PsychocompLA-2007

Psychocomputational Models of Human Language Acquisition

Abstracts from the 3rd Meeting of the Workshop

August 1, 2007
Nashville, Tennessee

In conjunction with the
29th Annual Meeting of the Cognitive Science Society

www.colag.cs.hunter.cuny.edu/psychocomp

Preface

Welcome to PsychoCompLA-2007 held as part of the 29th meeting of the Cognitive Science Society in Nashville Tennessee. This is the third meeting of the Psychocomputational Models of Human Language Acquisition workshop following PsychoCompLA-2004, held in Geneva, Switzerland as part of the 20th International Conference on Computational Linguistics (COLING 2004) and PsychoCompLA-2005 as part of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-2005) held in Ann Arbor, Michigan where the workshop shared a joint session with the Ninth Conference on Computational Natural Language Learning (CoNLL-2005).

Psychocomputational models of language acquisition are of particular interest in light of recent results in developmental psychology that suggest that very young infants are adept at detecting statistical patterns in an audible input stream. Though, how children might plausibly apply statistical 'machinery' to the task of grammar acquisition, with or without an innate language component, remains an open and important question. One effective line of investigation is to computationally model the acquisition process and determine interrelationships between a model and linguistic or psycholinguistic theory, and/or correlations between a model's performance and data from linguistic environments that children are exposed to.

It is our belief that this approach will not only inform developmental and theoretical linguistic research, but will also prove invaluable to research focused on cutting-edge computer-human language technologies that may soon fall victim to a *psychocomputational bottleneck* – in which machine learning techniques that are applied without consideration of how humans learn and process language see decreasing marginal success.

We would like to thank the invited speakers, both for agreeing to present and for reviewing the submitted abstracts, as well as Mike Schoelles, Kevin Gluck, Jan Hathaway, and Terry Love for their help and patience during pre-meeting organization and finally Hunter College, CUNY and the Cognitive Science Society for funding.

William Gregory Sakas
David Guy Brizan

New York, July 2007

Invited Presentations:

Shimon Edelman
Alex Clark
Robert C. Berwick, Michael Coen, Sandiway Fong, and Partha Niyogi
Robert Frank
Amy Prefors, Terry Regier & Josh Tenenbaum
Charles D. Yang
Elissa L. Newport

Organization:

The workshop was organized by CUNY CoLAG, City University of New York Computational Language Acquisition Group www.colag.cs.hunter.cuny.edu in conjunction with the 29th Annual Meeting of the Cognitive Science Society.

Workshop and Program Chairs:

William Gregory Sakas
Department of Computer Science, Hunter College
Ph.D. Programs in Linguistics and Computer Science, The Graduate Center
City University of New York
sakas@hunter.cuny.edu

David Guy Brizan
Ph.D. Program in Computer Science, The Graduate Center
City University of New York
dbrizan@gc.cuny.edu

Program Committee:

Submitted abstracts were reviewed by the invited speakers.

Sponsors:

The Cognitive Science Society (as part of its annual meeting) and Hunter College, City University of New York..

Table of Contents and Presentation Schedule[1]

[1] The page-ordering of the abstracts reflects nothing more than a best attempt at adjudicating a variety of submitted formats in as few pages as possible [WGS, DGB].

Statistical Language Learning: Computational and Maturational Constraints

Elissa L. Newport
George Eastman Professor of Brain & Cognitive Sciences
University of Rochester

In recent years a wide variety of studies have shown that infants, young children, and adults can successfully utilize the statistics of distributional linguistic information to find candidate words in a speech stream, form and alter phonetic categories, discover grammatical categories, and acquire simple syntactic structure in miniature languages in the laboratory. A major question we face, then, is how to think about the broader picture of statistical learning: How many kinds of statistical computations can learners perform? How are these computations organized, and how are they constrained? Must such mechanisms be combined with qualitatively different, more traditional mechanisms that form symbolic rules or set linguistic parameters?

I will address these questions by presenting findings from recent studies of statistical learning of syntax, examining the effects of complex multiple cues and also comparing child and adult learners given inconsistent input and (sometimes) forming rule-like generalizations. The results of these studies suggest the outlines of several distinct but suitably sophisticated candidate statistical learning mechanisms and raise questions for future research regarding how to develop a theory of statistical language learning.

Lexical Learning and Change

Charles D. Yang
University of Pennsylvania

Language learning is a remarkably robust process. The child is exceptionally good at recognizing systematic regularities even when faced with lexically and contextually restricted exceptions. In this talk, we present a model that recognizes productive processes and exceptions as such; accordingly, the learner can proceed to internalize these as different kinds of linguistic knowledge. Along the way we draw connections with recent work in artificial language experiments that explores how the learner derives linguistic generalization. Finally, the learning model is extrapolated into a model of language change, which may shed light on the interpretation of typological generalizations.

Progress in Unsupervised Language Acquisition

Shimon Edelman
Cornell University

To become full-fledged members of the linguistic community in which they are situated, language learners must be able (1) to capture the structural regularities in the stream of utterances to which they are exposed in the course of their interactions with other speakers, and (2) to put these regularities to generative use. In the past several years, statistical algorithms that recursively distill productive construction-like regularities from raw corpus data became available (Adriaans and van Zaanen, 2004; Solan, Horn, Ruppin, and Edelman, 2005; Sandbank, Edelman, and Ruppin, 2007). In particular, the ADIOS algorithm (Solan et al., 2005) achieved unprecedented recall, precision, and perplexity performance on a standard benchmark --- the Air Traffic Information System (ATIS) corpus. More importantly, it exhibited 50% recall and 63% precision on withheld test data in an experiment involving a large subset of the transcribed child-directed speech from English CHILDES corpora (Brodsky, Waterfall, and Edelman, 2007). Another algorithm, ConText (Sandbank et al., 2007), was recently shown to outperform ADIOS on ATIS.

Although these algorithms deal relatively well with two kinds of structural regularities, namely collocations and complementary distribution classes, they are not as effective in learning potentially long-range dependencies such as those that arise in gender and number agreement. They also do not work well with corpora that contain a large proportion of complex (multiple-clause) sentences. These limitations can be traced to certain computational characteristics common to these algorithms, such as insensitivity to the head structure of phrases and the lack of distinction between content and function words. Addressing these issues could bring unsupervised language acquisition close to the limit of what can be learned from corpus data alone. Beyond that, further progress would likely require embodied, socially aware learning, of the kind that human babies excel in.

References

Adriaans, P. and M. van Zaanen, Computational Grammar Induction for Linguists, Grammars 7:57-68 (2004).

Brodsky, P., H. Waterfall, and S. Edelman, Characterizing Motherese: On the Computational Structure of Child-Directed Language, Proc. Cognitive Science Society Conference (2007, in press).

Sandbank, B., S. Edelman, and E. Ruppin, From ConText to grammar: a step towards practical probabilistic context free grammar inference, Proc. Israeli Conference on Computational Linguistics (2007, in press).

Solan, Z., D. Horn, D., E. Ruppin, and S. Edelman, Unsupervised learning of natural languages. Proceedings of the National Academy of Science, 102:11629-11634 (2005).

Learnable Representations of Languages: Something Old and Something New

Alex Clark,
Department of Computer Science,
Royal Holloway, University of London

Introduction

Context free grammars are appealing as a linguistic representation because of their efficient, well understood parsing algorithms, the convergence with the class of push down automata and their natural description as trees. However there are no general, efficient algorithms for learning them from positive data, and given our current understanding of grammatical inference we will never be able to learn more than restricted subclasses. If we take the problem of language acquisition seriously, then this would appear to be a fatal flaw.

In this paper, I will describe some research into other representations of languages, that have both polynomial recognition algorithms, and can be learned from positive data alone using polynomial amounts of data and computation. The first, using ideas from the 1950s, is based on a formalisation of Zellig Harris's idea of substitutability and defines a learnable subclass of context free grammars, that coincides with the ideas of reduction system. The second approach, using more modern ideas from learning theory, is based on identifying hyperplanes in a Hilbert space defined by a string kernel. In this system we can represent a wide variety of languages including some mildly context sensitive languages that model phenomena that occur in natural language. In both cases, we are able to prove efficient learnability of the classes of languages concerned, and demonstrate in practice on small data sets that the algorithms work correctly.

Substitutability

Zellig Harris method of distributional learning is often appealed to in current learning algorithms, but it is normally used merely as a heuristic justification rather than as an algorithm in itself. Harris equivocates between two formal definitions: “Here as throughout these procedures X and Y are substitutable if for every utterance which includes X we can find (or gain native acceptance for) an utterance which is identical except for having Y in the place of X” Harris (1947). This definition can be interpreted in two ways: the first is identical to the syntactic congruence, a standard language theoretic definition: two strings X,Y are congruent with respect to a language L iff it is the case that for every pair of strings l,r, the string lXr is in L iff the string lYr is in L. The second interpretation is that we have *some* pair of strings l,r such that both lXr, and lYr are in the language (or some sample of the language); clearly a much weaker requirement. The substitutable languages are the subclass of context free languages where this latter weaker condition does in fact imply the former, stronger criterion. Clark and Eyraud (2005,2006) established that this

class of languages can be learned using a very simple algorithm, where the non-terminals of the target grammar correspond to the equivalence classes under this congruence. Though the class of languages the algorithm can learn perfectly is limited, the algorithm is still sufficiently powerful to learn the rule of auxiliary fronting in polar interrogatives from a small data set. Crucially, the prior assumptions required to get this result appear to be domain neutral.

Planar languages

A completely separate approach is to look to the theory of machine learning for representations of languages that are learnable. By mapping strings into points in a high dimensional feature space, we can consider languages being defined by regions of that space: a string is in the language if its image lies in some region of feature space. If that region is a plane, then the languages will be easy to learn, using standard techniques. The kernel trick allows us to use very large or infinite dimensional feature spaces, and using domain neutral string kernels, it transpires that we can represent and learn efficiently some context sensitive languages, including those classic examples, such as Swiss German cross serial dependencies, that established that natural languages are not weakly context free.

Conclusion

Finally, we will discuss how it is possible to integrate these two approaches, using representations that are sensitive to more complex properties of words, making further progress towards the long term research goal of identifying a large class of languages that contains the natural languages, and is efficiently learnable from positive raw data, using domain neutral algorithms.

Acknowledgements.

This paper describes joint work with Remi Eyraud, Chris Watkins and Christophe Costa Florencio. Some of this was supported by a grant from the EU funded Pascal network of excellence.

References

Clark, A and Eyraud, R (2005) "Polynomial identification in the limit of substitutable context free grammars" Proceedings of the conference on Algorithmic Learning Theory, Tokyo.

Clark, A and Eyraud, R (2006) " Learning Auxiliary Fronting with Grammatical Inference, Proceedings of the 8th Conference on Computational Natural Language Learning (CoNLL-X). New York, pp 125-32

Clark, A, Costa Florencio, C and Watkins, C (2006) "Languages as Hyperplanes" Proceeding of European Conference on Machine Learning, ECML 2006.

Harris, Z (1951) "Methods in Structural Linguistics", University of Chicago Press.

Indirect Evidence and the Poverty of the Stimulus

Amy Perfors, Terry Regier and Josh Tenenbaum (MIT)
MIT, University of Chicago and MIT

The Poverty of the Stimulus (PoS) argument holds that children do not receive enough evidence to infer the existence of core aspects of language. We argue that Bayesian methods of grammar induction and model selection can be useful in evaluating PoS arguments, and that these methods suggest a new approach to classic questions of innateness. Because they incorporate sophisticated statistical inference mechanisms that can operate over structured representations of knowledge (such as generative grammars), they allow us to rigorously explore a relatively uncharted region of the theoretical landscape: the possibility that genuinely structured knowledge is genuinely learned. We apply this approach to a specific version of the PoS argument, and show that a rational learner faced with typical child-directed input and without initial language specific biases could learn that linguistic rules depend on hierarchical phrase structure. This enables a learner to master aspects of syntax, such as the auxiliary fronting rule in interrogative formation, even without having heard the sort of data traditionally assumed to be necessary for learning. Our results suggest that it does not make sense to ask whether a specific generalization is based on innate knowledge when that generalization is part of a much larger system of knowledge (such as the grammar of a natural language) that is acquired as a whole. Abstract organizational principles can be induced based on indirect evidence from one part of a system and effectively transferred to constrain learning of other parts of a system.

The Great (Penn TreeBank) Robbery: When Statistics is not Enough

Robert C. Berwick, Michael Coen, Sandiway Fong, and Partha Niyogi
MIT, University of Wisconsin, University of Arizona, University of Chicago

Over the past 15 years, there has been increasing use of linguistically-annotated sentences collections such as the LDC Penn Tree Bank (PTB) for constructing statistically-based parsers. While these parsers have usually been built for engineering purposes, it has sometimes been suggested, either implicitly or explicitly, that such approaches will either solve or point the way to solutions for the problem of human language acquisition, particularly in the area of broad syntactic coverage. In this paper we examine this possibility critically, assessing how well such methods can actually approach human/child competence. We find that all approaches actually do quite poorly in this respect.

Our basic findings include these:

- Testing on standard linguistic *ungrammatical* sentences, reveals that such systems generalize extremely poorly, making errors that are never attested by children. For example, on the 314 test sentences drawn from Lasnik and Uriagereka (1988) as used by Fong (1990), many grammatical examples cannot be successfully parsed and many ungrammatical examples are readily parsed. Even a simple example such as "John continues stocks" is taken to be 'grammatical' with a probability score close to that of "John sold stocks." Perhaps more well-known is that empty categories are not really treated properly, thus often leaving the assignment of thematic roles incorrect. Thus there is a strong sense in which these trained systems have attained a very impoverished 'knowledge of

language,' indeed do not even begin to approach the basic competence that children attain by ages 3--5.

- While the addition of a statistical inference component has often been claimed as offering the promise of robustness in the face of noise and the idiosyncrasies of exceptions in the learning input, in fact the resulting parsers seem to be just as "fragile" as hand-built systems. For example, a single altered training example sentence out of the 39,832 in the training set can greatly alter basic system parsing performance, forcing PP attachment to be 'high' rather than 'low.' This kind of behavior, extremely sensitive to the specific distributional details of the input data, does not seem compatible with the robust character of child language acquisition.

- The claim of broad coverage and robust generalization, as opposed to the (supposedly) more narrow linguistically-based treatments also seems suspect. Many common systematic regularities that have syntactic reflexes, such as the verb alternation classes extensively discussed by Levin (1993) are not and probably cannot be 'discovered' by the current methodology. To take but one example from Levin out of many, "bounce" is a member of the causative/inchoative class. According to Levin, this class admits "Jane bounced the ball" but bars "Jane bounced at the ball." However, current systems assign the second, unacceptable sentence a *higher* probability score than the first. The same is true for middle constructions. Since these examples cause no seeming difficulty for human speakers, again there seems to be a large gap between the generalizations the systems can make and what people do.

- More generally and importantly, the reason for such failings as the one above seems to be that, contrary to the terminology often highlighted in conjunction with such systems, they are not, in fact "exicalized." Here we take the term "lexicalized" to mean that the information in the lexicon, viz., subcategorization, selectional information, and the like, is taken as primary, with phrase structure rules consequently being informationally redundant, as suggested since the late 1960s and as adopted by many current linguistic approaches. When examined closely, the statistical systems in fact cannot weight properly the particular, detailed lexical properties of 'bounce' vs. other verbs. Rather, they focus on the presence or absence a particular syntactic configuration, e.g., V-PPs, to the exclusion of other information. Given this fundamental limitation they cannot discriminate among lexically-grounded alternation classes.

- As has been pointed out by others, the PTB is both too large and too small for adequate generalization to mirror that of children or adults. On the one hand, the PTB has enormous redundancy, since it is based on 49,208 sentences from a syntactically repetitive source: Wall Street Journal articles. The resulting sparseness means that the that the training corpus does not even begin to cover the syntactic richness of English. Therefore, smoothing becomes critical and determines much of the resulting systems' behavior; some of the mismatches between system and human judgments are not artifacts of data, but artifacts of smoothing the data. Since the 'covering density' of examples is not adequate to span the space of English possibilities, constructions that are readily parsed by people without prior exposure are completely misparsed by such systems. The result again is that generalization is poor, not robust. One classic example are parasitic gap constructions, though there are many others. For example, in nearly-minimal pairs such as "Everyone that John knows Mary likes/Everyone that John knows likes Mary," such systems incorrectly parse the first example with "Mary" as the object of "likes" (rather than the subject of the main clause). On the other hand, the large size of the PTB, due to its repetitive syntactic constructions, means that in many cases patterns like V-NP-PP often dominate lexical details, as noted above.

- Such statistical systems easily learn *unnatural* languages just as readily and with just as much accuracy as natural languages. They easily learn languages with syntactically bizarre forms, and withjust as high precision and recall. For example, no known human language alternates its verbs as head-first in some clauses, and head-final in others, randomized in a 50-50 fashion. Yet the statistical systems trained on such non-natural languages learn and test just as well on this unnatural input. Similarly, interchanging arguments and adjuncts so that adjuncts are next to verbs, and arguments distant, contrary to natural language syntax, is just as "natural' for these systems – just as easily

learned and as accurately parsed. Combining this unnatural argument-adjunct reversal with randomized head-first/head-final patterns is also just as easily learned and parsed. Indeed, we have yet to find a pattern, however unnatural, that these systems *cannot* readily learn and parse. While for some this might be taken as a sign of engineering 'flexibility,' this behavior seems at odds with the restrictions found in human acquisition, which does not display such lability (see N. Smith 1988 and A. Moro 2007 among others for human experiments and fMRI studies confirming these restrictions). The bottom line is that the statistically-based inference systems appear to be driven by the vagaries of the distributional pattern of external data to an degree not attested by human language acquisition, similar to the findings of Yang (2002).

- The probability scores these systems assign to sentences after parsing often meant to reflect 'likelihood' as something akin to 'grammaticality' but often these values do not square with natural grammaticality judgments. Thus the simple equating of 'likelihood' with 'grammaticality,' often explicit or implicit in much of this work, does not in our view appear to hold. Some examples have already been given above, but there are a host of others. For example, the nearly minimal grammatical/ungrammatical pairs with radical case assignment violations, such as "I am proud of John/I am proud John" are assigned identical probability scores, even despite the general probabilistic penalty assigned to longer sentences (here the grammatical one) over shorter ones (here the ungrammatical one).

Semantic Heads for Grammar Induction

Andrew M. Olney
Institute for Intelligent Systems
University of Memphis

Abstract

Olney (2007) presents an unsupervised grammar induction model that uses semantic similarity to induce syntactic structure. A key element of this model is the operational definition of syntactic heads as being semantically substitutable for their phrases. This paper describes the history of this operational definition for heads and tests its validity with respect to four computational implementations. The paper concludes with implications for these results on the operational definition of heads proposed by theoretical linguists as well as the model presented by Olney (2007).

1 Introduction

Previous work on unsupervised grammar induction [1 , 3 , 11] has made use of the distributional hypothesis [9], which characterizes words by their contexts. Under the distributional hypothesis, a phrase that occurs in the same environments as a single word, i.e. has a similar distribution, is likely to be have the same syntactic function as the single word. Hierarchical descriptions of sentences can then be built by attaching such phrases to a higher level node (represented by the single substituting word) until only a single root node remains [1].

Olney (2007) recently proposed a semantically-oriented model based on the distributional hypothesis. This model is distinguished from previous models in that it does not make use of part of speech tags or nonterminal nodes, yet it still manages to significantly outperform a right branching baseline. A key element of this model is the operational definition of syntactic heads as being semantically substitutable for their dependents, as discussed in the theoretical linguistics literature [17 , 10 , 4]. This paper explores using semantic substitutability to determine syntactic heads and presents perhaps the first computational evaluation of this notion.

2 Semantic Heads

Heads are theoretically-motivated linguistic elements with a primary role in syntactic description. In X-bar theory, a phrase contains a single head which determines the syntactic type of the phrase [2]. In dependency

grammars, heads have a similar role, except that syntactic relations exists solely between words [16]. The importance of heads across these theories suggests that the proper identification of heads is of central importance to a language learner.

Despite the wide use of heads and head-like notions in syntactic theory, there has been much debate as to precisely what a head is [17 , 10 , 4]. However, some agreement exists on a loose semantic definition of head: X is a head of X+Y if X describes a kind of thing described by X+Y [17 , 10 , 4]. Beyond this initial agreement, differences emerge. Zwicky (1985) equates a kind of with semantic arguments. He argues that "green car" describes a kind of car, rather than a kind of green. This example is endocentric, since the head, "car" appears within the phrase "green car." Hudson (1987) makes the opposite claim, that a kind of refers to semantic functors, e.g. "on the desk" refers to a kind of location rather than a kind of desk. This example is exocentric, since the meaning "location" is not directly attributable to any particular word in "on the desk," but to the phrase as a whole. Croft (1996) unites these two perspectives by noting that semantic functors are heads for government relationships and semantic arguments are heads for modification relationships.

Given the theoretical importance a kind of has in identifying heads, it is worthwhile exploring machine learning methods that can acquire this kind of knowledge. In particular, Latent Semantic Analysis (LSA) is a vector space method capable of measuring the similarity between words and collections of words [5 , 6 , 12]. LSA has been shown to closely approximate vocabulary acquisition in children [12], grade essays as reliably as experts in English composition [7], and understand student contributions in tutorial dialogue [8 , 14]. These results are particularly impressive considering that LSA creates its knowledge representation without human intervention.

3 Methodology

We present four methods for identifying heads using LSA. The basic methodology is to create an LSA space and to compare the semantic similarity of a dependency pair's elements to the whole. For example, "green" and "car" would both be compared with "green car." Using the Penn Treebank [13], heads found using these four methods are compared with manually identified heads.

The four methods presented use this basic methodology along the dimensions +/- order and +/- endocentric. The ordered methods use unigrams and bigrams as basic elements, inherently preserving word order. The minus endocentric, or exocentric, methods do not compare a dependency element to the whole, but rather to the nearest unigram neighbor of the whole. For example "in bed" may have a nearest unigram neighbor, "sleepy," which is more similar to "bed" than to "in." Furthermore, the construction of the LSA spaces varied in terms of local and global context. Global context represents the traditional LSA calculation, in which $cell_{ij}$ denotes the number of times $term_i$ appeared in $document_j$. In local context, $cell_{ij}$ is the number of times $term_j$ occurred before the target $term_i$, and the value of $cell_{i(j+n)}$ is the number of times $term_j$ occurred after the target $term_i$, where n is the number of terms in the corpus. Both local and global spaces were constructed using both unigrams and bigrams as terms to preserve word order.

Table 1: Head Discrimination Results for WSJ10

Method	Local Context	Global Context
Ordered/Endocentric	Percentage Correct	Percentage Correct
-/-	42.3%	41.7%
-/+	42.3%	41.6%
+/-	56.8%	39.0%
+/+	57.3%	37.4%

4 Results & Discussion

Results in Table 1 show that only the ordered methods were significantly better than chance, and that unordered methods were significantly worse ($p = .05$). There was no significant difference between endocentric and exocentric methods ($p = .05$). These results suggest that LSA is capturing "a kind of "-like information on a more abstract level than endocentric and exocentric, which would make LSA similarity closer to the loose semantic definition of head described in the literature [17 , 10 , 4]. However, the low

overall discriminability of LSA, 57% in the best case, further suggests that semantic similarity is not the only factor in determining headhood. It appears likely that there is another element to determining headness that is missing from the discussion amongst theoretical linguists.

These results have similar significance to the model proposed by Olney (2007). It is somewhat surprising that this model can outperform a right branching baseline even though the method of determining headhood has a weak discriminability of 57%. It seems likely therefore that an improvement in the ability to determine heads will be a major source of improvement in this model.

References

[1] Eric Brill and Mitchell Marcus. Automatically acquiring phrase structure using distributional analysis. In Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, pages 155-160, Philadelphia, February 23-26 1992. Association for Computational Linguistics.

[2] Noam Chomsky. Remarks on nominalization. In R. Jacobs and P. S. Rosenbaum, editors, Readings in English Transformational Grammar. Ginn and Co., Waltham, Massachusetts, 1970.

[3] Alexander Clark. Unsupervised induction of stochastic context-free grammars using distributional clustering. In Walter Daelemans and R'emi Zajac, editors, Proceedings of the ACL 2001 Workshop on Computational Natural Language Learning, pages 105-112, Philadelphia, July 6-7 2001. Association for Computational Linguistics.

[4] William Croft. What's a head? In Laurie Zaring and Johan Rooryck, editors, Phrase structure and the lexicon, pages 35-75. Kluwer, Dordrecht, 1996.

[5] Scott C. Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard A. Harshman. Indexing by latent semantic analysis. Journal of the American Society of Information Science, 41(6):391-407, 1990.

[6] Susan Dumais. Improving the retrieval of information from external sources. Behavior Research Methods, Instruments and Computers, 23(2):229-236, 1991.

[7] P.W. Foltz, S. Gilliam, and S. Kendall. Supporting Content-Based Feedback in On-Line Writing Evaluation with LSA. Interactive Learning Environments, 8(2):111-127, 2000.

[8] A.C. Graesser, P. Wiemer-Hastings, K. Wiemer-Hastings, D. Harter, T.R.G. Tutoring Research Group, and N. Person. Using latent semantic analysis to evaluate the contributions of students in AutoTutor. Interactive Learning Environments, 8(2):129-147, 2000.

[9] Zellig Harris. Distributional structure. Word, 10:140-162, 1954.

[10] Richard A. Hudson. Zwicky on heads. Journal of Linguistics, 23:109-132, 1987.

[11] Dan Klein and Christopher D. Manning. A generative constituent-context model for improved grammar induction. In Proceedings of the 40th Annual Meeting of the ACL, pages 128-135, Philadelphia, July 7-12 2002. Association for Computational Linguistics.

[12] Thomas K. Landauer and Susan T. Dumais. A solution to plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. Psychological Review, 104:211-240, 1997.

[13] Mitchell P. Marcus, Mary A. Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics, 19(2):313-330, 1993.

[14] B. A. Olde, D. Franceschetti, A. Karnavat, and A. C. Graesser. The right stuff: Do you need to sanitize your corpus when using Latent Semantic Analysis? In Proceedings of the 24th Annual Meeting of the Cognitive Science Society, pages 708-713, Mahwah, NJ, 2002. Erlbaum.

[15] A. M. Olney. Latent semantic grammar induction: Context, projectivity, and prior distributions. In Proceedings of TextGraphs-2: Graph-Based Algorithms for Natural Language Processing, pages 45-52, Rochester, NY, April 2007. Association for Computational Linguistics.

[16] Lucien Tesniere. Elements de syntaxe structurale. Klincksieck, Paris, 1959.

[17] Arnold M. Zwicky. Heads. Journal of Linguistics, 21:1-29, 1985.

Modeling the Development of Bilingual and Second Language Reading

Nicole Sager[1,3], Seth Herd[2,3], Eliana Colunga[2,3]

[1]School of Education, [2]Department of Psychology, [3]Institute of Cognitive Science
University of Colorado Boulder

Computational models of bilingual language processing are in the early stages of investigating various issues such as degree of language independence, cross linguistic positive transfer and interference, contexts of acquisition, and modality (speaking, writing, comprehension) specificity (Thomas & van Heuven, 2005). Models of adult bilingual word recognition (Dijkstra & Van Heuven, 1998) and speech perception (Lewy & Grosjean, 1997) have been developed. In addition, Thomas, (1997) developed the BSN model which is a developmental distributed model of word recognition that includes mapping of orthography and semantics with a language context layer.

The model presented here simulates and compares the bilingual and monolingual reading development of native Spanish speakers learning English as a second language (L2). By varying the training contexts we explore differences in English reading outcomes. Contexts differed by: sequence of training, degree of English (L2) word learning training, and whether or not the model is trained in Spanish (L1) word reading before English word reading. Simulations 1 and 2 model simultaneous bilingualism. Simulations 3 and 4 model sequential bilingualism with complete L2 word learning training, and simulations 5 and 6 model sequential bilingualism with partial L2 word learning training.

The architecture consists of a phonology layer, an orthography layer and a semantics layers with hidden layers connecting all layers. The design assumes the two languages are stored in a single common representational resource and become distinguishable through language-specific information. Language separation is seen as emerging in the process of learning respective regularities in Spanish and English input. The input was created from Spanish and English phonemes accounting for phonotactic regularities and the degree of overlap between the two languages. Multisyllabic words were used which enabled the modeling of differences in word length and stress patterns of English and Spanish. Fifty items in each language were constructed using both vowel and stress centering.

Results for the simultaneous bilingual model suggest intramodality facilitation effects, as it did not take twice as long for the model to reach criteria in bilingual word reading as it did to reach criteria for English word reading only. Comparisons of simulations 3 and 4 to simulations 5 and 6 also showed positive transfer of Spanish reading to English reading even when levels of English word learning varied. Models receiving initial word reading training in Spanish reached English word reading criteria in half the number of training epochs regardless of whether English word learning was partial or reached criteria.

To our best knowledge, our model is the first developmental computational model of bilingual word reading that incorporates phonology, orthography and semantics without the

use of a language context layer. In addition, by first training the distributed network on word learning (phonology to semantics) and subsequently on word reading (orthography to phonology) we take the modeling of development one step further. The intent was to more realistically model the relationship between oral language skills and reading development. Consideration was also given to real-world situations in which English language learners find themselves in current educational contexts.

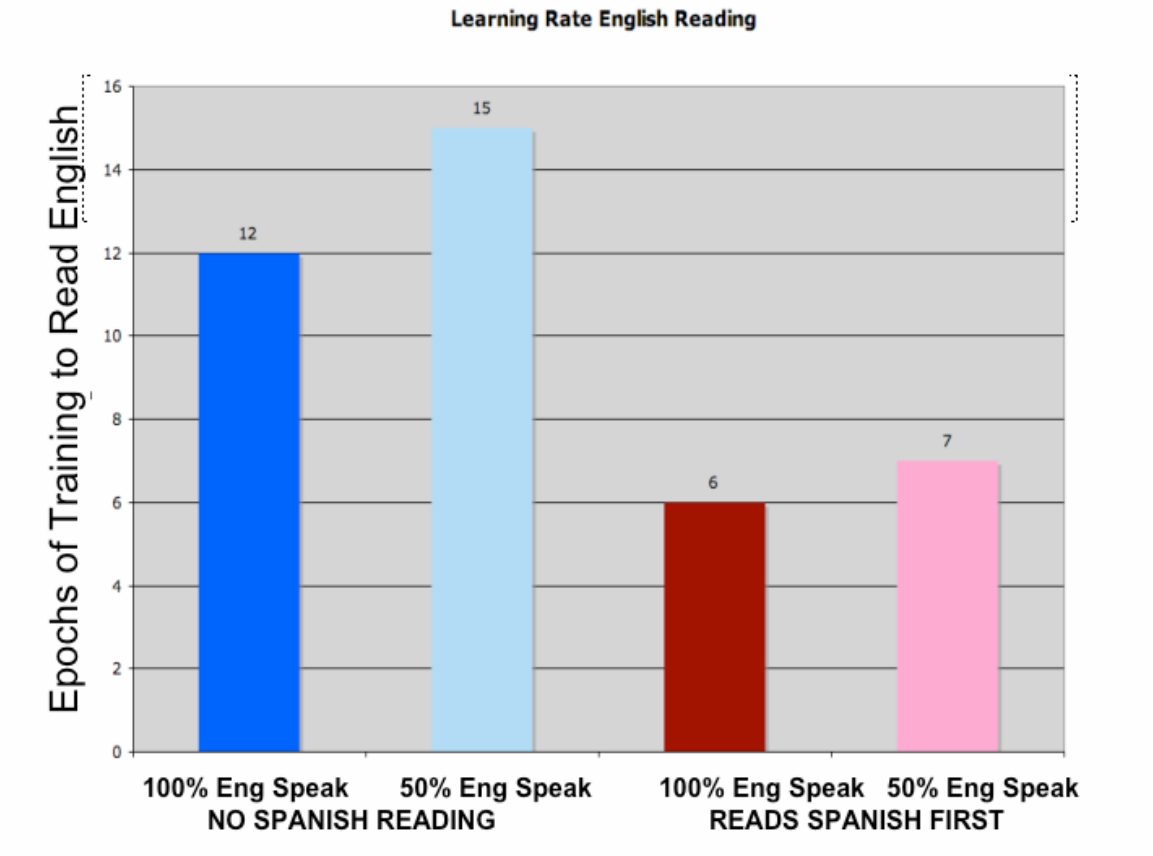

Figure 1: Comparison of Learning Rate

References

Dijkstra, A., & Van Heuven, W. J. B. (1998). The BIA model and bilingual word recognition. In J. Granger & A. M. Jacobs (Eds.), *Localist connectionist approaches to human cognition* (pp. 198-225). Mahwah, NJ: Erlbaum.

Léwy, N., & Grosjean, F. (1997). A computational model of bilingual lexical access. Manuscript in preparation, Neuchâtel University, Switzerland.

Thomas, M. S. C. (1997). Distributed representations and the bilingual lexicon: One store or two? In J. Bullinaria, D. Glasspool, & G. Houghton (Eds.) *Proceedings of the Fourth Annual Neural Computation and Psychology Workshop* (pp. 240-253). London: Springer.

Thomas, M. S. C., & van Heuven, W. J. B. (2005). Computational models of bilingual comprehension. In J. F. Kroll & A. M. B. de Groot (Eds.) *Handbook of bilingualism* (pp. 202-225). New York, NY: Oxford.

Algorithms for Learning the Raising/Control Distinction from Semantic Information

Misha Becker[1] and W. Garrett Mitchener[2]
[1]Linguistics Department, University of North Carolina Chapel Hill
[2]Department of Mathematics, College of Charleston

We developed an algorithm to model the learning of three verb classes: raising verbs (e.g., *seem*), control verbs (*try*) and ambiguous verbs that can be used as either (*begin*) (1a-c). These classes of verbs present an interesting learning problem because they are all used with *to*+infinitive complements, yet raising and control verbs have distinct syntactic and semantic properties. Any algorithm that attempts to classify an unknown verb by initially assuming the most restricted class (control) and passing to less restricted classes on the basis of positive evidence (such as use with an expletive subject) cannot distinguish ambiguous verbs from raising verbs. Thus, we developed an algorithm that uses semantic cues in the input.

Previous research (Becker, 2005) pointed to the usefulness of two cues found in sentences containing these verbs: animacy of the sentence subject, and eventivity of the predicate embedded under the main verb (2). Animate subjects are compatible with both raising and control main verbs (though they occur predominantly with control verbs) (2a), but inanimate subjects are compatible only with raising verbs (2b). Embedded stative predicates more commonly occur with raising main verbs (2c), while eventive predicates tend to occur with control main verbs (2d). However, to classify a verb, it is insufficient to only use the proportions at which a verb occurs in each of the four possible semantic frames (animate or inanimate subject plus eventive or stative predicate): Many raising verbs (e.g., *gonna*) occur with animate subjects so often that the crucially informative uses with inanimate subjects are relatively rare, particularly in child-directed speech.

We developed a learning algorithm that maintains numbers representing the verb's preference or aversion to each semantic frame. It automatically discards example sentences that merely reinforce its existing knowledge. This property, which was inspired by linear reward-penalty learning with batch (Yang, 2002), enables it to correctly classify raising verbs like *gonna* that occur frequently with animate subjects. After receiving all input sentences, the algorithm settles to one of twenty rest states. To test the algorithm, we counted the number of occurrences of a few common raising, control and ambiguous verbs with each of the four semantic frames in both the CHILDES database (child-directed speech; MacWhinney 2000) and an annotated version of the Switchboard corpus (adult-directed speech; Taylor et al., 2003, Bresnan et al., 2002). See Tables 1-2. We synthesized input sentences for each verb according to these proportions and fed them to the algorithm. The final states of many runs yield distinct patterns for the three verb classes. In addition, when learning a control verb, each run of the algorithm begins in a neutral state characteristic of raising verbs and drifts toward a state characteristic of control verbs. This effect is harmonious with child grammaticality judgments that vary with age (Becker, 2006): Younger children (age 3) are more likely to accept control verbs where only a raising verb is appropriate, and learn not to make this mistake over several years (by age 5).

(1) a. John seems to be clever.

b. John tried to win the race.
c. It began to rain. (raising)
c.' John began to write a novel. (control)

(2) a. Amy seems/tried to be a good waitress. (animate subject)
b. The truck seemed/*tried to roll down the hill. (inanimate subject)
c. Gordon seemed to be leaving/?leave. (raising verb with stative/?eventive pred)
d. Gordon tried to ?be leaving/leave. (control verb with ?stative/eventive pred)

Table 1. Numbers of Verb Classes with Animate/Inanimate Subject and Eventive/Stative Predicate, CHILDES

Verb subclass	Animate+Eventive	Animate+Stative	Inanimate+Eventive	Inanimate+Stative
Raising	1097	149	37	35
Control	604	110	2	0
Ambiguous	40	4	0	4

Table 2. Numbers of Verb Classes with Animate/Inanimate Subject and Eventive/Stative Predicate, Switchboard

Verb subclass	Animate+Eventive	Animate+Stative	Inanimate+Eventive	Inanimate+stative
Raising	273	241	40	127
Control	716	175	4	0
Ambiguous	673	173	18	32

References

Becker, M. (2005) Learning verbs without arguments: The problem of raising verbs, *Journal of Psycholinguistic Research*, 34, 165-191.

Becker, M. (2006) There Began to be a Learnability Puzzle. *Linguistic Inquiry*, 37, 441-456.

Bresnan, J., J. Carletta, R. Crouch, M. Nissim, M. Steedman, T. Wasow and A. Zaenen (2002) *Paraphrase analysis for improved generation, LINK project*, Stanford: HCRC Edinburgh-CSLI Stanford.

MacWhinney, B. (2000) *The Child Language Data Exchange System*, Mahwah: Lawrence Erlbaum Associates.

Taylor, A., M. Marcus and B. Santorini (2003) The PENN Treebank: An overview. In *Treebanks: The state of the art in syntactically annotated corpora*, ed. by A. Abeille, Dordrecht: Kluwer.

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

Selective Attention and Darwinised Data-Oriented Parsing

Dave Cochran

School of Computer Science, University of St. Andrews

Darwinised Data-Oriented Parsing (DDOP) is a new approach to unsupervised Data-Oriented Parsing (Bod 2006) which makes use of a hitherto unrecognised feature of Data-Oriented Parsing (DOP; Scha 1990, Bod 1992, 1998). Previous DOP algorithms have been static in character; they have been given their training data as a whole corpus which then remains static as the parser is run over it's training data. However, if it is run incrementally, feeding all new parses back into the training data for future reuse, it may be run as a Genetic Algorithm. DOP analyses novel strings by directly exploiting the statistical properties of a corpus of trees without producing any abstract representations of these regularities; it constructs novel parses by extracting any-depth tree-fragments from the training corpus, which are used to construct a Monte-Carlo sample of random derivations of parses; the output is the most frequent parse in the sample, taken as an approximation to the most probable parse. In an incremental DOP algorithm, subtrees are replicators; every time a novel input is parsed, the output parse will contain new copies of all the extracted subtrees used in the construction of it's derivations, which are then added to the training data for subsequent parses. Since more highly generalisable subtrees will be used and replicated more often, subtrees are subject to a selection pressure towards greater generalisability. DDOP exploits this by starting with an *empty* training corpus, and backing off to the use of randomly generated subtrees whenever a suitable subtree cannot be found in the training data.

However, true random subtree generation (i.e. sampling from the set of all subtrees of all possible trees over a given string) is computationally costly; the number of possible subtrees over a string of length n rises approximately exponentially with n. Furthermore, it is well documented that humans in fact find it difficult to produce true random behaviour, and attempts to do so will tend to be skewed by memory. Therefore, in this paper, we report on a number of tests with DDOP comparing training runs where the parser is set to simply ignore input strings over a certain length k until it has built up its training data, to runs where the parser if forced to parse all inputs of any length throughout the run. This is to test the hypothesis that DDOP training converges on highly generalisable tree-structures faster when it ignores longer stimuli in the early stages.

References

Bod, R. (1992). "A Computational Model of Language Performance; Data-Oriented Parsing". Proceedings COLING-92, Nantes, France

Bod, R. (1998), Beyond Grammar; An Experience-Based Theory of Language, Stanford, CA: Centre for the Study of Language and Information.

Bod, R. (2006a). "An All-Subtrees Approach to Unsupervised Parsing", *Proceedings ACL-COLING 2006*, Sydney.

Scha, R. (1990). "Taaltheorie en Taaltechnologie: Competence en Performance", in Q. de Kort and G. Leerdam (eds.), Computertoepassingen in de Neerlandistiek, Almere: Landelijke Vereniging van Neerlandici (LVVN-jaarboek).

Distributional Models of Syntactic Category Acquisition: A Comparative Analysis

Sharon Goldwater

Stanford University

Acquisition of syntactic categories is an important step in children's early language development. Recently, Mintz (2003) proposed that children may begin to acquire categories by attending to words that occur within *frequent frames* — pairs of words that occur frequently with a single intervening word. In this work, we analyze the kinds of words that are categorized by the frequent frames approach, and compare the results of this approach to two other models of syntactic category learning: the hierarchical clustering model of Redington et al. (1998) and the Bayesian model of Goldwater and Griffiths (2007). We provide each model with the same input corpus of child-directed speech and examine the resulting categorizations. Our analysis reveals several interesting facts. First, we show that all three models achieve similar high categorization accuracy when scored only on the words that occur within frequent frames. This suggests that under a variety of distributional categorization approaches, words that occur within frequent frames are indeed more easily categorized than other words. However, only a small percentage of words occur within frequent frames, and these words are overwhelmingly verbs and pronouns. In contrast, the clustering model and Bayesian model assign categories to nearly all word tokens, and are able to correctly cluster many nouns and other parts of speech as well as verbs and pronouns. The clustering model performs slightly better than the Bayesian model when evaluated on the full corpus, and we attribute this difference to the two models' treatment of *syntactic ambiguity*: the clustering model makes a simplifying assumption that every word type belongs to only one syntactic category, while the Bayesian model assigns categories on a token-by-token basis. We find that the Bayesian model overestimates the level of syntactic ambiguity in the corpus, which is actually quite low. Although the ability to learn syntactic ambiguity is surely necessary in the long run, our results suggest that strong constraints favoring unambiguous categorization are helpful in early acquisition.

References

Goldwater, S. and Griffiths, T. (2007). A fully Bayesian approach to unsupervised part-of-speech tagging. *Proceedings of the Association for Computational Linguistics*.

Mintz, T. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, 90, 91–117.

Redington, M., Chater, N. and Finch, S. (1998). Distributional information: a powerful queue for acquiring syntactic categories. *Cognitive Science*, 22, 425–469.

Transformational Networks

Robert Frank and Donald Mathis

Department of Cognitive Science
Johns Hopkins University

It is a age-old observation that the sentences of human languages exhibit hierarchical organization, and that this organization is implicated in the mapping being sentence types. Thus, passives are related to actives through the displacement of noun phrase constituents (among other things), and interrogatives are related to declaratives through the fronting of an auxiliary verb that follows the noun phrase subject constituent. How and why do language learners derive structural generalizations about the patterns of their languages? Chomsky's (1975) Argument from the Poverty of the Stimulus (APS) starts from the premise that the relevant data to distinguish between structural and linear generalizations are absent from the learner's input. As a result, the only explanation for the structural basis of language must come from an innate learning bias, which Chomsky argues takes the form of a template for grammatical rules. Though the precise nature of the innate bias has evolved as linguistic theory has developed, the idea that there is a language-specific bias for hierarchical representations has remained constant.

The APS has recently come under fire. On the one hand, Pullum and Scholz (2002) have disputed the degree to which the stimulus is actually impoverished, finding examples of . Yet it remains an open question whether the infrequent presence of examples like (1), which distinguish a structure sensitive generalization for question formation (i.e., front the *main* auxiliary verb) from a linear sensitive generalization (i.e., front the *first* auxiliary verb), is sufficient to drive successful learning (Legate and Yang, 2002).

(1) Is the bird that is singing lonely?

Lewis and Elman (2001) stage a more direct assault on the APS, arguing that even in the absence of examples like (1), learners without any innate grammatical bias can nonetheless induce a structure sensitive generalization for question formation. Specifically, they trained a Simple Recurrent Network (SRN) to perform the task of word prediction on a variety of declarative and interrogative sentence types, withholding examples of the form in (1). In prior work, Elman (1991) had shown that SRNs exhibit sensitivity during the word prediction task to the non-local dependencies involved in subject-verb agreement, and on that basis he argued that they induce a hierarchical representation of the sentence. Lewis and Elman demonstrate that the network they trained generalizes in an apparently structure sensitive fashion when tested on cases like (1): at the relative pronoun *that*, the network predicts the occurrence of an auxiliary verb, and at the end of the relative clause it fails to predict an auxiliary.

There are a number of reasons for skepticism, however. First of all, there is little reason to believe that Lewis and Elman's network represents the relationship between the declarative and interrogative forms of a sentence (or alternatively between the fronted and canonical positions of the auxiliary verb) as such knowledge is unnecessary for the prediction task. Yet the question of structure sensitivity arises only in the context of this relation. As a result this simulation simply doesn't bear on whether an innate structure sensitive bias is necessary. Secondly, as Reali and Christiansen (2005) show, the distinction between the grammatical (1) and its non-structure sensitive counterpart (2) can be predicted using a bigram language model.

(2) Is the bird that singing is lonely?

However, as Kam et al. (2005) demonstrate, as soon as one expands the empirical domain slightly, bigram statistics are no longer sufficient to distinguish between the structure sensitive and linear sensitive patterns. It is therefore possible that Lewis and Elman's network is achieving its success through simple means, which will not generalize beyond their original experiment.

target output					A	B	D	C	•
					C	D	B	A	•
input	A	B	D	C	IDENT				
					TRANS				

Figure 1: Training regimen for reversal network

To approach the question of structure dependence more directly, we moved away from the task of word prediction and focused instead the ability of a neural network to induce the kind of grammatical mappings that were the basis of Chomsky's original argument, namely structure-dependent transformations. There have been a number of previous attempts to get networks to learn structure sensitive mappings (Chalmers, 1990; Niklasson and van Gelder, 1994; Neumann 2002). However all of these works share the assumption that the network is presented with a representation that in some manner encodes the hierarchical structure of the input. Given such an input, it is the task of the network to learn a mapping between this representation of hierarchical structure and another. Yet the situation of language learning does not present a learner with hierarchical syntactic structure. Leaving aside the possible role of prosodic information, any hierarchical structure that is necessary to account for syntactic regularities must be imposed by the learner. Since SRNs have been touted as an instance of a system that can induce hierarchical structure from sequential input, we aimed to investigate their effectiveness in learning to transform sentences from one grammatical form to another.

Although past studies of SRNs have made great use of their ability to accept temporally ordered input, allowing them to take unboundedly long sentences as input, that work has not addressed the question of unbounded outputs of the sort that must be allowed as possible outputs of a grammatical transformation. Botvinick and Plaut (2006) provide a simple and elegant way to do this in their studies of short-term memory for serial order. Botvinick and Plaut demonstrated that when given a sequence of letters as input, an SRN can be trained, upon the presentation of a recall cue, to output the input sequence one element at a time. In order to assess the limits of this ability, we presented an SRN with a somewhat more complex task: instead of a single recall cue that triggered the identical sequence as output, we introduced an additional cue whose target output was a transformation of the original sequence. In the simulation, the input sequences (of which there were 72,000) were drawn from a set of four symbols {a, b, c, d}, and varied in length from 1 to 8, in equal numbers. Training consisted of the presentation of one of these sequences one symbol at a time, with no target output, followed the presentation of one of two recall cues (IDENT or TRANS) for a single time step, which triggered the target output sequence that was either the identity or reversal of the original. This is depicted in Figure 1. The input and output layers of the network contained 6 units, and the output contained 5 units, and these were used for localist representations of the input and output symbols. The hidden and context layers contained 100 hidden units. All units but the outputs used sigmoid activation functions, while the outputs used a soft-max activation function, so that activation was interpretable as the network's assessment of the probability of a particular unit as output. This network was trained for 120,000 weight updates with the Backpropogation Through Time algorithm, using a cross-entropy error function, a batch size of 50 examples, and initial random weights in the range [-.1,+.1]. As seen in Figure 2, this network is extraordinarily successful when tested on novel sequences. An output sequence was judged as correct only if each of the targets was the most active output unit at the appropriate time step.

	length 4 (n=1)	length 5 (n=604)	length 6 (n=4962)	length 7 (n=7870)	length 8 (n=8599)
IDENT	100%	99.8%	100%	99.3%	98.2%
TRANS	100%	99.8%	99.9%	99.3%	97.1%

Figure 2: Accuracy of reversal network on novel sequences

Figure 3: Network output for declaratives

Impressive as it is, the network's success at performing the reversal task does not guarantee success in a grammatical transformation task. As a transformation, reversal does not require sensitivity to any sort of structure in the sequence. We therefore attempted to train an SRN to perform a grammatical transformation on an input sentence, specifically the mapping from declarative to interrogative sentences. Following the design in Lewis and Elman (2001), we trained the network using simple sentences with both transitive and intransitive verbs, with and without auxiliary verbs, subject-verb agreement, and recursive modification of noun phrases by prepositional phrases and relative clauses. All inputs to the network were declarative sentences, while the target output sequence either consisted of the identical declarative (when triggered by a DECL cue) or an analogous interrogative (when triggered by a QUEST cue). Because the vocabulary of the network was larger, the number of input and output units was increased to 34 to allow for localist representations of the entire vocabulary. The number of hidden and context units remained at 100. The training data consisted of 100,000 stochastically generated sentence inputs (average length of 5.54 words, ≈15% including a prepositional phrase modifier, ≈8% with a relative clause modifier), with half of these were followed by a DECL cue, and the other half followed by a QUEST cue, with the appropriate declarative or interrogative sentence as the target output sequence after this point. In order to replicate Lewis and Elman's scenario, we withheld from training one class of training examples: those with a relative-clause-modified subject and a QUEST recall cue. This meant that although the network was exposed to sequences in the input with subjects modified by relative clauses, it was instructed on how to form questions from them. If the network had represented its knowledge of the question transformation in a structure sensitive fashion, so that it encoded a generalization about all kinds of noun phases in subject position, we should expect to find generalization to this held-out example type. In contrast, if the network has represented the question transformation as a mapping between linear sequences of

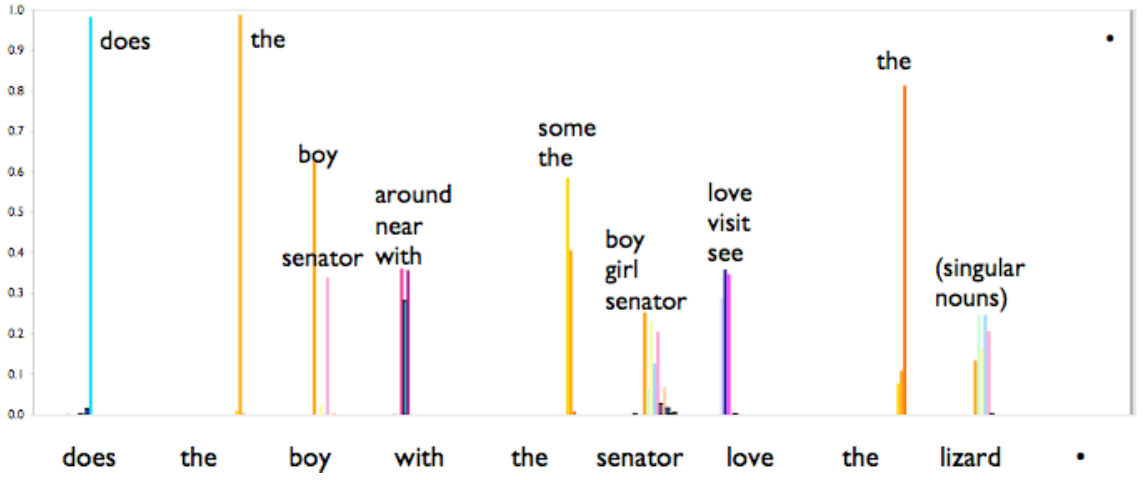

Figure 4: Network output for interrogatives

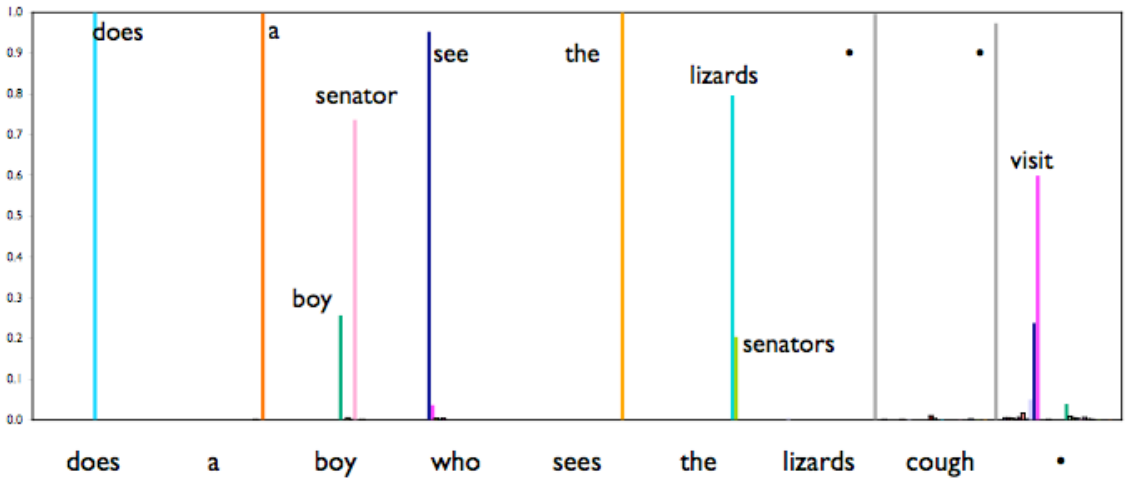

Figure 5: Network output for withheld interrogative

words, the absence of such a pairing in the training data would prevent the network from generalizing to the novel example type.

For the sentence types on which the network was trained, its performance was highly accurate, with the only errors being the substitution of one word by another within the same grammatical class. Examples of the network's outputs as compared to the target outputs for a declarative and interrogative output sentence is shown in Figures 3 and 4 (along the horizontal axis are the target words, vertical bars represent activation of lexical outputs). In contrast, the network was unsuccessful for its interrogative outputs for sentences of the type on which training was withheld. Indeed, the sequence of words output by the network never matched the target, even abstracting away from errors with word class. An example output is shown in Figure 5. Instead, the network's outputs almost always corresponded to output sequences of a type on which the network was trained. These erroneous outputs were not however random. Typically, though not always, they were well-formed questions of some sort, and they generally preserved one of two properties of the input sequence (or both):subsequences of lexical items from the input, or sentence length (the example in Figure 5 preserves the former property). We re-ran this simulation under a variety of conditions, changing the number of hidden units, batch size, learning rate, and vocabulary size, with the same qualitative result emerging in all cases.

The inability of the network to produce a question of the appropriate structure suggests that the network has not induced an abstract structural generalization about question formation that cuts across different instances of noun phrases in subject position. However, the fact that the network does not produce an output that is interpretable as the output of any coherent structural or linear transformation makes it difficult to determine just what sort of knowledge the network has acquired. Indeed, we suspect that the network induces some sort of "output grammar" on the basis of the sequences that it has been trained to output, and this grammar constrains the possible network outputs, even if the network's internal representation could be taken, in some sense, to represent the correct output of the transformation. We are investigating this possibility in ongoing work by considering whether other training regimens might allow the network to produce such output forms.

In spite of the possible presence of an output grammar, there is one way in which we might be able to diagnose the structure sensitivity of the network's knowledge. Consider, for instance, a sentence like the following:

(3) A boy who can love some lizards must cough.

We already know that our network will fail to produce an interrogative corresponding to this sentence. At the very first word of the output, however, the network will need to produce an auxiliary verb of some sort. If the network's transformation of (3) into a question is structurally-based, we should expect to find that the first word in its output will be *must*. In contrast, if the generalization is linearly-based, the first output will be *can*. A third possibility is that the network has learned a default

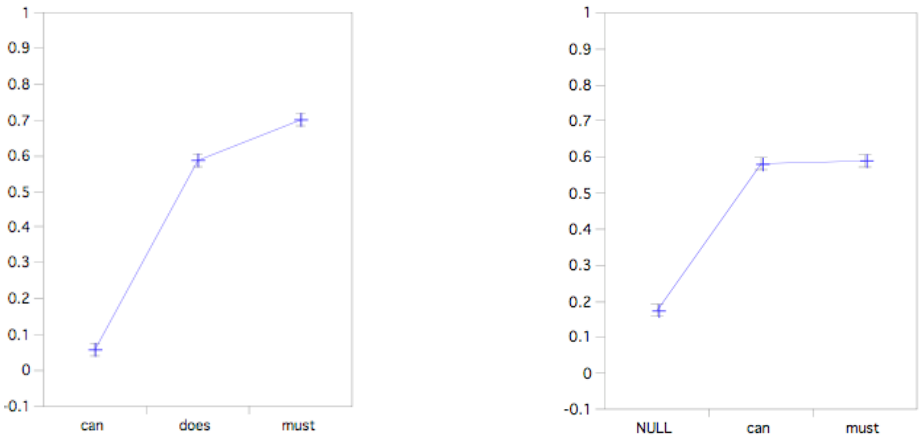

Figure 6: Activation of target auxiliary. Left graph gives activation as the target varies, Right graph gives activation on as the linearly first auxiliary (within the relative clause) varies.

of sorts, so that its output does not depend on the lexical content of the auxiliaries in the input. We can therefore find out something about the structure dependence of the network's behavior by observing the activation of this first word.[1] To test which of these possibilities characterized the network's behavior, we tested the network on a set of 161 sentences, all containing relative-clause-modified subjects, in which the bearer of verbal inflection in the main and relative clauses systematically varied among two modal verbs (*can* and *must*) and the main verb. We found that the mean activation of the correct auxiliary verb at the point immediately following the recall cue was .45. Average activation however varied by the target auxiliary: as illustrated in the left graph in Figure 6, the network's average activation when the target was *can* was virtually 0, while it was much higher for the other two possibilities. The right graph shows that the network's success in producing the correct auxiliary also varied depending upon the identity of the auxiliary in the relative clause modifying the subject (the linearly first auxiliary): when there was no auxiliary within the relative clause, indicated by the label NULL in the graph, as in a sentence like *a boy who loves some lizards can cough*, the network was quite unlikely to produce the correct auxiliary to start the question, but when the relative clause contained one of the modal verbs, the network was much more likely to correctly produce the correct auxiliary. Subsequent replications of this simulation, with different initial random weights, yielded qualitatively similar results.

The network's success in correctly producing the auxiliary verbs *does* and *must* does point to some sort of structural dependence in its knowledge of question formation. However, the network seems unable to put aside irrelevant non-structural factors, such as the identity of the auxiliary in the relative clause, in the formulation of its generalization concerning question formation. Concerning the first of these, it is possible that it is an accurate reflection of the path of child language acquisition. Santelman et al. (2002) *inter alia* have found that children vary in their success in producing correctly inverted questions depending upon the identity of the auxiliary verb, though they found worse performance with *do* than with modals. We leave for future work the question of whether this pattern might arise in a training set with more realistic distributions among the types of auxiliaries. Concerning the sensitivity to the linearly first auxiliary, we are unaware of evidence that children or adults show a similar pattern. We note, however, that we have found a similar inability of SRNs to attend to linearly-based generalizations in on-going work on the induction of anaphora. Contrary to what is sometimes assumed, then, the difficulty SRNs have in inducing grammatical generalizations does not reside in identifying structurally-based generalizations, but rather in ignoring linearly-based ones. Since it was precisely the ability to put aside such non-structural generalizations that was at the crux of Chomsky's APS, we contend that the argument still stands.

Acknowledgments

This work has been supported by NSF grant SBR-0446929. For help in running the simulations and useful discussion, we are grateful to Ebony Gussine, John Stowe, Manny Vindiola, and especially Esteban Buz. We have also benefited from the comments of colleagues at UCLA, University of

[1] Thanks to Bill Idsardi for this suggestion.

Delaware, AMLAP, and University of Maryland, where we have been fortunate to have the opportunity to present some of this work.

References

Botvinick, Matthew and Plaut, David C. 2006. Short-term memory for serial order: A recurrent neural network model. *Psychological Review*, 113:201–233.

Chalmers, David J. 1990. Syntactic transformations on distributed representations. *Connection Science*, 2(1& 2):53–62.

Chomsky, Noam. 1975. *Reflections on Language*. New York: Pantheon.

Elman, Jeffrey L. 1991. Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, 7:195–225.

Kam, Xuan.-Nga Cao, Stoyeshka, Iglika, Tornyova, Lidiya, Sakas, William G., and Fodor, Janet D. 2005. Statistics vs. UG in language acquisition: Does a bigram analysis predict auxiliary inversion? In *Proceedings of the Second Workshop on Psychocomputational Models of Human Language Acquisition*, pages 69–71, Ann Arbor. Association for Computational Linguistics.

Legate, Julie A. and Yang, Charles D. 2002. Empirical re-assessment of stimulus poverty arguments. *Linguistic Review*, 19:151–162.

Lewis, John D., and Elman, Jeffrey L. 2001. Learnability and the statistical structure of language: Poverty of stimulus arguments revisited. In *Proceedings of the 26th Annual Boston University Conference on Language Development*.

Neumann, Jane. 2002. Learning the systematic transformation of holographic reduced representations. *Cognitive Systems Research*, 3(2):227–235.

Niklasson, Lars F. and van Gelder, Tim. 1994. On being systematically connectionist. *Mind and Language*, 9:288–302.

Pullum, Geoffrey K. and Scholz, Barbara C. 2002. Empirical assessment of stimulus poverty arguments. *The Linguistic Review*, 19:9–50.

Reali, Florencia, and Christiansen, Morten H.. 2005. Uncovering the richness of the stimulus: Structure dependence and statistical evidence. *Cognitive Science*, 29:1007–1028.

Santelmann, Lynn, Berk, Samantha, Austin, Jennifer, Somashek, Shamitha, and Lust, Barbara. 2002. Continuity and development in the acquisition of inversion in yes/no questions: dissociating movement and inflection. *Journal of Child Language*, 29:813–842.