April 1993

An Information-Theoretic Solution to Parameter Setting*

Eric D. Brill
University of Pennsylvania

Shyam Kapur
University of Pennsylvania

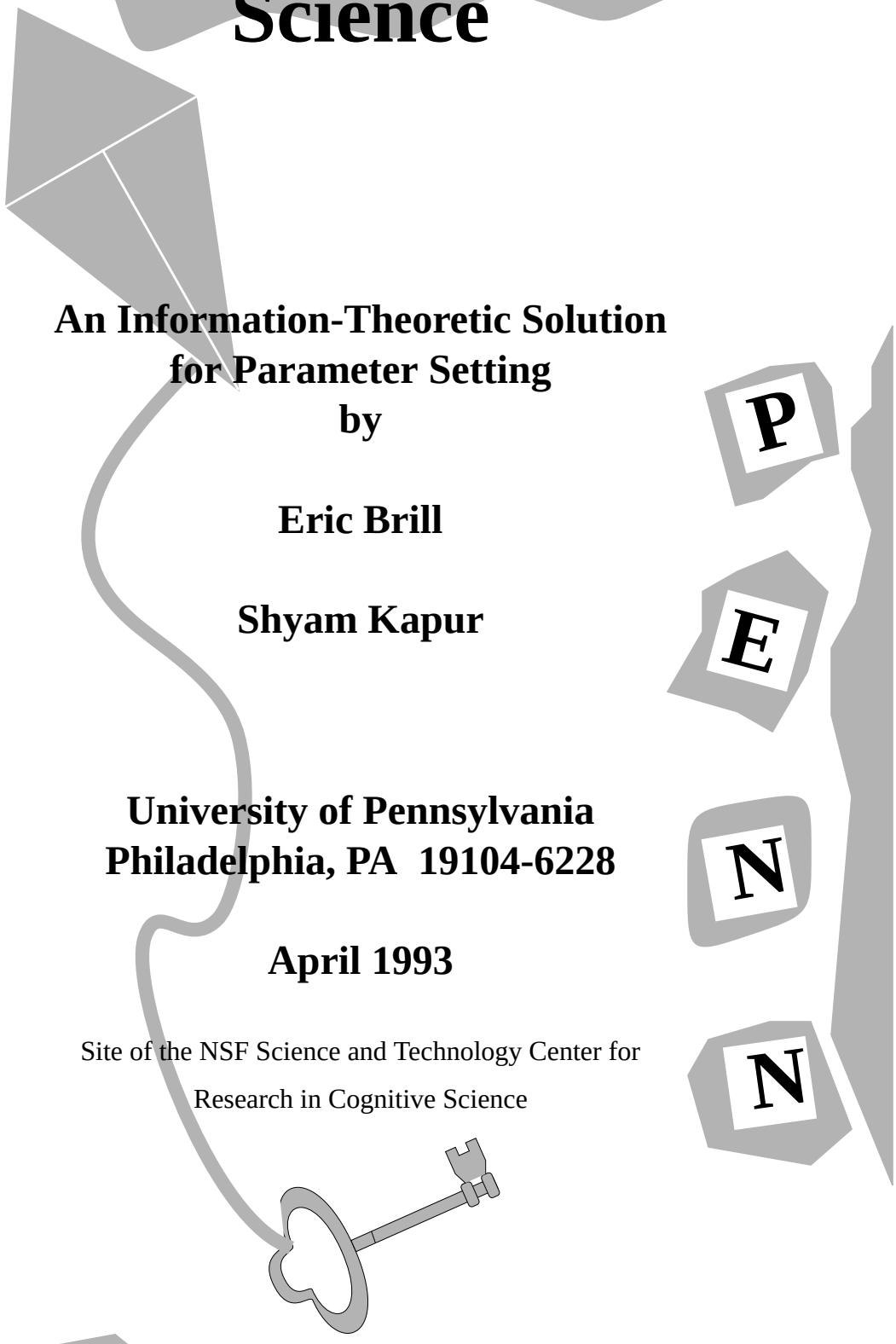
University of Pennsylvania Institute for Research in Cognitive Science Technical Report No. IRCS-93-07.

An Information-Theoretic Solution to Parameter Setting*

Abstract
In this paper, we point out a possible way by which the child could obtain the target values of the word order parameters for her language. The essential idea is an entropy-based statistical analysis of the input stream.

Comments
University of Pennsylvania Institute for Research in Cognitive Science Technical Report No. IRCS-93-07.

The Institute For Research In Cognitive Science



An Information-Theoretic Solution for Parameter Setting

by

Eric Brill

Shyam Kapur

University of Pennsylvania
Philadelphia, PA 19104-6228

April 1993

Site of the NSF Science and Technology Center for Research in Cognitive Science

University of Pennsylvania
Founded by Benjamin Franklin in 1740

IRCS Report 93-07

An Information-theoretic Solution to Parameter Setting*

Eric Brill
Department of Computer Science
University of Pennsylvania
130 Pender, Moore Building
Philadelphia, Pa. 19104
brill@unagi.cis.upenn.edu

Shyam Kapur
Institute for Research in Cognitive Science
University of Pennsylvania
3401 Walnut Street – suite 400C
Philadelphia, PA 19104
skapur@unagi.cis.upenn.edu

April 29, 1993

0 Introduction

In Government-Binding theory, the underlying word order variation of natural languages is accounted for by two parameters that determine the complement-head and specifier-head orders in the X-bar structure. It is clear that the child cannot even begin to apply any sophisticated mechanisms and innate knowledge towards acquiring the rest of the syntax of her language if she has not set these word order parameters to their target values. There is considerable evidence from psycholinguistic studies that children master the word order of their target language very early. By and large, their earliest productions are consistent with the word order of the language to which they are exposed. At the same time, there is considerable confusion in the input the child gets with regard to word order. For example, in some SOV languages such as Dutch and German, the finite verb in root clauses moves from its base position to the second position in the sentence, so that the child will get SVO forms. In fact, there is a preponderance of such forms shown below in (1).[1] In embedded clauses however, the verb remains in the final position and the SOV order is obtained, as shown in (2).

(1)	(Dut)	Ik	versta	je	niet		"I do not understand you"
		I	understand	you	not		
(2)	(Dut)	... dat	ik	je	niet	versta	"... that I do not understand you"
		that	I	you	not	understand	

When the verb appears in this second position, not only the subject but also objects can appear in the first position. Thus, the resulting OVS order is also possible, as shown in (3). In addition, adverbials and prepositional phrases can also occupy the first position (4).

(3)	(Dut)	Dat	weet	ik	niet		"I do not know that"
		That	know	I	not		
(4)	(Dut)	Dan	maken	ze	een	bootje	"Then they make a boat"
		Then	make	they	a	boat	

*This work was supported in part by ARO grant DAAL 03-89-C-0031, DARPA grant N00014-90-J-1863, NSF grant IRI 90-16592 and Ben Franklin grant 91S.3078C-1

[1] All the following examples are taken from the caretaker speech subcorpora of the CHILDES database (MacWhinney and Snow 1985).

This phenomenon is termed verb-second (V2). While linguists do not agree on the exact description and explanation of this phenomenon (for some suggestions, see Holmberg and Platzack 1990, Vikner and Schwartz 1991, Travis 1991, and Zwart 1991), it is clear that languages differ in whether or not they show V2 effects. Gibson and Wexler (1992) capture this variation by means of a V2 parameter and show that the resulting word order parameter space has local maxima, i.e., particular parameter settings different from the target from which the learner will never be able to escape. Given this type of logical problems in the acquisition of word-order (e.g., see Gibson and Wexler 1992, and, for relevant discussion, also see Frank and Kapur 1992), in addition to the apparent confusion in the input data, it is all the more surprising that the child appears to set the word-order parameters to her target value easily. In this paper, we point out a possible way by which the child could obtain the target values of the word order parameters for her language. The essential idea is an entropy-based statistical analysis of the input stream.

1 The Learning Algorithm

1.1 Introduction

In this section, we describe an algorithm which could be used to determine the proper setting of the V2 parameter. The parameter is set by using information-theoretic measures on a small[2] corpus of unannotated text. No structural analysis is carried out on the text. The only knowledge assumed prior to learning is a small list of words known to be verbs.[3] The first step in learning involves carrying out a distributional analysis to automatically learn a larger set of verbs. The V2 parameter is then set based upon a comparison of the distributional behavior of words at varying distances from the verbs.

1.2 Entropy

Entropy is a measure of randomness. In particular, the entropy of a random variable X, measured in bits, is $\sum_X p(x) \log p(x)$. To give a concrete example, the outcome of a fair coin has an entropy of $-(.5 * \log(.5) + .5 * \log(.5)) = 1$ bit. If the coin is not fair and has .9 chance of heads and .1 chance of tails, then the entropy is .5 bits. There is less uncertainty with the unfair coin—it is most likely going to turn up heads. Entropy can also be thought of as the number of bits on the average required to describe a random variable.

Conditional entropy is the entropy of a random variable given another random variable. Clearly, the conditional entropy is at most equal to the entropy of a random variable. For example, consider the random variable X which equals 1 if it rains on a particular day and is 0 otherwise. Consider another random variable Y which equals 1 if it is cloudy on a particular day and is 0 otherwise. We will be able to better predict the random variable X if we know the random variable Y (i.e., X conditioned on Y) than if we do not. Suppose another random variable Z is related to the outcome of a lottery. Clearly, the entropy of X conditioned on Z is the same as the unconditioned entropy of X.

1.3 Learning Verbs

For the V2 parameter-setting algorithm to work, a set of verbs must first be learned. In our experiments, we used 20 verbs. We could assume that the child has properly classified 20 words as

[2] 3,000 utterance

[3] 5 in this experiment.

verbs prior to setting the parameter, but this seems unreasonable and unnecessary. Rather, we can begin with a much smaller list of known verbs and use a learning algorithm to try to find a number of additional verbs. The learning algorithm is based upon the work of Zellig Harris (1951), in developing algorithmic methods for field linguists to determine word classes and class membership in an unfamiliar language.

First, a "distributional fingerprint" is built for the five representative verbs. The distributional fingerprint is a probability vector $P(W)$, where W is a random variable over words, indicating the probability of word $w \in W$ occurring before (after) any of the five verbs. Probabilities are estimated using a corpus of sample utterances. The distributional behavior of any word can then be compared to the approximated verb distributional behavior by comparing the distributional fingerprint of that word to that of the set of sample verbs. A number of similarity measures could be used for this comparison. In this work, we used relative entropy. Relative entropy is an information-theoretic measure indicating the amount of additional information (measured in bits) needed to describe the random variable X given the random variable Y. If $X = Y$, then $\textit{rel-entropy}(X, Y) = 0$. For every frequently occurring word in a sample corpus, we measured the relative entropy of its distributional fingerprint and the five verb fingerprint. If words are then sorted by this measure, the words on the top of the list, i.e., words with the smallest relative entropy, tend to be verbs. (See Brill and Marcus 1992 for a fuller description of distributional fingerprints and their use in computational linguistics.)

From this sorted list, we picked the first 20 words, most of which indeed turn out to be verbs. The chart below shows the accuracy of this method. It is not essential for the success of the learner that all the words claimed to be verbs are indeed verbs, just that a large fraction are.

Language	Size (K-words)	Number Correct (20 total)
Dutch	41	19 (95%)
English	314	20 (100%)
French	46	16 (80%)
German	14	17 (85%)
Italian	24	18 (90%)

1.4 Conditional Entropy in the Neighborhood of Verbs

Once the set of verbs has been learned, the next step is to analyze the distributional behavior of other words around these verbs. For each of the 20 verbs, we measure the entropy of the probability distribution of words occurring precisely one word to the left of the verb in the corpus. We also measure the entropy 2 and 3 words to the left of each verb, as well as 1,2 and 3 words to the right, obtaining a table of entropy(VERB, POSITION), where VERB=learned verbs and POSITION = [-3, -2, -1, +1, +2, +3] from verb. The position -1 is the closest to the verb on the left and +1 is the closest on the right. Then, we averaged the entropy at each of the positions for all the verbs so that we obtained entropy(POSITION), the average entropy of the word distributions at different distances from any of the 20 verbs.

We have now demonstrated how, for any language, provided with a small corpus of utterances that are not structurally annotated in any way, along with a very small list of words known to be verbs, we can determine the average positional entropy of the distribution of words at various distances from this set of words. We will next show how this information can be used in parameter setting.

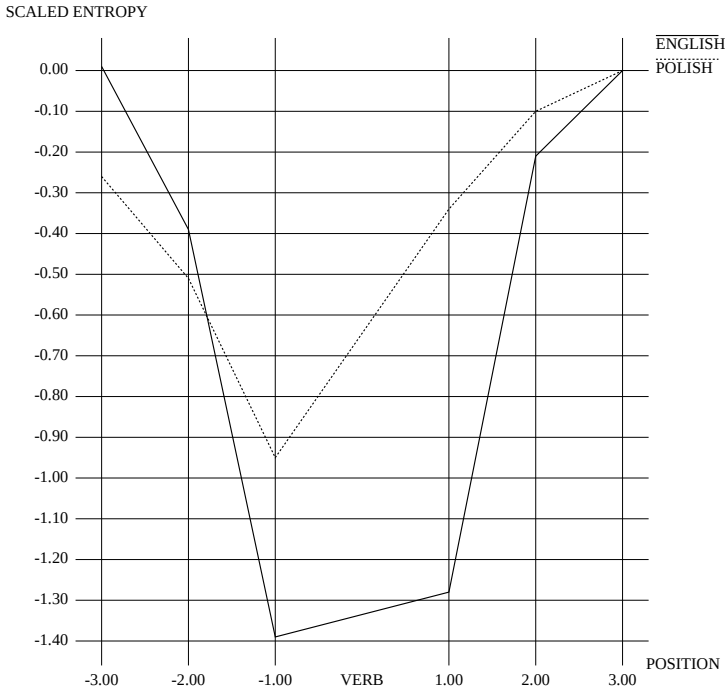


Figure 1: Entropy Conditioned on Position

2 The Main Results

Up to this point, we have shown how we obtain the entropy values for each of the three positions to the left and to the right of each of the verbs. At least for the languages we have considered so far, it is reasonable to assume that the third position to the right of the verb is not influenced by the verb to any measurable degree. In other words, with high probability, the conditional entropy of this position will approach the unconditional entropy because the verb will at most weakly "select" this position. With respect to this level of the entropy (which we call the base level), we expect the entropy to dip as the verb is approached from the right and then at or around the third position to the left of the verb the entropy should return to the base level. Our expectation is borne out as shown in Figure 1 for Polish and English where the entropy averaged over the 20 verbs is plotted at each position. Furthermore, certain aspects of word order seem to stand out. Polish has a much flatter graph than English, possibly a reflection of the relatively free word order in the former but not in the latter.

We next investigated the change of entropy values between adjacent positions. Recall that the positions marked -1 to -3 are the three positions to the left of the verb, -1 being the closest, while the positions marked +1 to +3 are to the right. All the languages behave similarly between every pair of adjacent positions except between position -1, the position immediately to the left of the verb, and the position +1, the one immediately to the right. For Danish, Dutch and German, the entropy of position -1 is considerably higher than that of position +1 (See Figure 2.). For all other languages, the reverse holds. (See Figure 1 and Figure 3.). Also see figure 4 for a compact representation of our results where predictability (inverse of randomness) is plotted along the y-axis for positions -1 and +1.) We believe that this distinction is precisely due to the V2 phenomenon.[4]

[4]It has been suggested to us that in a VS languages such as Irish and Welsh we would also observe that the entropy

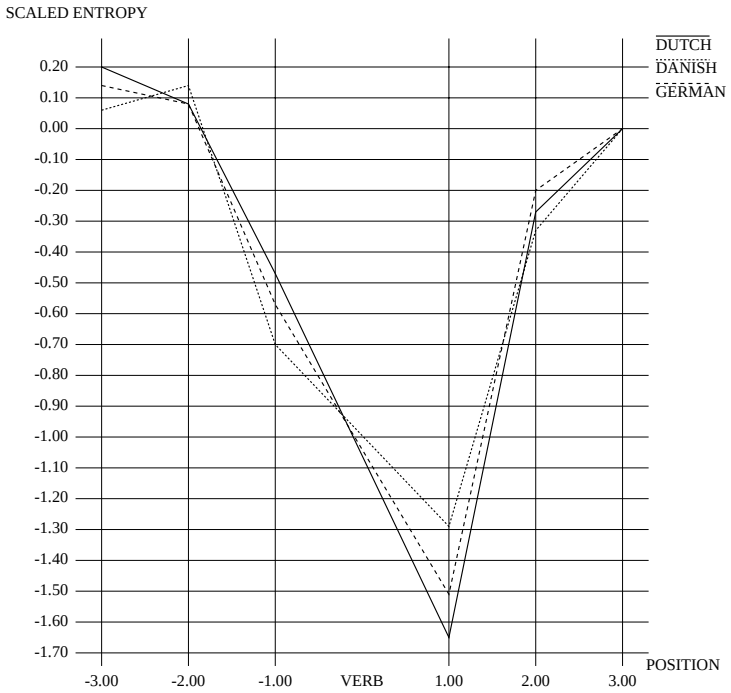


Figure 2: Entropy Conditioned on Position

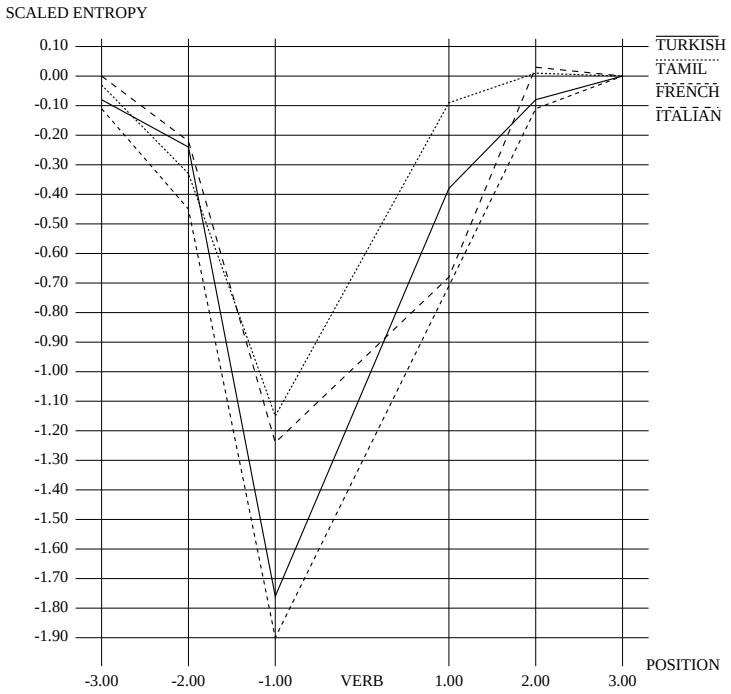


Figure 3: Entropy Conditioned on Position

The chart below shows the result of a 1-tailed paired t-test. Notice that except for English (possibly due to remnant V2 effects), all other results are significant (marginally so for Danish).

Language	**Significance %**
Danish	94.23
Dutch	99.99
English	66.75
French	99.98
German	99.84
Italian	98.77
Polish	98.02
Tamil	99.98
Turkish	99.92

Next we considered if something could be said about the similarity between the graphs of the various languages we had obtained. We used a simple measure of the distance between graphs, the mean square distance. Given two languages L_1 and L_2, we defined

Similarity(L_1, L_2) = square root of sum of square of difference in entropy(POSITION).

We then computed the possible partition of the languages into two sets such that the sum of the average similarity between all the languages in each of the groups is maximized. The optimal partition out of the 246 possible ones turns out to be the one in which the V2 languages Danish, Dutch and German are together. Thus, the graphs appear to be capturing an essential linguistic characteristic that the languages in this group share. However, as such, the graphs do not reflect the base word order since Danish is SVO while Dutch and German are SOV.

Recall that due to the V2 effect, the verb moves to the second position and then the first position may be occupied by any of a number of possible syntactic categories, i.e., noun phrases (both as subjects and objects), adverbials, and prepositional phrases. Whenever the subject is not in the first position, it is most likely in the first position to the right of the verb (as in (3) and (4) above). Thus, whether the basic word order of the language is SOV or SVO, the position to the left of the verb is likely to be much more random exactly in case the language shows V2. In our opinion, this is precisely reflected in the graph. We also believe that this is one plausible mechanism by which the child can determine whether or not her language has V2. It is also clearly the case that once the value of the V2 parameter is settled, the input is far more revealing with regard to the other word order parameters. The child can then set the remaining word-order parameters in a number of alternative ways, some of which may also involve information-theoretic criteria. On the other hand, variations of the simple trigger-based strategy of the kind Gibson and Wexler (1992) discuss might also suffice since once the value of the V2 parameter is settled, the remaining word-order parameter space is devoid of any local maxima. Since we have been constrained by the number of language corpora we have access to, we are not in a position to say anything definitive at this point.

---

of position -1 is much higher than that of position 1. We have been limited by the lack of the corpora we need to test this prediction.

3 Some Additional Observations and Results

3.1 Noise in the Input

Notice that our learning algorithm is remarkably robust. We did not clean the input, so that there are many instances of incomplete words as well as sentence fragments (e.g., "And that is a"). We did remove any annotations of the documenter and selected only those utterances whose length was at least two. We also ignored the sentence boundaries so that none of the punctuation marks are considered to be part of the input. Even though it is plausible that the child can introduce sentence boundaries and/or punctuation marks from speech systematically, we do not assume that she can. This is in keeping with our goal to establish upper bounds on both how much the child needs to know and what she needs to do in order to learn. To make the noise problem harder, we even introduced random noise so that about one third of the utterances had words altered randomly to noise. Remarkably, the results were unaffected.

3.2 Closed Class Words

The presence of articles could be making a difference to the entropy of certain positions in a non-uniform fashion. We considered whether removing some of them would allow us to obtain similar or better results. Since closed class words are often unstressed in caretaker speech and absent from the first productions of children, there is a rich ongoing debate in language acquisition about whether or not functional categories are available to the child from the outset. Keeping ourselves neutral on this issue, we obtained modified entropy graphs when certain words were effectively assumed to be absent from the input. In order to simulate a way in which the child could skip over the closed class words (without having to know them beforehand), we considered the possibility that the child ignores a certain number of words from among the most frequent ones. We have observed that most of the frequent words in any of the languages tend to be closed class. In English, for example, the six most frequent words—"you", "the", "a", "it", "to", and "what"—were ignored. In French, the words ignored were "tu" (you), "pas" (not), "cést" (it is), " a" (that), "le"(the (masc)) and "que" (that (compl.)). Our results shown in Figure 5, Figure 6 and Figure 7 continued to show the same pattern. In fact, we got strong similarity between the graphs for three pairs of languages—Dutch and German, Italian and French, and Tamil and Turkish. It is well known that the languages in each of these pairs are similar in a variety of respects including their word orders. Dutch and German are SOV with V2; French and Italian are SVO without V2 and Tamil and Turkish are both SOV also without V2.

3.3 Simplifying the Entropy Calculation

An objection to our approach can possibly be raised that the computational cost of calculating entropies of positions may be too severe on the child. Since entropy involves the transcendental function "log", in general, the child could at best be computing an approximation to it. We investigated this possibility by substituting a two-term Taylor approximation to the "log" function for the "log" function and approximated entropy in that way. The results were again unaffected by this change. Notice that the resulting computation only involves the four basic arithmetic operations. We also performed the same simplification in computing the verbs by similarity and obtained good results. Of course, there are other ways of answering this criticism. It is well known that there are measures comparable to entropy which involve only the basic arithmetic operations. We conjecture that the results would hold regardless of which of them is substituted for the real entropy.

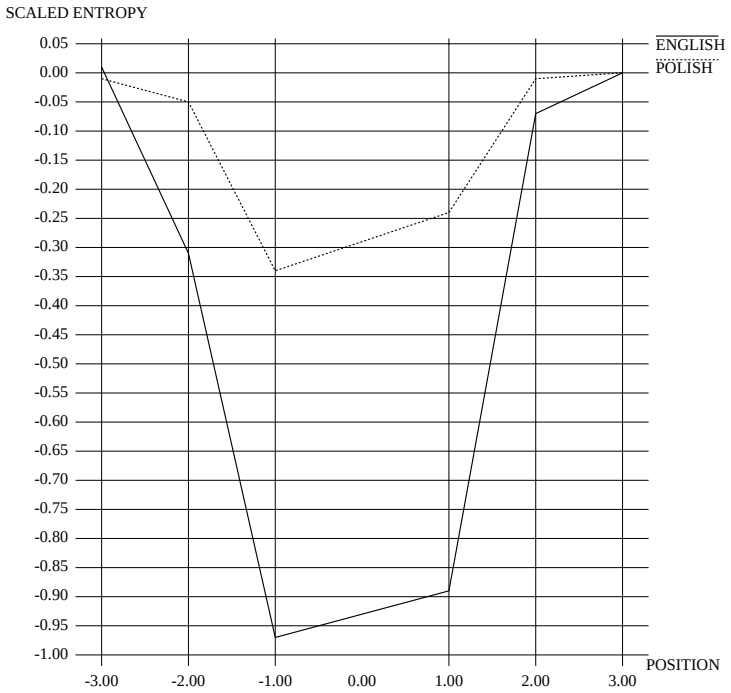


Figure 4: Entropy Conditioned on Position

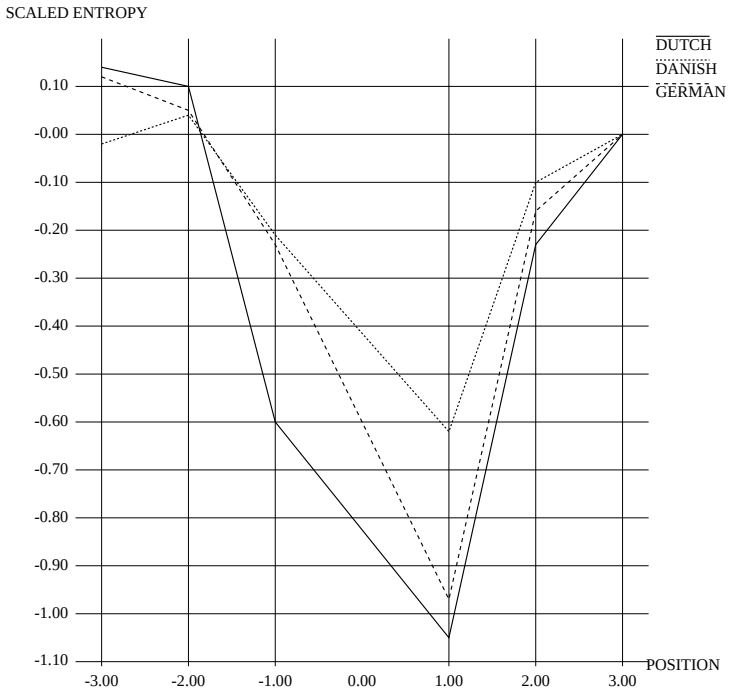


Figure 5: Entropy Conditioned on Position

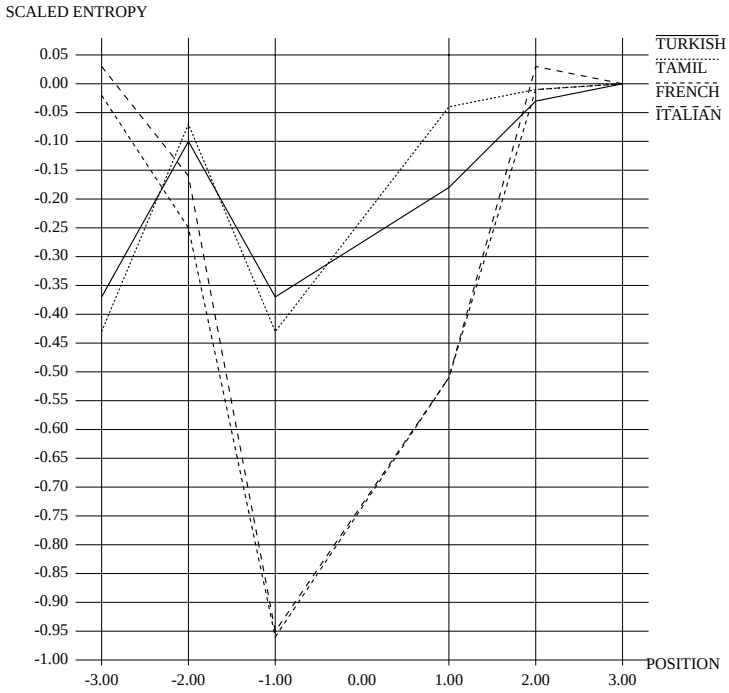


Figure 6: Entropy Conditioned on Position

3.4 Storage Requirements

Another possible objection to this approach is that the method of computing statistics over a corpus of utterances would require an unreasonable amount of memory in the child. First, it should be noted that the child need not store all past heard utterances in their entirety. Only word-pair co-occurrence statistics are stored. In the worst case, our method could require on the order of n^2 additional storage, where n is the size of the vocabulary, beyond that already employed to store the vocabulary. This is the storage required for keeping statistics on all possible word pair co-occurrences. However, due to Zipf's law (Zipf 1949), the actual storage requirements will be much less. In Brill (1993b), experiments show an empirical upper bound for the storage requirements at about $3 * n$, and since the only word pairs considered are those where one of the words is in the list of 20 verbs, this number will be much smaller.

3.5 Verbs rather than Words

We have been forced to assume that the child is able to distinguish certain word types which are all common verbs before it is able to acquire word order. We noticed that there was no systematic pattern in the graphs if instead of 20 words most of which were verbs we considered just any 20 words. Of course on the basis of what we know about natural languages, that verbs select their arguments etc., it is not surprising that the child needs to focus on them in order to obtain revealing information about the word order, in particular about the verb-complement order.

3.6 Corpus Size

We have shown results involving corpora containing 3000 utterances each. This number is not arbitrary but in fact it is the lower bound up to which we have been able to obtain the results we have reported (both with regard to verb extraction and the entropy of the positions around the verb). At 1500 utterances, it seems that the results are no longer systematic so that on different random corpora of that size for the same language we obtain graphs that look quite different from each other. On random corpora of size 3000 or larger we always obtain similar graphs. Depending on how we conceive the child to be learning, we could interpret these results in different ways. If the child does not need to commit herself to the value of the V2 parameter at any definite point, she could very well start off with a small corpus and at various stages continue to determine the value of the V2 parameter to the best of her ability. From around the point the child has seen about 3000 utterances, we believe the child would stabilize on the correct value of the V2 parameter. But of course she may never know (nor need to know) whether or not she has stabilized on the correct value.

In some learning proposals, it may be required that at some point the child must make a definite decision about the value of a particular parameter and never have to revoke this. Our learning strategy is consistent with such a learning proposal as well. The child could first figure out at which point the graph appears to be similar, for example, by comparing two or more fixed size corpora. She could then decide the value of the parameter based on a corpora of that size. Notice that there is a different kind of stability the child could seek. This would require that the graph on different corpora of some fixed size be almost identical. Clearly, the first kind of stability is a prerequisite for the second kind of stability. We conjecture that the second kind of stability may not be achievable given the constraints under which the child has to work. Furthermore, our results show that it is not necessary for the child to seek this second kind of stability in order to be successful at learning. Notice also that it is not at all necessary that stability be achieved at the same point for each of the languages. It could well be the case that the graph stabilizes for Turkish at 750 utterances while for English it only does so at 3000 utterances.

4 Properties of this Proposal

One feature of this proposal that stands out is that it is fully verifiable. The results for all the nine languages we have considered so far can easily be replicated. Furthermore, whatever claims we make about the information-theoretic properties of the input can be refuted or confirmed for other languages as soon as corpora of size 3000 utterances of caretaker speech are available for them. In this regard, the proposal is more scientific than many others considered in the literature.

Our proposal is also in line with some ideas based upon formal learning theory incorporated in some recent work (Kapur 1993a, Kapur 1993b). Here is a brief overview of this model of parameter setting. The parameters are subdivided into groups and the groups are ordered by their relative frequency of expression. The parameters which are expressed more frequently are assumed to be set first. In this proposal, learning is categorically not error-driven so that the motivation for the child to set parameters derives from efficiency considerations, i.e., the need to extract "meaning" from the input increasingly rapidly. Parser failure is not regarded to be of any importance in this process. Initially, all the parameters are unset and the parser is organized to obey only all the universal principles. At this stage, utterances from any possible natural language will be accommodated with equal ease, but no sophisticated structure can be built out of them. Input will be used to weigh support for each of the alternative settings of the parameters in the first group. Clearly, this evidence must be of a primitive nature, for the parser is incapable of anything but a rudimentary

analysis. The word order parameters of the kind discussed in this paper are very basic and expressed frequently so that they are likely to belong to this first group. It must be possible to set them, based on a very superficial analysis of the input.

We have demonstrated in this paper that in fact the word order parameters can be set, based on rather simple analysis of unstructured unannotated input. The method is tolerant of considerable noise in the input. Minimal processing requirements are assumed of the child; the child needs to semi-reliably find word boundaries in her speech stream but need not even find sentence boundaries, nor make use of any syntactic information. The few common verbs required initially only need to be acquired in the sense that their types are known but there is no need to have any knowledge of their subcategorization frames or their semantic categories. They are viewed as nonsense tokens but with some salience; all other words are viewed as plain nonsense tokens. Furthermore, the amount of input required—3000 utterances—is very reasonable. It is conceivable that the child uses prosodic information from the speech stream in order to acquire syntactic aspects of her language. Our work does not contradict this. It only shows that there is no logical necessity for the child to use prosodic information at least for determining the word-order parameters.

5 Future Work and Conclusions

There are a number of directions in which we plan to continue this work. For one, as soon as we can get access to corpora from other languages, we would like to verify that the relationships observed are maintained. For another, we still have to show how the child can determine the word-order parameters for other lexical and functional categories besides the verbal ones. We are quite optimistic that we will be able to obtain results along the lines developed in this paper.

Subsequently, we hope to develop our proposal to handle other parameters. It is not unlikely that for some other parameters we will need more sophisticated primitives and not just word tokens and their linear positions relative to words. For example, once the X-bar structure of the utterance can be built subsequent to the setting of word-order parameters, structural notions such as c-command would become available if needed. Just as the V2 parameter seems to reveal itself in a straightforward entropy characteristic, it is not unlikely that other parameters' values would also have consequences which are far more surface apparent than the phenomena themselves.

As far as we know, our work is the first effort to establish an interesting link between the traditionally theoretical generative notion of language that involves principles and parameters, and techniques that are similar to those used by the structuralists of the 1950s and 60s. (For other work in this spirit, see Brill 1991, Brill and Marcus 92a, Brill 93a, and Brill 93b.) If progress continues to be made along the direction we have embarked upon, there would be major ramifications for linguistics and the theory of language acquisition. For one, it would be conclusively demonstrated that the whole issue of the absence of negative evidence in the input has been overblown. Based on formal learning theory, we have shown elsewhere that stochastic nature of the input can be used to adequately compensate for the absence of negative evidence (Kapur and Bilardi 1992, Kapur 1993a, Kapur 1993b). The issue of overgeneralization and its purported solution, the Subset Principle (Berwick 1985, Manzini and Wexler 1987), are meaningless since the extensional relationship between various languages is not a parameter in our proposal. For example, two languages could well be subsets of each other, but there is adequate evidence to move from any one language to the other, if necessary. Assuming stochastic input, this evidence may require observation of non-occurrence in order to generate a canonical form of indirect negative evidence. In this paper, it more directly took the form of the entropy characteristic in the neighborhood of common verbs.

Our work would also establish that the trigger-based approach to learning is overly simplistic. A trigger need no longer be considered to simply be a single utterance or even a small set of them. Rather, it is the statistical properties of large portions of the corpus that trigger parameter values. Such a move is necessary if tolerance to noise and confusion in the input is to be satisfactorily established.

References

Berwick, Robert. *The Acquisition of Syntactic Knowledge.* MIT press, Cambridge, MA, 1985.

Brill, Eric. Discovering the Lexical Features of a Language Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics. Berkeley, Ca. 1991

Brill, Eric and Marcus, M. Automatically Acquiring Phrase Structure Using Distributional Analysis. Proceedings of the DARPA Speech and Natural Language Workshop; February, 1992. Harriman, N.Y.

Brill, Eric and Marcus, M. Tagging an Unfamiliar Text With Minimal Human Supervision. American Association for Artificial Intelligence (AAAI) Fall Symposium on Probabilistic Approaches to Natural Language, Cambridge, Ma. AAAI Technical Report. 1992.

Brill, Eric Automatic Grammar Induction and Parsing Free Text: A Transformation-Based Approach. Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics. Columbus, Ohio. 1993.

Brill, Eric A Corpus-Based Approach To Language Learning PhD Dissertation, Department of Computer and Information Science, University of Pennsylvania. 1993

Frank, Robert and Shyam Kapur. On the use of triggers in parameter setting. Technical Report IRCS-92-52, Institute for Research in Cognitive Science, 1992. Presented at the Boston University Conference on Language Development.

Gibson, Edward and Kenneth Wexler. Triggers. Presented at GLOW, April 1992.

Harris, Zellig. *Structural Linguistics.* Chicago: University of Chicago Press. 1951

Holmberg, A. and C. Platzack. On the role of inflection in scandinavian syntax. In W. Abraham, W. Kosmeijer, and E. Reuland, editors, *Issues in Germanic Syntax.* Mouton de Gruyter, Berlin, 1990.

Kapur, Shyam. (1993a) How much of what? Is this what underlies parameter setting? Presented at the 25th Stanford University Child Language Research Forum.

Kapur, Shyam. (1993b) Some applications of formal learning theory results to natural language acquisition. In Barbara Lust, Magui Suner, and Gabriella Hermon, editors, *Syntactic Theory and First Language Acquisition: Crosslinguistic Perspectives.* Lawrence Erlbaum Assoc. Presented at the 1992 symposium on 'Syntactic Theory and First Language Acquisition: Cross Linguistic Perspectives' at Cornell University.

Kapur, Shyam and Gianfranco Bilardi. Language learning from stochastic input. In *Proceedings of the fifth conference on Computational Learning Theory.* Morgan-Kaufman, 1992.

MacWhinney, B. and C. Snow. The child language data exchange system. Journal of Child Language, 12, 271-296. 1985.

Manzini, M. R. and Kenneth Wexler. Parameters, binding theory and learnability. *Linguistic Inquiry*, 18:413–444, 1987.

Travis, Lisa. Parameters of phrase-structure and verb-second phenomena. In R. Freidin, editor, *Principles and Parameters in Comparative Grammar.* MIT press, Cambridge, Mass., 1991.

Vikner, S. and B. Schwartz. The verb always leaves IP in V2 clauses. ms., University of Geneva, 1991.

Zipf, G. *Human behavior and the principle of least effort.* New York: Hafner Pub. Co. 1949.

Zwart, J. Clitics in Dutch: Evidence for the position of INFL. *Groninger Arbeiten zur Germanistischen Linguistic*, 33:71–92.