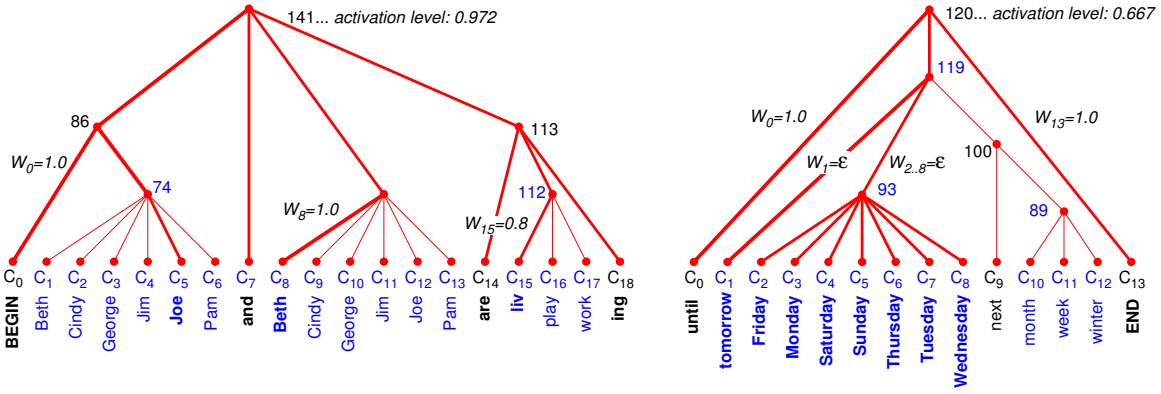


Figure 4: The two most active patterns responding to the partially novel input Joe and Beth are staying until 5pm. Leaf activation, which is proportional to the mutual information between input words and various members of the equivalence classes, is propagated upward by taking the average at each junction (Solan et al., 2003a).

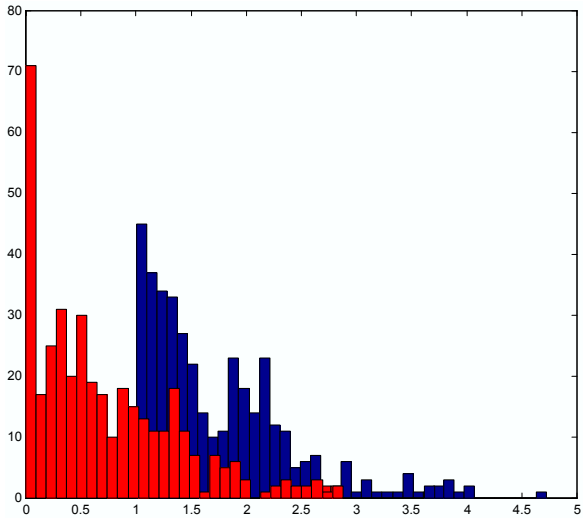


Figure 5: Grammaticality of perturbed sentences (CHILDES data). The figure shows a histogram of the input module output values for two kinds of stimuli: novel grammatical sentences (dark/blue), and sentences obtained from these by a single word-order permutation (light/red).

Gómez showed that the ability of subjects to learn a language L1 of the form $\{aXd, bXe, cXf\}$[1], as measured by their ability to distinguish it implicitly from L2=$\{aXe, bXf, cXd\}$, depends on the amount of variation introduced at X. We replicated this experiment by training ADIOS on 432 strings from L1, with $|X| = 2, 6, 12, 24$. The stimuli were the same strings as in the original experiment, with the individual letters serving as the basic symbols. A subsequent test resulted in a perfect acceptance of L1 and a perfect rejection of L2. Training with the original words (rather than letters) as the basic symbols resulted in L2 rejection rates of $0\%, 55\%, 100\%$, and 100%, respectively, for $|X| = 2, 6, 12, 24$. Thus, the ADIOS performance both mirrors that of the human subjects and suggests a potentially interesting new effect (of the granularity of the input stimuli) that may be explored in further psycholinguistic studies.

A developmental test. The CASL test (Comprehensive Assessment of Spoken Language) is widely used in the USA to assess language comprehension in children (Carrow-Woolfolk, 1999). One of its many components is a grammaticality judgment test, which consists of 57 sentences and is administered as follows: a sentence is read to the child, who then has to decide whether or not it is correct. If not, the child has to suggest a correct version of the sentence. For every incorrect sentence, the test lists 2-3 acceptable correct ones. The present version of the ADIOS algorithm can compare sentences but cannot score single sentences. We therefore ignored 11 out of the 57 sentences, which were correct to begin with. The remaining 46 incorrect sentences and their corrected versions were scored by ADIOS (which for this test had been trained on a 300,000-sentence corpus from the CHILDES database); the highest scoring sentence in each trial was interpreted as the model's choice. The model labeled 17 of the test sentences correctly, yielding a score of 108 (100 = norm) for the age interval 7-0 through 7-2. This score is the norm for the age interval 8-3 through 8-5.[2]

[1] Symbols $a - f$ here stand for nonce words such as pel, vot, or dak, whereas X denotes a slot in which a subset of 24 other nonce words may appear.

[2] ADIOS was undecided about the majority of the other sentences on which it did not score correctly.

		ADIOS		bi-gram	
benchmark	*#items*	*%correct*	*%answered*	*%correct*	*%answered*
Linebarger et al., 1983	26	**65**	**65**	42	92
Lawrence et al., 2000	70	**59**	**73**	38	63
Allen & Seidenberg, 1999	10	**83**	**60**	40	50
Martin & Miller, 2002	10	**75**	**80**	67	60
Göteborg/ESL	**100**	**58**	**57**	**45**	**20**
	216	**61**	**64**	42	34

Figure 6: The results of several grammaticality tests (the Göteborg ESL test is described in the text).

ESL test (forced choice). We next used a standard test developed for English as Second Language (ESL) classes, which has been administered in Göteborg (Sweden) to more than $10,000$ upper secondary levels students (that is, children who typically had 9 years of school, but only 6-7 years of English). The test consists of 100 three-choice questions, such as She asked me __ at once (choices: come, to come, coming) and The tickets have been paid for, so you __ not worry (choices: may, dare, need); the average score for the population mentioned is 65%. As before, the choice given the highest score by the algorithm won; if two choices received the same top score, the answer was "don't know". The algorithm's performance in this and several other tests is summarized in Figure 6 (these tests have been conducted with an earlier version of the algorithm (Solan et al., 2003a)). In the ESL test, ADIOS scored at just under 60%; compare this to the 45% precision (with 20% recall) achieved by a straightforward bi-gram benchmark.[3]

ESL test (magnitude estimation). In this experiment, six subjects were asked to provide magnitude estimates of linguistic acceptability (Gurman-Bard et al., 1996) for all the 3×100 sentences in the Göteborg ESL test. The test was paper based and included the instructions from (Keller, 2000). No measures were taken to randomize the order of the sentences or otherwise control the experiment. The same 300 sentences were processed by ADIOS, whose responses were normalized by dividing the output by the sum of each triplet's score. The results indicate a significant correlation ($R^2 = 6.3\%$, $p < 0.001$) between the scores produced by the subjects and by ADIOS. In some cases the scores of ADIOS are equal, which usually indicates that there are too many unfamiliar words. Omitting these sentences yields a significant $R^2 = 9.7\%$, $p < 0.001$; removing sentences for which the choices score almost equally (within 10%) results in $R^2 = 12.7\%$, $p < 0.001$.[4]

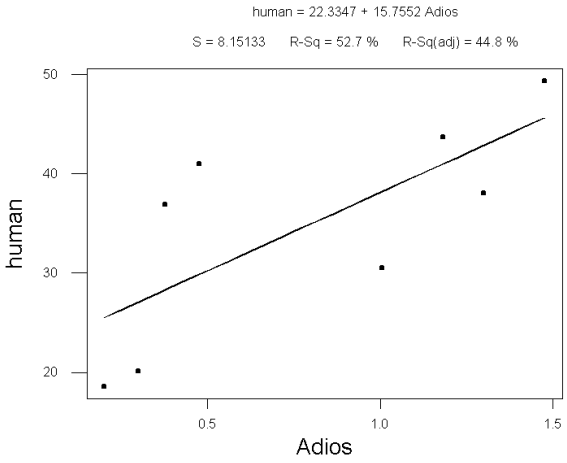


Figure 7: Magnitude estimation study from Keller, plotted against the ADIOS score on the same sentences ($R^2 = 0.53, p < 0.05$). The sentences (ranked by increasing score) are:
How many men did you destroy the picture of?
How many men did you destroy a picture of?
How many men did you take the picture of?
How many men did you take a picture of?
Which man did you destroy the picture of?
Which man did you destroy a picture of?
Which man did you take the picture of?
Which man did you take a picture of?

Modeling Keller's data. A manuscript by Frank Keller lists magnitude estimation data for eight sentences.[5] We compared these to the scores produced by ADIOS, and obtained a significant correlation (Figure 7). The input module seems capable of dealing with the substitution of a with the or of take with destroy, and it does reasonably well on the substitution of How many men with Which man. We conjecture that this performance can be improved by a more sophisticated normalization of the score produced by the input module, which should do a better job quantifying the *cover* (Edelman, 2004) of a novel sentence by the stored patterns. The limitations of the present version of the model became apparent when we tested it on the

[3] Chance performance in this test is 33%. We note that the corpus used here was too small to train an n-gram model for $n > 2$; thus, our algorithm effectively overcomes the problem of sparse data by putting the available data to a better use.

[4] Four of the subjects only filled out the test partially (the numbers of responses were 300, 300, 186, 159, 96, 60), but the correlation was highly significant despite the missing data.

[5] http://elib.uni-stuttgart.de/opus/volltexte/1999/81/pdf/81.pdf

52 sentences from Keller's dissertation, using his magnitude estimation method (Keller, 2000).[6] For these sentences, no correlation was found between the human and the model scores. One of the more challenging aspects of this set is the central role of pronoun binding in many of the sentences, e.g., The woman/Each woman saw Peter's photograph of her/herself/him/himself. Moreover, this test set contains examples of context effects, where information in an earlier sentence can help resolve a later ambiguity. Thus, many of the grammatical contrasts that appear in Keller's test sentences are too subtle for the present version of the ADIOS input module to handle.

Acceptability of correct and perturbed artificial sentences. In this experiment 64 random sentences was produced with a CFG. For uniformity the sentence length was kept within 15-20 words. 16 of the sentences had two adjacent words switched and another 16 had two random words switched. The 64 sentences were presented to 17 subjects, who placed each on a computer screen at a lateral position reflecting the perceived acceptability. As expected, the perturbed sentences were rated as less acceptable than the non-perturbed ones ($R^2 = 50.3\%$ with $p < 0.01$). We controlled for sentence number, for how high on the screen the sentence was placed, for the reaction time and for sentence length; only the latter had a significant contribution to the correlation. The random permutations scored significantly ($p < 0.01$) lower than the adjacent permutations. Furthermore, the variance in the scores of the randomly permuted sentences was significantly larger ($p < 0.005$), suggesting that this kind of permutation violates the sentence structure more severely, but may also sometimes create acceptable sentences by chance. Previous tests showed that ADIOS is very good at recognizing perturbed CFG-generated sentences as such, but it remains to be seen whether or not ADIOS also exhibits differential behavior on the adjacent and non-adjacent permutations.

Acceptability of ADIOS-generated sentences. ADIOS was trained on 12,700 sentences (out of a total of 12,966 sentences) in the ATIS (Air Travel Information System) corpus; the remaining 226 sentences were used for precision/recall tests. Because ADIOS is sensitive to the presentation order of the training sentences, 30 instances were trained on randomized versions of the training set. Eight human subjects were then asked to estimate acceptability of 20 sentences from the original corpus, intermixed randomly with 20 sentences generated by the trained versions of ADIOS. The precision, calculated as the average number of sentences accepted by the subjects divided by the total number of sentences in the set (20), was 0.73 ± 0.2 for sentences from the original corpus and 0.67 ± 0.07 for the sentences generated by ADIOS. Thus, the ADIOS-generated sentences are, on the average, as acceptable to human subjects as the original ones.

[6] We remark that this methodology is not without its problems. As one of our linguistically naive subjects remarked, "The instructions were (purposefully?) vague about what I was supposed to judge — understandability, grammar, correct use of language, or getting the point through...". Indeed, the scores in a magnitude experiment must be composites of several factors — at the very least, well-formedness and meaningfulness. We are presently exploring various means of acquiring and dealing with such multidimensional "acceptability" data.

5 Concluding remarks

The ADIOS approach to the representation of linguistic knowledge resembles the Construction Grammar in its general philosophy (e.g., in its reliance on structural generalizations rather than on syntax projected by the lexicon), and the Tree Adjoining Grammar in its computational capacity (e.g., in its apparent ability to accept Mildly Context Sensitive Languages). The representations learned by the ADIOS algorithm are truly emergent from the (unannotated) corpus data. Previous studies focused on the algorithm that makes such learning possible (Solan et al., 2004; Edelman et al., 2004). In the present paper, we concentrated on testing the input module that allows the acquired patterns to be used in processing novel stimuli.

The results of the tests we described here are encouraging, but there is clearly room for improvement. We believe that the most pressing issue in this regard is developing a conceptually and computationally well-founded approach to the notion of cover (that is, a distributed representation of a novel sentence in terms of the existing patterns). Intuitively, the best case, which should receive the top score, is when there is a single pattern that precisely covers the entire input, possibly in addition to other evoked patterns that are only partially active. We are currently investigating various approaches to scoring distributed representations in which several patterns are highly active. A crucial constraint that applies to such cases is that a good cover should give a proper expression to the subtleties of long-range dependencies and binding, many of which are already captured by the ADIOS learning algorithm.

Acknowledgments. Supported by the US-Israel Binational Science Foundation and by the Thanks to Scandinavia Graduate Scholarship at Cornell.

References

R. Bod. 1998. *Beyond grammar: an experience-based theory of language*. CSLI Publications, Stanford, US.

T. Cameron-Faulkner, E. Lieven, and M. Tomasello. 2003. A construction-based analysis of child directed speech. *Cognitive Science*, 27:843–874.

E. Carrow-Woolfolk. 1999. *Comprehensive Assessment of Spoken Language (CASL)*. AGS Publishing, Circle Pines, MN.

N. Chomsky. 1986. *Knowledge of language: its nature, origin, and use*. Praeger, New York.

A. Clark. 2001. *Unsupervised Language Acquisition: Theory and Practice*. Ph.D. thesis, COGS, U. of Sussex.

W. Croft. 2001. *Radical Construction Grammar: syntactic theory in typological perspective*. Oxford U. Press, Oxford.

P. W. Culicover. 1999. *Syntactic nuts: hard cases, syntactic theory, and language acquisition*. Oxford U. Press, Oxford.

S. Edelman, Z. Solan, D. Horn, and E. Ruppin. 2004. Bridging computational, formal and psycholinguistic approaches to language. In *Proc. of the 26th Conference of the Cognitive Science Society*, Chicago, IL.

S. Edelman. 2004. Bridging language with the rest of cognition: computational, algorithmic and neurobiological issues and methods. In M. Gonzalez-Marquez, M. J. Spivey, S. Coulson, and I. Mittelberg, eds., *Proc. of the Ithaca workshop on Empirical Methods in Cognitive Linguistics*. John Benjamins.

J. L. Elman, E. A. Bates, M. H. Johnson, A. Karmiloff-Smith, D. Parisi, and K. Plunkett. 1996. *Rethinking innateness: A connectionist perspective on development*. MIT Press, Cambridge, MA.

A. E. Goldberg. 2003. Constructions: a new theoretical approach to language. *Trends in Cognitive Sciences*, 7:219–224.

R. L. Gómez. 2002. Variability and detection of invariant structure. *Psychological Science*, 13:431–436.

M. Gross. 1997. The construction of local grammars. In E. Roche and Y. Schabès, eds., *Finite-State Language Processing*, pages 329–354. MIT Press, Cambridge, MA.

E. Gurman-Bard, D. Robertson, and A. Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language*, 72:32–68.

I. Guyon and F. Pereira. 1995. Design of a linguistic postprocessor using Variable Memory Length Markov Models. In *Proc. 3rd Int'l Conf. Document Analysis and Recogition*, pages 454–457, Montreal, Canada.

Z. S. Harris. 1954. Distributional structure. *Word*, 10:140–162.

P. J. Henrichsen. 2002. GraSp: Grammar learning form unlabeled speech corpora. In *Proceedings of CoNLL-2002*, pages 22–28. Taipei, Taiwan.

A. Joshi and Y. Schabes. 1997. Tree-Adjoining Grammars. In G. Rozenberg and A. Salomaa, eds., *Handbook of Formal Languages*, volume 3, pages 69 – 124. Springer, Berlin.

P. Kay and C. J. Fillmore. 1999. Grammatical constructions and linguistic generalizations: the What's X Doing Y? construction. *Language*, 75:1–33.

F. Keller. 2000. *Gradience in Grammar: Experimental and Computational Aspects of Degrees of Grammaticality*. Ph.D. thesis, U. of Edinburgh.

R. W. Langacker. 1987. *Foundations of cognitive grammar*, volume I: theoretical prerequisites. Stanford U. Press, Stanford, CA.

B. MacWhinney and C. Snow. 1985. The Child Language Exchange System. *Journal of Computational Lingustics*, 12:271–296.

A. Roberts and E. Atwell. 2002. Unsupervised grammar inference systems for natural language. Technical Report 2002.20, School of Computing, U. of Leeds.

J. R. Saffran, R. N. Aslin, and E. L. Newport. 1996. Statistical learning by 8-month-old infants. *Science*, 274:1926–1928.

Z. Solan, E. Ruppin, D. Horn, and S. Edelman. 2003a. Automatic acquisition and efficient representation of syntactic structures. In S. Thrun, editor, *Advances in Neural Information Processing*, volume 15, Cambridge, MA. MIT Press.

Z. Solan, E. Ruppin, D. Horn, and S. Edelman. 2003b. Unsupervised efficient learning and representation of language structure. In R. Alterman and D. Kirsh, eds., *Proc. 25th Conference of the Cognitive Science Society*, Hillsdale, NJ. Erlbaum.

Z. Solan, D. Horn, E. Ruppin, and S. Edelman. 2004. Unsupervised context sensitive language acquisition from a large corpus. In L. Saul, editor, *Advances in Neural Information Processing*, volume 16, Cambridge, MA. MIT Press.

M. van Zaanen. 2000. ABL: Alignment-Based Learning. In *COLING 2000 - Proceedings of the 18th International Conference on Computational Linguistics*, pages 961–967.

A. Wray. 2002. *Formulaic language and the lexicon*. Cambridge U. Press, Cambridge, UK.

Modelling Atypical Syntax Processing

Michael S. C. THOMAS
School of Psychology
Birkbeck College, Malet St.,
London WC1E 7HX
m.thomas@bbk.ac.uk

Martin REDINGTON
School of Psychology
Birkbeck College, Malet St.,
London WC1E 7HX
m.redington@ucl.ac.uk

Abstract

We evaluate the inferences that can be drawn from dissociations in syntax processing identified in developmental disorders and acquired language deficits. We use an SRN to simulate empirical data from Dick et al. (2001) on the relative difficulty of comprehending different syntactic constructions under normal conditions and conditions of damage. We conclude that task constraints and internal computational constraints interact to predict patterns of difficulty. Difficulty is predicted by frequency of constructions, by the requirement of the task to focus on local vs. global sequence information, and by the ability of the system to maintain sequence information. We generate a testable prediction on the empirical pattern that should be observed under conditions of developmental damage.

1 Dissociations in language function

Behavioural dissociations in language, identified both in cases of acquired brain damage in adults and in developmental disorders, have often been used to infer the functional components of the underlying language system. Generally these attempted fractionations appeal to broad distinctions within language. However, fine-scaled dissociations have also been proposed, such as the loss of individual semantic categories or of particular linguistic features in inflecting verbs. Here, we consider the implications of developmental and acquired deficits for the nature of syntax processing.

1.1 Developmental deficits

A comparison of developmental disorders such as autism, Downs syndrome, Williams syndrome, Fragile-X syndrome, and Specific Language Impairment reveals that dissociations can occur between phonology, lexical semantics, morpho-syntax, and pragmatics. The implications of such fractionations remain controversial but will be contingent on understanding the developmental origins of language structures (Karmiloff-Smith, 1998). These processes remain to be clarified even for the normal course of development.

In the area of syntax, Fowler (1998) concluded that a consistent picture emerges. Individuals with learning disabilities are systematic in their grammatical knowledge, follow the normal course of development, and show similar orders of difficulty in acquiring constructions. However, such individuals can often handle only limited levels of syntactic complexity and therefore development seems to terminate at a lower level.

While there is great variability in linguistic function both across different disorders and within single disorders, this cannot be attributed solely to differences in 'general cognitive functioning' (e.g., as assessed by problem solving ability). Syntax acquisition is therefore to some extent independent of IQ. However, adults with developmental disorders who have successfully acquired syntax typically have mental ages of at least 6 or 7, an age at which typically developing children also have well-structured language. The variability in outcome has been attributed to various factors specific to language, including verbal working memory and the quality of phonological representations (Fowler, 1998; McDonald, 1997). Most notably, disorders with different cognitive abilities show similarity in syntactic acquisition. The apparent lack of deviance across heterogeneous disorders has been used to argue for a model of language acquisition that is heavily constrained by the brain that is acquiring the language (Newport, 1990).

1.2 Acquired deficits in adulthood

One of the broadest distinctions in acquired language deficits is between *Broca's* and *Wernicke's* aphasia. Broca's aphasics are sometimes described as having greater deficits in grammar processing, and Wernicke's aphasics as having greater deficits in lexical processing. The dissociation is taken to support the idea that the division between grammar and the lexicon is one of the constraints that the brain brings to language acquisition.

Dick et al. (2001) recently argued that four types of evidence undermine this claim: (1) *all* aphasics

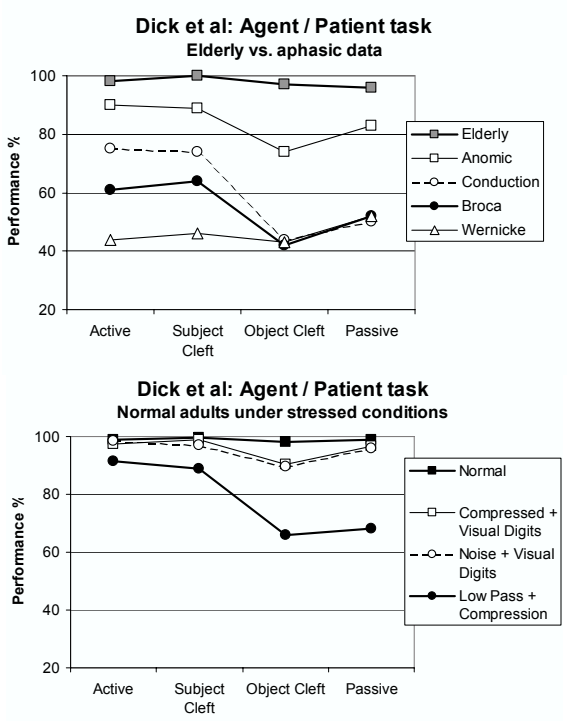
have naming deficits to some extent; (2) apparently agrammatic patients retain knowledge of grammar that can be exhibited in grammaticality judgements; (3) grammar deficits are found in many populations both with and without damage to Broca's area, the reputed seat of syntax in the brain; and (4) aphasic symptoms of language comprehension can be simulated in normal adults by placing them in stressed conditions (e.g., via manipulating the speech input or giving the subject a distracter task). Dick et al. pointed out that in syntax comprehension, the constructions most resilient in both aphasic patients and normal adults with simulated aphasia are those that are most regular or most frequent, and conversely those liable to errors are non-canonical and/or low frequency. Dick et al. (2001) illustrated these arguments in an experiment that compared comprehension of four complex syntactic structures:

- Actives (e.g., *The dog* [subject] *is biting the cow* [object])
- Subject Clefts (e.g., *It is the dog* [subject] *that is biting the cow* [object])
- Passives (e.g., *The cow* [object] *is bitten by the dog* [subject])
- Object Clefts (e.g., *It is the cow* [object] *that the dog* [subject] *is biting*)

The latter two constructions are lower frequency, and have non-canonical word orders in which the object precedes the subject. Dick et al. tested 56 adults with different types of aphasia on a task that involved identifying the agent of spoken sentences. Patients with all types of aphasia demonstrated lower performance on Passives and Object Clefts than Actives and Subject Clefts. Moreover, normal adults given the same task but with a degraded speech signal (either speeded up, low-pass filtered, or with noise added) or in combination with a distracter task (such as remembering a set of digits) produced a similar profile of performance to the aphasics (see Figure 1).

Dick et al. (2001) argued that the common pattern of deficits could be explained by the Competition Model (MacWhinney & Bates, 1989), which proposes that the difficulty of acquiring certain aspects of language and their retention after brain damage could be explained by considering *cue validity* (the reliability of a source of information in predicting the structure of a target language) and *cue cost* (the difficulty of processing each cue). Cues high in validity and low in cost, such as Subject-Verb-Object word order in English, should be acquired more easily and be relatively spared in adult breakdown. The proposal is that for a given language, any domain-general processing system placed under sub-optimal

Figure 1. Aphasic and simulated (human) aphasic data from Dick et al. (2001)



conditions should exhibit a similar pattern of developmental or acquired deficits. Thus Dick et al. predicted that a connectionist model trained on an appropriate frequency-weighted corpus would show equivalent vulnerability of non-canonical word orders and low frequency constructions under conditions of damage. In contrast to the inferences drawn from developmental deficits, the focus here is on attributing similarities in patterns of acquired deficits to features of the problem domain rather than constraints of the language system.

2 Computational modelling

Proposals that site the explanation of behavioural data in the frequency structure of the problem domain (here, the relative frequency of the construction types) are insufficient for three reasons: (1) language comprehension is not about passive reception. The language learner must do something with the words in order to derive the meanings of sentences. It is the nature of the transformations required that crucially determines task difficulty, which statistics of language input alone cannot reveal. (2) Whatever the statistics of the environment, such information must be accessed by an implemented learning system. This system may be differentially sensitive to certain features of the input, and it may find certain transformations more computationally expensive than others, further modulating task difficulty. (3) In the context of atypical syntax processing in developmental and acquired disorders, behavioural

deficits are caused by changes in internal computational constraints. Without an implemented, parameterised learning system, we can have no understanding of how sub-optimal processing conditions generate behavioural deficits in syntax processing. To date, this issue has been relatively under-explored.

The choice of learning system is evidently of importance here. In this paper, we explore the behaviour of a connectionist network, since these systems have been widely applied to phenomena within cognitive and language development (Elman et al., 1996) and more recently to capturing both atypical development and acquired deficits in adults (Thomas & Karmiloff-Smith, 2002, 2003).

3 Simulation Design

Our starting point is a set of models of syntax acquisition proposed by Christiansen and Dale (2001). These authors employed a simple recurrent network (SRN; Elman, 1990), an architecture that is the dominant connectionist model of sequence processing in language studies and in sequence learning more generally. As is typical of current connectionist models of syntax processing, the Christiansen and Dale (henceforth C&D) model focuses on small fragments of grammar and a small vocabulary. Nevertheless, it provides a useful platform to begin considering the effects of processing constraints on syntax processing.

The following models performed a prediction task at the word level. At each time step, the network was presented with the current word and had to predict the next word in the sentence. This component of the task induces sensitivity to syntactic structures. A localist representation was used, with each input unit corresponding to a single word. The artificial corpus consisted of 54 words and included 6 nouns, 10 verbs, 5 adjectives, and 10 functions words. Nouns and verbs had inflected forms represented by separate word units (N: stem, pluralised; V: stem, past tense, progressive, 3rd person singular).

C&D investigated the effect of several cues on syntax acquisition, such as prosody, stress, and word length. *Prosody* was represented as utterance boundary information that occurred at the end of an utterance with 92% probability. The utterance boundary cue was represented by an additional input and output unit.

Distributional cues of where words appeared in various sentences, along with utterance boundary information, were available to all networks. We refer to the networks that received only these cues as the "basic" model. We also tested a second set of "multiple cue" networks that also received cues about *word length* and *stress*. *Word length* was encoded with thermometer encoding, with one to three units being activated according to the number of syllables in the input word. In English, longer words tend to be content words. This was reflected in the vocabulary items that were selected for the grammar. *Stress* was encoded as a single unit that was activated for content words, which are stressed more heavily. The word length and stress units were present both as inputs and outputs, so that multiple cue networks had 59 input and output units to represent the words and cues.

3.1 The materials

The input corpus was a stochastic phrase structure grammar, derived from the materials used by C&D (2001). The grammar featured a range of constructions (imperatives, interrogatives and declarative statements). Frequencies were based on those observed in child-directed language. We added passives, subject and object cleft constructions to the grammar, which is illustrated in Figure 2.

Figure 2. Stochastic phrase structure grammar, including the probabilities of each construction

```
S -> Imperative [0.1] | Interrogative [0.3] | Declarative [0.6]

Declarative -> NP V-int [0.35] | NP V-tran NP active [.28] |
               NP V-tran NP passive [0.042] |
               subject cleft [0.014] |
               object cleft [0.014] | NP-Adj [0.1] |
               That-NP [0.075] | You-P [0.125]

NP-ADJ -> NP is/are adjective
          That-NP ->        that/those is/are NP
          You-P ->          you are NP

Imperative ->     VP
Interrogative ->  Wh-Question [0.65] | Aux-Question [0.35]
          Wh-Question ->  where / who / what is/are NP
                          [0.5] | where / who / what do /
                          does NP VP [0.5]
          Aux-Question -> do / does NP VP [0.33] |
                          do / does NP wanna VP [0.33] |
                          is / are NP adjective [0.34]

NP ->   a / the N-sing / N-plur
VP ->   V-int | V-trans NP
```

The four sentence types appeared with the following frequency: (Declarative) Active: 16.8%, Subject Cleft: 0.84%, Object Cleft: 0.84%, Passives: 2.52%. This gave a Passive-to-Active ratio of roughly 1:7, and ratio of OVS to SVO sentences of 1:21. Dick and Elman (2001) found that for English, the Passive-to-Active ratio ranged from 1:2 to 1:9 across corpora and that subject and object clefts appear in less than 0.05% of English sentences. They found that the relative frequency of word orders depended on whether one compares the passive OVS against transitive (SVO) or intransitive (SV) sentences and reported ratios that varied from 1:5 to 1:63 depending on corpus (spoken or written). The simulation frequencies were therefore an approximate fit, with the Subject

and Object Clefts slightly higher than in English due to the requirement to have at least a handful appear in our training corpus.

We generated a corpus of 10,000 sentences from this grammar as our training materials for the network, and a set of 100 test sentences for each of the active, passive, subject cleft and object cleft constructions.

3.2 Simulation One

The Dick et al. (2001) task consisted of presenting participants with a spoken sentence, and two pictures corresponding to the agent and patient of the sentence. The participant's task was to indicate with a binary choice which of the pictures was the agent of the sentence. For example, for sentences such as 'the dog is biting the cow', participants were asked to "press the button for the side of the animal that is doing the bad action".

Our next step was to implement this task in the model. One approach would be to train the network to output at each processing step not only the next predicted word in the sentence but also the thematic role of the current input. If the current input is a noun, this would be agent or patient. Joanisse (2000) proposed just such a solution to parsing in a connectionist model of anaphor resolution. We will refer to the implementation of activating units for agent or patient (solely) on the same cycle as the relevant noun as the "Discrete" mapping problem of relating nouns to roles.

The mapping problem adds to the difficulty of the prediction task. We can assess the extent of this difficulty by measuring performance on the prediction component alone, against the metrics of two statistical models. The bigram and trigram models are statistical descriptions of the sentence set that predict the next word given the previous two or three words of context, respectively, and these were derived from the observed frequencies in the training set.

Lastly, for the purposes of this simulation, we do not distinguish between the syntactic roles of subject and object, and semantic roles of agent and patient, even though a more complex model may separate these levels and include a process that maps between them. Although these simulations conflate the syntactic and semantic categories, we use the terms agent / patient for clarity in linking to the Dick et al. empirical data.

3.2.1 Method

For Simulation 1, we added two output units to the C&D network. The network was trained to activate the first extra unit when the current input element was the subject / agent of the sentence, and to activate the second extra unit when the object / patient of the sentence was presented. For all other inputs, the target activation of both units was zero. Thus, the number of input and output units was 55 and 57 respectively for the basic model, and 59 units and 61 units for the multiple-cue model.

The network's ability to correctly predict the next word was measured over the 55 word output units using the cosine between the target and actual output vectors. On novel sentences, a perfect network will only be able to predict the next item probabilistically. However, over many test items, this measure gives a fair view of the network's performance and we followed C&D (2001) in using this measure.

We initially chose our parameters based on those used by C&D (2001). Our learning rate was 0.1, and we trained the network for ten epochs. We performed a simple search of the parameter space for the number of hidden units to establish a "normal" condition (see Thomas & Karmiloff-Smith, 2003, for discussion of parameters defining normality). Eighty hidden units, the number used by C&D, gave adequate results for both models. This value was used to define the normal model.

We first evaluate normal performance at the end of training, then under the developmental deficit of a reduction in hidden units in the start state, and finally under the acquired deficit of a random lesion to a proportion of connection weights from the trained network.

3.2.2 Results

On the prediction component of the task, both models demonstrated better prediction ability than the bigram model, and marginally less prediction ability than the trigram model. This is in contrast to C&D's original prediction-only SRN model, which exceeded trigram model performance. It shows that the requirement to derive agent and patient roles increased the complexity of the learning problem, interfering with prediction ability.

The role-assignment component of the task was indexed by the activation of the agent and patient units *when presented with the second noun of the sentence*. At presentation of the first noun, there was no information available in the test sentences that would allow the network to distinguish between the possible interpretations of the sentence. At the second noun, the most active of the two units was assumed to drive the interpretation of the sentence and subsequent picture identification in the Dick et al. task. Therefore, the network's response was "correct" for Active and Subject Cleft sentences if the "patient" unit had the highest activation, and for Passive and Object Cleft sentences if the "agent"

Figure 3

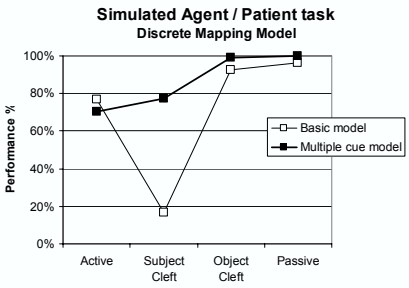


Figure 8

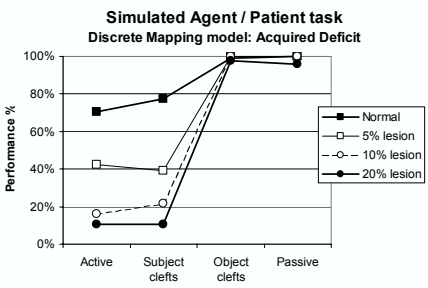


Figure 9

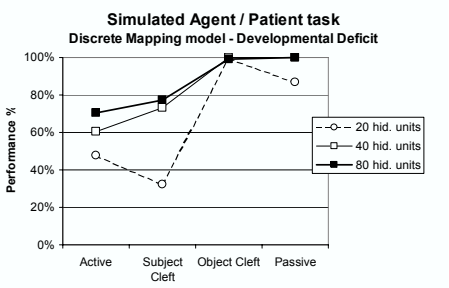


Figure 4

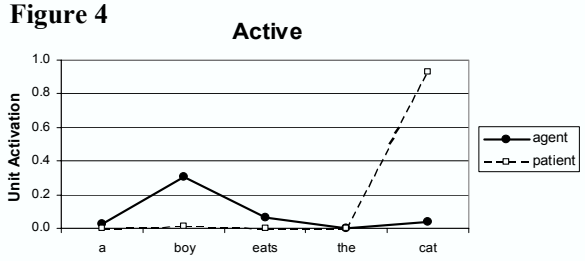


Figure 6

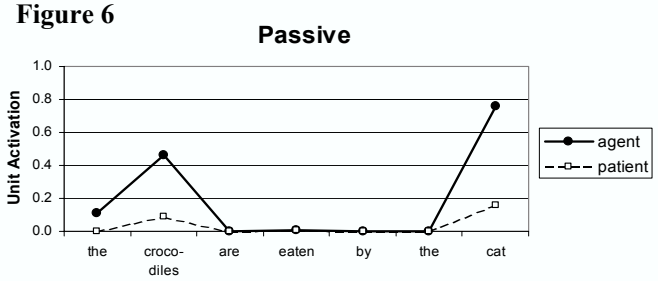


Figure 5

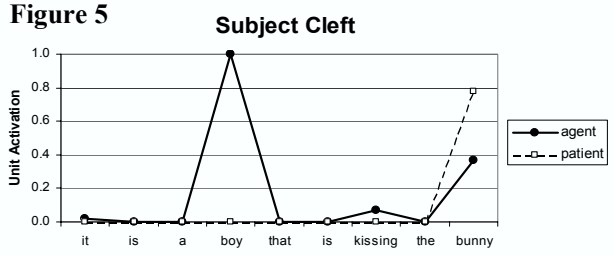
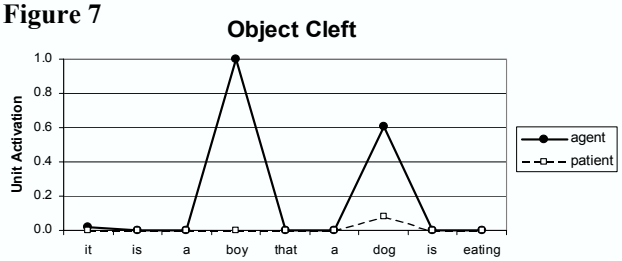


Figure 7



unit had the highest activation. The scores, measured in terms of the proportion of correct interpretations for the test sentences for each construction are shown in Figure 3.

Somewhat surprisingly, both the basic and multiple-cue models exhibited better performance on the Passive and Object Cleft sentences than on Active and Subject Cleft sentences. (These differences were statistically reliable.) The main difference between the two models was lower performance on Subject Cleft in the basic model, implying that cues to content-word status help to disambiguate the two cleft constructions.

Examining the profiles of performance for each sentence type gives some insight into the dynamics of the networks. Figures 4 to 7 show the activation of the agent and patient units for the multiple-cue model during the processing of examples of each construction, selected at random. The Subject Cleft sentence shown in Figure 5 is typical of the pattern for both Active and Subject Cleft sentences. That is, agent unit activation is close to 1.0 at the first noun, while patient unit activation is close to zero. At the second noun, the network is usually able to correctly distinguish the patient, but some agent unit activation also occurs. Therefore, using our decision criteria, the network is not always able to correctly identify the patient, and scores on Active and Subject Cleft sentences are not perfect.

In contrast, in the example Passive and Object Cleft sentences, the network incorrectly activates the agent unit at presentation of the first noun. At this point, the network has no information that could possibly allow it to distinguish between the two different kinds of sentence, and so its response is driven by the relative frequency of the constructions. However, for the second noun (the agent), although the patient unit does show some activation, the agent unit is clearly favoured.

Generally, the advantage of the agent unit for the Passive and Object Cleft sentences is greater than the advantage of the patient unit for the Active and Subject Cleft sentences. This can be explained by a general bias in the network in favour of the agent unit. In the training set, agents (subjects) occur much more frequently than patients (objects). All of the interrogatives and imperatives only have agents, and these comprise 30% of the training sentences. Thus, paradoxically, the network suffers when attempting to produce activation on the patient unit, and this impacts on the Active and Subject Cleft performance, despite the much greater frequency of these constructions.

Figures 8 and 9 illustrate the affects of initially reducing the numbers of hidden units in the network and of lesioning connections in the

endstate. In both cases, non-optimal processing conditions exaggerated the pattern of task difficulty, with Actives and Subject Clefts failing to be learned or showing greater impairment after lesioning. Object Clefts are the most easily learnt and most robust to damage, despite their non-canonical word order and low frequency. With the task definition of responding "agent" to the second noun, this construction gains most from the prevalence of the agent status of nouns in the corpus.

This interpretation of the Dick et al. agent-identification task does not provide an adequate fit to the human data, either for normal or atypical performance. Why not? This implementation of the task requires that the network keep track of two roles at the same time and assign those roles at the correct moment. It is therefore driven by the independent probability of a noun being an agent or a patient at multiple time points through the sentence. The result is a de-emphasis of global sequence information and an emphasis on local lexical information, leading to a relative advantage of responding 'agent' to *any* noun.

In the Dick et al. task, the participant is asked to make a single decision based on the entire sentence, rather than continously monitor word-by-word probabilities. Responses occurred between 2 and 4 seconds after sentence onset, with words presented at around 3 words-per-second. In the next section, we therefore provide an alternate implementation of the task based on a single categorisation decision for the whole sentence. But Simulation 1 serves as a demonstration that the statistics of the input set alone do not generate the task difficulty. It is the mappings required of the network. Moreover, we might predict that a modification of the Dick et al. study to encourage on-line monitoring of roles would alter the pattern of task difficulty. Thus, the four options might be presented as pictures (each noun twice, once as agent, once as patient), and the participants' eye-gaze direction recorded as the sentence unfolds.

3.3 Simulation Two

An alternate implementation of the Dick et al. task is that the network should be required to make a single categorisation on the whole sentence as to whether the agent precedes the patient, or the patient precedes the agent. This implementation follows the assumption that task performance is driven by higher-level sentence-based information rather than lexically-based information. A single unit can serve to categorise the input sentence as agent-then-patient or patient-then-agent. During training, the target activation for the unit is applied continuously throughout the entire utterance. We therefore call this the Continuous Mapping problem for sentence comprehension. Like the Discrete Mapping problem, the Continuous version has also been employed in previous connectionist models of parsing (Miikkulainen & Mayberry, 1999). (Note that Morris, Cottrell & Elman, 2000, used an implementation that combines Discrete and Continuous methods, providing a training signal that is activated when a word appears and is then maintained until the end of the sentence). The Continous method generates a training signal for comprehension. It does not constrain on-line comprehension, which may be subject to garden-pathing and dynamic revision.

3.3.1 Method

A single output unit was trained to produce an activation of 1 for sentences with Subject-Object word order (active and subject cleft constructions), and 0 for Object-Subject word order (passives and object cleft constructions). Apart from this difference, the basic and multiple-cue models were identical in all other respects, with 55 input and output units in the basic model, and 59 units in the multiple cue model. As before, we trained the network on 10,000 sentences generated by the stochastic phrase structure grammar, and tested the trained network on sets of 100 Active, Passive, Subject Cleft and Object Cleft sentences. One hundred and twenty hidden units were required to define the 'normal condition' for these simulations.

3.3.2 Results

As with Simulation 1, the prediction ability of both basic and multiple-cue models suffered due to the burden imposed by the mapping task. Although the networks' performance reliably exceeded a bigram prediction model, the trigram statistical model was slightly superior.

The network's ability to correctly "interpret" the test sentences was measured as follows. If the semantic output unit's activation at the time of second noun presentation was greater than 0.5, then the response was assumed to indicate that the sentence had Subject-Object word order and the agent was the first noun. If the activation was less than or equal to 0.5, then the response was assumed to indicate that the sentence had Object-Subject word order and the agent was the second noun. Although the target output for the network was consistent throughout each sentence, we selected the presentation of the second noun as our point of measurement, as this was where the network's discrimination ability was greatest. Figure 10 depicts performance on the four constructions.

On Active, Subject Cleft, and Passive sentences the basic model showed appropriate performance,

Figure 10

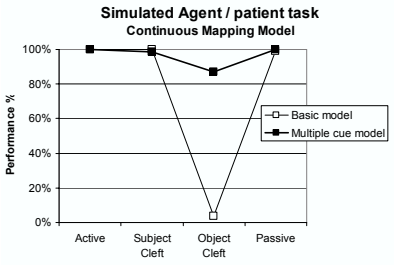


Figure 15

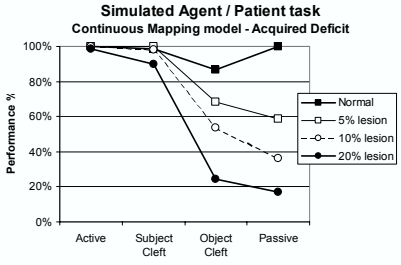


Figure 16

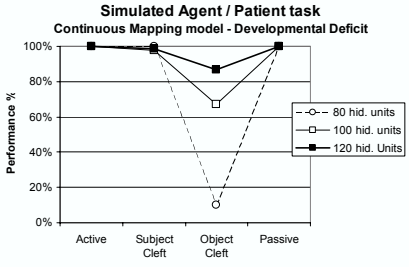


Figure 11

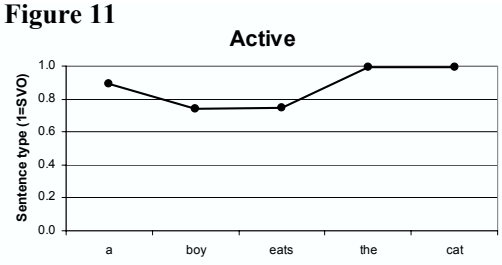


Figure 13

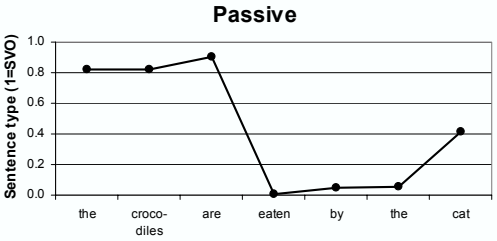


Figure 12

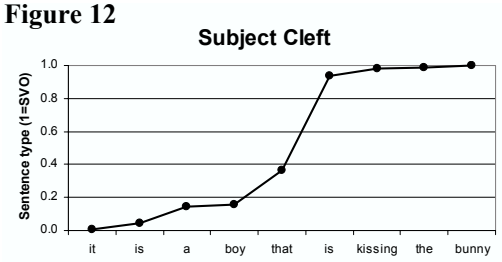
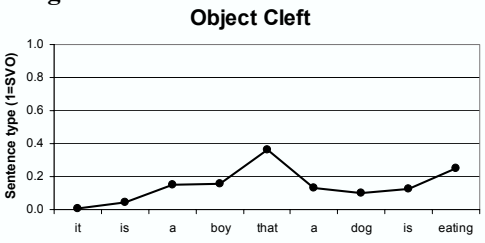


Figure 14



but it failed to correctly distinguish the Object Cleft sentences. Doubling the hidden units did not markedly alter this pattern. The multiple-cue model showed a much better fit to the human data, performing at close to ceiling for the Active, Passive and Subject Cleft constructions, and scoring in excess of 85% correct on Object Cleft constructions. The content-word cues provided in the multiple-cue model again appeared important in disambiguating the cleft constructions.

Focusing on the multiple-cue model, Figures 11-14 show the activation of the network's semantic output unit over a random sentence from each of the four test constructions. For the Active sentence, the network maintains a fairly constant high level of activation throughout the sentence. That is, it starts with the "assumption" that sentences will have a Subject-Object word order, and becomes more certain of this result (as shown by rising output activation) as the sentence proceeds.

For the Passive sentence, again, the network starts out assuming that the sentence will have the more frequent Subject-Object word order. But on seeing "eaten by", the network reverses its original diagnosis. However, the influence of this cue noticeably fades as the sentence proceeds. It persists enough that by the second noun, the network (just) manages to indicate correctly that the sentence has Object-Subject word order.

The Cleft constructions show a very different pattern. For the Subject Clefts, the network begins with a low output value from the semantic unit. This increases slightly as the first determiner and noun are presented, but the most valuable cue arrives with the words "that is kissing". These provide a perfect indicator (in this context) that the sentence has Subject-Object word order, and the activation of the semantic unit jumps dramatically, staying near ceiling for the rest of the sentence.

Finally, examining the Object Cleft sentence, output activation again starts low and rises only modestly during presentation of the first noun. However, the presence of a second noun following immediately after the first pulls the activation back down, to correctly indicate that the sentence has Object-Subject word order. Notice that, as with the Passive sentence, as the distance increases from the cue that marks the (less common) status of the Object Cleft sentence, so the activation level of the semantic unit tends to drift back to the default of the more frequent constructions.

Figures 15 and 16 illustrate, respectively, the effects of reducing the initial numbers of hidden units in the network and of lesioning connections in the endstate. In the case of acquired damage, non-optimal processing conditions exaggerate the

pattern of task difficulty, with Passives and Object Cleft's showing greater impairment after lesioning in line with the empirical data in Figure 1. Interestingly, in the case of the developmental deficit, the pattern is subtly different. While Object Clefts show increased vulnerability, Passives are far more resilient to developmental damage.

We carried out further analysis of this difference. Using the examples in Figs. 13 and 14, the cues predicting Object-Subject order for Passives turned out to be the inflected verb 'eaten' followed by 'by', i.e., two *lexical* cues (the second redundant). For Object Clefts, the cue for Object-Subject order was *sequence-based* information: in this construction, two nouns are not separated by a verb. This is marked by the arrival of a second noun prior to a verb, that is, the words 'a' and 'dog'. While both lexical and sequence cues are low frequency by virtue of their constructions, they differ in that the Passive cue comprises lexical items unique to this construction, while the Object Cleft cue involves a particular sequence of lexical items that also appear in other other constructions. Examination of activation dynamics reveals that both low frequency cues are lost after acquired damage. However, the network with the developmental deficit retains the ability to learn the lexically-based cue that marks the Passive, but has insufficient resources to learn the sequence-based cue that marks the Object Cleft construction.

Three points are evident here. First, the model makes a strong empirical prediction that when developmental deficits are compared to acquired deficits, passive constructions will be relatively less vulnerable. This renders the model testable and therefore falsifiable. Second, the model demonstrates the differential computational requirements of tasks driven by local (lexically-based) and global (sequence-based) information in a parsing task. Third, the model reveals the distinction between acquired and developmental deficits, with compensation possible in the latter case for cues with low processing cost (see Thomas & Karmiloff-Smith, 2002, for discussion).

4 Discussion

Implemented learning models are an essential requirement to begin an exploration of the internal constraints that influence successful and atypical syntax processing. Our model necessarily makes simplifications to begin this exploration (e.g., the distribution and frequency of lexical items across constructions is not in reality uniform; cleft constructions may have different stress / prosodic cues). A precise quantitative fit to the empirical data must await models that include those factors.

However, the current model is sufficient to demonstrate the importance of the mapping task in specifying difficulty (over and above the statistics of the input); how internal processing constraints influence performance; and how local and global information show a differential contribution to and vulnerability in sequence processing in a recurrent connectionist network.

5 Acknowledgements

This research was supported by grants from the British Academy and the Medical Research Council (G0300188) to Michael Thomas.

References

Christiansen, M. & Dale, R. 2001. Integrating distributional, prosodic and phonological information in a connectionist model of language acquisition. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (p. 220-225). Mahwah, NJ: LEA.

Dick, F. & Elman, J. 2001. The frequency of major sentence types over discourse levels: A corpus analysis. *CRL: Newsletter*, 13.

Dick, F., Bates, E., Wulfeck, B., Aydelott, J., Dronkers, N., & Gernsbacher, M. 2001. Language deficits, localization, and grammar: Evidence for a distributive model of language breakdown in aphasic patients and neurologically intact individuals. *Psychological Review*, 108(3): 759-788.

Elman, J. 1990. Finding structure in time. *Cognitive Science*, 14, 179-211.

Elman, J., et al., (1996). *Rethinking innateness*. Cambridge, Mass.: MIT Press.

Fowler, A. (1998). Language in mental retardation: Associations with and dissociations from general cognition. In J. Burack et al., *Handbook of Mental Retardation and Development* (p.290-333). Cambridge, UK: CUP.

Joanisse, M. 2000. *Connectionist phonology*. Unpublished Ph.D. Dissertation, University of Southern California.

Karmiloff-Smith, A. (1998). Development itself is the key to understanding developmental disorders. *Trends in Cognitive Sciences*, 2(10): 389-398.

MacWhinney, B. & Bates, E. 1989. *The cross-linguistic study of sentence processing*. New York: CUP.

McDonald, J. 1997. Language acquisition: The acquisition of linguistic structure in normal and special populations. *Annu. Rev. Psychol*., 48, 215-241

Miikkulainen, R. & Mayberry, M. 1999. Disambiguation and grammar as emergent soft constraints. In B. MacWhinney (ed.) *Emergence of Language*. Hillsdale, NJ: LEA.

Morris, W., Cottrell, G., and Elman, J. 2000. A connectionist simulation of the empirical acquisition of grammatical relations. In S. Wermter & R. Sun (eds.), *Hybrid Neural Systems*. Heidelberg: Springer Verlag.

Newport, E. 1990. Maturational constraints on language learning. *Cognitive Science*, 14, 11-28.

Thomas, M.S.C. & Karmiloff-Smith, A. (2002). Are developmental disorders like cases of adult brain damage? Implications from connectionist modelling. *Behavioural and Brain Sciences*, 25(6), 727-788.

Thomas, M.S.C. & Karmiloff-Smith, A. 2003. Modelling language acquisition in atypical phenotypes. *Psychological Review*, 110(4), 647-682.

Combining Utterance-Boundary and Predictability Approaches to Speech Segmentation

Aris XANTHOS
Linguistics Department, University of Lausanne
UNIL - BFSH2
1015 Lausanne
Switzerland
aris.xanthos@ling.unil.ch

Abstract

This paper investigates two approaches to speech segmentation based on different heuristics: the utterance-boundary strategy, and the predictability strategy. On the basis of former empirical results as well as theoretical considerations, it is suggested that the utterance-boundary approach could be used as a preprocessing step in order to lighten the task of the predictability approach, without damaging the resulting segmentation. This intuition leads to the formulation of an explicit model, which is empirically evaluated for a task of word segmentation on a child-oriented phonemically transcribed French corpus. The results show that the hybrid algorithm outperforms its component parts while reducing the total memory load involved.

1 Introduction

The design of speech segmentation[1] methods has been much studied ever since Harris' seminal propositions (1955). Research conducted since the mid 1990's by cognitive scientists (Brent and Cartwright, 1996; Saffran et al., 1996) has established it as a paradigm of its own in the field of computational models of language acquisition.

In this paper, we investigate two *boundary-based* approaches to speech segmentation. Such methods "attempt to identify individual word-boundaries in the input, without reference to words per se" (Brent and Cartwright, 1996). The first approach we discuss relies on the *utterance-boundary* strategy, which consists in reusing the information provided by the occurrence of specific phoneme sequences at utterance beginnings or endings in order to hypothesize boundaries inside utterances (Aslin et al., 1996; Christiansen et al., 1998; Xanthos, 2004). The second approach is based on the *predictability* strategy, which assumes that speech should be segmented at locations where some measure of the uncertainty about the next symbol (phoneme or syllable for instance) is high (Harris, 1955; Gammon, 1969; Saffran et al., 1996; Hutchens and Adler, 1998; Xanthos, 2003).

Our implementation of the utterance-boundary strategy is based on n-grams statistics. It was previously found to perform a "safe" word segmentation, that is with a rather high precision, but also too conservative as witnessed by a not so high recall (Xanthos, 2004). As regards the predictability strategy, we have implemented an incremental interpretation of the classical successor count (Harris, 1955). This approach also relies on the observation of phoneme sequences, the length of which is however not restricted to a fixed value. Consequently, the memory load involved by the successor count algorithm is expected to be higher than for the utterance-boundary approach, and its performance substantially better.

The experiments presented in this paper were inspired by the intuition that both algorithms could be combined in order to make the most of their respective strengths. The utterance-boundary typicality could be used as a computationally inexpensive preprocessing step, finding some true boundaries without inducing too many false alarms; then, the heavier machinery of the successor count would be used to accurately detect more boundaries, its burden being lessened as it would process the chunks produced by the first algorithm rather than whole utterances. We will show the results obtained for a word segmentation task on a phonetically transcribed and child-oriented French corpus, focusing on the effect of the preprocessing step on precision and recall, as well as its impact on

[1] To avoid a latent ambiguity, it should be stated that *speech segmentation* refers here to a process taking as input a sequence of symbols (usually phonemes) and producing as output a sequence of higher-level units (usually words).

memory load and processing time.

The next section is devoted to the formal definition of both algorithms. Section 3 discusses some issues related to the space and time complexity they involve. The experimental setup as well as the results of the simulations are described in section 4, and in conclusion we will summarize our findings and suggest directions for further research.

2 Description of the algorithms

2.1 Segmentation by thresholding

Many distributional segmentation algorithms described in the literature can be seen as instances of the following abstract procedure (Harris, 1955; Gammon, 1969; Saffran et al., 1996; Hutchens and Adler, 1998; Bavaud and Xanthos, 2002). Let S be the set of phonemes (or segments) in a given language. In the most general case, the input of the algorithm is an utterance of length l, that is a sequence of l phonemes $u := s_1 \dots s_l$ (where s_i denotes the i-th phoneme of u). Then, for $1 \leq i \leq l-1$, we insert a boundary after s_i iff $D(u,i) > T(u,i)$, where the values of the *decision variable* $D(u,i)$ and of the *threshold* $T(u,i)$ may depend on both the whole sequence and the actual position examined (Xanthos, 2003).

The output of such algorithms can be evaluated in reference to the segmentation performed by a human expert, using traditional measures from the signal detection framework. It is usual to give evaluations both for word and boundary detection (Batchelder, 2002). The *word precision* is the probability for a word isolated by the segmentation procedure to be present in the reference segmentation, and the *word recall* is the probability for a word occurring in the true segmentation to be correctly isolated. Similarly, the *segmentation precision* is the probability that an inferred boundary actually occurs in the true segmentation, and the *segmentation recall* is the probability for a true boundary to be detected.

In the remaining of this section, we will use this framework to show how the two algorithms we investigate rely on different definitions of $D(u,i)$ and $T(u,i)$.

2.2 Frequency estimates

Let $U \subseteq S^*$ be the set of possible utterances in the language under examination. Suppose we are given a corpus $C \subseteq U^T$ made of T successive utterances.

The absolute frequency of an n-gram $w \in S^n$ in the corpus is given by $n(w) := \sum_{t=1}^{T} n_t(w)$ where $n_t(w)$ denotes the absolute frequency of w in the t-th utterance of C. In the same way, we define the absolute frequency of w *in utterance-initial position* as $n(w|\mathrm{I}) := \sum_{t=1}^{T} n_t(w|\mathrm{I})$ where $n_t(w|\mathrm{I})$ denotes the absolute frequency of w *in utterance-initial position* in the t-th utterance of C (which is 1 iff the utterance begins with w and 0 otherwise). Similarly, the absolute frequency of w in utterance-*final* position is given by $n(w|\mathrm{F}) := \sum_{t=1}^{T} n_t(w|\mathrm{F})$.

Accordingly, the relative frequency of w obtains as $f(w) := n(w)/\sum_{\tilde{w}\in S^n} n(\tilde{w})$. Its relative frequencies in utterance-initial and -final position respectively are given by $f(w|\mathrm{I}) := n(w|\mathrm{I})/\sum_{\tilde{w}\in S^n} n(\tilde{w}|\mathrm{I})$ and $f(w|\mathrm{F}) := n(w|\mathrm{F})/\sum_{\tilde{w}\in S^n} n(\tilde{w}|\mathrm{F})$ [2].

Both algorithms described below process the input incrementally, one utterance after another. This implies that the frequency measures defined in this section are in fact evolving all along the processing of the corpus. In general, for a given input utterance, we chose to update n-gram frequencies first (over the whole utterance) before performing the segmentation.

2.3 Utterance-boundary typicality

We use the same implementation of the utterance-boundary strategy that is described in more details by Xanthos (2004). Intuitively, the idea is to segment utterances where sequences occur, which are typical of utterance boundaries. Of course, this implies that the corpus is segmented in utterances, which seems a reasonable assumption as far as language acquisition is concerned. In this sense, the utterance-boundary strategy may be viewed as a kind of learning by generalization.

Probability theory provides us with a straightforward way of evaluating how much an n-gram $w \in S^n$ is typical of utterance endings. Namely, we know that events "occurrence of n-gram w" and "occurrence of an n-gram in utterance-final position" are independent iff $p(w \cap \mathrm{F}) = p(w)p(\mathrm{F})$ or equivalently iff $p(w|\mathrm{F}) = p(w)$. Thus, using maximum-likelihood estimates, we may define the *typical-*

[2] Note that in general, $\sum_{\tilde{w}\in S^n} n(\tilde{w}|\mathrm{F}) = \sum_{\tilde{w}\in S^n} n(\tilde{w}|\mathrm{I}) = \tilde{T}$, where $\tilde{T} \leq T$ is the number of utterances in C that have a length greater than or equal to n.

ity of w in utterance-final position as:

$$t(w|\mathrm{F}) := \frac{f(w|\mathrm{F})}{f(w)} \quad (1)$$

This measure is higher than 1 iff w is more likely to occur in utterance-final position (than in any position), lower iff it is less likely to occur there, and equal to 1 iff its probability is independent of its position.

In the context of a segmentation procedure, this suggest a "natural" constant threshold $T(u,i) := 1$ (which can optionally be fine-tuned in order to obtain a more or less conservative result). Regarding the decision variable, if we were dealing with an utterance u of infinite length, we could simply set the order $r \geq 1$ of the typicality computation and define $d(u,i)$ as $t(s_{i-(r-1)} \dots s_i|\mathrm{F})$ (where s_i denotes the i-th phoneme of u). Since the algorithm is more likely to process an utterance of finite length l, there is a problem when considering a potential boundary close to the beginning of the utterance, in particular when $r > i$. In this case, we can compute the typicality of smaller sequences, thus defining the decision variable as $t(s_{i-(\tilde{r}-1)} \dots s_i|\mathrm{F})$, where $\tilde{r} := \min(r,i)$.

As was already suggested by Harris (1955), our implementation actually combines the typicality in utterance-*final* position with its analogue in utterance-*initial* position. This is done by taking the average of both statistics, and we have found empirically efficient to weight it by the relative lengths of the conditioning sequences:

$$D(u,i) := \frac{\tilde{r}}{\tilde{r}+\tilde{r}'}\, t(w|F) + \frac{\tilde{r}'}{\tilde{r}+\tilde{r}'}\, t(w'|I) \quad (2)$$

where $w := s_{i-(\tilde{r}-1)} \dots s_i \in S^{\tilde{r}}$, $w' := s_{i+1} \dots s_{i+\tilde{r}'} \in S^{\tilde{r}'}$, $\tilde{r} := \min(r,i)$ and $\tilde{r}' := \min(r, l-i)$. This definition helps compensate for the asymmetry of arguments when i is either close to 1 or close to l.

Finally, in the simulations below, we apply a mechanism that consists in incrementing $n(w|\mathrm{F})$ and $n(w'|\mathrm{I})$ (by one) whenever $D(u,i) > T(u,i)$. The aim of this is to enable the discovery of new utterance-boundary typical sequences. It was found to considerably raise the recall as more utterances are processed, at the cost of a slight reduction in precision (Xanthos, 2004).

2.4 Successor count

The second algorithm we investigate in this paper is an implementation of Harris' successor count (Harris, 1955), the historical source of all predictability-based approaches to segmentation. It relies on the assumption that in general, the diversity of possible phonemes transitions is high after a word boundary and decreases as we consider transitions occurring further inside a word.

The diversity of transitions following an n-gram $w \in S^n$ is evaluated by the *successor count* (or successor *variety*), simply defined as the number of different phonemes that can occur after it:

$$succ(w) := |\{s \in S | n(ws) > 0\}| \quad (3)$$

Transposing the indications of Harris in the terms of section 2.1, for an utterance $u := s_1 \dots s_l$, we define $D(u,i)$ as $succ(w)$ where $w := s_1 \dots s_i$, and $T(u,i)$ as $\max[D(u,i-1), D(u,i+1)]$. Here again a "backward" measure can be defined, the *predecessor count*:

$$predec(w) := |\{s \in S | n(sw) > 0\}| \quad (4)$$

Accordingly, we have $D'(u,i) = predec(w')$ where $w' := s_{i+1} \dots s_l$, and $T'(u,i) := \max[D'(u,i-1), D'(u,i+1)]$. In order to combine both statistics, we have found efficient to use a composite decision rule, where a boundary is inserted after phoneme s_i iff $D(u,i) > T(u,i)$ *or* $D'(u,i) > T'(u,i)$.

These decision variables differ from those used in the utterance-boundary approach in that there is no fixed bound on the length of their arguments. As will be discussed in section 3, this has important consequences for the complexity of the algorithm. Also, the threshold used for the successor count depends explicitely on both u and i: rather than seeking values higher than a given threshold, this method looks for *peaks* of the decision variable monitored over the input, whether the actual value is high or not. This is a more or less arbitrary feature of this class of algorithms, and much work remains to be done in order to provide theoretical justifications rather than mere empirical evaluations.

3 Complexity issues

It is not easy to evaluate the complexity of the algorithms discussed in this paper, which consist mainly in the space and time needed to store

and retrieve the necessary information for the computation of n-grams frequencies. Of course, this depends much on the actual implementation. For instance, in a rather naive approach, utterances can be stored as such and the memory load is then roughly equivalent to the size of the corpus, but computing the frequency of an n-gram requires scanning the whole memory.

A first optimization is to *count* utterances rather than merely store them. Some programming languages have a very convenient and efficient built-in data structure for storing elements indexed by a string[3], such as the frequency associated with an utterance. However, the actual gain depends on the redundancy of the corpus at utterances level, and even in an acquisition corpus, many utterances occur only once. The time needed to compute the frequency of an n-gram is reduced accordingly, and due to the average efficiency of hash coding, the time involved by the storage of an utterance is approximately as low as in the naive case above.

It is possible to store not only the frequency of utterances, but also that of their subparts. In this approach, storing an n-gram and retrieving its frequency need comparable time resources, expected to be low if hashing is performed. Of course, from the point of view of memory load, this is much more expensive than the two previous implementations discussed. However, we can take advantage of the fact that in an utterance of length l, every n-gram w with $1 \leq n < l$ is the prefix and/or suffix of at least an $n+1$-gram w'. Thus, it is much more compact to store them in a directed tree, the root of which is the empty string, and where each node corresponds to a phoneme in a given context[4], and each child of a node to a possible successor of that phoneme in its context. The frequency of an n-gram can be stored in a special child of the node representing the terminal phoneme of the n-gram.

This implementation (tree storage) will be used in the simulations described below. It is not claimed to be more psychologically plausible than another, but we believe the size in nodes of the trees built for a given corpus provides an intuitive and accurate way of comparing the memory requirements of the algorithms we discuss. From the point of view of time complexity, however, the tree structure is less optimal than a flat hash table since the time needed for the storage or retrieval of an n-gram grows linearly with n.

[3]This type of storage is known as *hash coding*.

[4]defined by the whole sequence of its parent nodes

4 Empirical evaluation

4.1 Experimental setup

Both algorithms described above were implemented implemented in Perl[5] and evaluated using a phonemically transcribed and child-oriented French corpus (Kilani-Schoch corpus[6]). We have extracted from the original corpus all the utterances of Sophie's parents (mainly her mother) between ages 1;6.14 and 2;6.25 (year;month.day). These were transcribed phonemically in a semi-automatic fashion, using the BRULEX database (Content et al., 1990) and making the result closer to oral French with a few hand-crafted rules. Eventually the first 10'000 utterances were used for simulations. This corresponds to 37'663 words (992 types) and 103'325 phonemes (39 types).

In general, we will compare the results observed for the successor count used on its own ("SC alone", on the figures) with those obtained when the utterance-boundary typicality is used for preprocessing[7]. The latter were recorded for $1 \leq r \leq 5$, where r is the order for the computation of typicalities. The threshold value for typicality was set to 1 (see section 2.3). The results of the algorithms for word segmentation were evaluated by comparing their output to the segmentation given in the original transcription using precision and recall for word and boundary detection (computed over the whole corpus). The memory load is measured by the number of nodes in the trees built by each algorithm, and the processing time is the number of seconds needed to process the whole corpus.

4.2 Segmentation performance

When used in isolation, our implementation of the successor count has a segmentation precision as high as 82.5%, with a recall of 50.5%; the word precision and recall are 57% and 40.8%

[5]Perl was chosen here because of the ease it provides when it comes to textual statistics; however, execution is notoriously slower than with C or C++, and this should be kept in mind when interpreting the large differences in processing time reported in section 4.4.

[6]Sophie, a French speaking Swiss child, was recorded at home by her mother every ten days in situations of play (Kilani-Schoch and Dressler, 2001). The transcription and coding were done according to CHILDES conventions (MacWhinney, 2000).

[7]Results of the utterance-boundary approach alone are given in (Xanthos, 2004)

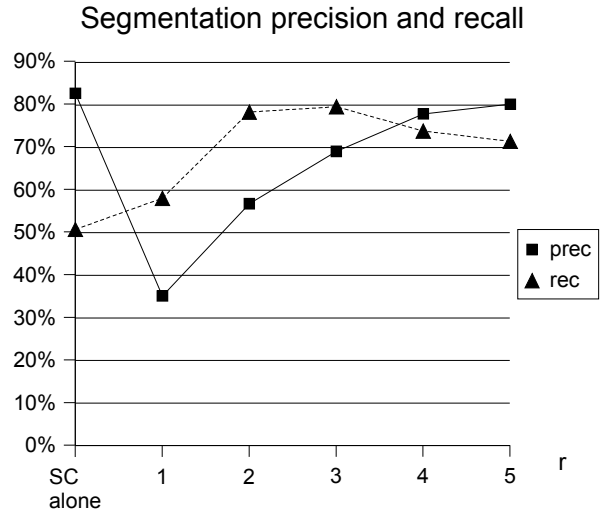


Figure 1: Segmentation precision and recall obtained with the successor count alone and with utterance-boundary preprocessing on n-grams.

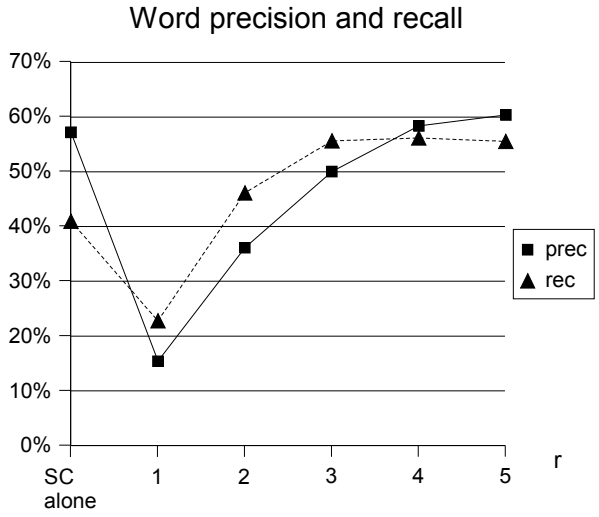


Figure 2: Word precision and recall obtained with the successor count alone and with utterance-boundary preprocessing on n-grams.

respectively. For comparison, the highest segmentation precision obtained with utterance-boundary typicality alone is 80.8% (for $r = 5$), but the corresponding recall does not exceed 37.6%, and the highest word precision is 44.4% ($r = 4$) with a word recall of 31.4%. As expected, the successor count performs much better than the utterance boundary typicality in isolation.

Using utterance-boundary typicality as a preprocessing step has a remarkable impact on the performance of the resulting algorithm. Figure 1 shows the segmentation performance obtained for boundary detection with the successor count alone or in combination with preprocessing (for $1 \leq r \leq 5$). The segmentation precision is always lower with preprocessing, but the difference dwindles as r grows: for $r = 5$, it reaches 79.9%, so only 2.1% are lost. On the contrary, the segmentation recall is always higher with preprocessing. It reaches a peak of 79.3% for $r = 3$, and stays as high as 71.2% for $r = 5$, meaning a 20.7% difference with the successor count alone.

Concerning the detection of whole words, (figure 2), the word precision is strictly increasing with r and ranges between 15.2% and 60.2%, the latter being a 3.2% increase with regard to the successor count alone. The word recall is lower when preprocessing is performed with $n = 1$ (-18.2%), but higher in all other cases, with a peak of 56% for $n = 4$ (+15.2%).

Overall, we can say the segmentation performance exhibited by our hybrid algorithm confirms our expectations regarding the complementarity of the two strategies examined: their combination is clearly superior to each of them taken independently. There may be a slight loss in precision, but it is massively counterbalanced by the gain in recall.

4.3 Memory load

The second hypothesis we made was that the preprocessing step would reduce the memory load of the successor count algorithm. In our implementation, the space used by each algorithm can be measured by the number of nodes of the trees storing the distributions. Five distinct trees are involved: three for the utterance-boundary approach (one for the distribution of n-grams in general and two for their distributions in utterance-initial and -final position), and two for the predictability approach (one for successors and one for predecessors). The memory load of each algorithm is obtained by summation of these values.

As can be seen on figure 3, the size of the trees built by the successor count is drastically reduced by preprocessing. Successor count alone uses as many as 99'405 nodes; after preprocessing, the figures range between 7'965 for $n = 1$ and 38'786 for $n = 5$ (SC, on the figure)[8]. However, the additional space used by the n-grams

[8]These values are highly *negatively* correlated with the number of boundaries–true or false–inserted by preprocessing ($r = -0.96$).

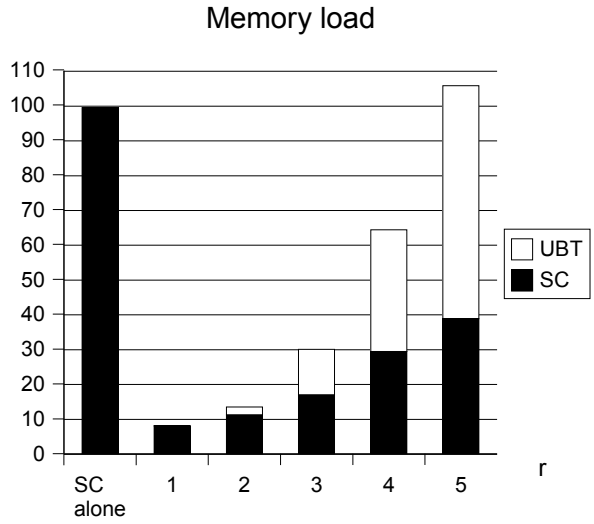


Figure 3: Memory load (in thousands of nodes) measured with the successor count alone and with utterance-boundary preprocessing on n-grams (see text).

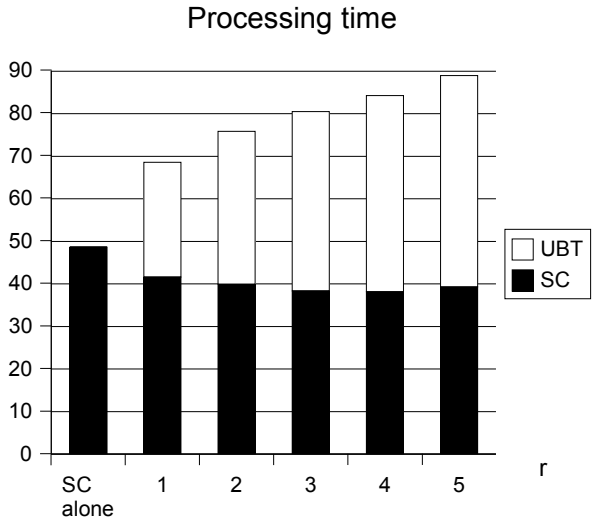


Figure 4: Processing time (in seconds) measured with the successor count alone and with utterance-boundary preprocessing on n-grams.

distributions needed to compute the utterance-boundary typicality (UBT) grows quickly with n, and the total number of nodes even exceeds that of the successor count alone when $n = 5$. Still, for lower values of n, preprocessing leads to a substantial reduction in total memory load.

4.4 Processing time

It seems unlikely that the combination of the two algorithms does not exhibit any drawback. We have said in section 3 that storing distributions in a tree was not optimal from the point of view of time complexity, so we did not have high expectations on this topic. Nevertheless, we recorded the time used by the algorithms for the sake of completeness. CPU time[9] was measured in seconds, using built-in functions of Perl, and the durations we report were averaged over 10 runs of the simulation[10].

What can be seen on figure 4 is that although the time used by the successor count computation is slightly reduced by preprocessing, this does not compensate for the additional time required by the preprocessing itself. On average, the total time is multiplied by 1.6 when preprocessing is performed. Again, this is really a consequence of the chosen implementation, as this factor could be reduced to 1.15 by storing distributions in flat hash tables rather than tree structures.

[9] on a pentium III 700MHz

[10] This does not give a very accurate evaluation of processing time, and we plan to express it in terms of number of computational steps.

5 Conclusions and discussion

In this paper, we have investigated two approaches to speech segmentation based on different heuristics: the utterance-boundary strategy, and the predictability strategy. On the basis of former empirical results as well as theoretical considerations regarding their performance and complexity, we have suggested that the utterance-boundary approach could be used as a preprocessing step in order to lighten the task of the predictability approach without damaging the segmentation.

This intuition was translated into an explicit model, then implemented and evaluated for a task of word segmentation on a child-oriented phonetically transcribed french corpus. Our results show that:

- the combined algorithm outperforms its component parts considered separately;
- the total memory load of the combined algorithm can be substantially reduced by the preprocessing step;
- however, the processing time of the combined algorithm is generally longer and possibly much longer depending on the implementation.

These findings are in line with recent research advocating the integration of various strategies for speech segmentation. In his work on

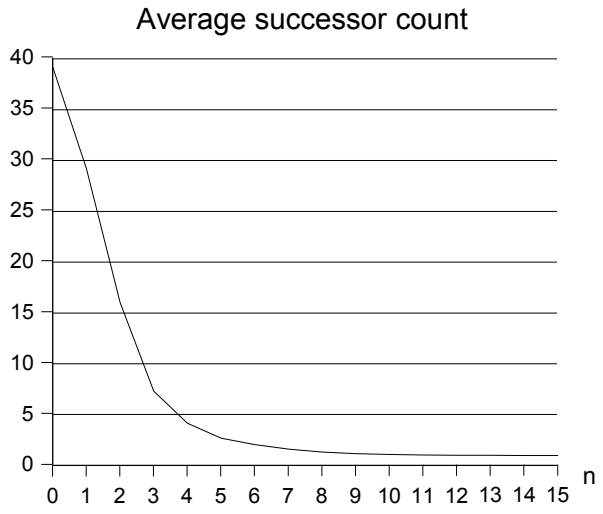


Figure 5: Average successor count for n-grams (based on the corpus described in section 4.1).

computational morphology, Goldsmith (2001) uses Harris' successor count as a means to reduce the search space of a more powerful algorithm based on minimum description length (Marcken, 1996). We go one step further and show that an utterance-boundary heuristic can be used in order to reduce the complexity of the successor count algorithm[11].

Besides complexity issues, there is a problem of *data sparseness* with the successor count, as it decreases very quickly while the size n of the context grows. In the case of our quite redundant child-oriented corpus, the (weighted) average of the successor count[12] for a random n-gram $\sum_{w \in S^n} f(w)\, succ(w)$ gets lower than 1 for $n \geq 9$ (see figure 5). This means that in most utterances, no more boundary can be inserted after the first 9 phonemes (respectively before the last 9 phonemes) unless we get close enough to the other extremity of the utterance for the predecessor (respectively successor) count to operate. As regards the utterance-boundary typicality, on the other hand, the position in the utterance makes no difference. As a consequence, many examples can be found in our corpus, where the middle part of a long utterance would be undersegmented by the successor count alone, whereas preprocessing provides it with more tractable chunks. This is illustrated by the following segmentations of the utterance [il ɛm pɑ le karɔt papa] (*Daddy doesn't like carrots*), where vertical bars denote boundaries predicted by the utterance-boundary typicality (for $r = 5$), and dashes represent boundaries inferred by the successor count:

SC	[il-ɛmpɑlekarɔt-papa]
UBT ($r = 5$)	[ilɛmpɑ|lekarɔt|papa]
UBT + SC	[il-ɛm-pɑ|le-karɔt|papa]

This suggests that the utterance-boundary strategy could be more than an additional device that safely predicts some boundaries that the successor count alone might have found or not: it could actually have a *functional* relationship with it. If the predictability strategy has some relevance for speech segmentation in early infancy (Saffran et al., 1996), then it may be necessary to counterbalance the data sparseness; this is what these authors implicitely do by using *first-order* transition probabilities, and it would be easy to define an n-th order successor count in the same way. Yet another possibility would be to "reset" the successor count after each boundary inserted. Further research should bring computational and psychological evidence for or against such ways to address representativity issues.

We conclude this paper by raising an issue that was already discussed by Gammon (1969), and might well be tackled with our methodology. It seems that various segmentation strategies correlate more or less with different segmentation levels. We wonder if these different kinds of sensitivity could be used to make inferences about the hierarchical structure of utterances.

6 Acknowledgements

The author is grateful to Marianne Kilani-Schoch and the mother of Sophie for providing the acquisition corpus (see p.4), as well as to François Bavaud, Marianne Kilani-Schoch and two anonymous reviewers for useful comments on earlier versions of this paper.

References

R.N. Aslin, J.Z. Woodward, N.P. Lamendola, and T.G. Bever. 1996. Models of word segmentation in fluent maternal speech to infants. In J.L Morgan and Demuth K., editors, *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Language Acquisition*, pages 117–134. Lawrence Erlbaum Associates, Mahwah (NJ).

E. Batchelder. 2002. Bootstrapping the lexicon: A computational model of infant speech segmentation. *Cognition*, 83:167–206.

[11] at least as regards memory load, which could more restrictive in a developmental perspective

[12] The predecessor count behaves much the same.

F. Bavaud and A. Xanthos. 2002. Thermodynamique et statistique textuelle: concepts et illustrations. In *Actes des 6è Journées Internationales d'Analyse Statistique des Données Textuelles (JADT 2002)*, pages 101–111.

M.R. Brent and T.A. Cartwright. 1996. Distributional regularity and phonotactics are useful for segmentation. *Cognition*, 61:93–125.

M.H. Christiansen, J. Allen, and M. Seidenberg. 1998. Learning to segment speech using multiple cues. *Language and Cognitive Processes*, 13:221–268.

A. Content, P. Mousty, and M. Radeau. 1990. Brulex: Une base de données lexicales informatisée pour le français écrit et parlé. *L'Année Psychologique*, 90:551–566.

E. Gammon. 1969. Quantitative approximations to the word. In *Papers presented to the International Conference on Computational Linguistics COLING-69*.

J. Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27 (2):153–198.

Z.S. Harris. 1955. From phoneme to morpheme. *Language*, 31:190–222.

J.L. Hutchens and M.D. Adler. 1998. Finding structure via compression. In *Proceedings of the International Conference on Computational Natural Language Learning*, pages 79–82.

M. Kilani-Schoch and W.U. Dressler. 2001. Filler + infinitive and pre- and protomorphology demarcation in a french acquisition corpus. *Journal of Psycholinguistic Research*, 30 (6):653–685.

B. MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk. Third Edition.* Lawrence Erlbaum Associates, Mahwah (NJ).

C.G. de Marcken. 1996. *Unsupervised Language Acquisition*. Phd dissertation, Massachusetts Institute of Technology.

J.R. Saffran, E.L. Newport, and R.N. Aslin. 1996. Word segmentation: The role of distributional cues. *Journal of Memory and Language*, 35:606–621.

A. Xanthos. 2003. Du k-gramme au mot: variation sur un thème distributionnaliste. *Bulletin de linguistique et des sciences du langage (BIL)*, 21.

A. Xanthos. 2004. An incremental implementation of the utterance-boundary approach to speech segmentation. To appear in the Proceedings of Computational Linguistics in the Netherlands 2003 (CLIN 2003).

Author Index