

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

COMPUTATIONAL EVIDENCE
FOR THE USE OF FREQUENCY INFORMATION IN
DISCOVERY OF THE INFANT'S FIRST LEXICON

by

ELEANOR OLDS BATCHELDER

A dissertation submitted to the Graduate Faculty in Linguistics
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy, The City University of New York

1997

UMI Number: 9807905

**Copyright 1997 by
Batchelder, Eleanor Olds**

All rights reserved.

**UMI Microform 9807905
Copyright 1997, by UMI Company. All rights reserved.**

**This microform edition is protected against unauthorized
copying under Title 17, United States Code.**

UMI
300 North Zeeb Road
Ann Arbor, MI 48103

© 1997

ELEANOR OLDS BATCHELDER

All Rights Reserved

This manuscript has been read and accepted for the Graduate Faculty in Linguistics in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

16 Sept. 1997
Date

Virginia Teller
Chair of Examining Committee

Sept 16, 1997
Date

Charles E. Cain
Executive Officer

Martin S. Chodorow

Virginia Valian
Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

COMPUTATIONAL EVIDENCE
FOR THE USE OF FREQUENCY INFORMATION IN
DISCOVERY OF THE INFANT'S FIRST LEXICON

by

Eleanor Olds Batchelder

Adviser: Professor Virginia Teller

My thesis is that statistical characteristics of language are a significant source of information to prelinguistic infants in isolating meaning-based chunks from ambient speech. The principal demonstration is a cross-linguistic corpus study, using the O-B computational algorithm to locate word boundaries in a wide variety of language corpora. Additional evidence that such a strategy is cognitively plausible is cited from research in both linguistics and psychology.

First, a review of the linguistics literature shows that the use of distributional information as one of several combined cues to segmentation of the speech stream would fill an explanatory gap.

Then the O-B algorithm, a modification of work by Olivier (1968), takes a language text with word-delimiting spaces removed and successfully re-divides it into most of the original words. The algorithm proceeds incrementally, segmenting each utterance as it is presented by using frequency information gathered from all preceding utterances, and building a lexicon as it goes along. A number of experiments on spoken and written texts in English, Spanish, and

Japanese show that such distributional cues are inherent in all the texts examined and are effective in segmentation. Specific characteristics of texts which facilitate segmentation are identified:

1) More language is better.

2) Short utterances, short words, and high lexical repetition make learning faster, but are not absolutely necessary. In the corpora studied, these were more typical of spoken than written texts, and of speech to young children than speech to adults.

3) No consistent difference is found between various coded representations, including standard alphabetic spelling, phonemes, and Japanese hiragana (mora), suggesting that children can use idiosyncratic categorical representations in this early period.

A review of results from cognitive psychology show that the encoding of frequency data is an implicit, automatic, innate process, and that infants and adults actually do use frequency information to segment language-like stimuli in laboratory experiments.

Finally, I compare the O-B algorithm with five other computational models of speech segmentation, concluding that the O-B algorithm is comparable in quantitative performance and superior in cognitive plausibility.

Acknowledgments

The Internet has been essential to this dissertation in many roles. All my corpora came from the Internet, as well as phoneme conversion programs and special fonts. The Web has often been my library, providing research papers to download, and definitions and discussions of specialized terms in various fields. E-mail serves up a virtual community of scholars—I received timely news of conferences and publications, and corresponded with many far-flung strangers who were kind enough to answer my questions about their work. Telnet enabled me to do the bulk of this research on a Unix computer in New York while I was actually living in Tokyo.

The eight years that I have been studying linguistics have been absorbing and eventful ones. I cannot review everyone who has influenced me during this long process, so I apologize to many who deserve thanks (and a few who deserve worse) for omitting to record them here.

I have been spurred forward over the rough patches by the knowledge that my father, my mother, and my daughter all preceded me down this doctoral path, and by the faith placed in me by my mother, Victoria Olds, and by my partner of ten years, Fumiko Ohno.

This dissertation leans heavily on the elegant work of Donald Olivier, whose algorithm and whose prose have inspired me to try and do likewise.

I am most directly indebted, however, to my committee—Martin Chodorow and Virginia Valian—and, above all, to Virginia Teller, without whose guidance and support over many years this research and this document would not have been possible.

Table of Contents

Abstract	iv
Acknowledgments	vi
Introduction	1
1 The Infant Segmentation Problem	5
1.1 What and when? The results of infant segmentation	6
1.1.1 Evidence from production	8
1.1.2 Evidence from comprehension	9
1.1.3 Modeling the outputs of segmentation	12
1.2 Possible paths to proto-lexical segmentation	14
1.2.1 The influence of motherese	16
1.2.2 Isolated words and utterance boundaries	18
1.2.3 Suprasegmental cues	20
1.2.3.1 Prosody and bracketing	21
1.2.3.2 Prosody and clustering	22
1.2.3.3 The “trochaic bias” in English	25
1.2.4 Phonetic cues and phonotactic constraints	28
1.2.5 Distributional cues	32
1.2.6 Summary of segmentation hypotheses	36
1.3 Modeling the input: Theory and data	37
1.3.1 Phonological theories	39
1.3.2 Experimental data	43
1.3.3 Implications for a segmentation model	45
2 Computational Models and Metrics	47
2.1 Models and metrics in computational linguistics	47
2.1.1 Information theory, entropy and mutual information	49
2.1.2 Probabilistic language models	52
2.1.3 Recall, precision, and F-measure	56
2.1.4 Bracketing measures	60

2.2	Statistical approaches to segmentation	62
2.2.1	Segmentation into words	63
2.2.2	Segmentation of words into morphemes	66
2.2.3	Discovery of words in Japanese and Chinese	68
2.3	Connectionist models of word segmentation	74
2.3.1	University of Rochester	76
2.3.2	University of Edinburgh	78
2.3.3	University of Southern California	80
2.3.4	Comparison of network results	82
2.4	Probabilistic approaches to word segmentation	87
2.4.1	Olivier and Wolff	87
2.4.2	de Marcken	88
2.4.3	Brent & Cartwright	91
2.4.4	Redlich	95
2.5	Summary	96
3	The O-B Algorithm: Methods	98
3.1	The language input	98
3.1.1	Introduction to six text corpora	99
3.1.2	Statistical characteristics of corpora	101
3.1.3	TTR and G: measures of repetition	102
3.2	Metrics to evaluate segmentation performance	104
3.2.1	Evaluation by types of segmentation	104
3.2.2	Adjusting for the random baseline	107
3.2.3	Counting divisions rather than words	110
3.2.4	Significance and correlation measures	112
3.3	The O-B algorithm	113
3.3.1	Building the lexicon	114
3.3.2	The parsing optimization	117
3.3.3	Controlling word length with the optlen parameter	119
3.3.4	Do utterance boundaries contribute information?	128
3.3.5	Training, testing, and held-out data	132
3.3.6	Cognitive implications of the O-B algorithm	135

4	The O-B Algorithm: Segmentation Results	137
4.1	Quantity of text matters	137
4.1.1	Quantity of text for training	137
4.1.2	Length of line as parse span	141
4.2	Symbol set sometimes matters	143
4.2.1	English: standard orthography vs. phonemes	144
4.2.2	Spanish: more or less phonetic representations	147
4.2.3	Japanese: hiragana vs. roman characters	150
4.2.4	Summary and comment on cognitive issues	154
4.3	Register does matter	156
4.3.1	English	157
4.3.2	Spanish	160
4.3.3	Japanese	162
4.3.4	Conclusions	166
4.4	Language doesn't matter	167
4.5	Conclusions	176
4.5.1	Computational implications	176
4.5.2	Cognitive and linguistic implications	178
5	Experimental Evidence	181
5.1	Experiments in speech segmentation	182
5.1.1	Hayes & Clark (1970)	182
5.1.2	Goodsitt, Morgan & Kuhl (1993)	185
5.1.3	Saffran, Newport & Aslin (1996)	188
5.1.4	Saffran, Aslin & Newport (1996a,b)	191
5.2	Frequency and cognition	193
5.2.1	Implicit learning	194
5.2.2	What is counted? Salience and redundancy	197
5.2.3	What is useful? Cue reliability	201
5.2.4	Cue reliability in corpus experiments	203

5.3	Segmentation experiments redux	208
5.3.1	Cue reliability in speech segmentation experiments	208
5.3.2	Experiments in syntactic segmentation	213
5.3.2.1	Green (1979)	214
5.3.2.2	Valian & Coulson (1988)	218
5.4	Summary	220
6	Cognitive Models	222
6.1	The infant's cognitive development	223
6.1.1	Frequency and clustering	223
6.1.2	Starting small	226
6.1.3	Trial and error	229
6.1.4	Frequency and rules	231
6.2	Computational algorithms as cognitive models	233
6.2.1	Criteria for cognitive models	234
6.2.2	Strategy and goals: dividing or clustering?	236
6.2.3	Timing	239
6.2.4	Constraints	242
6.2.5	Processing resources: mind vs. machine	246
6.2.6	Representation of language	247
6.2.7	Quantitative comparison of algorithms	251
6.3	Summary	256
7	Conclusions about Language	258
7.1	The effect of input on O-B segmentation	259
7.2	Modeling the input	262
7.3	The future of statistics in language	266
	Appendix A. Parsing Algorithm	270
A.1	Building the Lexicon	273
A.2	Comparing the Olivier and O-B Algorithms	274
A.3	Forward parse of: youthisawhim	275
A.4	Backward parse of: youthisawhim	279

Appendix B. Corpus Descriptions	283
B.1.0 Introduction	283
B.2.0 About CHILDES	284
B.3.1 EngV	287
B.3.2 EngW	292
B.4.1 SpaV	296
B.4.2 SpaW	301
B.5.1 JapV	303
B.5.2 JapW	309
Appendix C. Summary Table of Corpus Characteristics	313
References	316

Introduction

The problem of the nature of language can be approached from many points of view—historical change vs. synchronic variation, the individual and the community, learning vs. processing, and so forth. The many subdisciplines within linguistics testify to the complexity of the phenomenon of human language.

This paper is concerned with one small part of the process of language acquisition by the prelinguistic child. As infants listen to the language about them and somehow make it their own, one of their first tasks is to break up the continuous streams of speech into more manageable chunks. By the end of the first year, we observe the first “word comprehension,” or “the child’s dawning appreciation of some of the conventional meaning units of the adult language” (Vihman 1996:122). This achievement of locating “words” is a logically necessary precursor to the more complex learning which follows, but until very recently it had not been given much attention.

This dissertation will argue that the child bootstraps a first lexicon by combining several kinds of information in the language signal, and that one of these is the cooccurrence frequencies of the various sound classes. Our demonstration will be twofold: that such information would be of substantial assistance in solving the segmentation problem; and that it is likely that infants actually notice and make use of such information. Evidence of utility will be by computer simulation, and evidence of plausibility will come from human experiments. This demonstration will therefore straddle several boundaries into which scholarship is traditionally segregated, borrowing traditions and techniques from three

fields of cognitive science: linguistics, computer science, and psychology. Because we cannot assume that readers have specialist knowledge in each of these fields, we will provide minimal background information as necessary.

First, the computer simulation will be motivated and constrained by the facts of child language learning. The problem addressed in this thesis is a puzzle posed by child language learning: How does the young child first break the flow of continuous sound into pieces small enough to be associated with meaning? We present in Chapter 1 a review of the relevant research in developmental psychology and linguistics, to show that our appeal to the statistical structure of language will combine with other sources of information to provide a complete account of infant behavior. We also begin to address the thorny issue of cognitive plausibility by discussing how the ambient language can be best represented as input in a computer simulation.

In the second chapter, we present a second context for this work. Technologies developed by computational linguistics to enable computers to process human language by exploiting its quantificational characteristics—an activity called, rather ambiguously, “natural language processing.” The assumptions and conventions necessary to provide a foundation for understanding techniques and results in this field will be briefly surveyed. Then we review previous computational approaches to language segmentation in general, and finally those studies which are most relevant to the infant segmentation problem will be discussed in some detail. Explicit comparison of the approaches, however, will be deferred until the concluding chapter.

This dissertation research is a computational study designed to learn how much information about word segmentation is contained in the language signal. Chapter 3 describes the approach and materials used in this study, and introduces the metrics which will be used to report results. The computer algorithm (program) used here was chosen for its simplicity and efficiency. We wanted to be able to modify its operation easily so that we could find the optimum set of parameters. We also wanted it to extract the relevant statistical regularities from text as completely and as rapidly as possible, so that we could compare a wide variety of language samples. The final version of the program is called the O-B algorithm. It is a self-organizing (unsupervised) process which uses only language units such as letters, phonemes, etc. and utterance boundary indicators. It begins with single units as “words,” and makes larger and larger clusters, preferring as words those clusters which occur most frequently. The product of the algorithm is a lexicon and a segmented version of the input.

Chapter 4 presents the quantificational results of a number of experiments, or “runs,” each consisting of the application of the algorithm to a different corpus. These experiments explore the effect on the segmentation results of various corpus characteristics such as size, representational code, whether the text was originally spoken or written, and the language—English, Japanese, and Spanish.

After demonstrating that this algorithm is successful in segmenting the speech stream, we consider its plausibility as mental process in the infant (Chapter 5). There is evidence that the recording of event frequencies is a computation which the human brain is well equipped to do and that it is commonly observed

in other cognitive processes. Further, experiments have been done on both adults and infants, using linguistic stimuli, which directly suggest that frequency information is used for language segmentation. We will critique these experiments and ask what statistic in particular was cognitively effective.

Lastly, we consider the implications of treating this computer simulation as a cognitive model, and we compare it with other simulations of infant segmentations with respect to its cognitive plausibility. While none of the computational approaches can be truly called “realistic,” a close comparison of them with each other, and with the infant’s cognitive characteristics, brings to light important issues in our attempt to better understand language development (Chapter 6).

The work concludes with a summary of what we have learned about the nature of language from this study. Looking to the future, we argue that the statistical structure of language should be considered as an important attribute and considered as an possible influence in all areas of linguistic research.

Chapter 1. The Infant Segmentation Problem

The general progress of the infant's language behavior is well documented. After several stages of babbling development (Oller & Lynch 1992:513), at about a year old, or perhaps as late as 18 months, "recognizable morphemes begin to be produced by normal infants" (Menyuk 1984:197). By this time, the receptive vocabulary is already about 50 words (Benedict 1979; Stoel-Gammon 1992b). Children continue to add new productive vocabulary slowly, still generally producing single-item utterances. By the middle of the second year, they have a receptive vocabulary of about 250 words and a productive repertoire of 50 words or so (Ingram 1992:434). They begin to form multi-word utterances, slowly adding a few grammatical morphemes, and at about this time also there is a "vocabulary burst," with new words being learned at a much more rapid rate (Macken 1979). By about two years old, they begin to show evidence of learning syntactic structure.

The question we consider here is how the baby begins the process of word-learning by first extracting from the speech stream discrete chunks of sound to which meaning can be associated. Between about 9 and 18 months of age, the child somehow comes to understand that the stream of vocal sound is meaningful and makes a good start on discovering the necessary sound-meaning pairings.

Our first question is how learners pick out 'pieces of the acoustic-linguistic world' from the sound stream, how they detect and represent the sound. Notoriously, the phones, words, phrases, and sentences of everyday speech

run together so complexly that their proper segmentation would appear to be very difficult to accomplish. (Gleitman et al. 1988:152)

This chapter has three sections and three goals, all intended to set the stage for our computational presentation by addressing pre-existing assumptions that the reader may entertain. First, to orient readers who are less familiar with the child language literature, we provide a brief review of available descriptions of the child's "first words." We will see that the outcome of the child process is less crisply defined than the outcome of a computer process, and we emphasize that this and other differences between the child and the computer need to be kept in mind throughout.

Second, we review a number of explanations of this first segmentation process which have been put forward elsewhere. We will argue that some are inappropriate for infant bootstrapping, and others are useful as partial cues to word boundaries, but that there is nevertheless reason to think that distributional facts also play a significant role, as one of several information sources. We will thus propose to isolate distributional information from other cues in order to estimate just how much of the task it could account for.

In the third section, we consider how we might best represent the language input to our computer algorithm. While we want to get as close as possible to the infant's mental representations, there are conflicting theories and little hard evidence about what that representation is.

1.1 What and when? The results of infant segmentation

What are the contents of the baby's first lexicon, and when does it begin? In this section we will see what kinds of evidence we have about this process. But

first let us briefly consider some terminology. First, the term **lexicon** is shorthand for “the mental lexicon,” or the mental store of linguistic forms which are arbitrarily and irregularly associated with meaning in any particular language and, since they cannot be generated by rule or analogy, must be committed to memory.

We conventionally treat the adult lexicon as filled with **words**—units that generally have distinct meanings and independent forms, and serve as the primitive units of grammatical structure. One source says:

Although elusive, and barely definable at best, the word plays an important role in the phonologies of many of world’s languages. For whatever reason, language users tend to build utterances out of grammatical/semantic units known as words, which in turn have effects on the sound structures of languages... It has been approached in a number of different ways, with definitions always falling somewhat short of perfection. It nonetheless maintains great intuitive appeal... (Hyman 1978:445)

Here, we will interpret words as lexical items, pieces of linguistic form that are associated with a meaning. These may be morphemes, compounds, or phrases, and we do not rule out the possibility of a redundant lexicon which contains both larger groups and their components (Peters 1983).

The term **segmentation** is used in the literature to refer to several different concepts. Here, we mean segmentation into lexical units/words, but in other contexts it can refer to division into sub-word units like morphemes or syllables, or to division of text into phrases or topical sections. Another type of segmentation, which we will not discuss in detail, is a process used by the adult in understanding spoken language. While there may be a developmental continuum between the infant’s first segmentation and the adult processing, still the adult always

takes into account an extensive mental lexicon, while the infant has no mental store of words to begin with. The initial bootstrapping segmentation creates the child's first lexicon, so it can be called "prelexical" or "proto-lexical" segmentation.

1.1.1 Evidence from production

The most direct evidence about children's linguistic knowledge comes from what they say. Though such evidence is neither necessary nor sufficient, it is the most frequently relied upon by both lay persons and professionals (Curtiss 1982). The term "language acquisition" itself often refers to speech production, implicitly ignoring comprehension. "Learning words" often means producing them and one-year-olds are said to be "acquiring the first word," despite the fact that they demonstrate understanding of a number of words by then.

One problem with studying speech is that there are severe limitations to the infant's production abilities. Some particular sounds cannot yet be articulated, and an undeveloped working memory makes possible only short and simple utterances. These limitations make it difficult to translate what the child "says." Early speech often does not resemble the adult forms very much, if at all, and indeed may sound more like the babbles which precede and overlap it. Moreover, there is great variability, even for the same child at the same time (Schwartz 1988; Ferguson & Farwell 1975), so it is not easy to determine just what the "first words" are. One researcher described her criteria:

A word was accepted as produced if it was used spontaneously with meaning, the meaning being determined from the context of use. The sounds used by the child did not need to duplicate adult sounds but the sound used had to be a consistent form for the child. (Benedict 1979:186)

In studies of 9- to 18-month-olds, the words most often reported are those which seem to be the same as adult “words”—many nouns and names, followed by other single words, but also common phrases of several words, such as *out there* (Vihman & McCune 1994).

But is this limited output really an accurate gauge of what the child “knows”? Are these “production constraints” across-the-board cognitive limits, or do they merely affect the child’s speech? Perhaps a wider definition of “utterance” would be a fairer approximation. It seems reasonable that all children enhance their limited speech ability by using intonation, gesture, and facial expression. Thus, when a child says [jy] while stretching a cup-holding hand toward the refrigerator, it is not hard for even a stranger to understand “I want some juice!” or “Please give me more juice.” In a sense, this child’s single syllable can be considered a “word,” and the single word has the pragmatic force of an utterance, so one could say that the syllable, the word, and the utterance at this early stage of development are identical.

But what then should be considered the end product of segmentation—the (assumed) word or phrase, or the (pronounced) syllable? Is “shoe” the output token for a mental “These are my shoes”? This more inclusive view of “production” does not give us a well-defined result of segmentation.

1.1.2 Evidence from comprehension

Since we are really interested in what the child understands, measures of comprehension would seem to be more appropriate. Benedict (1979) is an often-cited study which shows that comprehension develops earlier than production,

and that it develops much more rapidly.¹ She studied eight children from 0;9, using a variety of tools including parental diaries and checklists, direct observation, and explicit testing procedures, and then compiled separate lists for each child of the first 50 “words” produced and the first 50 comprehended, further broken down by word class. The following examples of the “action words” class (p. 196) show that some “words” are usually considered to be phrases (the numbers in parentheses refer to the number of individual child lists that contain the word in each mode, showing the degree of individual variation at this stage):

<u>Both Compr. & Prod.</u>	<u>Comprehension only</u>	<u>Production only</u>
bye-bye (7, 5)	come here (7)	woof woof (5)
peek-a-boo (5, 4)	wait a minute (2)	there she is (1)
nite-nite (3, 4)	so big (3)	read (1)
get (7, 2)	give (8)	out (3)
see (2, 5)	kiss (5)	stop it (1)
no (7, 6)	may I have (2)	eat (1)
get down (4, 3)	look at (6)	walk (1)

If we envision a redundant lexicon, we can make the unconventional assumption that these frequently heard phrases are also entries in the adult lexicon, along with their component words (Pawley & Syder 1983). In the young child’s lexicon, however, it may be that there is as yet no redundancy: Perhaps the phrasal elements are not yet broken down, but are simply holistic units. The results of early segmentation may be less uniform and more coarse-grained than would be suggested by simply assuming standard “words.”

¹ “The results showed, in addition, a gap of approximately 5 months between the attainment of the 50-word level in comprehension and attainment of the same level in production. These results support the long-standing claim that comprehension develops before production...” (Benedict 1979:190). In detail, by the time they could produce 10 words, children understood 50-60 words on average, with individual children ranging from 30 to 182. The ages at which production reached 10 words ranged from 0;11 to 1;6, with a mean at 1;1.21.

Recently, subtler experimental techniques have given even better indications of the baby's receptive knowledge. Though simple recognition is not the same as comprehension, nevertheless it seems logical that segmentation and at least some sort of stored mental representations are prerequisites to word recognition. Hallé & Boysson-Bardies (1994) presented two-syllable words in isolation to French infants of 12 months of age, comparing their responses to words likely to be familiar and those unlikely to be. In a head-turn preference procedure, with no prior training period and in the absence of any situational or contextual cues, the infants were found to prefer the familiar words. The authors conclude that "the ability to recognize familiar words... appears to be based on stable word representations" (Hallé & Boysson-Bardies 1996:464).

Taking a different tack, Myers, Jusczyk, Kemler Nelson, Charles-Luce, Woodward & Hirsh-Pasek (1996) placed one-second pauses in samples of fluent speech, a pause for every 3-4 words. They found that 11-month-old infants, but not 9-month-olds, consistently listened significantly longer to passages with such pauses placed between words rather than between syllables within words. Further, low-pass filtering² removed this distinction, so it was presumably not based on suprasegmental information alone. In additional trials, it was shown that the different response did not depend solely on using familiar words, nor on stress patterns, since both strong-weak and weak-strong words were treated the same. They hypothesized that younger English-learning infants may be locating prosodic units like trochaic feet (two syllables with strong+weak stress pattern), but by

² As discussed below in Section 1.2.3, low-pass filtering removes phonetic details from the speech signal, leaving only prosodic information like pitch, stress, and intonation.

11 months they can use multiple sources of information to locate “true words.”

1.1.3 Modeling the outputs of segmentation

What do these studies tell us about the child’s first lexicon? What are the implications of the interaction of production and comprehension? Does the child segment as a function of comprehension, and then adapt these word-like representations for output? Or is the child attempting to produce an entire utterance, beginning with the most salient syllable, so that segmentation is an implicit by-product of production?

The connection between comprehension and production in the very young child is still a mystery (Stoel-Gammon 1992a). It might be the case that the mental representations used in production and those used in comprehension are not as closely linked in the young child as in the adult, which leads to proposals that the child has two distinct lexicons. The “two-lexicon model,” though it has shortcomings as “a cutting-edge research-level model,” is still an “intuitively satisfying... way of organizing many kinds of data” (Menn & Matthei 1992:214).

Perhaps if we view early language learning as involving effort on the part of the child (Elbers & Wijnen 1992), what looks like two lexicons can be seen as evidence that this effort proceeds in two directions at once during the first year. Production, with its more limited span, concentrates on starting with small units and expanding them, while comprehension may focus on trying to chop up somewhat larger chunks into smaller ones. Without wishing to suggest a rigid division of labor, we will see that the cooperation of these two strategies—both “chopping” the speech stream into sections and “plucking” out salient pieces—

will have echoes in the segmentation hypotheses that we consider below.

In summary, infant segmentation is not deterministic, but neither is it completely undefined:

- Two kinds of segmentation can be observed: individual syllables and words isolated from the speech stream, and short phrases recorded as all or part of frequent utterances. In both cases, the boundaries are such as to preserve meaning, or at least not violate it. That is, even when some standard word boundaries are ignored in “fused phrases,” the boundaries which *are* created generally coincide with some standard word or morpheme boundaries.
- Segmentation does not always happen all at once, but may be a gradual process. A frequent utterance like ‘What’s that?’ may be learned as a whole ([wəsæ] from Vihman 1996:245), and then later segmented into words over a period of time.³
- Though length and phonetic content may differ between production and comprehension, there is no evidence that the boundaries of lexical items conflict between the two modes. For instance, it is possible that “dyawannabanana?” is a lexical item for comprehension at some stage, while “na” is the equivalent production version at the same time. But we would not expect to find mutually exclusive boundaries, such as “dyawa,” “nabanana” as lexical items in one mode and “dyawanna” and “banana” in the other.

When we present our computer algorithm in Chapter 3, we will need some way to measure its results against those of the child. Since the baby will eventual-

³ Two further examples of phrasal “words” are from Vihman (1996:256): *dans l’eau* in the water’ [dalo] and *de l’eau* ‘some water’ [dəlo].

ly arrive at the adult lexicon, perhaps we might be justified in simply measuring the algorithm against that ultimate goal. The characteristics summarized above lend some support to that plan, but they also suggest that a graded series of criteria might give a more finely grained evaluation, and we will propose such a series in Chapter 3.

1.2 Possible paths to proto-lexical segmentation

The question of initial segmentation has not been addressed as often as it would seem to have been relevant:

...theories of syntax acquisition commonly adopt the assumption that, prior to acquiring structural knowledge of their language, children can represent input utterances as strings of words. A few theorists have worried over how children might attain proper word-level representations... Most often, however, this assumption has been adopted uncritically. (Morgan & Demuth 1996a:7)

Many studies which discuss early “words” simply do not comment on how children discover these as linguistic units. Statements like “Words are salient in the input” are more a description of the problem than a solution to it. A recent book by a prominent researcher on *The Lexicon in Acquisition* says all she has to say on the topic in a footnote:

Children acquiring words have a mapping problem. They must isolate the word-forms of their language; they must create potential meanings; and then they must map the meanings onto the forms. All three tasks are critical to the solution of the mapping problem in lexical acquisition. For present purposes, I presuppose that children can isolate the relevant forms. [Footnote: This task is not an easy one... However, it will not be covered further in the present discussion.] (Clark 1993:43)

Of the various proposals that we will review here, none claims to account for the child’s success single-handedly. Currently it is uncontroversial that

segmentation is the result of several strategies or information sources operating in concert (Morgan 1996b). Further, most researchers allow for a goodly amount of individual variation in acquisition strategies, and thus implicitly or explicitly characterize the language-learning child as actively seeking solutions to the linguistic puzzles that confront her. A memorable example, from a somewhat older child, illustrates the process (Peters 1983:43):

Christine is a 4-year-old girl; Steven is an adult male. They are riding in the back seat of a car. Christine is acting rowdy. Steven tells Christine she "must behave" if she wants Steven to read her a book. He is, however, paying more attention to a cassette tape that is playing music than he is to Christine. A couple of minutes later:

- C: Steven I am /heyv/.
 S: What? You hate? What do you hate?
 C: /heyv/. I am /heyv/.
 S: You hate? You hate me? The music? What?
 C: No, I am /heyv/. /heyv/.
 S: I don't know what you are talking about.

Silence. A bit later:

- C: I /heyv/.
 S: You hate me?
 C: (shakes her head no)
 S: Who do you hate?

Silence. A bit later:

- C: I am behaving.

In this example, Steven's repeated lack of understanding twice forced Christine to try and reanalyze the word "behave." It was only when she reached the correct (adult) analysis that Steven became aware of the nature of her difficulty.

The general cognitive strategy which children apply to the language learning task is illustrated by an influential series of "operating principles" proposed by Slobin (1973, 1985) and Peters (1983). For instance, Peters (1983:36-37) proposes that children memorize in their entirety many of the utterances that they hear and then try to segment them into smaller units, using strategies like:

- Segment off the last syllable of a unit from the rest.
- Segment off the first syllable of a unit from the rest.
- Segment off a stressed syllable from the rest.
- Segment units at rhythmically salient places.
- Segment units at intonationally salient places.
- Segment off sub-units that are repeated (in segmentals of meter or melody) within the same unit.

While clearly these are not conscious “rules,” the child’s linguistic development does seem to be shaped by some such guidelines. An approach to language learning which hypothesizes general guidelines and strategies in this way is difficult to quantify or to model, but it may be as close as we can get to precision, given the variable nature of the data and the many different patterns cross-linguistically.

So we see that the first segmentation process is the result of a number of information sources and strategies, and the particular mix may vary from child to child (Peters 1977) and from language to language (Peters 1981). What are the possible cues that the child might use?

1.2.1 The influence of motherese

A number of studies of “motherese” (also called “child-directed” or “infant-directed” speech: CDS/IDS) have documented characteristics that seem likely to facilitate language acquisition, among them: slower articulation, short utterances, higher pitch, exaggerated pitch excursions, pauses occurring more reliably at phrase and clause boundaries, and less segmental variability (Gleitman et al. 1988; Bernstein Ratner 1996; Cuñer 1994; van de Weijer 1997). The logical prob-

lem with these claims, however, is that one does not wish to make “motherese”-type input a necessary precondition of language acquisition.

...[T]here exist cultures whose child-rearing practices have the effect of reducing the opportunity to hear this speech style (Schieffelin 1979), not to mention neglected or abused children in American homes. So far as we know these individuals learn to talk all the same. (Gleitman et al. 1988:153)

It seems likely that children will learn to talk, as they learn to walk, under all but the most extreme conditions. They are successful even in times of war and social chaos, for instance, and in conditions of severe neglect—though not absolute deprivation, as shown by the case of Genie (Curtiss 1977). While there are children whose language development is slower than others, virtually all children eventually acquire adult competency (Leonard 1992). If this were not the case,

...then mothers who, out of belief or distraction, neglect to speak in this special way would be responsible for their child’s failure to learn language. (Mehler & Christophe 1993:13)

The IDS/CDS register or style is best seen as including “prosodic modifications [that] facilitate speech processing, and in this way could ease children’s entry into language” (Werker, Lloyd, Pegg & Polka 1996:430), perhaps contributing to earlier or more rapid language development:

...[T]he motherese register is neither necessary nor (in the end) explanatory of language learning... So the question of category learning remains to be explained even after it is acknowledged that most children solve the problem in the presence of ‘the best possible’ input data base. (Gleitman et al. 1988:153n)

In Chapter 4, we will see that the O-B algorithm produces a better segmentation on texts that have one of the characteristics of “motherese”—short utterances—

but prosodic and phonetic variations are not represented in our corpora, so our results do not comment upon those possible influences.

1.2.2 Isolated words and utterance boundaries

Since the utterance—speech bounded by silence—is salient and universal, it is the least controversial candidate for a meaningful unit, particularly when it is short. One simple hypothesis that has been put forth is that infants use one-word utterances to bootstrap an early vocabulary, and then use that rudimentary vocabulary to discover more words, and so forth (e.g., Brent 1996). Such a theory assumes not only that one-word utterances occur with sufficient frequency and variety in the environment, but also that children have some way to identify them as such. How are infants to know which short utterances consist of one vs. several “words”?

In addition, the success of an isolated-word strategy depends on their frequency in the input. Generally the statistic reported is the percent of utterances which contain only one word, but we would also like to know the percent of word tokens which appear in isolation, and the proportion of word types. Here, for instance, here are the figures reported for various corpora of speech addressed to small children:

<u>Source</u>	<u>Isolated words as % of</u>		
	<u>Utterances</u>	<u>Word tokens</u>	<u>Word types</u>
Aslin et al. (1996) ⁴	28%	—	—
Bernstein Ratner (1996:142)	22%	—	—
Our corpus	21%	6%	26%
Morgan (1996b:75)	18%	5%	—
Morgan et al. (1996:277)	<10%	—	—

⁴ Reported in Morgan et al. (1996).

If most one-word utterances are instances of a very few types, the child cannot build even a rudimentary lexicon this way (Chiat 1979). Morgan (1996b:75) determined that the majority of the one-word utterances in their data were "exclamations such as *hey!* or *whoa!*" When these were excluded, less than 5% of utterances and 2% of tokens were one-word. For our corpus shown above, containing 2,000 utterances (33,000 words), the number of types represented seems fairly high, but half of these occur only once in the isolated form, and another one-fifth appear only twice. Thus, a more realistic percentage might be the 8% of word types that appear alone three times or more. In sum, it seems that one-word utterances are not a very fruitful source of information for the word-seeking infant.

Much more helpful would be to pay attention to the beginnings and ends of utterances, taking advantage of the information provided by the existence of each utterance boundary. This echoes the "operating principles" of Slobin (1973, 1985) and Peters (1981, 1983) that tell learners to focus extra attention on the beginnings and endings of things. Two kinds of information are available. The short chunks of speech at the beginning and end of each utterance might be salient as word candidates. Aslin et al. (1996) reported that parents tend to place important words at the beginning and ends of utterances in both English and Turkish (a verb-final language). If this is so, then children who choose to learn "boundary words" first will be good learners.

A second kind of information is the distributional information of sounds and sound clusters at the utterance boundaries, reasoning that "properties occurring at the boundaries of utterances may also signal boundaries of units within utterances" (Morgan, Shi & Allopenna 1996:279). This in fact is the

approach taken by several of the computational models that we will look at, and is based on the assumption that there are phonotactic regularities which are correlated with word boundaries. We will discuss phonotactic information further in Section 1.2.4 below.

To summarize, isolated words (one-word utterances) are only a minimal source of information for word segmentation, but the infant may get considerable help from observing the pieces of language which lie next to utterance boundaries. This confirms the usefulness of shorter utterances in infant-directed speech that we noted in the previous section, since shorter utterances provide more utterance boundaries from which to extract information.

1.2.3 Suprasegmental cues

The speech signal consists of both segmental and suprasegmental information. Segments are discrete units, “the vowels and consonants... the ‘verbal’ aspect of speech” (Crystal 1987:169). The suprasegmental, or prosodic, features of language are contributed by variation in pitch (fundamental frequency, f_0), loudness (amplitude, intensity), and duration (speed, length). These “-etic” components of prosody combine in language-particular ways to produce “-emic” constructs at the word or phrase level—generally termed intonation, stress, and rhythm. The least complicated prosodic element is a pause, or silence, which we discussed above.

Prosodic components of speech can be separated from segmental ones by the technique of low-pass filtering, which removes phonetic detail and thus allows experimenters to distinguish between effects of prosody alone vs. prosody

plus segmental information. Experiments with infants in the first few months of life have shown that they are sensitive to prosodic features of language from birth (deCasper & Fifer 1980; Mehler et al. 1988). There is even evidence that prosody can be heard in the womb and is thus the first language stimulus (deCasper & Spence 1986).

In the following discussion, we return to the cooperating strategies of “chopping” and “plucking” that we introduced earlier in this chapter. Here we argue that prosodic information is useful to both strategies—providing cues for effective bracketing and making salient certain clusters.

1.2.3.1 Prosody and bracketing

Just as pauses bracket the flow of speech into “utterances,” so other prosodic features of the speech stream serve to break up the utterances into even smaller chunks. Thus, prosodically marked phrasing can assist word discovery in the same ways as do utterance boundaries. Useful prosodic cues are universally available:

Across languages, prosodic cues drawn from a limited set comprising pitch discontinuities, lengthening of segment duration, variation of sentence rhythm, and pausing universally indicate phrase boundaries. Although the manifestations of these cues may differ across languages, all languages recruit cues from a small set of possibilities (Lehiste 1970) [citation in original]. Moreover, the multiple prosodic cues to phrase boundaries within any given language tend to be correlated with one another. (Morgan, Meier, and Newport 1987:502f.)

Even though these prosodic characteristics are language-specific abstractions, they are apparently salient enough to be learned by simple observation within the first year. There have been a number of demonstrations that infants are sensi-

tive to such sub-utterance phrasing, for example:

- Jusczyk, Hirsh-Pasek et al. (1992) showed that infants of 9 months, but not 6 months, listened longer to speech with pauses at major syntactic boundaries than speech with pauses at non-boundary positions. Trials with low-pass filtering got the same results, suggesting that the effect is due to prosody.
- Hayashi et al. (1996) extended these results to Japanese-learning infants, who were sensitive to acoustic cues of clausal units at 10 months but not at 6 months.
- Gerken, Jusczyk & Mandel (1994) showed that 9-month-old infants could discriminate prosodic phrases even when they were not also syntactic phrases. Since boundaries of phrases at either linguistic level will also lie at word boundaries, for our purposes their recognition of either prosodic or syntactic phrases, or both, will supply more short “utterances” for further processing.

These studies show that before the first lexicon is achieved, infants have discovered how to use prosody to break up the speech signal into more manageable chunks, and that these subutterance chunks do not conflict with boundaries of meaning-based units.

1.2.3.2 Prosody and clustering

Aspects of prosody have also been suggested as helpful to a “bottom up” or clustering strategy. With a small amount of working memory, infants are trying to build mental representations of the speech they hear, so that they can compare one occurrence with another and determine whether they are “the same.” As described in Miller’s classic “The magical number seven” (1956), we can envision

mental representations as formed in incremental chunks, starting small and gradually incorporating more material as new input is received and digested. So we ask, where will the child begin? What small pieces of language will be the most memorable, the most salient in the input, and serve as the nucleus of a larger construct?

We saw above that one candidate for salience are the edges of longer strings which have been prosodically delineated. Other aspects of prosody may more directly highlight some pieces of language over others, depending on the particular language. Several studies found cross-linguistic evidence that prosodic salience (termed “stress,” although the specific components may vary cross-linguistically) significantly influences which parts of the speech signal are learned first:

- Vihman (1981) studied speech errors of both children and adults, finding that for both the most salient features of words were the number of syllables and the location of stress. There were no significant differences in these respects across child data from English, French, and Estonian, despite somewhat different uses of stress in the three languages.
- Pye (1983) showed that stress is an important determinant of which part of a complex word children at the one-word stage will produce. In Quiché, verbs are morphologically complex forms which resyllabify and change stress configuration depending on their place in the sentence. Children’s use of these forms reflects the prosodic aspects of the verb rather than the semantic ones. They consistently select the most perceptually salient syllable, as defined by stress, syllabicity, and utterance-final position, which is sometimes the syllable containing the verb stem and sometimes the inflection.

- Gleitman & Wanner (1982), reinterpreting a study of 2- to 4-year-olds in English, Turkish, Serbo-Croatian, and Italian by Slobin (1982), claimed that the influence of stress would account for much of Slobin's data. Their analysis of the data led them to hypothesize that the child begins by extracting particularly salient syllables from the speech stream and storing them, ignoring the surrounding unanalyzed noise. These stored syllables are equated with the early "word," they proposed. Gleitman et al. (1988) cited additional evidence from several languages for early learning of stressed elements at the expense of unstressed ones, even when they are not the most important for semantics or syntax.

- Echols & Newport (1992) looked at rates of syllable omission and segmental error by children in the one-word stage (ages 1;5 to 1;11). They found that, in multi-syllable words uttered spontaneously, stressed syllables were retained and pronounced correctly significantly more often than unstressed syllables, and nonfinal unstressed syllables had the greatest rates of omission and error.

In summary, there is a variety of evidence that prosodic cues act to focus attention on small pieces of language for early learning. This is the "plucking" mechanism we alluded to earlier, and we conclude that prosodic salience is an important influence on early word location. However, because stress can vary considerably from one instance of an utterance to another and because a large number of elements receive stress, we argue that stress could not be the sole determinant of words but will work in concert with other cues.

1.2.3.3 The “trochaic bias” in English

An influential proposal regarding word segmentation in English deserves to be reviewed here, even though it is language-specific, since it has had a considerable effect on the literature. As a second motivation, the experiments conducted to test the proposal incidentally offer further evidence for the degree to which infants are sensitive to stress cues.

Using new techniques, assessments of infants’ perception of stress (sometimes also called “rhythm” in this context) have become possible. Perhaps inspired by a model of adult segmentation from Cutler and associates⁵, several studies focused on a possible “trochaic bias”—favoring trochaic, or strong–weak patterns (S-W) as opposed to iambic (W-S) in English-learning infants. These studies, all on English-learning infants, show a clear developmental trend:

- Jusczyk, Cutler & Redanz (1993) tested whether infants of 6 and 9 months could discriminate different stress patterns. Stimuli were novel isolated words of two syllables, with strong–weak stress in one condition and weak–strong in the other. The 9-month-olds were found to prefer (look longer in the direction of) the strong–weak words, while the 6-month-olds showed no preference. A trial with low-pass filtering for the 9-month-olds continued to show the same prefer-

⁵ Cutler & Norris (1988) and Cutler (1994) proposed that English-speaking adults use a stress-based heuristic that “words” (candidates for lexical access) begin with stressed syllables. A detailed explanation and quantification of this model (Cutler & Carter 1987) is quite convincing, showing that, for a short sample passage of text, 80% of such cues will be fruitful. In a larger corpus, they estimate that a stress-related strategy would “successfully locate the initial boundaries of 90% of the lexical items” (p. 138). However, their algorithm critically depends upon having a lexicon. It is a “preprocessor” which supplies candidates that are then accepted or rejected based upon the results of lexical access. For the infant with no lexicon, a “rule” that each strong syllable begins a word would not be productive, and no one has proposed this.

ence, confirming the hypothesis that the effect was due to the different stress patterns.

- However, a study by Crowhurst and Echols reported in Echols (1996) used longer stimuli of 5 or 6 syllables and obtained results that differ from those of Jusczyk, Cutler & Redanz (1993) in a puzzling way. As predicted by the earlier study, the 9-month-olds in the Crowhurst & Echols experiment were sensitive to penultimate vs. final-syllable stress at the end of the phrase-like sequence, but they preferred (looked longer at) the final-stress version. Moreover, differences in rhythm elsewhere in the sequence did not cause significant differences in response. This echoes our discussion earlier of the salience of material at utterance boundaries.

- Echols, Crowhurst & Childers (1997) habituated infants to a weak–strong–weak three-syllable sequence and then tested them with two versions of an inserted pause: W#SW and WS#W. The 9-month-old infants found the W#SW version more familiar, while the 7-month-olds showed no preference. In other words, by 9 months of age, there has come to be a tendency to group a stressed syllable with the following unstressed one as a mental unit—a tendency “to preserve the coherence of the trochee” (p. 202).

- Morgan and colleagues (Goodsitt, Morgan & Kuhl 1993; Morgan 1994; Morgan & Saffran 1995; Morgan 1996a) used several techniques to investigate what kinds of mental chunks infants were constructing, finding evidence that trochaic sequences were more “cohesive” than iambic ones, and that this strategy became more influential as infants matured over the first six to nine months of life. (These experiments also involved issues of distribution and frequency, so

they will be discussed in more detail in Chapter 5.)

This evidence indicates that, at least for a time, the English-learning infant is influenced by the tendency of English to favor trochaic units. Between 9 and 11 months, the baby may find that the easiest words to locate in the speech stream are those that have two syllables with strong–weak stress patterns. By the end of the first year, however, these strategies are getting more help from distributional information, as we shall in Section 1.2.5 below. Again, we emphasize that even in the “prime” period of its influence, stress is undoubtedly combined with other sources of information.

Of course, since the usefulness of trochaic stress is not universal across all languages, we must ask how English-learning children first discover it. One possibility is that many longer English words have penultimate stress, the end of which is a trochee (Echols 1996), and thus infants observe them frequently at the end of utterances and other bracketed groupings. Another hypothesis comes from Jusczyk (1997:108): Most babies’ names or nicknames in English are two-syllable strong–weak words (*Jimmy, Donna, Baby, Binky*), as well as many names in the environment (*Mommy, Daddy, doggy, kitty*). The frequent repetition of these names in isolation could help account for an early habituation to the trochaic form. According to Mandel, Jusczyk & Pisoni (1995), at 4.5 months babies can distinguish their own names, said in isolation, from other names which are similar in stress pattern (e.g., *Joshua* vs. *Agatha, Brandon* vs. *Kevin*) and prefer them as shown by significantly longer looking times.

To sum up this discussion of prosody, we have said that prosody is likely to have a significant effect on bracketing, creating smaller chunks for further pro-

cessing. Prosodic characteristics are also implicated as making certain small pieces of language more salient for early learning. Thirdly, we can take the case of the English “trochaic bias” as perhaps a representative case of the way a child can make use of information about the most frequent patterns of the ambient language. Each of these sources of information are candidates for the child’s “bag of tricks” in early word discovery.

1.2.4 Phonetic cues and phonotactic constraints

Allophones (systematic differences in pronunciation of segments) and phonotactic constraints⁶ are two more sources of information that might help differentiate word-internal positions from word beginnings or ends. While these two sets of possible cues are not the same, they are certainly related. The acoustic signals are the “-etic” or directly perceivable stimuli, while the phonotactics are the more abstract “-emic” derivatives of these physical differences. Since infants cannot have much phonological knowledge yet, the two levels may be fairly close together in early development. Before assessing how useful these cues might be, we sample the considerable evidence that infants in the first year are sensitive to both phonetic detail and phonotactic regularities.

- With respect to the infant’s general ability to discriminate speech sounds, an enormous amount of research in the past fifteen years on infant perception of speech has yielded the following results (from Werker & Pegg 1992:285f., but many other reviews exist):

⁶ The term “phonotactics” refers to the structural characteristics of phoneme sequences, “essentially a description of the distribution of phonemes, once they have been identified” (Hill 1958:68). Hill also calls it the “clustering habits” of phonemes, including suprasegmentals like juncture and tone as well as vowels and consonants.

1. Young infants can discriminate nearly every phonetic contrast on which they have been tested—including those that do not occur in their language-learning environment.
2. There is a profound change across age in the ease with which people discriminate some non-native contrasts.
3. This developmental change is evident within the first year of life.
4. The change involves a reorganization in perceptual biases rather than a loss of initial auditory capabilities.

Thus, the infant loses fine phonetic discrimination and gains a more abstract and experience-dependent ability to recognize those sound classes which are significant in the ambient language(s).

- Church (1983) argued forcefully that allophonic variation should not be considered just interfering “noise,” but as carrying information which can be useful. He showed that there are small but regular phonetic distinctions between phonemes in different positions; for example, /t/ is aspirated word-initially but not word-finally. Brown (1973) used as an example of such information the difference in pronunciation between *night rate* and *nitrate*.⁷

- Looking for evidence that acoustic markers exist at word boundaries, Hohne & Jusczyk (1994) showed that 2-month-old infants could tell the difference between *night rate* and *nitrate*.

- Christophe, Dupoux, Bertoncini & Mehler (1994) found that three-day-old infants discriminated between tokens of two syllables spliced from the middles of long French words and the (phonemically) same two syllables spoken in a context where they span a word boundary. They concluded that word boundaries

⁷ Elman (1986) claims that people don't ignore irrelevant acoustic variation, but rather assign it to separate spheres of meaning (lexical meaning, pragmatic meaning, speaker identity or underlying intent, etc.).

are acoustically marked, and that infants can perceive these markings. They noted, however, that there is a possible confound with the word-final stress typical of French.

- Friederici and colleagues (Friederici & Wessels 1993; Jusczyk, Friederici et al. 1993) showed that infants of 9 months, but not those 6 months or younger, preferred isolated words which were phonotactically legal in their language to those containing illegal sequences. To rule out the possibility that the experimental effect was due to prosody, they repeated the trials using low-pass-filtered stimuli and found no difference between legal and illegal stimuli. The authors pointed out that the data could be interpreted as “a sensitivity to the frequency of occurrence of certain syllables in certain positions” (p. 294).

- Jusczyk, Luce & Charles-Luce (1994) made a similar demonstration for forms with phoneme transitions of high probability and those where phoneme transitions are improbable, showing that “infants are very sensitive to the frequency with which certain patterns appear in the input” (Jusczyk 1997:159).

Reviewing this evidence, we argue against more than a minimal role for each of these information sources in bootstrapping of the first lexicon. First, the allophonic information has only been shown to be discriminable to very young infants. We would want to see these results replicated with older infants, since the interpretation crucially depends upon assuming the ability to make fine phonetic discriminations, an ability which declines during the first year, as pointed out above.

Second, it has not been demonstrated that these allophonic cues would be effective for word discovery. How many word types have reliable cues of this

sort, and to what extent are the cues dependent on the specific phonetic context?

We have only informed opinion, not data, to guide us:

Although recent work suggests that there might be more markings for word boundaries than so far estimated... it is certain that segmentation cues in the signal are not extremely numerous and not unequivocal. (Flores d'Arcais 1988:100)

There is evidence that these cues are correlated with syllables and prosodic feet rather than words:

Perhaps many of the so-called "word boundary phenomena" are really "syllable boundary phenomena." (Church 1983:21)

For example, "...an alternation in aspiration co-occurs with the change in stress in word pairs like *atomic* and *atom*" (Myers et al. 1996:5).

As for phonotactic information, the case is considerably stronger. Clearly this is information which children will eventually master, and the observation of the sounds which lie at utterance boundaries (or at other bracketed sections) is very informative about those at word boundaries, as we will show in a later chapter. However, the evidence shows only that the 9- to 12-month-old infant has a relative familiarity with phonotactic regularities, better described as "preferences" than as "rules." Although linguists may talk about "phonotactic constraints," infants know only frequency of occurrence—either they hear something a lot, or a little, or not at all. It would be a very foolish baby who would decide that some combination of sounds is "not a possible word" and reject it!

We conclude that neither allophonic nor phonotactic information can be seen as primary in the first lexicon, principally because the developmental arc is too early in the first case and too late in the second to make their use plausible.

1.2.5 Distributional cues

Finally, we come to the information source which is the subject of this dissertation, which has been much less widely treated than any of the factors discussed above. One does occasionally encounter in the child language literature casual references to “distributional” information, such as:

...the child can then build up from that syllable, perhaps with the assistance of other perceptual–attentional biases, or of distributional cues, to achieve a unit corresponding to the adult word. (Echols & Newport 1992:190)

But it is rare to find more than the bare mention of the process.

The term **distribution** generally refers to the relative frequency of occurrence of each member of some set of elements or events, or more often their cooccurrences, either as a contiguous sequence or within some local range. Information about distribution thus presupposes (a) perception of the elements or events in question and (b) their relative sequential order, plus (c) recording the frequencies (or at least the relative frequencies) of both. We have reviewed above infants’ perception of speech events, and in Chapter 5 we will discuss in detail their propensity to notice and record frequencies. In this section we look at evidence that information about the sequence and distribution of speech sounds are used in early language processing.

The logical problem of language learning suggests that order information will be important. Speech and language involve a very precise temporal ordering of events. In the case of speech, the events are acoustic; in the case of language, the events are categorical objects at the several levels of linguistic analysis.

Every language is manifested in the linear form of utterances which represent what is sometimes called the ‘spoken chain’. This linear form of

spoken language derives in the last resort from its vocal character: vocal utterances are necessarily produced in time and are necessarily perceived by the ear in succession. (Martinet 1960:25)

The many dimensions of structure that we find in language are all founded upon linear order, on the single dimension of time. So the language-learning child needs to be skilled at noticing and reproducing this dimension.

A dramatic demonstration of the sensitivity of infants to serial order is Mandel, Kemler Nelson & Jusczyk (1996). They habituated 2-month-old infants to a short four-word sentence recorded by a speaker with normal sentence intonation ("Cats jump wood benches."), while a different group of infants heard the same four words recorded as two separate two-word sentences: *Cats jump wood benches.* vs. *Cats jump. Wood benches.* where the speaker treated the two-word sentence as the answer to a question, like "Do cats jump or run? Cats jump." After a two-minute delay, both groups heard slightly different versions of the stimuli, but with the second and third words exchanged (*Cats would jump benches.* vs. *Cats would. Jump benches.*). The group hearing the four-word sentence noticed the shift in word order (as indicated by a change in sucking rate using the high-amplitude sucking procedure), but the group hearing two short sentences did not.

Thus, even at two months of age, infants are able to remember the order of spoken words when they are embedded within the coherent prosodic structure of a single well-formed sentence. (Mandel et al. 1996:181)

Jusczyk and his colleagues have actively studied the ability of infants to locate words, in particular the influence of stress and of distributional information, as the following review summarizes:

- Jusczyk & Aslin (1995) showed that babies who had been familiarized

with two monosyllabic words (*feet, bike* for one experimental group and *cup, dog* for another) would then tend to listen longer to a short passage which contained these words in sentences than to similar passages with the alternate words, on which they had not been trained. The same recognition effect was found if the infants were first familiarized with the passages, and then tested on the words. However, there was no significant difference in the same task when the infants were trained and tested using two sets of monosyllabic nonwords which rhymed with those used in the previous experiment: *zeet, gike* or *tup, bawg*. These effects were all found with 7.5-month-old infants, but not with those 6 months old.

- [A]⁸ A continuation of this inquiry by Jusczyk and colleagues reported in Jusczyk (1997) switched to two-syllable words (*doctor, hamlet*), with similar results on 7.5-month-old infants. The babies trained on isolated words seemed to “recognize” the words in the passage, and babies trained on the passages recognized the words. But the effect disappeared when the training used only the first stressed syllable (*dock, ham*) and the test used the two-syllable trochaic word (*doctor, hamlet*), or vice versa. This showed that the babies were not just recognizing the stressed syllable contained in the word, but were matching up both syllables with the text.

- [B] Next, iambic words were tried (*guitar, surprise*), but 7.5-month-old babies did not listen longer to passages containing the two-syllable words after familiarization with the isolated versions. When they were familiarized with *tar* and *prize*, however, they listened longer to a passage containing the whole word.

⁸ In order to facilitate reference to a series of experiments reported but not otherwise identified in Jusczyk (1997), we assign a letter identifier to each experiment.

"It was as if the infants perceived the 'tar' from 'guitar' as initiating a new word when it occurred in a fluent speech context" (Jusczyk 1997:102).

- [C] The next experiment used a passage in which the word *guitar* was always followed by *is*, inducing 7.5-month-old babies to construct the pseudo-word *taris*, and indeed the infants listened longer to the string *taris* presented as an isolated word during the test phase, while treating the isolated syllable *tar* as unfamiliar.

Jusczyk and his colleagues hypothesized that the distributional properties of the sentential contexts may have been a key factor. In particular, whenever a strong/ weak word appears in a sentential context, the strong syllable for the word is always followed by the same weak syllable (namely, the one that belongs to the word). However, this is not necessarily true for the strong syllable of a weak/strong word. Thus, the "tar" of "guitar" might be followed by "is" on one occasion, by "has" on another, by a sentence boundary on another, and so on. These differences across the various contexts in which the word appears may help to signal a word boundary at end of "guitar." (Jusczyk 1997:102)

- [D] When 10.5-month-olds were tested on *guitar* and *surprise*, they had no trouble recognizing them in the contextual passage. Further, when the older infants were familiarized with *tar* and *prize*, they did not listen significantly longer to the passage with *guitar* and *surprise*, and they were not fooled by the context into looking for *taris*. These results suggested that the stress-based strategy is a temporary stage, or at least that other strategies develop which come to modify its effects.

Morgan & Saffran (1995) also investigated the relative influences of prosody and distribution on infants' speech segmentation. Following directly on the

work of Goodsitt et al. (1993) and Morgan (1994),⁹ they varied two characteristics of a language-like stream of speech syllables—vowel length and sequential order. Vowel lengthening is one component of stress in English, while changes in sequential order produce different distributional characteristics. Morgan & Saffran (1995) studied the performance of 6- and 9-month-old infants using two different experimental techniques and showed that at 6 months, the prosodic characteristics are the most influential in the infants' discrimination of the syllables, but that the 9-month-old infants give equal weight to distributional information.

We can summarize the implications for our hypothesis of these experiments as suggesting a shifting mix of cues as the infant matures: At 6 months, prosody is most important, then by 7.5 months stress and distribution information combine, and this continues in 9-month-old infants. By 10.5 months allophonic cues—such as the difference between *night rate* and *nitrate* appear able to override distributional information. If so, this would indicate that bootstrapping proper is largely complete by 10.5 months, and the infant at that age already has enough information about word boundaries to have begun learning their phonotactic and allophonic correlates.

1.2.6 Summary of segmentation hypotheses

We have seen that there are several characteristics of speech which might plausibly be part of the child's segmentation strategy:

⁹ Since Morgan & Saffran (1995) replicated the results of Morgan (1994), I do not discuss Morgan (1994) separately here. Several other experiments which directly test the ability of adults and children to perform segmentation, including Goodsitt et al. (1993) and the later work of Saffran, will be discussed in some detail in Chapter 5.

- Prosodic cues, including actual pauses in the speech flow, provide initial segmentation of utterances (“bracketing”). The beginnings and ends (ends perhaps more than beginnings) of sections so identified are particularly salient as potential words, and can also be used to collect phonotactic and distributional information about word boundaries.
- Syllables or syllable-like units which are given emphasis by language-specific systems of accent or stress may, if the emphasis is sufficiently consistent across utterances, be memorized by the child and become the nucleus of a meaningful element (“clustering”).
- Distributional cues seem certain to be part of the child’s strategy.

Bootstrapping segmentation, like other major language-learning tasks, is accomplished using multiple cues, and the relative contribution of each cue and its timing will vary from one individual child to another. Perhaps it is this dynamic aspect of the process—its redundancy and indeterminacy—that makes it a virtually fail-safe, “highly buffered mechanism” (Lieberman 1991) so that children in widely different situations are able to learn language.

1.3 Modeling the input: Theory and data

A computational model of infant segmentation must have inputs and outputs, and the closer they are to the babies’ inputs and outputs, the more plausible is the simulation. In general, we know that these inputs are the speech sounds that they hear, but how exactly do babies “hear” speech? Are early mental representations composed of segments, or syllables, or some other unit?

It is clear that the infant has somehow to transform the acoustic speech signal from a “continuous and... temporally overlapping” stream (Strange 1995:7) to linguistic categories which are abstract, discrete and interrelated (Brown 1958).

The ability to perceive a physically varying and multi-dimensional speech signal as a single abstract object, or category, has been called “perceptual constancy” (Kuhl 1980), and is one that machines cannot imitate, or not so far. Despite a great deal of research on artificial speech recognition, no instrumental technique can successfully decode normal continuous speech (Wode 1992). All speech recognition systems to date depend upon restricted input, such as pauses between words, or a limited vocabulary.

These problems have their root in the ubiquitous *variation* which characterizes normal speech. The same phrase pronounced by one and the same speaker at different times has different acoustic properties. And the same phrase pronounced by different speakers, even if they speak the same dialect, differ widely. The situation gets even more complicated when we take into consideration dialect variation and nonlinguistic factors such as noise. (Halvorsen 1988:214)

Children somehow learn not only that there is a language system to be discovered, but that the first step depends upon converting this stream of sound into a sequence of recurring units drawn from a limited set. The child’s first such set may be different from the adult’s, but by the end of the first year some process of chunking and classifying has taken place, as we saw in Section 2.2.4 above.

What kinds of chunks and classes have they discovered? What is the child’s “segmentation unit, or prelexical atom” (Mehler & Christophe 1993:19)? For our modeling task, the choice of representational units may constrain the segmentation possibilities. For instance, if we knew that children represent speech as a sequence of syllables, then we conclude that segmentation must come between syllables. On the other hand, if segments are the minimal unit, then the possible

points of segmentation are more numerous.

As Vihman (1996:67) says, “Persuasive evidence favoring one [basic unit] over the other has proven elusive.” Data are hard to interpret as evidence about what happens inside the infant mind, and reasonable opinions may sharply disagree. A given set of results can always be analyzed as the effect of quite different mental representations, combined with appropriate mental processes.

...[I]t will always be possible for a theory subscribing to one view of representations to mimic the predictions of a proposal adopting a different perspective, so long as the appropriate assumptions about such processes are made... Accordingly, only representation–process pairs are tested by behavioural data and any number of such pairs will be consistent with any set of such data. (Atkinson 1985:292)

Because none of the evidence is conclusive (and it is hard to say what kind of evidence *could* be conclusive), discussion usually takes place in the realm of competing theories and how well they explain the available data, both experimental and observational.

First we will briefly review the various theories, and then the experimental evidence, both of which are inconclusive but will give us a range of possibilities to consider.

1.3.1 Phonological theories

The dominance of **segments** (phones and phonemes) in phonological theory usually makes them the default choice for the infant’s mental “atom” in the absence of strong evidence to the contrary.

Infants can and do perform an initial segmentation of the wave form, to derive phonetic segments. (Gleitman et al. 1988:154)

More than just a default, segmental representations are actively preferred over

“undifferentiated syllables” by Kuhl (1986) and Stemberger (1992), among others:

There are no compelling arguments *against* segmental representations from the beginning [of language acquisition]... and evidence *for* syllable units cannot be construed as evidence *against* the additional presence of segments. I assume that the child’s representations are identical to those of adults (barring occasional misperceptions). (Stemberger 1992:171f.)

The second most popular choice is the **syllable**. Babbling and early utterances largely consist of CV groups (sequence of one consonant and one vowel), which are called “canonical syllables,” defined as “a sound sequence including a nucleus (a highpoint of energy) and at least one margin (an adjacent lowpoint of energy)” (Oller & Lynch 1992:514). Of course, a CV group is only one version of the adult syllable, which can range from a single vowel to the “bulging monosyllable” of Southeast Asian languages, with not only consonant clusters and vowel diphthongs, but suprasegmental features such as lexical tone, glottalization, and nasalization (Matisoff 1989). The fact that a lone vowel can be called both a segment and a syllable increases the classification problem.

At a finer level of discrimination, acoustic **features** have been found an appealing way to encode characteristics of sound that the baby might be noticing. Ingram (1992:434f.) saw the use of distinctive features in the apparently systematic patterns of substitution in “babbling sequences such as [bi], [ba], [bu], rather than [bi], [mu], [ta].” However, Macken (1979) argued against them and in favor of phonemes:

...[T]he ‘phoneme’ is a basic building block of the acquisition process... evidence for the feature is considerably weaker... [T]here is little evidence that once having mastered particular features, she could generalize this knowledge to the acquisition of a new phoneme. For example, the ‘acqui-

tion' of the features voicelessness and frication ($|f, s|$) and velarity ($|k|$) were not sufficient for the acquisition of $/x/$.¹⁰ (Macken 1979:44f)

Although features are traditionally associated with segments, they have also been used to describe non-concatenative phonologies, where features are organized around a larger unit. Waterson (1971) was one of the first to treat infant language as a system of its own, organized differently from the adult system. She presented a persuasive hypothesis that the child incorporated into its representations salient phonetic features of the whole percept, but not necessarily in the same "sequence and combinations." The following example, though lengthy, gives a good sense of the kind of evidence offered by data-driven theory creation:

[Hildegard's] form $[lɔɪf]$ for 'story' puzzled Leopold so much that he doubted the interpretation... The adult form $[stɔːrɪ]$ has the features (checked) sibilance and friction, $[st]$, liquid and continuance, $[r]$, more open vowel followed by more close, $[ɔː]$ and $[ɪ]$, and rounding and backness in the first syllable and non-rounding and frontness in the second, with voiceless onset and voiced ending of the word... The child's form has sibilance and friction, $[f]$, liquid and continuance, $[l]$, more open vowel followed by more close, $[ɔː]$ and $[ɪ]$, rounding and backness in the first syllable and non-rounding and frontness in the second; all these features are shared with the adult form, but the child's form has voiced continuant onset and voiceless sibilant ending, which is a different sequence from the adult's but is required by the child's system... viz. continuant initial and sibilant final. This form is thus considered to be completely regular. (Waterson 1971:205f.)

Another hypothesis is that for the year-old child the "**word**" or meaning unit is associated with a chunk of sound which has no internal structure. It is the whole pattern, the combined effect of its various characteristics, that is "recognized" in its entirety by the child. In trying to fully understand this idea, we can compare it to musically untrained adults who can recognize a tune and repro-

¹⁰ Here, $|...|$ refers to the child's phonemes, while $/.../$ is used for phonemes of adult Spanish (Macken 1979:13).

duce it even when they cannot perceive any of its components individually (Bever 1975).

The most influential proposal for words as minimal units is Ferguson & Farwell (1975), who took as primary data the very short utterances of a (roughly) one-year-old child. As sound units, these are generally CV, CVC, or CVCV units, often with vowel and/or consonant harmony. The authors showed that individual segments were not treated consistently across different lexical items, and even syllables that were the same in several adult words were not necessarily the same in the child versions of those words.

Some researchers resolve the conflicting demands of complex data by combining two or more representations, like the word-based features suggested above by Waterson (1971). Another solution is the “two-lexicon” hypothesis we saw in Section 1.1.3 above, or a proposal that the infant moves through several representational systems as development proceeds. One such version of developmental progression suggests that segments come to replace more holistic representations in response to the early growth of the vocabulary:

Early vocabulary *growth* may necessitate a restructuring of the lexicon... the need to discriminate a growing number of lexical items quickly and accurately should... serve as a major impetus for the *emergence* of more fine-grained, sublexical units of representation, including the phoneme. (Walley 1993:291)

According to this reasoning, the developmental shift to a more analytic representation would take place after vocabulary building is well underway, beyond the narrow window of development we are interested in here.

1.3.2 Experimental data

It is difficult to present experimental results in this area, since there is no consensus on their interpretation. We will describe each experiment from the point of view of its authors, and the reader should not assume agreement with that interpretation by either the scholarly community nor by the present writer. Most of these intend to contrast the possibility of segmental vs. syllabic mental representations. First, one which argued for segmental representations (though it could have implicated features instead):

- The results of Jusczyk & Aslin (1995), discussed above, in which 7.5-month-old infants responded quite differently to stimuli like *cup* and *tup* that differed by one segment, have been said to show that babies' mental representations are segmental.

Then, several experiments interpreted as being for syllables and against segments:

- Bertoncini & Mehler (1981) hoped to demonstrate that, almost from birth, humans combine consonants and vowels to form syllable-like mental representations. Their experiment showed that infants less than 2 months old could discriminate slightly different CVC units, and even VCCCV units, but ignored changes in CCC groups. They interpret these results to mean that vowel-less sequences are not perceived as speech by the babies, and thus do not result in mental representations that would permit discrimination.
- Jusczyk & Derrah (1987) presented 2-month-old infants with an assortment of /bV/ syllables (/bi bo ba/) and then added a new syllable (/bu/ or /du/ or /da/) to the mix, with a control condition of no change. They found

that the babies noticed the change—there was a difference between the new/no-change conditions—but there was no significant difference between hearing a new syllable which reused a familiar consonant and/or vowel and one which had no segment in common with previous stimuli. The authors say (p. 652): “The data provide no support for the view that the infants were representing the syllables as [b] plus a particular vowel.”

- Werker & Pegg (1992) became discouraged trying to get infants to make minimal-pair distinctions between a known word and an unfamiliar word, like *dog* and *bog*. This was successful with 19-month-old babies, but not with younger ones, despite the fact that the younger babies could successfully distinguish between two actual words that they already knew. The authors take this as evidence against segmental representations.

- In work by Jusczyk, Jusczyk, Kennedy, Schomberg & Koenig (1995), 2-month-old infants were habituated to two-syllable trochaic (S-W) stimuli which contained either common syllables: [bazi], [balo], [bamIt]; or common segments: [zabi], [labo], [mabIt]; or neither: [pæzi], [nelo], [komIt]. Only the infants exposed to the common syllables could detect a change in the stimulus sets after a two-minute delay period.

And one which is against segments, but argues for a mental unit larger than the syllable:

- Hallé & Boysson-Bardies (1996) presented as stimuli to French 11-month-olds isolated two-syllable words of three different types: familiar words, rare (unfamiliar) words, and familiar words with an altered segment (as *bonjour* to *ponjour*). In general, they found that the infants preferred the familiar to the rare

(replicating their 1994 work), but in most cases showed no preference between a familiar word and its altered variant. The authors say (p. 458): “Infants recognize words from rather global representations, ignoring certain phonemic variations they are, in theory, sensitive to (as we know from perceptual experiments using meaningless syllables).”

1.3.3 Implications for a segmentation model

Faced with a smorgasboard of choices for the infantile mental unit, we will turn the question around. Instead of asking what form of input is best, we will test various inputs and ask which one produces the best results, thus making the issue of representational units one of the outcomes of this research rather than one of its starting assumptions. A clear advantage for one unit over another in the critical task of segmentation would be an important piece of evidence. We will see, however, that performance in our experiments does not depend upon the unit of representation.

Of the units we have discussed, we will not further consider words or syllables. It is not practical to use words as an input if we wish to evaluate segmentation of words as an output. The problem with syllables as an input is that they offer either too much information or too much disinformation. The syllable is not a much more rigorously defined concept than the word (Ladefoged 1982), and the relationship of syllables and words is very language-dependent. In Spanish, for instance, it is often not possible to predict syllable boundaries without prior knowledge of word boundaries; in the following examples, sequences are underlined to show segment similarities: *de la casa* vs. *del acaba*; *se*

levanta vs. *grandes elefantes*. In English, resyllabification across word boundaries occurs in normal speech and there is a good deal of ambisyllabicity, so that words and syllables form two distinct and largely unrelated structural systems.

...it is not always easy to determine whether such processes are to be attributed to the syllable or to the word... syllables and words do not always line up (e.g. English can have a *syllable* beginning with [ʒ], e.g. *he sees you* [hiy.siy.ʒə])... (Hyman 1978:447)

Thus, even hand coding a language sample of English into syllables would be very difficult; some scheme of overlapping constituency like Wickelphones might be necessary to deal with ambisyllables. And even if it were practical, syllabic input would either unrealistically overdetermine the segmentation problem, by building in some word boundary information as in the case of Spanish, or make segmentation impossible in some cases by creating indivisible units which straddle word boundaries.

Of the remaining possibilities, in Chapter 2 we will review several connectionist experiments in segmentation which have used segmental features as a way to allow for greater (or perhaps just more efficient) generalization between segments, and we will discuss some implications of that choice in Chapter 6. In Chapter 4, we will compare the results of segmentation experiments using the O-B algorithm on corpora using various representations:

- Alphabetic orthography, which approaches a phonemic representation (more closely in some situations than others)
- Phonetic and phonemic segmental representation schemes
- Japanese hiragana, a moraic representation

Chapter 2. Computational Models and Metrics

The problem of infant segmentation of the speech stream has been addressed directly only very recently. In Chapter 1, we reviewed data and hypotheses about the problem from linguistics, speech technology, and psychology. In this dissertation, we claim that an important source of information used for segmentation is the relative frequencies and distribution of speech units. To support this hypothesis, Chapters 3 and 4 will present computer simulations to show that such information exists in language to a degree that would be helpful in this task. The computer simulations use a computer algorithm, computer-readable corpora, and information-theoretic metrics in the tradition of computational linguistics. In this chapter, we review some key assumptions of this tradition, and describe previous computational research in speech segmentation and related areas.

2.1 Models and metrics in computational linguistics

The use of computers in linguistics has from the beginning taken two main paths. One branch is toward cognitive models, intended to help understand the ways language is processed by the human mind. The other branch, “natural language processing” (NLP) is an engineering approach in which “models” are intended to imitate only the results of human language processing, by whatever methods work. Neither of these areas has had very much to say about segmentation of language into word-like units, in general considering the “word” as the atom upon which the system is built.

Natural language processing has been dominated for thirty years by approaches from artificial intelligence (AI), which emphasize linguistic rules and world knowledge as the basis for decisions and discriminations (Charniak 1993). AI simulations provide a logical sequence of processes which are designed by human ingenuity after close consideration of the problem at hand. Thus, rule-based programs have a lengthy development time and, because they are tailored to a more or less particular situation, they generally have to be altered when some aspect of the situation changes, or when our understanding of it changes.

The “knowledge-based” approach to segmentation uses a lexicon of known words and morphemes. Most NLP systems do this, including most speech-recognition systems (e.g. Briscoe 1989). This technique, however, is not available to infants, who must create their own lexicon. Thus, we will be interested in techniques for self-organization of language, or for bootstrapping the discovery of linguistic units using some information source which is available to the infant.

The other two approaches to computational modeling of language are connectionist and probabilistic systems. While rule-based models are often quite abstracted from real data, they do try to come to grips with the interaction of a number of complex variables. The other two types are in one sense less ambitious, generally focusing on narrower and more specific problems. On the plus side, they rely more on standardized techniques and try to avoid hand-fitting or “heuristics.” Thus, when they work, they are much more widely applicable at less cost.

In this section we will provide some background on information theory and probabilistic modeling, and introduce the most widely used metrics in these domains.

2.1.1 Information theory, entropy and mutual information

Claude Shannon (1948) developed a “mathematical theory of communication,” which is now called Information Theory (IT). He was interested in messages and their efficient and safe transmission over a communication channel, and he formalized certain mathematical principles which he showed govern that process. These principles have been applied to problems involving communication in a more general sense and to codes defined more broadly, including natural language. In this section, we will review some concepts from this theory that will be referred to later.

Shannon said that a message is transmitted from one station to another by a code which is prearranged to designate one of a number of messages. That is, the message will not create knowledge, but will merely refer to that which is already known by both the sender and the receiver. Thus, “One if by land and two if by sea” was a code whose meaning was prearranged by both parties. The code had only two states and thus only two possible messages could be sent. By extension, a telegram may be considered a very complicated code which involves not only the alphabet, but the dictionary and the grammar. In other words, part of the “prearrangement” necessary here is that both sender and receiver have sufficient knowledge of the natural language involved.

When a transmission is received and its “meaning” is deciphered, we say colloquially that we have received “information.” Shannon gave the term a technical meaning which related the amount of **information** to the number of possible messages. For instance, the communication of one out of two possible messages involves the transfer of less information than if one of a thousand mes-

sages is designated. The amount of information increases directly with the number of possibilities or choices. To express this information quantity, Shannon chose a logarithm to the base 2, calling the resulting units “bits.” Thus, 2^N choices can be expressed as containing N bits of information ($\log_2 2^N = N$). This quantity was termed **H**, or **entropy**. Entropy is thus a measure of choice: the more choice, the higher the entropy.

When the various possibilities do not occur equally often, the calculation of entropy must take into account their distribution:¹

$$H = - \sum p(x) \log p(x)$$

A perfectly even distribution will have the maximum entropy for its codeset. Skewed distributions will have a lower entropy, since the choice is biased. It has also been said that lower entropy represents more structure and less disorder or randomness. Another meaning of entropy is the total number of messages which can be sent given the code structure and message length. This “message entropy” is usually normalized² to bits per character, so that texts of different lengths can be compared.

When we speak of messages (and texts and corpora are treated as simply long messages), we use the term **n-gram** to refer to combinations of codes. The most primitive symbols are termed “unigrams,” a combination of two unigrams

¹ In this and later formulas, p refers to the probability of an event, generally estimated by its frequency in a particular trial. Since probability is assumed to be $0.0 < p < 1.0$, its logarithm will be a negative number which, for convenience, is converted to a positive number H . Logarithms throughout are to the base 2.

² Here and throughout, the term “normalize” is used to mean a standardization of several measures with respect to a common base; it does not refer to the Gaussian, or normal, distribution.

is a “bigram,” and so forth. The generic term is “n-gram.” Note that a “symbol” or primitive unit is in some contexts a character, but in other contexts it is a word or even a phrase. Another necessary distinction is “types” vs. “tokens,” where a **type** is a unique configuration, and a **token** is an instance of a type. The sentence “to be or not to be” contains six word tokens and four word types; looking at word bigrams, there are five tokens and four types.

A further standardization is **relative entropy** and its derivative, **redundancy**. Shannon defined the concept of redundancy as the ratio of the potentiality of the code and a particular use to which it is put. Relative entropy is the entropy of a particular message or text divided by the maximum entropy for the codeset ($\log_2 N$, where N is the number of unigram codes), and redundancy is 1.0 minus relative entropy (often expressed as a percentage). For instance, a transmission of 3 instances (tokens) drawn from a codeset of 4 symbols (types) can designate one of 64 different messages ($4 \times 4 \times 4 = 4^3 = 64$). However, if the number of messages to be sent is only 24, the codeset is too rich; only 3 symbol types would be necessary for this purpose ($3 \times 3 \times 3 = 3^3 = 27$). The excess “power of expression” here is the redundancy of the code with respect to the possible messages, or 24% (relative entropy = $\log_2(24) / \log_2(64) = 0.764$; redundancy = $1.0 - 0.764 = 0.236$).

Mutual information is a related measure used to filter bigrams which we will see in some of the research cited below. Intuitively, it is the amount of information one outcome gives us about the other (Church, Gale, Hanks & Hindle 1991; Charniak 1993). If we take information of x to be $f(x) = -\log p(x)$, then the amount of information that event y gives us about event x is $f(x) - f(x|y) = -\log$

$p(x) + \log p(x|y)$ and, where $x = \text{"pancake"}$ and $y = \text{"syrup"}$,³

$$\text{MI}(x,y) = \log \frac{p(x,y)}{p(x)p(y)} = \log \frac{p(\text{"pancake syrup"})}{p(\text{"pancake"}) p(\text{"syrup"})}$$

Limitations on these mathematical rules are significant. Shannon (1948) specified that they apply to sources which are **discrete**—those with inherently categorical representations—and **ergodic**—roughly speaking, statistically homogeneous. Any sufficiently large sample of an ergodic source will be statistically identical to any other sample of the same size. Written texts of natural language may, for these purposes, be considered discrete if they are treated as limited to the letters and spaces in which they are represented (omitting, for instance, intonation and pauses, etc.). It is difficult, however, to assume that natural language is more than minimally ergodic, since even two halves of the same text, written as a coherent whole by a single author, are not interchangeable with respect to their statistical characteristics. They will usually have different words, different word frequencies, and even somewhat different letter frequencies. However, much of interest has been discovered about natural language by ignoring this limitation.

2.1.2 Probabilistic language models

Probabilistic (also called statistical or stochastic) language models are based on information theory and have recently become much more practical due to the availability of powerful computers and large text corpora (Church & Mercer 1993). Perhaps due to the stigma of behaviorism, or because the gap between

³ The complete derivation uses the probability formula $p(a,b) = p(a)p(b|a)$ or $p(b|a) = p(a,b) / p(a)$. Then, $f(x) - f(x|y) = -\log p(x) + \log p(x|y) = -\log p(x) + \log p(x,y) - \log p(y) = \log [p(x,y) / p(x)p(y)]$.

theoretical principles and statistics is too great, there has been greater resistance to probabilistic approaches than to either rule-based or connectionist ones, from the fields of both linguistics and of artificial intelligence.

Shannon (1948) showed that a language text can be approximated as a **Markov model**—a system of transitions from state to state, together with a probability for each transition. Each state is associated with the production of a unit of language, and the text is produced as a result of following a particular path through the network. The units of language produced are the primitive symbols for that model.

Instead of literally following a path, the text may be generated by randomly producing symbols in accordance with their Markov probabilities, and the primitive symbols may be expanded to strings of linguistic units like letters or words. As Shannon (1948) demonstrated, the accuracy of a language model increases with the length of the units: A word-based model is more powerful than a character-based one, and word bigrams or trigrams are more accurate still. The Markov model can thus be represented by a **distribution of n-grams**, with increasing accuracy as n increases. Creating such a model involves two tasks:

- Choosing the symbol set which best represents the data (“selecting a feature set”).

The vocabulary, or lexicon, of the model is usually letter or word n -grams, but may also include semantic features, etc. The symbol set may be determined in advance according to linguistic theory or intuition, or it may be selected from such a predefined set under control of a computer algorithm.

- Computing a probability of occurrence for each symbol or state, all sum-

ming to unity (“estimating the parameters”).

This process is always the result of a computer algorithm. It is often combined with selecting the feature set, by computing probabilities for a large number of possible features with the expectation that many of them will ultimately be assigned zero probability and thus not be included in the final lexicon which the algorithm builds.

In the usual case, a large number of possible features and parameters makes the search extremely computationally intensive, so a good deal of research effort consists in finding more efficient and effective algorithms to discover the correct combination. The next level of complexity considers a family of possible models and uses even more sophisticated algorithms to select the best one.

Generally at least part of the model-creating process involves observing a sample of text and using information gleaned from it to determine the features and/or the probabilities. This is called a “data-oriented” model, and we speak of its creation as a “training” process in which decisions are guided by some metric, or “objective function.” The final model will be the one which maximizes the objective function. It has extracted or generalized certain characteristics from the training text, so then it can serve as a decision machine, or probabilistic classifier, for other texts with similar characteristics.

A simple example is a trigram model, where the symbol set is all possible combinations of three letters (or words, or grammatical classes). The probability assigned to each trigram is generally calculated by the **Maximum Likelihood Estimate (MLE)**, or the ratio of its frequency of occurrence in the training text (number of tokens of trigram ÷ total number of tokens in text). Those trigrams

which do not appear in the training text—generally a large proportion of those possible—are assigned zero probability.

A more complex example is an algorithm like the one we will present in Chapter 3, which alternates two processes—first it calculates probabilities and then it uses the probabilities to make decisions, the results of which in turn change the probabilities. This iteration in effect creates a series of continuously improving models of the same family, until some predefined ending condition is reached, often (though not in this case) convergence according to some criterion.

Model-creating algorithms are often known by the objective function which guides them. Ristad & Thomas (1995) add features based on a **Minimum Description Length (MDL)** evaluation criterion, which looks for the shortest combined length of both the lexicon and the text encoding in terms of the lexicon. We will consider segmentation programs built along these lines in Section 2.4 below.

Compared to the amount of research which uses corpora to generate and confirm hypotheses about language, there has been less research exploring the effect of the particular text itself on these results. Few studies compare the results of an algorithm on different corpora, except to ascertain that it performs equally well on all corpora. Any differences in results are not analyzed in any depth. Also, there has not been much work on the amount of text necessary for various purposes. As we have seen, some studies use quite small amounts of data, while others use enormous corpora (100 million words and more).

Often a model is characterized by the **cross entropy**⁴ between the model and a test text (Church & Mercer 1993):

$$H(\text{test}, \text{model}) = - \sum p(x_i) \log p(x_m)$$

That is, for each symbol type in the codeset of the model, compare its probability in the text with the model's prediction. This will be a measure of the number of bits required to encode the test data in the structure specified by the model.

However, a successful computation requires that every x in the text be represented by some non-zero probability in the model. When, as is often the case, the model has failed to predict some of the observations in the text, a way must be found to reserve some of the probability space for these "unseen words." A number of schemes for this "**smoothing**" process exist, but all assume that the proportion of unseen types will be very small. In our project, there is such a wide gap between the lexicons of the models and the test texts that a cross-entropy metric turns out not to be useful.

2.1.3 Recall, precision, and F-measure

Two ratios from the field of Information Retrieval (IR) are widely used in computational linguistics. The information retrieval problem is to find objects (usually small texts such as newspaper articles) that most closely fit a particular profile (query). Assuming that the notion of "fit" can be precisely defined, the ideal is to retrieve all objects that do fit and none that don't. This is almost never possible, since the criteria are inevitably incomplete and fuzzy, but the ratios

⁴ "Cross entropy" and "relative entropy" (Cover & Thomas 1991) are also sometimes used in information theory to refer to Kullback-Leibler distance (KLD), a similar calculation, which can be confusing.

called recall and precision serve as a dual measure of the results of a particular query. Perhaps because IR as a field is very results-oriented, it has paid closer attention to problems of evaluation than has computational linguistics, including comparing the effects of various algorithms and various data bases (corpora).

Precision is the proportion of those objects retrieved that fit the query (correct retrievals $\div$ all retrievals). **Recall** is the number of retrievals that were appropriate, relative to all those that were defined as correct in the standard. These relationships are illustrated in the diagram below:

	Model = Yes	Model = No	
Standard = Yes	A Hits	C Misses	Recall = $\frac{A}{A+C}$
Standard = No	B False Alarms	D Correct Negatives	Precision = $\frac{A}{A+B}$

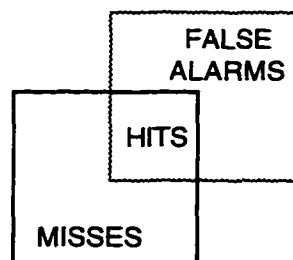
Generally, as precision improves, recall suffers, and vice versa. This is because the way to find more of the correct items (i.e. raise recall) is to make the search criteria wider, but this will also pull in more that turn out to be inappropriate (i.e. lower precision). If the search is narrowed, fewer inappropriate objects will be returned, but also fewer appropriate ones.

When the model retrieves about the same number of items as there exist correct ones (# Model Yes = # Standard Yes; or $A+C=A+B$), then recall and precision will take roughly the same value. If the model over-performs, retrieving more items than exist correct ones, then precision will be lower than recall. If it under-performs, presenting too few instances, then precision will be higher than recall.

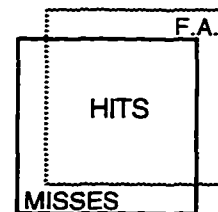
Consider the following example: If the model simply selects every object, then it will have successfully located all the desired objects, but in a rather meaningless way, since no classification has been attempted. In such a case, recall would be 100%, but precision would be the same as the proportion of correct objects in the set. The opposite strategy, that of picking only the few which are most easily classified, will lead to a very low recall, but a precision of 100%.

The following Venn diagrams⁵ show schematically the relation between recall and precision. The standard is in solid lines and the test algorithm performance (the model) in dotted lines:

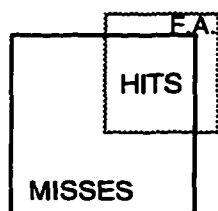
25% recall, 25% precision



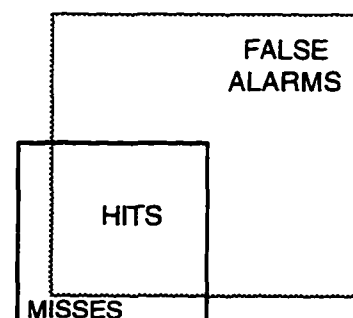
75% recall, 75% precision



25% recall, 75% precision



75% recall, 25% precision



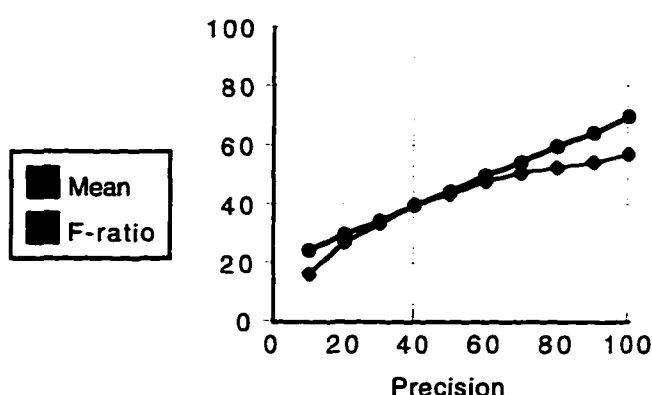
⁵ We are indebted to Brent & Cartwright (in press:18) for this way of presenting the concept. We have, however, reversed the meanings of the solid and dotted squares.

In an imperfect world, the trade-off between recall and precision is generally unavoidable. Depending on the situation, some value must be assigned to each measure and the best combination selected. Occasionally, the arithmetic mean of recall and precision is used to approximate a single measure (Sproat, Shih, Gale, & Chang 1996). Another balancing formula, and the one we will use extensively in Chapter 4, is the **F-measure** (Hatzivassiloglou 1996):

$$F = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}}$$

This formula has the effect of penalizing precision and recall values which are very dissimilar. When the two measures are close together, the F-measure is only slightly less than the mean, but it is much less than the mean when they are far apart, as shown below.

Mean vs. F-ratio for recall=40 and precision=X-axis



In applying recall and precision to the segmentation problem, we can use either divisions (cuts) or the stretches between divisions (words) as units. Divisions are a more direct measure, since each correct boundary decision is simply tallied, whereas each correct "word" requires two correct and coordinated bound-

aries. Thus, counting cuts gives credit to every correct boundary, even if the neighboring boundary is missed, and the algorithm's performance is more accurately gauged.

On the other hand, while counting cuts is a more exact measure than counting words, it is less intuitively satisfying, since quite a number of correct boundaries can be located without necessarily resulting in many correct words. For instance, if exactly one-half of the correct boundaries are found in a hypothetical text, one-half are missed, and an equal number are false-alarmed, it is logically possible that not a single word will be correctly identified despite a calculated F-measure of 50. But we will consider cuts as a source of additional information nonetheless.

2.1.4 Bracketing measures

From the subfield of computational syntactic parsing comes a group of measures known as crossing or bracketing measures. Syntactic parsing works with words as primitives, rather than letters (characters). It segments words into phrase groupings and needs to evaluate the success of these groups with respect to some standard grouping. In these respects the situation is like word segmentation, but parsing differs in two ways. First, each bracketed group of words generally is assigned a label ("non-terminal"), while we do not use labels. Second, the syntactic bracketings are hierarchical: Each group of words may become a unit in the next higher group of words. In our segmentation, we have only a single level (though some word segmentation programs do retain a hierarchy, as we will see below).

Some of the metrics developed for use with parse trees (Black et al. 1991; Goodman 1996) are useful for word segmentation also. A “bracketing” for us is simply a “word,” with no problem of dealing with nested groups. The **Bracketed Match** measure (B) counts the number of bracketings (words) for which the model version exactly matches the standard. The **Crossing Brackets** measure (C) counts not only exact matches to standard words, but also those model words which are exact groupings of standard words (like “phrases”) and exact divisions of standard words (like “morphemes”). The following hypothetical example illustrates these two counts (the ratios are discussed below):

Standard: to the lighthouse for a tea party [=7 words]

<u>Model versions</u> [=word count]:	<u>#B</u>	<u>#C</u>	<u>B ÷ Std</u>	<u>C ÷ Model</u>
i) to the light house for a tea party [=8]	6	8	0.86	1.00
ii) tothe lighthouse fora teaparty [=4]	1	4	0.14	1.00
iii) to thelight house for atea party [=6]	3	4	0.43	0.67
iv) to the ligh thouse fo rate a party [=8]	3	5	0.43	0.63

Either of these counts can be expressed as a ratio, divided by either the number of word tokens in the standard or in the model. Usually, the Bracketed Match measure is divided by the standard words; this ratio is then the same as recall based on words.

Crossing Brackets is treated a little differently by each author. Goodman (1996) uses the model word token count as a base, but de Marcken (1995:15) uses the standard word token count, and Nagata (1994) uses number of sentences. Frequently, crossing-brackets is expressed as the number of violations rather than the number of correct words. Every pair of words, one standard and one

model, that partially overlap in such a way that neither is completely contained in the other, can count as a violation (Black et al. 1991). Or, sometimes, only the number of model words that contain such an overlap are counted, or only the number of standard words involved in an overlap. Since the percentages of crossing violations which are cited in the literature are generally very small, the details of the calculations are perhaps not critical. In Chapter 4, we will use the proportion of standard words which contain a crossing violation.

2.2 Statistical approaches to segmentation

Until very recently, there were no attempts to create a computational model of infant segmentation. Computer models aimed at that particular problem will be presented below, but first it may be of interest to look at other computational problems which share some features with the infant segmentation problem, and to consider whether techniques used there might be of use to us.

Most of the approaches discussed here use, as does the O-B algorithm, an unsupervised learning procedure—one in which the classification to be achieved is not provided as input, but the algorithm is expected to extract it from the data. This is also called “self-organizing.” A supervised procedure is “trained” by observing the distinction to be learned and extracting regularities from the data which will enable it to predict the same distinction in new and unmarked data. In our case, supervised training provides word boundary information, while unsupervised does not.

2.2.1 Segmentation into words

Because European languages have an alphabetic orthography and word spacing, NLP applications can easily adopt the “word” as the basic unit. Using blank space and various punctuation marks as delimiters, it is fairly simple to mechanically arrive at conventional units which can be matched with machine-readable lexicons. While there still remain problems with rare and novel words, proper names, hyphenations, and so forth, the great majority of word tokens are trivial to extract.

Some NLP applications do not start with orthographic representations, such as speech recognition, or do not end with them, such as text-to-speech conversion. The dominant methodologies for these, however, also use lexicons, and in practice succeed by modeling the phonological processes which work against word identity, such as elision and resyllabification across word boundaries. Techniques like phonetic lattices are used, which allow for the matching of the speech stream to various possible word combinations retrieved from the lexicon and modified by rule-like representations of common phonological regularities. Thus, the notion of word units is again an input to the analysis rather than an output, and the main strategy is a search strategy.

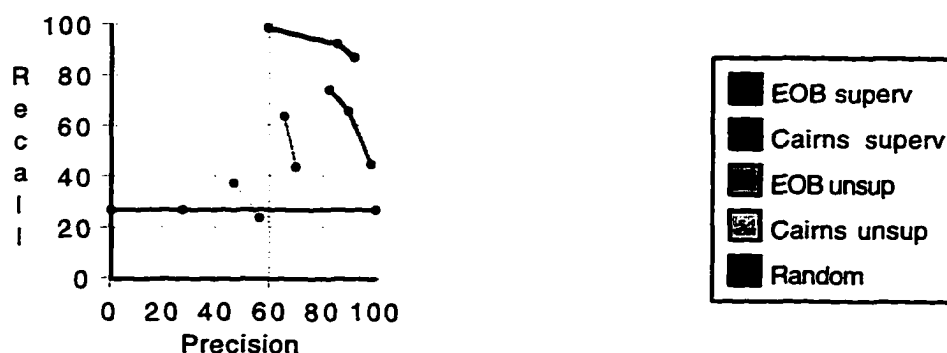
Despite the paucity of evidence, there is a persistent rumor—perhaps based partly on intuition—that phoneme and/or letter n-grams can yield significant amounts of word boundary information. Certainly, there are some sounds that never or only occur word-finally (“ng” and “h” in English, for example). The differential frequency of occurrence of certain phonemes in certain combinations, or at word (or syllable) boundaries—language-particular phonotactics—have been

shown to be statistically significant (Church 1983; Brent & Cartwright 1996; among others).

Cairns, Shillcock, Chater & Levy (1997) attempted to fill a conspicuous gap in the evidence chain by testing these claims for phoneme n-grams. They used their own phonological retranscription of the London-Lund Corpus of English conversation (Shillcock, Hicks, Cairns, Levy & Chater in press), which has a set of 45 phoneme symbols and an average word length of 3.7 symbols. They used a supervised learning procedure which for each phoneme bigram calculated the ratio of occurrences across a word boundary vs. those within a word, and simply postulated word divisions between all bigrams whose ratio exceeded a certain threshold. Categorizing each such cut as a hit or a false alarm, they showed the results as a Receiver Operating Characteristics (ROC) curve⁶ across a range of thresholds. They also tested an unsupervised procedure which uses counts of phoneme n-grams and places a word boundary between the most infrequent bigrams, again using a ROC diagram to show the results across a range of threshold parameters. They found that the unsupervised procedure performed fairly poorly, not much above chance. We will present results for similar procedures on our corpus in Chapter 5.2.4. As shown in the graph below (not ROC in this case, but recall vs. precision), our results were somewhat better, perhaps indicating significant differences between corpora, even when both have the same random baseline, as in this case.

⁶ The Receiver Operating Characteristics curve is an evaluation procedure from Signal Detection Theory (Green & Swets 1966) which shows the tradeoff between false negatives and false positives (misses and false alarms, or Type I and Type II error). Generally, as the number of successful hits increases, so does the proportion of false alarms, and the two ratios are plotted on a coordinate graph.

Supervised and unsupervised bigram models, cut recall and precision



Statistically speaking, the difference between letter n-grams and word n-grams is not merely a quantitative one. Letter and phoneme code sets are small and fixed, while word inventories are large and unbounded. Thus, the problem with word n-grams is always “sparse data”—never enough instances of a category to generalize adequately. The more text one processes, the more types accumulate, so it is difficult to increase the depth of coverage. With letter n-grams, however, more text translates pretty directly to more instances of all the categories. The problem with letter n-grams is of a qualitatively different sort. At this low level the number of categories that can be discriminated from the data is much smaller than the number of phenomena of interest in the world. Thus, for instance, the combination “er” occurs in various linguistic contexts (*worker, erase, severely, one ring*), but in an n-gram algorithm they all collapse into a single data bin.

Finally, one application which has been popular lately, and which might seem to overlap with our goals, is building lexicons by adding phrases that have non-compositional meanings. These are extracted as groups of orthographic

words from text corpora, generally using a combination of filters. The first set of criteria uses a measure of strength of association between the words, such as frequency of cooccurrence (Merkel et al. 1994) or mutual information (Smadja 1993; Kita et al. 1994). Since natural language texts contain large amounts of repetition, these frequency-based criteria produce an overwhelming number and variety of candidates that must then be filtered down. The second set of criteria is generally one based on linguistic principles, such as a list of “empty” stop words or syntactic structure. However, the differences discussed above between words and letters make it impossible for us to use these techniques.⁷

2.2.2 Segmentation of words into morphemes

The segmentation of words into morphemes, which would seem to be similar in some ways to segmentation into words, is a problem even for languages with clearly defined orthographic words. Morphological segmentation is necessary for various applications, such as syntactic analysis, since grammatical prefixes and suffixes can influence the syntactic representation independently of their hosts, and information retrieval, where terms with a common lexical stem should be treated together. Since the most efficient and effective method of morphological analysis is generally thought to be a lexicon for stems and a list of grammatical morphemes, there is very little interest in other approaches. Anderson (1994) argued convincingly that morphology is so much less regular than syntax that even these hand-crafted techniques will have difficulty. So the three studies we look at below are perhaps motivated more by curiosity than any real

⁷ Ever stubborn, I did try to use mutual information between letters, but to no avail.

application.

Juola, Hall & Boggs (1994) built a morphological lexicon by storing word types found in a corpus and then dividing them into morphemes based upon local entropy peaks and valleys, calculated to represent the number of possible continuations at each letter of the word. Where there are many possible continuations, it is posited that a morpheme boundary exists. The results were moderate—good enough to show that the hypothesis had some merit, but not good enough to warrant further work.⁸

Two systems for “morphological segmentation without a lexicon” were described in Klenk & Langer (1989), one for Spanish developed by Klenk and one for German by Langer, which are mixtures of rule-based and frequency-based approaches. The Spanish system used decision rules to determine where the stem and inflection are joined. The rules were hand-coded based on inspection of letters and letter combinations occurring at the ends of words in Spanish text, and arranged in an ordered performance sequence, very much like decision lists (Rivest 1987). Thus, though the end result was a hard-coded system of rules, they were grounded more in frequency than in linguistic principles. Using these decision rules, a sequence of binary decisions about characteristics of the instance ultimately led to a classification decision. Since the strongest evidence was tested first, a decision could be made that ignores weaker evidence which may be contradictory. As Yarowsky (1994) pointed out for decision lists, such a system differs from both probabilistic and rule-based approaches; it combines various

⁸ The same intuition seized me, and I experimented with dividing words by observing entropy rise and fall over letter strings; I also concluded that the method was not completely without merit, but did not seem very promising, either.

sources of information, both strong and weak, to reach a final verdict.

The second system described in this paper segmented German words into morphemes based on frequency counts of letter bigrams drawn from a training corpus. The relative frequencies of each of three cases for each bigram—a morpheme boundary occurs just before the bigram, between the two letters, or just after the bigram—formed the basis for determining the average “tendency to be divided” between any two letters in a word. The points in the word with the highest such tendency were prime candidates for a morpheme boundary. However, this frequency information was tempered by a number of other rule-based heuristics. Also, certain graphemes in the test data were replaced by less ambiguous phonemic representations, based on external information. In all, though they reported success rates in the mid-90%, these two remained “toy” systems.

2.2.3 Discovery of words in Japanese and Chinese

The long tradition of using spaces to separate written words in English probably accounts for at least a part of our knowledge of what constitutes a word. However, word separation is not a universal feature of writing systems. Even in English and other European languages, word spacing did not become general until about the tenth century (Sampson 1985). Of modern languages, Japanese and Chinese are two important ones which do not delimit written words. For various applications, such as machine translation, keyboard input, or information retrieval, it is necessary to segment a written character stream, and techniques used in these situations may be relevant here.

There are three conditions which limit the usefulness of this literature. First,

most segmentation techniques, as in alphabetic languages, rely on lexicons and hand-coded rules (e.g. Sakai & Phillips 1993; Makino 1985); an excellent review is Wu & Tseng (1993). Second, the codeset is usually not small and finite like an alphabet, but has more in common with English words. While technically not an open set, in both Chinese and Japanese the number of available Chinese characters (kanji) are large enough that adding more text nearly always means encountering previously unseen characters, just as with English words. From a cognitive point of view, too, kanji are phonetically opaque and thus even farther from the infant's representational problem than is alphabetic text. In this respect, the single Chinese character is more like a word in English and multi-character compounds more like our phrases. So we need to be sure that the techniques used will apply to the smaller codeset problem of speech sounds.

The third major problem with segmenting "words" in these languages is that there is no general standard for the definition of a word, since the writing system does not provide word boundaries. Almost every paper refers to this:

... 'word boundary' [is] a perennial issue in synchronic Chinese linguistics... for in very many cases there simply is no clear word boundary. Instead there is a gradation of word-hood. (Chen 1993:1)

Actually, there does not exist any clear definition of a *word* in Chinese. The differences between words and morphemes, between simple words and compound words, between compound words and phrases, and between free morphemes and bound morphemes are not clear. (Chang & Krulee 1991:294n.)

Some studies of statistical segmentation use the results of a traditional lexical-lookup algorithm as the standard, and others employ human beings to create a segmented standard of the test text, or to evaluate directly the results of statisti-

cal segmenting. Perhaps surprisingly, humans show considerable variations in judgment: Sproat et al. (1996) cited a range of 69-89% agreement among humans vs. 64-84% for the human/computer comparison. Huang, Ahrens & Chen (1993) hypothesized a gradation of associativity from simplex words, to compounds, to phrases, suggesting that binary decisions on acceptability may not capture the reality of lexicalization.

As in English, there have been few studies in Japanese or Chinese which use purely unsupervised techniques. Even probabilistic language models are created by supervised training on a previously segmented and/or tagged corpus, despite the fact that such corpora are even harder to obtain than for word-spaced languages. The only study we have found that uses an unsupervised training process for Japanese is a recent one which creates a language model for speech recognition. Ito & Kohda (1996) compared their results to traditional models based on character and word n-grams, using a metric of model perplexity.⁹ As described previously, such language models consist of a lexicon (symbol set) with a probability estimate for each lexical element. In a character bigram model, for instance, probabilities are determined for each contiguous pair of characters in the training set, and then the cross-entropy of this model with the test set is computed and expressed as perplexity.

Ito & Kohda (1996) created a lexicon of the character strings occurring most frequently in the training corpus, which could be of any length, and then constructed a language model using bigrams and trigrams of these. They found that

⁹ Where information = entropy = H , being the $\log_2$ of the number of choices for each unit of interest (word, character, etc.), perplexity is the number of choices itself, or 2^H .

the perplexity of such a model was better than that of character n-gram models for the same corpus, which in turn were better than word n-gram models.¹⁰ The best result was for the character-string trigram model, but this was only slightly better than the one using character trigrams. However, the underlying unit of the character-string model was on average 1.22 characters, while that of the character model was of course only 1.00 characters, so there was an immediate gain in modeling power for the string version. (The word model had an average unit length of 1.65 for the same corpus.)

This model is of considerable interest, though the authors themselves point out that the resulting string patterns are not necessarily intuitive ones. It has echoes of earlier research in Japanese (Kita et al. 1994; Nagao & Mori 1994) which focused on frequently occurring strings for the purpose of extracting terminology and collocations. Ito & Kohda (1996), however, integrated the technique with a parsing algorithm to produce a complete language model.

Two other studies in Japanese are interesting, but do not present conclusive results. The first, Teller & Batchelder (1994), bypassed the problem of a large character set to focus on the information available from the fact that Japanese writing uses four different types of characters. According to Nagata (1994), "To handle unknown words, most Japanese morphological analyzers use character type heuristics, which is 'a string of the same character type is likely to constitute a word.'" To investigate how much information about word boundaries was contained in character-set shifts, Teller & Batchelder (1994) segmented Japanese

¹⁰ The Japanese corpora which Ito & Kohda (1996) used for training and testing were divided into words manually to construct the word n-gram.

text based on a combination of heuristic rules and a statistical measure similar to mutual information, applied to pairs of hiragana characters, and pairs consisting of each hiragana character and another character type, including katakana, kanji, punctuation, roman letters and arabic numbers. Their algorithm, trained on a corpus of 1.17 million characters, achieved recall of 98% and precision of 94% on a very small text corpus of 2,200 characters, versus a hand-coded standard. This work was, however, of a “toy” nature and was not extended to a larger test set.

Nobesawa et al. (1994,1996) also relied on quantitative measures exclusively. They devised a score of the association between two characters that takes into account all cooccurrences that lie within d characters of each other. They called this statistic a “ d -bigram,” and used it to compute a “linking score” for each pair of contiguous tokens (characters) in a corpus, an MI-like measure of strength of association. They segmented a sentence wherever the linking score fell below a certain level, representing a point of weak association. After training on a newspaper corpus of 7,502 sentences and testing on a held-out portion of 302 sentences, they computed the percent of division points which failed to retain a minimal meaningfulness, that is, oversegmentation error. They commented that the result is not really a “linguistic” unit, but a new, “computational linguistic” unit which reflects statistical relationships rather than linguistic ones, and may thus be useful for various computational purposes, such as building lexicons. They report 9.7% such errors, compared to a simple bigram MI procedure with 7.4% error. This seems a promising approach, but we are not given enough information to be sure. They give no examples of the extracted “linky strings” and we cannot tell from this report just how different they are from morphemes or words. Perhaps

future work will be more fully documented.

For Chinese, the problem is somewhat different. Using only one character set—the up to 13,000 ideographs¹¹ called “hanzi” in Chinese—there are not even character-set distinctions to mark word boundaries. In addition, Chinese has a relatively large amount of productive word-formation, including “a specific rule in Chinese that derives new lexical items from longer corresponding forms... called *Suoxie* ‘abbreviation’” (Huang, Ahrens & Chen 1993:62). In this environment, there seems to be even less interest in non-lexical approaches. Sproat et al. (1996:399) comment:

The major problem for our segmenter, as for all segmenters, remains the problem of unknown words... [T]here will [always] remain a large number of words that ...simply have to be added to the dictionary. This implies, therefore, that a major factor in the performance of a Chinese segmenter is the quality of the base dictionary, and this is probably a more important factor—from the point of view of performance alone—than the particular computational methods used.

Sproat, Shih, Gale, & Chang (1996) combined lexical lookup and probabilistic techniques in segmenting Chinese text. Although they began with a previously compiled lexicon, they used an iterative process to recalculate the frequencies of each lexical item based on a training corpus of 20 million words. To briefly summarize: All words from a machine-readable dictionary were entered into the lexicon, plus all possible single characters. Probabilities were initialized with the raw string frequency for each item, and then the entire corpus was parsed based on this lexicon. The frequencies resulting from this parse were recorded, and the parsing and recording cycles were iterated until convergence. The final

¹¹ Depending on the computer code standard, from 6,000 to 13,000 characters may be available (Lunde 1993).

lexicon was then augmented by several additional algorithms to take into account unseen words, proper names, and foreign loan words (together accounting for 2.5% of the types in the test corpus). Performance of the augmented lexicon on the test corpus was close to that of a panel of human segmenters (76.4% average of precision and recall between humans vs. 75.2% for the human/computer comparison).

Luo & Roukos (1996) added the logical next step to the above procedure, using a parse algorithm which allowed some non-zero probability for any unseen word. After each parse iteration, therefore, in addition to modifying frequency counts of listed words, some number of previously unseen words were added to the lexicon. In one test, an unsegmented corpus of 13 million characters was processed (about 6.5 million words at the usual rate of 2 characters per word) and 1500 unseen word types were added to the lexicon. Of these, a sample of 20 picked randomly were examined and 16 were found to be “reasonably good words.” The segmentation results also compared very favorably with more traditional models and with human raters. But the low number of new words added to the lexicon forces us to realize that the large symbol set in Chinese shares the data sparseness problem which plagues statistical approaches based on alphabetical words.

2.3 Connectionist models of word segmentation

A competing paradigm for computational models is connectionism, also called artificial neural networks (ANN) and parallel distributed processing (PDP). While these models are also based on statistical relationships and mathe-

mathematical calculations (Elman 1993; Hinton 1989), they are different in several interesting ways. First of all, both input and output data can be coded so as to allow the recognition of more complex relationships. Indeed, this is one of the main advantages of distributed representations, that is, elements which are defined as patterns of activation rather than as discrete symbols. For example, language can be represented not as orthographic or phonemic symbols, but as their component features, permitting generalization between phonemes on the basis of common characteristics. This is frequently called a “sub-symbolic” representation.

A second difference is that the precise nature of the path by which input is transformed to output need not be specified—or even known—by the designer beforehand. A standard “feed-forward” network has input and output units linked by a number of hidden units. Depending on the architecture, two or more input units will be mathematically combined in a hidden unit, and two or more hidden units will together determine an output unit. With iteration and back-propagation of error, experience gradually alters the state of the hidden units, and thus a way is eventually discovered to link particular input values with particular output states.

Three groups of researchers have created connectionist nets to study segmentation from the point of view of developmental cognition, and two of the three got results that demonstrate substantial success in the task. A drawback that networks share with other approaches (both computational and experimental) is that it is difficult to control all the variables sufficiently. On the other hand, the lack of a priori rules can lead to a much finer-grained inspection of the results than is common in other computational approaches. We meet here for the first

time several key issues which will come up again later.

We will discuss the overall research project for each group and then, in a final subsection, compare their results; the reader will find a summary table in Section 2.3.4. This discussion will omit many technical details which influence network operation (learning rates, etc.), and the reader should be aware that conclusions reached here are thus potentially incomplete.

One technical matter that may cause confusion is how the input is represented to the network. Three input coding schemes for network input can be distinguished:

- Component **features**, where the underlying characteristics of an element are analyzed and presented to the net as a cluster of separate elements.
- Random coding, where each element is assigned an arbitrary bit configuration, or number, as A=#1=0001, B=#2=0010, C=#3=0011, etc.
- Local coding, where the number of input units is the same as the number of possible input values and each unit is associated with one value, as when 26 input units represent the English alphabet.

In discussing each network, we will point out which system is used and the possible consequences of this decision.

2.3.1 University of Rochester

Aslin, Woodward, LaMendola & Bever (1996) of the University of Rochester chose to use data representing speech to very young children (under a year old), since they suspected that it is "different in important respects from speech to 2- and 3-year-olds" (Aslin et al. 1996:127). They used the speech of mothers to their

infants collected by Aslin's group in a previous experiment. The orthographic transcription of these sessions was converted to phonemes by the DECTalk automated text-to-speech synthesizer, and thence to features.

The network they used was a simple feed-forward network, which was trained to predict utterance boundaries from phoneme and pause information. Each phoneme was coded as 18 articulatory features of a type now traditional in linguistics (sonorant, coronal, round, etc.), and the input stream was presented as moving triples—phonemes 1 2 3, then phonemes 2 3 4, etc. There were thus 55 input units—three phonemes ($3 \times 18 = 54$), plus a unit indicating the end of an utterance or pause between phrases. The output of the network was a single unit that indicated the presence or absence of an utterance boundary. They discuss their rationale as follows:

Reinforcement, or feedback to the model when a word boundary has occurred, was limited to the simplest source of information available to the infant; namely, when the end of a sentence or the end of a major phrase was detected. That is, we assumed that the infant had the ability to recognize the end of a word if it was followed by a substantial (>1 second) period of silence.... our goal was to provide the model with the most limited source of reinforcement to determine if this minimal level of information was sufficient for learning to take place. (Aslin et al. 1996:126)

The amount of activation of the output unit at various points was used as an indication of the network's learning, and it was found that the highest activation was at utterance boundaries, as expected, but also that there was greater activation on average at lexical (word) boundaries than at word-internal phonemes. This meant that the network was "learning" something about word boundaries by observing the beginnings and ends of words at utterance boundaries. Assuming a word boundary wherever output activation exceeded the mean activation

across all phonemes, the recall rate for word boundaries (not the resulting words) was 62%, and precision of 74%, a respectable performance.¹²

When the experiment was repeated using a random coding scheme, this learning of word boundaries failed to occur. In another variation, utterance boundaries were placed every 6-8 words instead of every 3-5 words, and learning did not take place. Aslin et al. (1996) concluded that both featural representation and short utterances were necessary for this network to learn. However, in a personal communication, Aslin agreed that phonemes might perform as well as the lower-level features if comparably more data were available; that is, the net itself might perform the featural analysis if sufficient training were given.

2.3.2 University of Edinburgh

A group at the University of Edinburgh (Cairns, Shillcock, Chater & Levy (1994; 1997); Shillcock, Hicks, Cairns, Levy & Chater (in press)) has been experimenting for several years using an unsupervised training procedure and a different network architecture. The simple recurrent network (SRN) is based on the work of Elman (1990), who first used this procedure in research on language segmentation. Elman wanted to represent language as a stream of events moving forward in time. The SRN has, in addition to input, hidden, and output units, another set of units generally called "context units." The hidden units are fed backward (copied) to the context units, and at the next cycle both new inputs and the context units (holding the previous cycle's hidden-unit values) feed forward. Thus, in each cycle the network combines the new input with a "memory"

¹² In this brief report of their work, Aslin et al. (1996) do not make absolutely clear how they computed these figures, but those we have cited are plausible.

of its own preceding state.

The effect of the SRN architecture is that the network can discover patterns in the sequence of inputs (syntagm) as well as in the internal composition of various inputs (paradigm). Elman (1990) had demonstrated this with a small artificial corpus consisting of 200 short sentences (4-9 words) generated from a lexicon of 29 words, yielding a stream of 1,270 words.¹³ The corpus was presented one letter at a time to the network, with no breaks between words or sentences. The network's task was to predict the next letter, which it learned to do with varying degrees of success. Elman plotted the error value for the trained network at each letter and observed a pattern in the error fluctuations across a continuous ("graded") range of values:

...[A]t the onset of each new word, the error is high. As more of the word is received, the error declines, since the sequence is increasingly predictable. The error provides a good clue as to what the recurring sequences in the input are, and these correlate highly with words. The information is not categorical, however. The error reflects statistics of cooccurrence, and these are graded. (Elman 1990:192)

While the letter prediction task is learned by supervised training, there is no input or feedback about word boundaries, so any segmentation information that the network gives us (indirectly, via the error measure) must be the result of an unsupervised process, or self-organization. Elman goes on:

The simulation makes the simple point that there is information in the signal that could serve as a cue to the boundaries of linguistic units which must be learned, and it demonstrates the ability of simple recurrent networks to extract this information. (Elman 1990:193)

¹³ The tables and text in this paper are generally not in sync, and the sample shown in the table for this experiment does not match the description in the text. We present the text description, but cannot therefore give any samples of the corpus.

In contrast to the very small amount of child-directed speech used by Aslin et al. (1996), the Edinburgh group worked with the large London-Lund Corpus of adult speech, converted from standard orthography to a phonemic representation, and then to a feature-based representation, described in more detail in Shillcock, Hicks, Cairns, Levy & Chater (in press). As in Elman (1990), evaluation of results after training is based on observation of the error measure for a test corpus, and word boundaries were assumed when the error measure rose above a certain cutoff point. The best results were not very good—measured with respect to word divisions, recall was 21%, precision was 60%. They conclude that “phonotactics provide a fairly weak source of information for the bootstrapping of segmentation” (Cairns et al. 1994:139) and “...the network does not segment more whole words from the test stretch than it would by chance” (Cairns et al. 1997). Why this study failed to replicate the results of the other studies¹⁴ will be discussed further below.

2.3.3 University of Southern California

A group at the University of Southern California (Christiansen, Allen & Seidenberg, in press) wanted to test the hypothesis that “constellations of overlapping, partially predictive cues” could account for language learning. They set out to show that sequential phonotactic structure, utterance boundaries, and word stress information could serve as combined cues to word segmentation.¹⁵

¹⁴ Since none of the Edinburgh papers mention the work of the Rochester or USC groups, we do not know how they account for the discrepancy.

¹⁵ We will not discuss their results on the contribution of stress information, since it is not directly comparable to our own work, and also because we do not feel prepared to evaluate how representative of natural speech are the sources of stress information in the Christiansen et al. (in press) study.

A fairly large corpus of speech directed to infants of 6-16 weeks of age from the CHILDES database was used, phonemically translated according to the MRC Psycholinguistic Database, and edited to avoid words not listed there (about one-quarter of the text was deleted for this reason). Utterance boundaries were represented as a single binary unit in both input and output. The training corpus was passed once through the SRN under supervised training, and then the network's predictions on held-out data were evaluated.

The network showed a much higher mean activation level of the boundary unit at lexical boundaries than word-internally. The results were very similar to those of the Rochester group.¹⁶

<u>Mean boundary unit activation at:</u>	<u>Christiansen & al.</u>	<u>Aslin & al.</u>
Word-internal phoneme	.05	.08
Lexical boundary	.25	.18
Utterance boundary	.34	.35

Note that the difference between the low and high activation here is about the same in both studies, but in one the "lexical boundary" level is closer to "utterance boundary" and in the other it is closer to "word-internal." Since there was no lexical boundary information given to the net, the activation of the boundary unit at those points can only be a generalization from the utterance boundary information which *was* given. Christiansen et al. (in press) thus claim that their net has a greater ability to generalize than was found in Aslin et al.'s study.

Using a threshold of mean activation of the boundary unit across the whole

¹⁶ These values include the stress component, and are estimated from Figure 4 in Christiansen et al. (in press:23). The comparison column is estimated from Figure 8.5 in Aslin et al. (1996:129).

corpus as the criterion for word boundary decision, the network achieved 71% recall and 66% precision on cuts, and 40% and 37% respectively for whole words.¹⁷ These levels are very similar to the Rochester results.

2.3.4 Comparison of network results

One conclusion we can draw from these studies is that at least two of the three demonstrate that there is structure in the speech signal which can be extracted by statistical techniques. However, we will argue that connectionist techniques are not necessarily the best way to investigate this structure.

First of all, in terms of raw performance, the best results here are barely competitive with the techniques to be discussed below. Second, it is by no means clear what the determinants of success are in a connectionist trial. We can introduce some key issues here, asking why the Edinburgh experiments did not match the results of the Rochester and USC groups.

The following table presents a summary of the important differences in design and implementation among the three, and adds the early Elman experiment as well:

¹⁷ These results exclude stress. The network trained on stress information in addition to phonemic and utterance boundary information achieved slightly higher levels: 74% recall and 70% precision on cuts, 45% recall and 43% precision on words.

	<u>Elman</u>	<u>Rochester</u>	<u>Edinburgh</u>	<u>USC</u>
<u>Corpus</u> of speech directed to:	—	child	adult	child
Segments per word	3.92	~3.75	3.30	3.00
Segments per utterance	25	10-15	—	9
Word TTR	85	≤10	≤25	30
<u>Coding:</u> # segments	~26	~44	45	36
# binary features (bits)	(random 5)	18	9	11
<u>Training:</u> Size in segments	4963	<5000	1,000,000	73,947
Iterations	10	2-3	2	1
Total segments input	49,630	≤15,000	2,000,000	73,947
Total bits input	248,150	≤270,000	18,000,000	813,417
Utterance boundary training	unsup.	superv.	unsup.	superv.
<u>Net architecture</u>	SRN	window of 3	SRN	SRN
Task: Predict segment?	yes	yes	yes	yes
Predict boundary?	no	yes	no	yes
Input units	5	54+1	9	11+1+2
Hidden units	20	30	60	80
Context units	20	—	60	80
Output units	5	1	27	36+1+2
<u>Performance:</u> Cuts recall	?	62	21	71
Cuts precision	?	74	60	66
Words recall				40
Words precision				37

Note 1: ~ indicates an approximate estimate when the paper is silent or imprecise.

Note 2: USC inputs and outputs include 1 boundary unit and 2 stress units; they use a local representation for output of phonemes (one bit for each phoneme).

There are a number of differences between the poorly-performing Edinburgh study and the two more successful ones:

- Edinburgh used a large adult corpus, while the others used small child-directed ones. The actual effect of these variables is not known: Does a larger corpus require more iterations, or fewer? Since the word size and repetition rate (TTR) are not different, perhaps the child/adult distinction

doesn't translate into any effective difference.

- Boundary information was not supplied to the Edinburgh net, but was input to the other two. In other words, the Edinburgh net received unsupervised training for word boundary detection, while the others' training was supervised. While it is true that the Elman net was also unsupervised in just this way, we do not have performance data on the Elman study, only impressionistic results. Perhaps it too did only "better than chance" on predicting boundaries.
- Edinburgh added "noise" to the input by randomly flipping feature bits.¹⁸ How much difference did that make?
- Edinburgh used features from the theory of Government Phonology, while the other two used more mainstream Halle-type articulatory/acoustic features. In addition, we can see from the more complete description of their corpus (Shillcock, Hicks, Cairns, Levy & Chater in press) that the distribution of these features is much less skewed than is typical of linguistic stimuli, with a redundancy of less than 3%. (The unigram (letter or phoneme) redundancy of the corpora used in this study are in the range of 9-18%.) By hypothesis, it is easier to learn data which has salient distributional peaks and valleys than data which does not, so we should be more interested in the feature distribution than the phoneme distribution. We do not have comparable facts about the feature distribution of the Rochester and USC

¹⁸ Cairns et al. (1997:133) said: "Noise is added to the input by flipping features from 0 to 1 (or vice versa) with a certain probability, in order to encourage the network to rely on sequential information (i.e., if the identity of the current segment is ambiguous or unclear, then the net will have an incentive to use the local phonetic context to recover its identity)."

data, however.

We need more information, and more comparative studies, to answer these questions. One team that has tried to compare some of these variables comments on the problems of doing so:

One of the disadvantages of applying a connectionist approach to any empirical problem, is that the designer of the simulations is confronted with a large search space formed by alternative architectures (number of layers, size of hidden layer, feedback connections from output to input or from hidden to input), training data selection and presentation methods, learning algorithms and activation functions, learning rate and momentum parameters, and data encoding schemes. As the simulation of massively parallel algorithms on serial machines is notoriously slow, we have been able to explore only a small region in this search space. (Bosch & Daelemans 1992:185)

Bosch & Daelemans (1992) compared the SRN with a standard feed-forward net on the task of segmenting Dutch words into syllables, which required both phonological and morphological information. They found that the feed-forward architecture and Elman's SRN performed about the same. They then compared orthographic with phonemic representations (using 35 phonemes), both using random coding on 5 or 6 units, and found that the phonemes did slightly better. A coding of phonemes by sonority value class (0.1 to 1.0 by steps of 0.1) did a little better than the random-coding version.

In the feed-forward approach, they supplied context with a moving window of various sizes, finding that a window of five units (two left-context and two right-context) did best for phonemes, while seven units (3-1-3) was optimal for the graphemes. They speculated:

A possible explanation for the counter-intuitive result that a smaller window size produces better generalization, would be that in case of a larger

window size, more ambiguous patterns... are present in the training material, or that all areas of the larger space of possible training examples is not sufficiently covered by the training material. It may well be possible that the morphological constraints (internal word boundaries) are so unconstrained that they act as noise. (Bosch & Daelemans 1992:189)

This suggests that there are very subtle trade-offs in the connectionist process—like a nonlinear dynamic process where small changes in input or control parameters can have disproportionate effects on the results. In other words, more is not necessarily better. However, the range of difference across all of these variations was not great—only 6.3% to 16.5% incorrect decisions in the test set.

This study is unusual in the number of different parameters that were varied. One can regret that they did not attempt a feature-based coding as well, since the number of possible interpretations of the results increases when a network uses a feature-based representation. For instance, does the net generalize the relationship of particular features between one phoneme and the next, e.g. [+cor +cor] is less likely than [+cor +round]? Or are particular configurations of features within a phoneme being identified, resulting in less contrast between feature coding and random coding?

A larger question is whether finding word boundaries by using distributional information is a process of learning particular words, or of learning word-typical patterns (phonotactics), or of some combination of these two. As an example, does segmenting “thebunny” into “the | bunny” require or imply that “the” or “bunny” have been identified as words? Or is it a case of knowing that “theb” and “thebu” are atypical combinations (less frequent linguistic events) than “bun” or “bunny”? We will return to this question in later chapters.

2.4 Probabilistic approaches to word segmentation

2.4.1 Olivier and Wolff

Olivier (1968) was the first to create a working probabilistic segmentation routine. In a dissertation entitled *Stochastic grammars and language acquisition mechanisms*, he was concerned both with psychological reality and empirical effectiveness. His algorithm inserted word-separating spaces into “collapsed” English text—text from which spaces had been purposely removed. The algorithm built a dictionary incrementally, parsing the text section by section, each time using the dictionary resulting from all previous passes. Beginning with single letters, the dictionary was expanded by adding contiguous pairs of parsed “words” as new lexical items, so there was a tendency for words to get bigger and bigger as the process advanced. Although Olivier (1968) is widely cited in later research, the effect of his work until now seems to be mainly as a curiosity. However, we have adopted his algorithm and will demonstrate its effectiveness in Chapters 3 and 4, as the O-B algorithm. Olivier himself seems to have done no further work in this direction.

Wolff (1975, 1977) used a very similar approach in his dissertation entitled *The acquisition of linguistic structures*. Working at the University of Wales, Olivier’s Harvard dissertation was known to him only by a report in Brown (1973), and he apparently developed the idea independently, inspired by the artificial-language-learning experiments of Hayes & Clark (1970). Wolff also began with single letters and entered in the dictionary pairs of parsed “words,” with the difference that, while Olivier entered all possible pairs, Wolff entered only those

pairs that passed some frequency threshold. Wolff's results, both lexicon and parsed text, highlighted the various levels of structure in the language. The following parsed phrase is a sample from Wolff (1977):

(HE) (ACH) ((ED) (AND)) (SOME) (WHERE) (IT) (HURT) (HIM)

The lexicon also gave a count of the number of text letters which had been traversed when the item was added, which can be seen as an indicator of structural level, or strength of association. In fact, Wolff was less concerned with the results of parsing or the "words" found in the final dictionary than he was with the relative strengths of association. He showed that the general trend of his routine was not only to discover boundaries, but to rank them in terms of strength, with intra-morpheme connections stronger than morpheme boundaries, which in turn are stronger than word boundaries, words stronger than phrases, and so on. He saw the program as a model of a number of aspects of language learning, including the psychological reality of the word, a child's vocabulary learning spurts, and the increasing length of words learned.

Because we have adopted the Olivier algorithm for further development, and Wolff's work does not differ in relevant ways, we will not say more about Wolff's work.

2.4.2 de Marcken

Recently, several approaches to word segmentation have been based on Minimum Description Length (MDL), a technique from information theory which is currently held in high respect by the computational community. Originally designed as a powerful technique for data compression, the MDL hypothe-

sis is that the most compact representation of a body of information (here, a text or corpus)—which has been made compact by removing as much redundancy as possible—“is the theory that best explains the data.” In this case, the evaluation metric is the combined size of the data and its analysis, that is, the text and the lexicon derived from it.

Using the MDL metric, Carl de Marcken (1995,1996a,b) created a lexicon with probabilities by iterating four processes:

In stage 1, [parse and update] the probabilities of the words in the current dictionary. In stage 2 new words are added to the dictionary if this is predicted to reduce the combined description length of the dictionary and input. Stage 3 is identical to stage 1, and stage 4 words are deleted from the dictionary if this is predicted to reduce the combined description length. Stages 2 and 4 are a means of increasing the likelihood of the input by modifying the contents of the dictionary rather than the probability distribution over it... (de Marcken 1995:10)

He tested his procedure on a variety of corpora, and the results were quite remarkable as a testimony both to the amount of structure in language and the ability of the MDL algorithm to extract it. De Marcken presented as a figure “The top-level segmentation of the first ten sentences of the test set...” (de Marcken 1995:13), from which we repeat the following examples:

| a bit of| trouble| |".| the| campaign|
| williams| said he|, too,| was| subjected to| | anonymous| calls|

(In this test, punctuation and spaces were not removed, but other tests were made on “collapsed” text with similarly good results.)

The term “top-level” is necessary since his algorithm discovered structure at various levels, as we saw in Wolff’s work, and there might be different amounts of hierarchy in individual stretches of text. For example (de Marcken undated):

```
[s[h(or)]t][c[ut]]
[[[ju][st]]_][a[s_]]
[m(er)][e[[ly]_]]
```

As de Marcken commented, this can be an advantage for some purposes, but it makes evaluation of segmentation more difficult. One demonstration of the MDL algorithm was on a corpus of speech to children, converted mechanically from orthographic text to phonemes and then word spacing removed. After unsupervised training, de Marcken cited the recall rate at 96.2%, and a crossing rate of 0.9%. However, he himself noted the difficulty of interpreting these results:

...the *recall rate* of our algorithm... 96.2% of the [words] in the [standard] are exactly spanned by a single word at *some* level of our program's hierarchical segmentation. ... Comparing to ...other similar results... is difficult: at first glance our recall figure seems dramatically better, but this is partially because of our multi-level representation, which also renders accuracy [=precision] rates meaningless. (de Marcken 1995:15-16)

In a subsequent paper, de Marcken (1996a) reported a segmentation of the Brown corpus, with spaces and punctuation removed, with similar results: recall 90.5% and crossing-brackets 1.7%. A large Chinese corpus was also segmented, with recall of 96.9% and crossing 1.3% when compared to "another computer segmentation program that had access to a (human-created) lexicon" (de Marcken 1996a:338). The disclaimer here was:

Of course, the hierarchical lexical representation does not make a commitment to what levels are "true words" and which are not; about 5 times more internal nodes exist than true words. (de Marcken 1996a:339)

These experiments are tantalizing, but we are left with the suspicion that these scores are the result of simply including an enormous number of words in the lexicon. Although the MDL algorithm is advertised to create a "compact"

lexicon, “5 times more internal nodes” doesn’t sound very constrained, and we have no precision figure to reassure us.

2.4.3 Brent & Cartwright

The work of Brent and Cartwright (1996, in press; Brent 1997), also using the theory of Minimum Description Length, differs in several key respects from that of de Marcken. They explicitly give a cognitive rationale for using MDL, a compression algorithm:

At an intuitive level, the need to represent an infinite language in finite memory can be viewed as a functional pressure to discover regularities that minimize the amount of memory needed... (Brent & Cartwright in press:12f.)

Second, their algorithm worked differently from de Marcken’s, although it was at least as computationally intensive. The algorithm began with only utterance boundaries in the text. It added word boundaries slowly, either one or two at a time but never more, evaluating the results at each change by computing the MDL objective function.

We use the term *major step* to describe the process during which the algorithm decides whether to add one new boundary or two, and where to add the new boundary or boundaries.... Specifically, all the segmentations that result from adding either one or two new boundaries to the segmentation selected at the previous major step are evaluated.... Eventually, the search reaches the segmentation in which each pair of adjacent phonemes is separated by a word boundary, so no more boundaries can be inserted. At that point, it outputs the segmentation on which f [the evaluation function] has the smallest value, among all those evaluated in the entire search. (Brent & Cartwright 1996:107f.)

The objective function used for evaluation was the combined length of the lexicon and the coded representation of the text. That is, a particular segmentation of the text would result in a lexicon which contained one copy of each

“word” and a numerical index assigned to each word. The text could then be re-presented as a sequence of index numbers. As Brent & Cartwright spelled out, this objective function resulted in “words” which frequently occurred as a unit, and which occurred in a variety of contexts. They called this the Distributional Regularity strategy (DR strategy), which aimed to maximize four criteria at once (Brent & Cartwright 1996:104f.):

- a. Fewer types (least number of lexical entries)
- b. Shorter types (shortest sum of the lengths of all lexical entries)
- c. Fewer tokens (least number of units in the segmented text)
- d. Minimum type entropy (information is concentrated in a few high-frequency words)

For instance, a segmentation down to the smallest units (characters) would favor criteria (a), (b), and (d), but not (c), while a segmentation to very large units (utterances) would favor only (c). Obviously, there are trading relationships among these criteria, so the interactions will vary depending on the weights assigned to each by the particular evaluation formula.

They ran experiments on “naturally occurring, child-directed English” from the Bernstein-Ratner CHILDES data (Bernstein Ratner 1996). Utterance boundaries were implicitly word boundaries and were also used to derive information about word-initial and word-final phonotactics (Brent & Cartwright 1996). The data were transcribed into “broad phonetic classes”—for instance, “label” became ‘lebL’ and “bird” was ‘bRd’ (Cartwright & Brent 1994). Careful attention was paid to giving each orthographic type a consistent transcription each time it

occurred, and onomatopoeia and interjections (e.g. *uh* and *oh*) were eliminated.

The results reported by Brent & Cartwright (1996) were based on nine transcripts (nine mother-child pairs with the children aged 1;1 to 1;9), and the measures were averaged across the nine runs. Brent & Cartwright used much less data than did de Marcken, presumably because their algorithm was even more computationally intensive and could not have handled a much larger corpus:

Each transcript was truncated to about 1350 segments [= phonemes], not counting those that immediately precede an utterance boundary. Truncation was necessary both because simulations on longer input samples would have required excessive computer time, and because the lengths of the original transcripts varied widely... The average number of words per transcript was 527 and the average number of utterances was 170. (Brent & Cartwright 1996:109)

Each of the nine transcripts thus contained about 1520 phonemic symbols and consisted of, on average, 1350 possible word boundaries (1520 less 170, since utterance boundaries were “given” as word boundaries). They reported recall and precision¹⁹ with respect to the segmented units (“words”) rather than boundaries and compared them to a segmentation procedure which made the same number of cuts as are in the standard, but at random locations.

Brent & Cartwright then augmented the purely probabilistic approach with a rule-based component. They constrained the results of the MDL iteration process by two “phonological constraints” which were learned from utterance and pause boundaries:

- a. Every word must contain a vowel (heading “DR+V” in the table below).

¹⁹ Brent & Cartwright (1996) use the terms “completeness” and “accuracy” instead of the more usual “recall” and “precision.”

- b. Word-initial and word-final consonant clusters are each restricted to a finite set²⁰ (heading “DR+V+BC” for “boundary clusters”).

The results are shown below, with standard deviations in parentheses:

	<u>Random</u>	<u>DR</u>	<u>DR+V</u>	<u>DR+V+BC</u>
Recall (tokens)	13.4 (1.9)	47.3 (9.1)	60.1 (6.8)	69.2 (8.1)
Precision (tokens)	13.4 (1.9)	41.3 (12.7)	68.2 (5.1)	76.3 (5.0)

In this experiment, adding boundaries stopped when no more legal boundaries could be placed. As can be seen from the table above, adding phonotactic information to the algorithm greatly improved performance.

Brent & Cartwright (1996) reported that several variations of the search procedure did not work as well. Specifically, when only one boundary was added at a time, rather than one or two, results were poorer. Also, the reverse procedure—starting with all possible boundaries in place and removing them one or two at a time—failed, probably because this eliminated the information provided by utterance boundaries.

Both de Marcken and Brent & Cartwright are firmly in the computational tradition. De Marcken made no claim to be modeling cognitively realistic processes, and Brent & Cartwright waffled on the issue:

Although this model task may differ from the one children face, we believe it is similar enough to provide insight into the natural task. (Brent & Cartwright 1996:106)

...[O]ur implementation is intended as a general model of the use of distributional information, not as a detailed proposal about the computations children carry out as they process the utterances they hear. Indeed, our

²⁰ The consonant clusters were culled by hand from a larger superset of the Bernstein-Ratner data, which consisted of 10,883 utterances, compared to 1530 utterances in the training data.

current implementation reads in an entire input corpus before outputting a single lexicon, whereas children add to their lexicons incrementally as new input becomes available. (Brent & Cartwright in press:22)

However, we can agree with both of these studies that “distributional regularity” serves as an important source of information:

...[S]ound sequences that occur frequently and in a variety of contexts are better candidates for the lexicon than those that occur rarely or in few contexts (Brent & Cartwright 1996:93)

2.4.4 Redlich

A rather offhand experiment in segmentation carried out by Redlich (1993), as a trial and demonstration of techniques to be applied later in visual image processing, is an interesting combination of elements we have seen previously. It uses entropy as the evaluation criterion—specifically, the redundancy of the word distribution relative to the character distribution. As Redlich says:

To demonstrate the power of the present approach, I apply it to strip the redundancy from English text—to learn the text’s statistical properties. This example is used because we all know a fairly good solution to the problem: transform from the letter representation to the word representation. (Redlich 1993:291)

Like Olivier and Wolff, he starts by assuming that each letter is a word and gradually combines them to make larger and larger words. But like the MDL techniques, his technique creates a lexicon by optimizing the representation of a particular and finite corpus. While the MDL calculation uses the number and length of types and the number of tokens in addition to word entropy, Redlich’s technique uses word and character entropy plus number of tokens, with the goal of maximizing the following redundancy formula:

$$L = 1 - \frac{H_w/c}{H_c}$$

where:

H_w/c = word entropy per character
 = word entropy ÷ average word token length
 H_c = character entropy (unigram entropy)

Redlich points out that this measure rewards letter clusters which show a good combination of both high mutual information and high frequency.

He does not report quantitative results, but shows the effect of successive cycles on a small sample of text. From this sample, we can compute word recall and precision of 47%. When the L measure shown above is applied to our own results, it generally accords with other measures, but not always. Here are results for O-B runs on a single corpus using four different sets of parameters:

<u>Condition</u>	<u>Recall</u>	<u>Precision</u>	<u>Average Word Length</u>	<u>H_w</u>	<u>H_w/c</u>	<u>L</u>
Standard	—	—	2.885	7.40	2.56	0.50
I	38.9	41.4	3.067	8.84	2.88	0.43
II	49.4	48.0	2.802	8.56	3.05	0.40
III	57.8	51.8	2.583	8.28	3.21	0.37
IV	57.4	42.5	2.137	7.67	3.59	0.29

These show that Condition III was best with respect to word recall and precision, but Condition I has the best (highest) redundancy L score of the trials, showing that this measure is perhaps overly sensitive to the average token length.

2.5 Summary

This chapter has provided the context of computational linguistics for our computer simulation, to be described next. After we have presented the results of the O-B algorithm, we will discuss in more detail the other simulations which

are closest to it—the three connectionist experiments and the work of de Marcken and Brent and Cartwright—and compare them to the O-B algorithm. In Chapter 6, we will consider both quantitative results and qualitative issues of psychological plausibility.

Chapter 3. The O-B Algorithm: Methods

We saw in the previous chapter various ways of modeling language using computers, including some studies concerned directly with early speech segmentation. In this chapter, we will introduce the O-B algorithm, a simple computational algorithm which can discover many of the word boundaries in computer-readable text corpora. We claim that the algorithm is extracting information from observing cooccurrence frequencies in language, much as the human mind might be doing it.

We will first introduce the various corpora that will provide the language input, and then describe the quantitative measures which we will use to evaluate performance on the segmentation task.

3.1 The language input

Although linguistic research using text corpora presumably derives its value from the assumption that results are not specific to the corpus being tested, this assumption is rarely tested in any principled fashion. Specifically, the extent to which the results will differ on different materials has not been a subject of inquiry. Here, in the interests of generating hypotheses about how language input affects human cognition, we will focus attention on how the O-B segmentation algorithm will perform on different texts. We will argue that differences in performances can be hypothesized to apply to human segmentation of actual language streams.

3.1.1 Introduction to six text corpora

In order to study several linguistic variables, we use six texts, two each for English, Spanish, and Japanese, and for each pair one written and one spoken text (identified as xxxW and xxxV below). These are all described in detail in Appendix B, so they will be identified only briefly here.

- ENG V: Speech of adults to children, and of adults to other adults, from the Bernstein corpus (Bernstein Ratner 1996) in the CHILDES collection (MacWhinney 1995).
- ENG W: The novel *Of Moon and Sixpence* by W. Somerset Maugham, from the Gutenberg archives of electronic texts.
- SPA V: Recorded Spanish speech from the CHILDES corpus—the datasets contributed by Deuchar & Clark (1992) and Marrero.
- SPA W: A series of short stories in Spanish by Alejandro Dolina, an Argentinian writer, from an Internet site.
- JAP V: Recorded speech of adults to children from the CHILDES corpus (Miyata 1995).
- JAP W: A series of children's picture books about science were made available in hiragana from Prof. Syun Tutiya of Chiba University in Japan.

None of these corpora is extremely large. The biggest (ENG W) is less than 80,000 words and 500,000 bytes. On the other hand, none is smaller than the smallest ENG V corpus, at 4,746 words (13,690 bytes), which we saw in Chapter 2 as the Brent-Cartwright data.

All the corpora were “preprocessed,” to a greater or lesser extent, by a combination of visual inspection and computer processing, to remove characters

which might “mislead” the training process. For instance, in the CHILDES corpora, “xxx” and “yyy” are used to indicate unintelligible speech; these were removed for our purposes. Throughout, punctuation and numerals were removed, and capitals were made lower-case, to induce as much uniformity as possible. Appendix B contains a more detailed description of each of the six corpora and the preprocesses used for each, plus samples of each text.

There is a paradox involved in using samples of actual human language in a computer simulation. Human beings have a boundless ability to abstract across surface variation to discover meaningful categories. Computers do not. So model builders must find a balance between unnaturally controlled transcripts and the chaos of lifelike language. For instance, it is often claimed that speech is more “natural” than written texts, and an important part of this claim is its greater variability:

...[S]pontaneous conversation is characterised by frequent disfluencies, hesitation, filled pauses, false starts and repetitions, all of which compound the problems of segmentation and lexical access, while being a normal feature of spoken language. (Shillcock et al. in press:5)

It is therefore particularly ironic that in order to take advantage of the greater “naturalness” of a “spoken corpus,” the speech must be transcribed into a “written” form, or discrete computer-readable code like standard orthography or phonemes. This is a major blow to the “naturalness,” since it means removing or transforming many aspects of the speech signal, such as variation in pronunciation, prosodic features, and even vocabulary:

The philosophy behind our transcription was to preserve all phonemic distinctions while minimizing the number of subjective judgments [by the tran-

scriber] and the amount of labor required. Accordingly, every word was transcribed the same way every time it occurred, regardless of context.... Syllabic consonants were distinguished from the corresponding non-syllabic consonants.... Onomatopoeia (e.g., *bang*) and interjections (e.g., *uh* and *oh*) were eliminated, for three reasons: They occur in isolation much more frequently than ordinary words, and hence would have skewed the scores toward better performance; their frequency is highly variable from speaker to speaker, so they would have increased the variance in the scores; and there is no standard spelling or pronunciation for many of them, so we could not tell from the orthographic transcript what sound was actually uttered. (Brent & Cartwright 1996:108f.)

Utterances which included a word that was not found in the dictionary were deleted from the corpus... Many of the word types left out are repetitions of baby babble (e.g., *agooo* and *boobababoobaba*), examples of “baby talk” (e.g., *cupsy* and *ticky*), interjections (e.g., *huh* and *ooooh*) as well as onomatopoeia (e.g., *beep* and *boom*). (Christiansen et al. in press:15f.)

This paradox seems unavoidable, however. The variation in language input that we would like to capture for the sake of “reality” may overwhelm the limited ability of a computer model to locate underlying regularities. As computers become able to imitate more aspects of human performance, it will become possible to use input that is closer to what humans hear. For now, the naturalness of language input is necessarily limited by the less-than-human cognitive capacities of the computer models.

3.1.2 Statistical characteristics of corpora

As will be discussed throughout the next chapter, the corpora differed in a variety of ways, many of which are hypothesized to have influenced the results of the training algorithm. Further, we manipulated some of them to create variant corpora to test, by changing the line length, modifying the orthography, and so forth. Rather than assuming, as is often done, that “text” is like an ever-flowing river, one piece indistinguishable from another, we will aim to character-

ize the differences between different corpora.

Each corpus will be characterized statistically by a group of measures that have been found to correlate with one or more aspects of performance in the O-B algorithm:

- Number of symbol types and tokens (characters, or unigrams)
- Number of word tokens
- Average word (token) length, in characters
- Number of word types (omitted in favor of the TTR and G, described below.)
- Number of lines (omitted in favor of the average line length.)
- Average line length, in characters

Corpus quantitative characteristics will naturally vary depending on the particular corpus and the research goal of the project. For instance, although we have deliberately used Japanese texts with word divisions inserted, word spacing is not natural to Japanese writing, so a Japanese corpus description would not normally include a word count. Similarly, we are using a parser which works on one “line” at a time, so the length of a line turns out to be an important determiner of performance; this might not be true in other situations.

The discussion of each corpus group in Chapter 4 will include a short table of summary statistics, and these are collected in a reference table in Appendix C.

3.1.3 TTR and G: measures of repetition

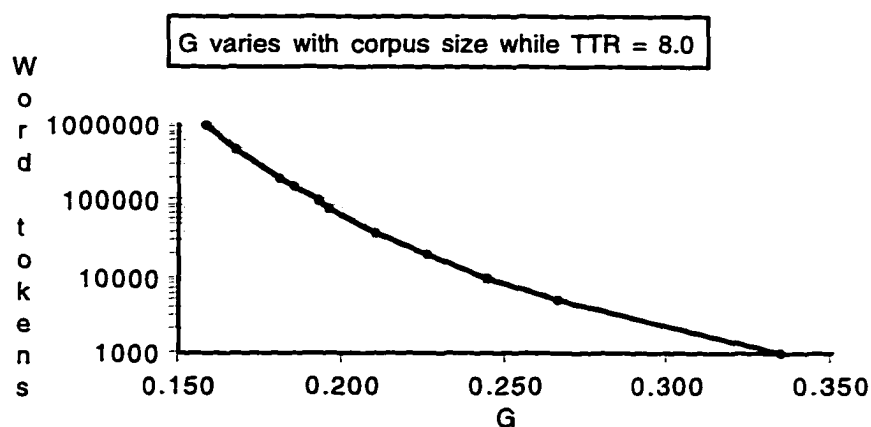
The variety of the vocabulary of a text, or the amount of word repetition, turns out to be an important factor in our results. Typically, a collection of texts

from various sources on various topics will have less repetition, or more word types for the same number of word tokens, than will a single coherent text. The most widely used measure of word repetition is the **Token/Type Ratio (TTR)**.¹

A variation of TTR is **G** which, according to Ejiri & Smith 1993:195), is “a better measure of constraint (size of the domain expressed by the text)”:

$$G = \frac{\log_2 \left(\frac{\#tokens}{\#types} \right)}{\log_2 (\#tokens) - 1}$$

G attempts to normalize for the size of the text (in tokens), so that G increases as TTR increases, but the increase is relatively less as the size of the text increases. This allows for the fact that larger texts are expected to have a higher TTR, irrespective of variety of vocabulary. Looked at another way, TTR is a linear relation between tokens and types while G is a geometric one. The graph below shows how G changes with the size of the corpus, keeping TTR at a fixed level of 8.0:



¹ The Token/Type Ratio is always greater than 1. Sometimes the inverse ratio is used—the Type/Token Ratio, which is always less than 1—but we prefer the >1 version. For instance, 10 tokens and 5 types will have a Token/Type Ratio of 2.0 and a Type/Token Ratio of 0.5; unfortunately, both versions are called TTR.

Other density measures of vocabulary density that have been used in the literature are the ratio of *hapax legomena* (types of frequency 1) to the total number of word types, and the proportion of word types that accounts for one-half the word tokens, when taken in order of highest type frequency. We have not explored these measures here.

3.2 Metrics to evaluate segmentation performance

In Chapter 2, we discussed the widely used measures of recall, precision, and the F-measure, as well as the bracketing metric from parsing applications. These will be used in Chapter 4 to evaluate performance in the O-B algorithm. In this section, we will present additional measures which we have created and which will also appear in Chapter 4.

3.2.1 Evaluation by types of segmentation

We propose an approach to evaluation that is richer in information than simple recall and precision. This method classifies the various kinds of segmentation errors that are made. In order to make comparisons more easily from one run of a corpus to the next, we base our classification on standard words rather than parse words. As shown in the following example, five different types of results are identified:

	P	P	M					C	L	C	W																
Std:	t	o	t	h	e	l	i	g	h	t	h	o	u	s	e	f	o	r	a	t	e	a	p	a	r	t	y
Parse:	t	o	t	h	e	l	i	g	h	t	h	o	u	s	e	f	o	r	a	t	e	a	p	a	r	t	y

W = "word" = the standard word and parse word exactly match

Of course, this is the best situation.

P = "phrase" = two or more standard words make one parse word

Second best, often matching an intuitive unit.

M = "morphs" = one standard word divided into two or more parse words

The standard word *lighthouse* has been broken into two units here that are actually not very intuitive. When a word is broken along morpheme boundaries, we feel it is not too bad, but it is difficult to gauge this mechanically, so we consider all these one-to-many situations as a middle level of error.

C = "crossing error" = a standard word that contains a parse boundary, but does not completely contain the resulting parse words

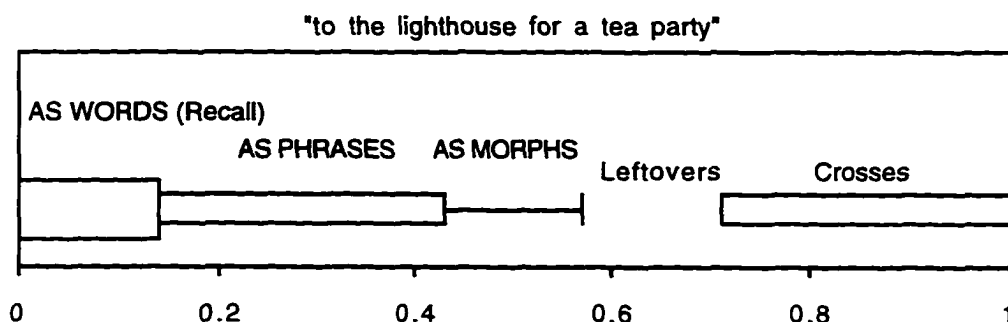
This is the very worst kind of error, chopping words into pieces and then combining the end of one word with the beginning of the next to make a completely non-intuitive unit.

L = "leftover" = any standard word not defined otherwise, being the residue of crossing errors

The score for this example is thus:

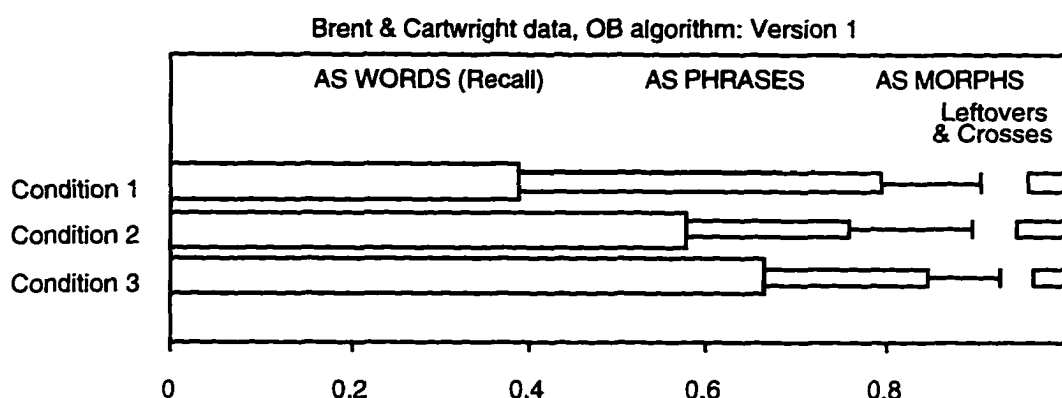
	<u>Standard words</u>	<u>Percentage</u>
words	1	14
phrases	2	29
morphs	1	14
crossing errors	2	29
leftovers	1	14
Total	7	100

We propose to express these percentages graphically, as:



The best condition has the fattest bar, on the left, and the next best a narrower bar, and the minimal acceptable condition is a line. Thus, the full extension from the left is the proportion of standard words which were divided in an acceptable way. The crossing errors are a bar extending from the right edge of the graph, and the empty space between the left line and the right bar is the amount of "leftovers."

Of course, this sample is a very short piece of text. Each chart in Chapter 4 will show these results for a entire corpus. Here we develop this concept by showing the comparable figures for the O-B algorithm's segmentation of the Brent & Cartwright corpus, discussed above, under three different sets of run conditions:



This graph shows a more usual case than the toy example above, with very small amounts for leftovers and crosses. We can easily see here the increasing percentages of the recall figures reported above; these are the same as the “exact words” type in the segmentation classification.

We can see that having a richer picture of the results can make it more difficult to determine which is the better performance. Condition 3 is the best by all measures shown here, but we can see from this graph that Condition 2 is better than Condition 1 with respect to recall, but worse by word+phrases, word+phrases+morphs, and crossings. This presentation gives much more information about the segmentation performance than simple recall.

To add a more balanced measure to the graph, we can combine this “typology graph” with an indicator of the F-measure, which mediates between recall and precision. First, however, we will consider the issue of the random baseline.

3.2.2 Adjusting for the random baseline

The problem of how to account for a varying baseline has not yet been addressed. On the one hand, it does not seem reasonable to compare directly the performance of a simple problem with that of a hard problem, or a very skilled performer and a beginner. Some sort of “handicap” is needed to set the two at some equal starting point. In the case of a segmentation algorithm applied to various corpora, we can evaluate the success of each run by some objective measure and then use this to rank-order the corpora.

But we also need to assess the inherent “segmentability” of each corpus, and this can be done by running a procedure which uses no “intelligence” to

segment the unbroken text into words, but instead simply chooses points at random. Following Brent and Cartwright, we have run such a procedure² many times and averaged the results to get a “random baseline” for segmentation on each corpus. Unsurprisingly, we found that different corpora have different baselines. While it seems clear that the effect of the random segmenter can depend only on the lengths of lines and of words, it is not possible to predict the results of the random procedure for a particular corpus; one just has to run it. Our procedure has been to average the results of 500 random runs.

We will adjust measures for the random baseline by the procedure for calculating the kappa coefficient of agreement, as described in Siegel & Castellan (1988:285):

$$K = \frac{p(x) - p(c)}{1 - p(c)} \quad \begin{array}{l} p(x) = \text{observed probability} \\ p(c) = \text{chance probability} \end{array}$$

Example: Recall after random segmentation = 13%

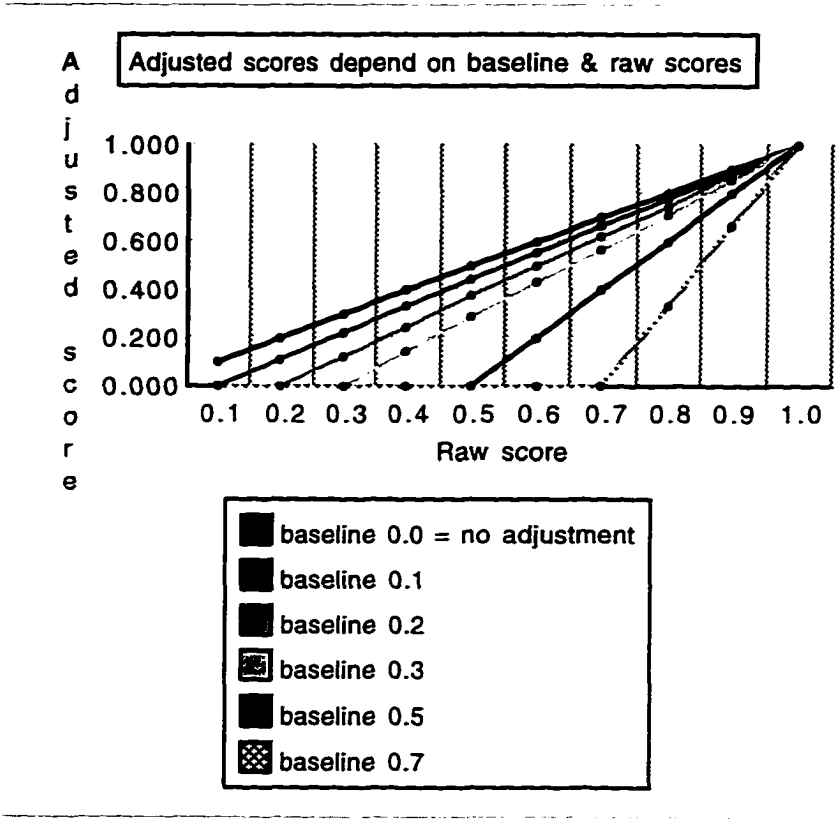
Recall after O-B segmentation = 39%

$$\text{Adjusted recall} = \frac{.39 - .13}{1.0 - .13} = \frac{.26}{.87} = .30$$

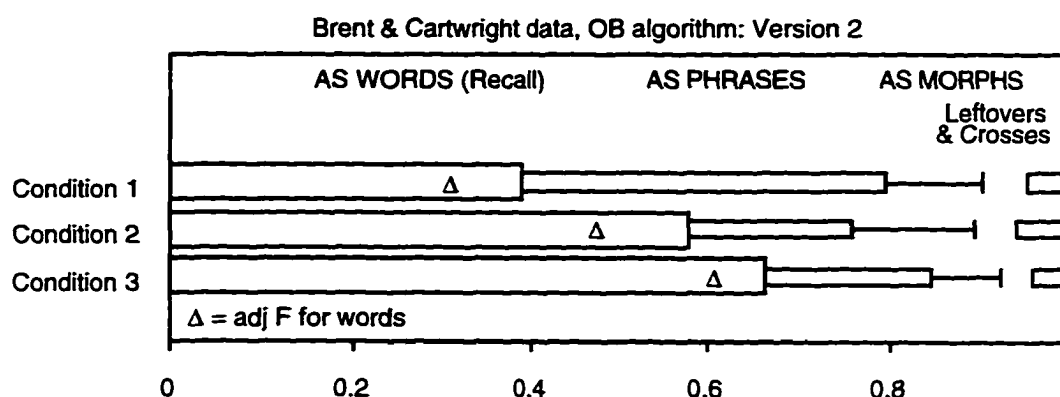
This calculation is similar to a simple subtraction of the random results from the results of the more sophisticated algorithm, thus “netting out” the algorithm’s performance. It lowers all the scores, but it lowers those with a higher random baseline more than those with a low baseline. The following graph gives a sense

² Our procedure was to create an array with a cell for 0 (no boundary) or 1 (boundary) for each character in the corpus. It was initialized as an image of the “collapsed” text, with boundaries only at the end of each line. Then random numbers were generated and each was interpreted as a boundary to follow the *n*th character in the corpus. If that character was already marked for a boundary, no action was taken. If it was not marked for a boundary, one was added. This process continued until the number of boundaries in the image array was the same as the number in the standard version of the corpus. The resulting randomly segmented corpus was then evaluated by the usual procedure, and the results of 500 such cycles of segmentation and evaluation were averaged.

of the effect, showing also that as the raw score is higher, the reduction is somewhat less:



Below we show the typology graph for the Brent-Cartwright data again, this time adding the F-measure for word recall and precision, adjusted for the random baseline for this measure. While normalizing is not strictly necessary in this case, since all the data in this graph is for the same corpus, we introduce it here to illustrate the procedure that will be used extensively in Chapter 4. We show the adjusted F-measure as a triangle on the same scale with the typology bars.



Here is the component information that has been expressed by the triangles:

<u>Baseline word-F = 13</u>	<u>Percentages for words</u>			
	<u>Recall</u>	<u>Precision</u>	<u>Calc F</u>	<u>Adj F</u>
Condition 1	39	41	40	31
Condition 2	58	52	55	48
Condition 3	66	66	66	61

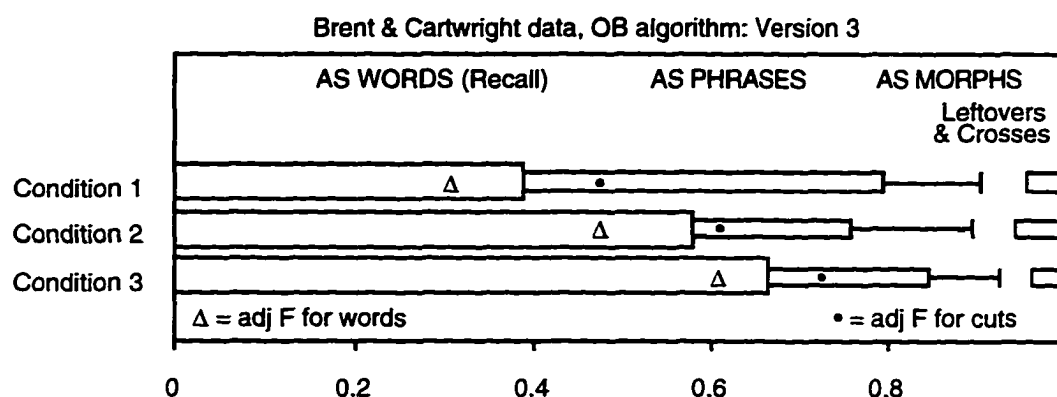
In this case, the normalized F-word measure does not give us much additional information, but we will see that in some cases it does.

3.2.3 Counting divisions rather than words

To reflect the more direct evaluation based on number of correct cuts, we can use the same measures, recall, precision, and F, calling them “cut recall,” “cut precision,” and “cut F” to distinguish them from the measures based on words. Further, we will include the line-breaks in the calculations, although these are “free,” being supplied and not discovered by any algorithm. However, the normalization shows the true performance above the random baseline (which also receives the line-breaks for free), and it gives a measure which is more realistically related to the goal of finding words. For the Brent-Cartwright data, the following values apply (rounded to the nearest 1%):

<u>Baseline cut-F = 50</u>	<u>Percent for cuts only</u>		<u>Percentages for cuts + line-breaks</u>			
	<u>Recall</u>	<u>Precision</u>	<u>Recall</u>	<u>Precision</u>	<u>Calc F</u>	<u>Adj F</u>
Condition 1	58	64	72	76	74	48
Condition 2	78	67	85	77	81	61
Condition 3	81	76	87	86	86	72

We add this normalized F-measure for cuts+lines to our graph as a black circle:



We cannot directly compare the normalized measures (white triangles and black dots) with the typology bars because the typology values are not normalized but are the algorithm's raw measures—the results at face value. However, when we compare the two normalized measures, we see that cut effectiveness, as measured by the normalized cut-F measure, is higher than the normalized word-F measure. This is to be expected, since more than a single correct boundary (here including both cuts and line-breaks) is required to identify a correct word.³ Over an entire corpus, the number of correct boundaries necessary to create N correct words is indeterminate, but will always lie between N and 2N.

³ More than one correct cut is necessary, but generally less than two on average. That is, two correct cuts in a row, with no intervening misses or false alarms, will make one word (2.0 cuts per word). But a series of three correct cuts will create two words (1.5 cuts per word); a series of four cuts will create three words (1.3 cuts per word); and so forth, with the average number of cuts declining as the sequence of good cuts lengthens.

3.2.4 Significance and correlation measures

Since cognitive computer models, or simulations, aim to imitate human behavior, it is natural that measures of performance will be borrowed from the methods used to study actual humans. As in psychological experiments, it is necessary to determine in advance what the criteria of evaluation will be and to structure the experiment accordingly, even if sometimes this results in rather artificial tasks or arbitrary evaluation. For an exact science, purely impressionistic results are not acceptable. When dealing with segmentation, it is usually the case that we evaluate the model (the experiment, or trial) against some standard (the “correct” segmentation).

One problem in applying metrics from psychology to computer studies is that of significance. Standard statistical tests for significance used in psychology use the number of subjects to determine degrees of freedom. But how many “subjects” do we have in a computer run? First of all, most computer trials are completely deterministic: No matter how many times they are run, the answers will be the same if the inputs are the same, unless a randomizing factor is deliberately inserted. (The O-B algorithm is deterministic in this sense.) Secondly, computers don’t get tired, so the number of questions that can be asked of a computer is much greater than for a human subject. Does this change the basis of the statistics?

The approach taken in this work has been to score each run in sections, ten to twenty for each corpus, and to compare the scores for the same section across two or more runs under different conditions. If the direction of the difference between two runs is the same for all, or almost all, sections, then we consider that

the difference is significant. If more than one or two sections differ in the opposite direction, then we conclude that the effect seen in the total for the corpus may simply be due to chance.

Another notion from psychology that is used frequently in cognitive modeling is that of validity and reliability. In the social sciences, a **valid** measure is one that has a good correlation with the behavior or trait it is taken to represent. A **reliable** measure is one that will give more or less the same results again and again. In computational approaches, these two characteristics are harder to distinguish from one another, and they often seem to be used interchangeably. For instance, we may ask whether the cluster “nth” is a good predictor of the end of a word, since we can imagine strings like *tenth in line* but also some like *often thinking*. To test whether “nth” is a **valid** cue for the end of a word, we would compare its occurrence at the end of a word with all its occurrences. To test whether “nth” is a **reliable** cue for end of word, we would do exactly the same thing.

3.3 The O-B algorithm

The computer algorithm used in this study is a very simple exercise in self-organization. It collects cooccurrence statistics from “collapsed text” from which the word-boundary spaces have been removed, and then uses this information to infer where the word boundaries belong. The algorithm used here is borrowed from the work of Olivier (1968). Because Olivier’s work was the first computational approach to this problem, it is widely cited in the literature, but the actual algorithm has recently been either ignored, or disparaged as relying upon an “ad-hoc mechanism” which “fails to suggest general principles of language

learning” (Brent 1996:100). However, Olivier grounds his work solidly in computational theory and discusses its cognitive plausibility as well. We will return to these arguments later.

3.3.1 Building the lexicon

Olivier (1968) segments text into words using a “word grammar”: “a stochastic grammar whose language is finite; it is thus equivalent to a finite dictionary of strings, each with an associated probability” (p. 30). He builds the grammar by an iterative process which builds the dictionary and computes probabilities incrementally, and does both together. The process is as follows:

- a) Create a starting dictionary containing each basic symbol that occurs in the corpus (the “alphabet” of letters or other graphs), and give each a minimal frequency.
- b) Remove all word-spaces from the text, and divide it into sections. Olivier used sections of exactly 480 characters; we used lines or utterances of varying lengths.
- c) Divide the first section into words, using only those words which are found in the current dictionary, and selecting the “best” division by applying some measure. (We will discuss the parsing process separately.)
- d) Update the frequency counts in the dictionary by adding one count for each occurrence of a word in the section just parsed.
- e) Augment the dictionary list by considering each pair of contiguous words in the section just parsed as a potential new word. For each such pair that is not already in the dictionary, add it with a minimal frequency.

- f) Repeat Steps (c), (d), and (e) to the end of the text, each time using the dictionary as just modified.

The effect of this procedure is illustrated by Olivier:

... if a section included the string 'abcdefgh' parsed as four words 'a bc de fgh' the three new words 'abc', 'bcde', and 'defgh' would be added to the dictionary... A principal advantage of this system for adding words to the dictionary is that it swells the dictionary with large numbers of words rather quickly. (Olivier 1968:67)

Consequently, a large part of the dictionary will consist of such pairs which have been added as "potential words," but are never actually used in a parse. Olivier had to be concerned with deletions as well as additions to the dictionary, since he had a very small computer, but we can simply allow the dictionary to expand as it will.

The operation of the O-B algorithm is not easy to show in a small space, since its success depends upon seeing repetitions of the same chunk in many different contexts. However, some sense can be gained by the following example—a corpus consisting of these four "utterances" repeated four times:

a big pet a cute dog
the pet the dog
a dog a pet
the big dog the cute pet

Here is the parsed version as it emerges line by line during training (lines are numbered for reference in the following discussion). On the left is the utterance as it is parsed based on the current lexicon. On the right are the new word candidates added to the lexicon as the result of this parse, by combining contiguous

pairs. Those word candidates which later appear as words in at least one parse are underscored.

0	-	a b c d e f g h i j k l m n o p q r s t u ...
1	a b i g p e t a c u t e d o g	ab <u>bi</u> ig gp <u>pe</u> et ta ac <u>cu</u> ut <u>te</u> <u>ed</u> do <u>og</u>
2	t h e p e t t h e d o g	th <u>he</u> epe <u>pet</u> tt th hed edog
3	a d o g a p e t	ad <u>dog</u> oga <u>apet</u>
4	t h e b i g d o g t h e c u t e p e t	<u>the</u> hebi <u>big</u> bdog dogt the hecu <u>cute</u> tepet
5	a b i g p e t a c u t e d o g	<u>abig</u> bigpet <u>peta</u> acute <u>cutedog</u>
6	t h e p e t t h e d o g	<u>thepet</u> petthe <u>thedog</u>
7	a d o g a p e t	adog <u>dogapet</u>
8	t h e b i g d o g t h e c u t e p e t	<u>thebig</u> bigdog <u>dogthe</u> thecute <u>cutepet</u>
9	a b i g p e t a c u t e d o g	<u>abigpeta</u> <u>petacutedog</u>
10	t h e p e t t h e d o g	<u>thepetthedog</u>
11	a d o g a p e t	<u>adogapet</u>
12	t h e b i g d o g t h e c u t e p e t	<u>thebigdogthe</u> <u>dogthecutepet</u>
13	a b i g p e t a c u t e d o g	<u>abigpetacutedog</u>
14	t h e p e t t h e d o g	-
15	a d o g a p e t	-
16	t h e b i g d o g t h e c u t e p e t	<u>thebigdogthecutepet</u>

The algorithm begins (line 1) with single letters as “words,” and then (#2) finds some pairs that reoccur in the second line, so their higher frequency in the lexicon and greater length promote them to “words” in the parse of the second line. Since longer forms are always preferred over their separate components, the algorithm will create bigger and bigger units, up to the point where the new candidate is never repeated. In this toy corpus, everything is repeated, so ultimately each utterance will be a lexical item.

As discussed in Chapter 2, a probabilistic language model consists of the lexicon (a list of strings) and the probabilities assigned to each lexical item. In the O-B algorithm, the probability of a lexical item is the maximum likelihood

estimate (MLE)—its current count divided by the sum of all such counts in the dictionary. The procedure for determining these frequency counts is another difference between Olivier's procedure and the O-B algorithm. After parsing a section, Olivier increments the count in the lexicon for each word token in the parse, and also for each occurrence of a pair of contiguous words in the parse.

But Olivier says about his choice of a method to add new words to the dictionary:

The main justifications for this procedure are first, that it is simple; and second, that... a word of any length can find its way into the dictionary reasonably quickly... There are many reasonable alternative systems and we will make no attempt to decide among them. (1968:67f.)

So we used his pairs-plus-words approach ("P+W" below), but also tried two other variations: tallying only pairs and only words. Across a large number of trials, we found that tallying both pairs and words produces results virtually indistinguishable from tallying words only, and that tallying pairs alone is worse than either of them. Therefore, we will discuss only the word-tally version.

Details of all three procedures can be found in Appendix A.

3.3.2 The parsing optimization

The selection of the best parse for each text section uses not only the current list of lexical "words," but also each word's current probability, defined as its current count divided by the sum of all such counts in the dictionary. Both the Olivier and O-B algorithms use a "maximum-likelihood" segmentation (MLE = probability estimate based on frequency in the corpus) to parse the text sections into words. For each word token, the probability of its word type is multiplied

into the product, and the division which maximizes the product of the probabilities of the component words is chosen.

In order to select the best parse, a search through the possible parses is necessary. Because we wanted to do a large number of trials, using different corpora and conditions, speed was an important factor. After experimenting with an exhaustive search procedure, which took from several hours to several days to execute, depending on the number of conditions involved in the scoring algorithm, we selected a compromise technique which compared well with the exhaustive runs and required less than an hour in the most complex case. (We later noticed that Olivier had used the same optimization.)

This shortcut technique proceeds by requiring that a parse for each substring must be built upon the parse for some previous substring. The space of possible parses for the string $\{1, i+1\}$ can be limited to continuations of the i parses already identified. This considerably reduces the number of parses that must be explicitly evaluated, since for an n -symbol string there are 2^{n-1} possible parses, but the shortcut search, using the results of substring parses, only checks $1+2+3+4 + \dots + n = n(n+1)/2$ of them:

Length of string (in chars.)	4	5	6	8	10	20	40
Number of parses considered:							
Exhaustive	8	16	32	128	512	524,288	549,755,813,888
Optimized	10	15	21	36	55	210	820

The difference is that the optimized search evaluates each partial parse and then uses that evaluation to bypass a number of its extensions; the exhaustive search simply evaluates all full parses. When the strings are very short, the extra work

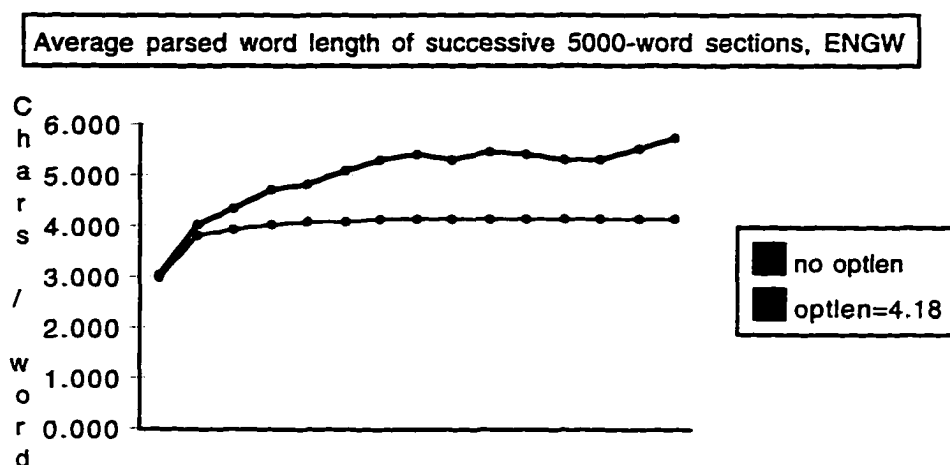
of the partial evaluations will not pay off, so the optimized method will be a longer path.⁴

As a test of whether the optimized technique yields the same results, we computed the optimized parse both from the beginning of the string forwards, and from the end backwards, and found that the score of both paths is the same. (The parses themselves may be very slightly different, due to different defaults in case of a tie.) The details of these procedures are described and illustrated in Appendix A.

3.3.3 Controlling word length with the *optlen* parameter

The Olivier procedure by its nature tends to prefer longer and longer parse units as the dictionary grows, since each new word creates even longer word pairs which are entered in the dictionary as candidates for future words. The following figure shows that as the algorithm proceeds through the text unconstrained, the average parsed word tends to get longer unless it is constrained. To prevent overlong parse units the O-B algorithm adds an “optimum-length” parameter, which is described in this section.

⁴ For instance, a string “a b” will have just two possible parses in the exhaustive method: a+b, ab. The optimized method, however, will evaluate three substrings, including one “partial parse”: a, a+b, ab.



The maximum-likelihood parse criterion greatly favors a longer unit over two shorter ones. This is because the probabilities of all the word tokens in a parse are multiplied together, so when the (fractional) terms are roughly equal, two multiplied together will always be smaller than one. For instance, here is an example from a corpus of 74,951 word tokens:

		<u>Occurrences</u>	<u>Parse Likelihood</u>
<i>of</i>		1839 times	$1839 \div 74951 = 0.0245$
<i>the</i>		3184 times	$3184 \div 74951 = 0.0424$
<i>of + the</i>	(2 words)	(product)	$0.0245 \times 0.0424 = 0.0010$
<i>ofthe</i>	(1 word)	393 times	$393 \div 74951 = 0.0052$ **

This example demonstrates that, on the pure mathematics of it, *ofthe* as a single word will outscore the standard two-word version. The word *the* would have to represent about 21% of the corpus rather than 4% in order for the two-word combination to win ($.0245 \times .22 = .0053$). In fact, nearly all combinations will outscore their components under maximum likelihood procedures. As Olivier says, "the estimated frequency of a word is much larger than the product of the estimated frequencies of its parts" (p. 59).

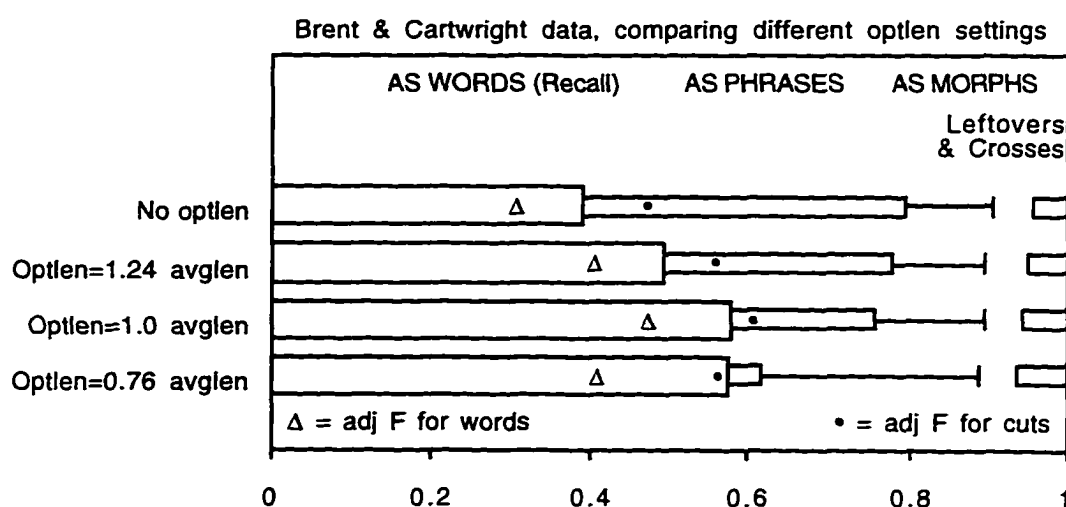
A typical run thus tends to under-divide the text, creating words which are on average longer than those in the standard. To correct this, the O-B algorithm uses an **optimum-length parameter**—a target for the average word token length. Using this value, parse evaluation proceeds not only by maximum likelihood, but also by how closely the average word length of that parse candidate approaches the “**optlen**.” By varying “optlen,” the number of divisions produced by the algorithm can be fairly well controlled. In general, setting optlen equal to the average word length (in characters) of the standard version of the text will produce a number of parsed words roughly equal to the number of word tokens in the standard.

To illustrate, we show the same “dog” corpus used earlier as it is parsed using an optlen equal to its average word token length of 2.7 characters:

```
a b i g p e t a c u t e d o g
t h e p e t t h e d o g
a d o g a p e t
t h e b i g d o g t h e c u t e p e t
a b i g p e t a c u t e d o g
t h e p e t t h e d o g
a d o g a p e t
t h e b i g d o g t h e c u t e p e t
a b i g p e t a c u t e d o g
t h e p e t t h e d o g
a d o g a p e t
t h e b i g d o g t h e c u t e p e t
a b i g p e t a c u t e d o g
t h e p e t t h e d o g
a d o g a p e t
t h e b i g d o g t h e c u t e p e t
```

It is apparent that the words do not keep getting longer, since the optimum length calculation penalizes parses with an average word length longer than optlen, so that parses using shorter words win the competition.

Forcing shorter words by using decreasing values of optlen tends to improve performance up to a point, but the best performance is generally achieved by a value of optlen close to the standard average word length. The figure below shows the corpus presented in Section 3.2.1 above, where “no optlen” was called Condition 1, and “optlen=1.0 avglen” was Condition 2.⁵



Here we can see that word recall improves as the optlen value decreases, but the normalized F-word measure gives us additional information about the performance of these runs. It shows that, despite the similar recall values for the third and fourth runs, the performance above the baseline is much greater for the third run, indicating that, as we would expect, the best optlen value is that of the

⁵ Three values are standard here: the average value is that same as the average word token length in the corpus itself. The large value is 1.24 times the average length, and the small value is 0.76 times the average length. These values were initially assigned rather arbitrarily and then retained, but there is no reason to think that other representatives of a range would not serve equally well. Also, it is possible that using the standard average word length is not the best way to get comparable results across different corpora.

standard average word length.

How does the optlen parameter work to affect performance? The optlen constraint controls the number of word tokens created and thus the average word length in the parsed text, and these in turn have an influence on the number of word types. Even the following tiny parses, one and a half lines taken from a series of runs of another corpus, illustrate the influence of optlen on average token length and number of tokens:

Standard (average token length 4.125; 16 tokens, 13 types, TTR 1.23):

she went to the house ascended the few steps that led to the
verandah and entered

Optlen = 2.18 (average token length 2.129; 31 tokens, 25 types, TTR 1.24)

she w ent to the h ous e as cend ed t he few s t ep s that l ed to the
ver and a h and ent er ed d

Optlen = 3.18 (average token length 3.14; 21 tokens, 17 types, TTR 1.23):

she went to the house as cend ed the few steps that led to the
ver andah and ent er ed

Optlen = 4.18 (average token length 4.4; 15 tokens, 14 types, TTR 1.07):

shewent tothe house as cend edthe few steps that led tothe
ver andahand enter ed

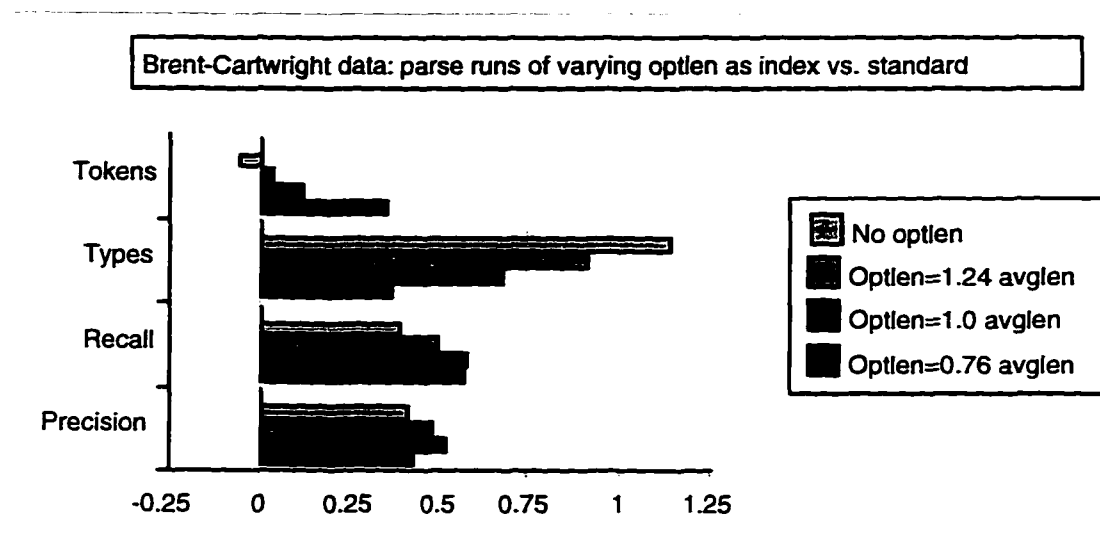
Optlen = 5.18 (average token length 5.5; 12 tokens, 12 types, TTR 1.00):

shewent tothehouse as cend edthe few steps thatled tothe
verandah and entered

We can also see from these samples that TTR increases as tokens (and types) get shorter. Over a larger sample, this increase in repetition rate will result in fewer types overall as token length decreases and number of tokens increases. Generally, there is a trading relation between the number of tokens and the number of

types. This can be understood if we consider that a dictionary consisting mainly of two-character words would be fairly small, since the number of two-character combinations is limited. As the length of words gets longer, there are potentially more of them, and the chance of "reusing" the same types gets smaller.

Returning to the Brent-Cartwright corpus, the following chart shows the relation of tokens and types for the same corpus and runs shown in the figure above. As optlen increases across the four runs, there is a monotonic increase in number of tokens (from fewer than the standard to about 35% more) and a monotonic decrease in number of types (from more than twice as many types as in the standard to about 35% more). However, the performance indicators—recall and precision—are not monotonic. The precise way that number of tokens and types affect performance is not possible to state, since that is not just a question of the number of boundaries chosen but of finding the appropriate locations. In the following figure, the bars represent an index vs. the standard; that is, 0.25 represents a 25% increase over the number in the standard segmentation. Although there are obviously more tokens than types by actual count, types vary much more than tokens in terms of this index measure.

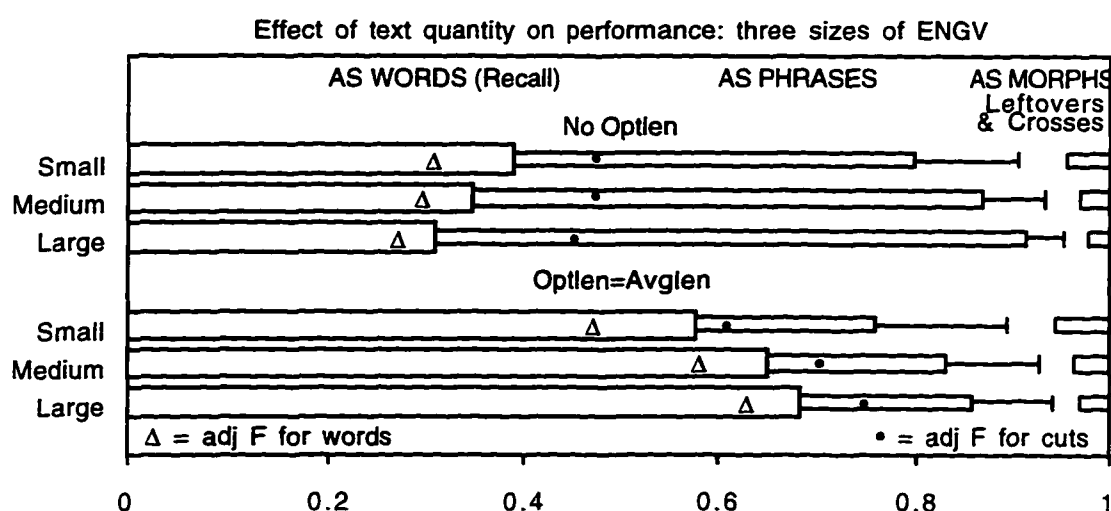


It is also possible to see from this figure that there is an inverse connection between two ratios: When there are more word tokens in the parsed version than in the standard, the word recall will be greater than precision. When there are fewer tokens in the parsed version, word precision will be greater than recall, as is the case only for the no-optlen run. It should also be clear that when a parse creates longer words, it will also tend to create more "phrases" (clumps of standard words). On the other hand, when the average parsed word is close to the standard in length, this will tend to be the result of a length distribution similar to the standard, which in turn will tend to facilitate more matching words.

Looking again at the typology graph two pages back, it is apparent that there is a trading relationship between the number of standard words which are clumped ("phrases") and those which are divided ("morphs"). This is of course due to the shift in the number of divisions which are made. The more cuts, the more word tokens, and the smaller the average segmented unit ("word"). When no length goal is provided, the words tend to get longer, and more clumping

than dividing takes place. At the other extreme, when the length goal is very small, the words get shorter and shorter, clumping declines and the incidence of (over)dividing increases. So the recall rate, which is the same as the "exact word" percentage in this more complex picture, is not a perfect evaluator. As recall goes up, relatively more standard words are divided than clumped, whereas we would prefer to have more clumping than dividing. We are forced to choose between a large number of exact words or a large number of words+phrases. The best we can say here is that this range of results perhaps mirrors the cognitive problem of the child, who also may be able to range across units of varying lengths and experiment with the results, using other dimensions of reality to determine the final choice, or even allowing several different segmentations to coexist for different purposes, much as we adults may be aware of morpheme, word, and phrase levels of cohesion.

The Brent-Cartwright corpus is a very small one. A comparison of larger corpora will show that the performance differential between the no-optlen and optlen runs becomes more pronounced the longer the training process continues:



We see that when no optlen is used, and the dictionary is free to grow longer “words,” the word matches get worse, but the total of words and phrases and the crossings get better. But when the optlen constraint is used, it prevents excessive clumping, allowing words to be better recognized, but also interferes to some extent with the process of learning, so that more decisions appear in the less desirable “morph,” leftover, and crossing categories. However, all of the adjusted F measures improve in the constrained condition.

Unfortunately, the particular way that the optlen value is used in the parse calculations (described in detail in Appendix A) has introduced some undesirable complexity. Using a compound calculation combining maximum likelihood and some function of optlen, rather than just the simple MLE criterion as in the no-optlen situation, we were not able to devise a scoring system for which the shortcut version of the parse algorithm was exactly equivalent to an exhaustive search. Thus, while there were no cases in the optlen version of the algorithm for which the forward and backward parse yielded different results, when optlen calculations are included the number of such cases ranges from a few to almost all the parses. The proportion of parses which are different forward and backward varies depending on the corpus, and it is also larger for runs using lower optlen values.⁶ Until a more effective calculation is found, we will have to be satisfied with empirical evidence of the good results that the O-B algorithm achieves rather than upon a theoretical justification.

⁶ Since the optlen calculation (see Appendix A) uses the difference between the average length of the parsed word, which tends to get large, and the optlen value, it is to be expected that smaller optlen will lead to a greater effect.

3.3.4 Do utterance boundaries contribute information?

A further difference between the O-B procedure and Olivier's is the bounding of a section of text to submit for parsing. Olivier divided his "collapsed" corpus (that is, after removing spaces) into arbitrary chunks of exactly 480 characters, ignoring sentence, utterance, and even word boundaries in so dividing it. That is, a section for parsing might well end in the middle of a standard word. Olivier considered that whatever difficulties this introduced would be ironed out as the algorithm proceeded and, in any case, were not unlike any number of anomalies that complicate real language processing.

The O-B algorithm, instead of counting characters to form a parse unit, uses a line-end character as a signal to begin parsing. Each corpus studied is divided into lines of fewer than 150 characters, with the principle of division differing somewhat from corpus to corpus. Some have lines based on utterances (the CHILDES corpora), and others are correlated with "lines" originating as page-width (or screen-width) lengths of printed text, which are not necessarily correlated with utterance or sentence boundaries, though they do respect word and paragraph divisions.

It is reasonable to ask, to what extent is performance enhanced by having parse sections aligned with linguistic units? Is Olivier correct that it doesn't make much difference? To answer these questions, we ran a couple of tests.

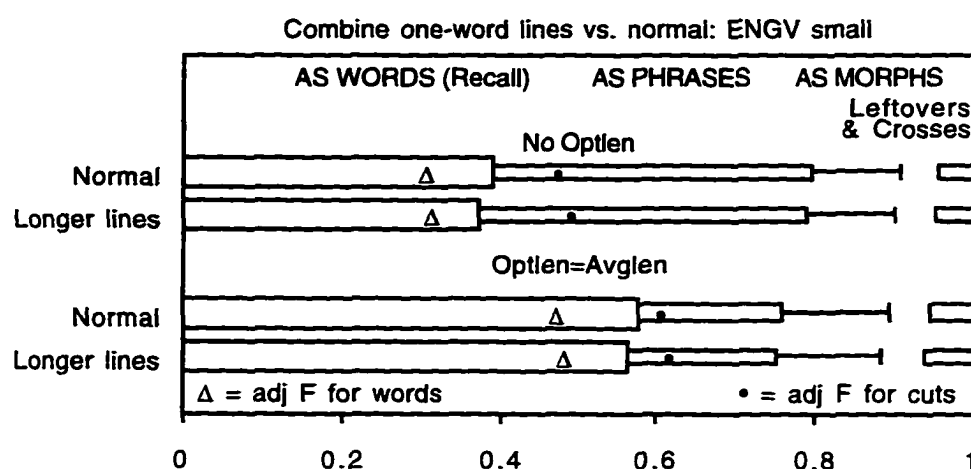
One way to measure the contribution of utterance boundaries is to reduce the number of boundaries in a given text and then compare the results before and after such reduction. In the small text we have been looking at, there are 364 one-word lines (utterances) out of 1533 lines total. We removed the line-end

following each of these very short utterances, joining each one-word utterance with the following utterance. Thus, sequences like those on the left were transformed as shown on the right:

<u>Original</u>	<u>Lines Combined</u>
9 TINK It wll kAm Qt	9 TINK It wll kAm Qt
lEts si	lEts si
y&	y& pUI It Qt
pUI It Qt	WAts It
WAts It	IUk IUk
IUk	WAts D&t
IUk	
WAts D&t	

This resulted in a corpus with only 1240 lines⁷ and an increase in the average line length from 8.9 to 11.0 characters. Despite this seemingly substantial change, there is only a very small difference, as shown in the following figure. There is a slightly less accurate segmentation for the modified corpus, across all the categorical measures, but slightly higher scores as adjusted for the random baseline, which is substantially lower for the modified corpus:

⁷ Some of the one-word lines were contiguous, so not all were joined to a line of more than one word.



Corpus and Version	#Tokens		#Types	Avglen(chs)		Repetition		F% baselines	
	Char	Word	Char	Line	Word	TTR	G	Words	Cuts
ENGv small (normal)	13,690	4,746	50	8.9	2.88	9.7	0.292	13.4	50.2
ENGv longer lines	13,690	4,746	50	11.0	2.88	9.7	0.292	10.2	46.9

For another trial, we created another version of the same corpus with line-ends mechanically shifted one word to the left:

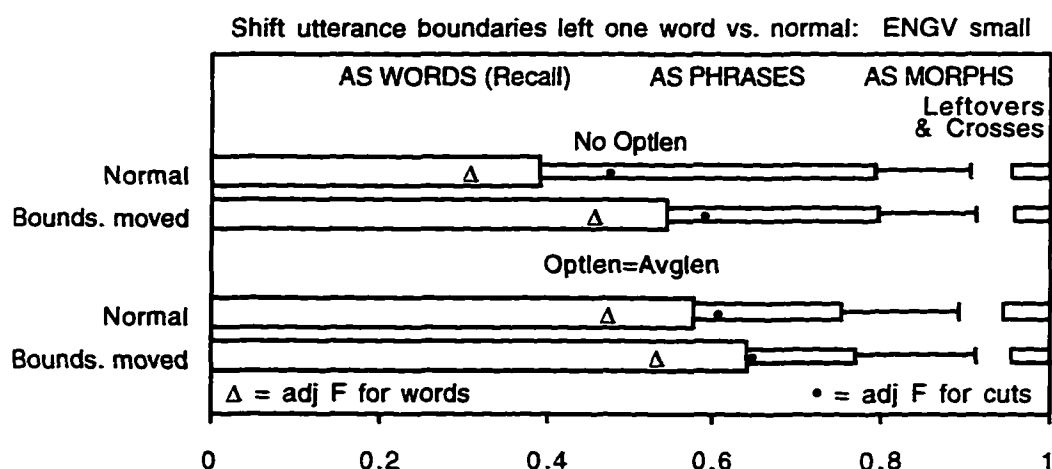
Original

yu want tu si D6 bUk
lUk D*z 6 b7 wIT hlz h&t
&nd 6 dOgi
yu want tu lUk &t Dls
lUk &t Dls

Line Boundaries Moved

want tu si D6 bUk lUk
D*z 6 b7 wIT hlz h&t &nd
6 dOgi yu
want tu lUk &t Dls lUk
&t Dls h&v

The results of the O-B algorithm on this modified corpus reveal that, if anything, the modified corpus produced better results—the misaligned boundaries seem to actually help the process of segmentation. As can be seen in the figure below, the improvement in performance is fairly dramatic in the no-optlen case, and more modest in the optlen run.



Corpus and Version	#Tokens		#Types	Avglen(chs)		Repetition		F% baselines	
	Char	Word	Char	Line	Word	TTR	G	Words	Cuts
ENG V small (normal)	13,690	4,746	50	8.9	2.88	9.7	0.292	13.4	50.2
ENG V bounds. moved	13,688	4,745	50	8.9	2.88	9.7	0.292	13.8	50.2

How can we understand this seemingly counterintuitive result? Note that the improvement lies mainly in the increase of words which are well-bounded, at the expense of phrases. We suspect that moving the utterance boundary serves to break up frequently-occurring word pairs, which tended to be identified as a single unit in the first version, but are then diagnosed as separate words in the modified corpus. This hypothesis assumes that there are fewer frequent word pairs across a word boundary than at the beginning of an utterance. It would also explain why there is not so great improvement in the number of words found when optlen is used, since then more words are found in both corpora anyway: The length goal prevented the parser from treating the first two words of utterances as a unit in the first place.

These trials show that utterance boundaries are not the uncomplicated source of information that might be thought. Clearly, the utterance boundary gives unambiguous information about one boundary of a word, but this informa-

tion is only one piece of the puzzle, and must be interpreted in the context of a number of other facts and influences.

The second issue, of whether word boundaries make a difference, is impossible to test in the O-B framework, since line boundaries are implicitly also word boundaries. That is, the original text with spaces is used to evaluate the parsed version, into which spaces have been reinserted, but by definition line boundaries are identical. Therefore, moving a line boundary implicitly also moves the word boundary, and we cannot create two versions of the “same” corpus for comparison. If we assume that lines represent utterances, and that utterances are objectively present in the speech signal (defined by pauses, intonation, etc.), then it makes sense to assume that an utterance boundary is always a word boundary.⁸

3.3.5 Training, testing, and held-out data

When working with a probabilistic language model, a distinction is typically made between a **training run** and a **test run**. The practice in the field of computational linguistics is to train the model (the lexical entries and their probabilities) on a body of text, and then test the model on a different text. In order to have a similar but not identical test corpus, generally about 10% of a large corpus is reserved for testing, and the balance serves as the training corpus.

In the present study, most runs are what would be called training runs: The lexicon is built as the run progresses, and each successive parse during the run references a slightly different version of the lexicon. When all text has been

⁸ This section has dealt with two differences between Olivier (1968) and the O-B algorithm: Olivier used parse sections of exactly the same length, which resulted in boundaries that didn't always correspond to word boundaries. A third difference is that his parse sections were much longer; the effect of length of parse span will be discussed in Section 4.1.2 in the next chapter. A complete comparison of the Olivier and O-B algorithms is found in Appendix A.

exhausted, the lexicon reaches its end state, and training is complete. This final lexicon can then be input to a “parse-only” or test run, during which all parses reference the same unchanging lexicon.⁹ Lexical entries are neither added nor removed in the course of the test run, nor are any frequencies altered from their beginning state. We will term such static runs **dictin** runs (“dictionary is input”).

Of course, as the process of building the lexicon goes forward in a training run, the parsing generally gets better as the number of lexical possibilities increases and the frequency scores represent more experience. The early parses of a training run are pretty primitive, while the final parses are much more successful, but the overall score of a training run reflects the average of all parses, not just the final state. A dictin run, on the other hand, parses the entire corpus using the final lexicon, so it generally scores much higher than the training run which created the lexicon (though not necessarily higher than the final training parses).

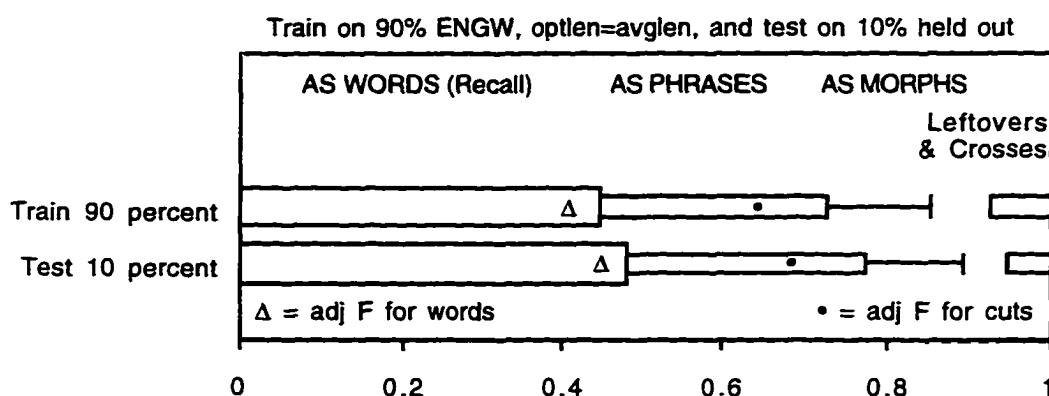
However, this tendency to improve is limited by the tendency for lexical entries to get longer. In a dictin run, as in training runs, the parsing process may be constrained by the *optlen* parameter or not. When no *optlen* parameter is used in the dictin parsing process, the result is only slightly better than the training run. Thus, there are two variables in a dictin run: the lexicon used (that is, the parameters used in the training run which produced it—*optlen* or not, etc.), and the *optlen* value during the parsing run itself. The results called “Condition 3” in Section 3.2.1 were the results of a dictin run with *optlen*=*avglen*, where the

⁹ In practice, the frequency-1 pairs (word candidates that were never used in a parse) are removed from the lexicon before it is input to a parse-only run. This saves a lot of run time, and in fact seems to produce a somewhat better parse.

dictionary was trained using optlen equal to the standard word average length for that corpus, and then the dictionary was input to a dictin parsing run using the same optlen value.

Unlike most research in computational linguistics, we have not followed the practice of training and testing on different pieces of text. Our goal here is to show that there is substantial information about segmentation in the cooccurrence frequency of language, and we are using a self-organizing algorithm to demonstrate that. The results of the self-organizing process can be measured directly by comparing them with the standard word division. We need not test the resulting model against another text to evaluate it. If we can demonstrate the self-organizing character of language, it is superfluous to show that the resulting words can also be located in a new text (or an unseen portion of the training text).

However, just to be sure there are no weasels lurking in the woodpile, we did one run with the standard protocol: we trained using 90% of the ENGW text, and then did a dictin run on the remaining 10% (from the end of the file). Considering that dictin runs by their nature do better than training runs, it is not surprising that in this case also, the test run did better than the training run on all measures, even though it contained previously unseen text. So we can be reassured that the learning taking place is not based on simply “memorizing” a text.



3.3.6 Cognitive implications of the O-B algorithm

Our goal in exploring this computer algorithm is to demonstrate that there is information available in the cooccurrence frequencies of language which could materially assist the infant in initial segmentation. We are proposing that this information could be helpful in forming hypotheses about appropriate units, which would then be tested against other knowledge sources, such as stress, single-word utterances, etc.

The focus of this study is on explaining a human mental process—language learning—rather than the engineering goal of producing usable results in practical applications. Human infants are in fact not “training,” but are engaged in an active learning process on many fronts at once, not only trying to figure out the form-meaning pairs in the speech stream, but also discovering that blocks are hard and pillows are soft, and experimenting with how best to get people to pick them up (a smile? a gurgle? a word?), and so forth. Infants will subject their discoveries about language to a variety of further tests. We want to show that frequency properties of speech contain useful cues which could assist this process.

The use of the optlen parameter to constrain the search may at first appear

to be an *a priori* source of information that the infant could not possess. Of course, infants do not know the average length (in any units) of the standard word. On the other hand, if the infants' search for meaning-based segments proceeds from small to large, testing hypotheses as they goes along, it seems reasonable that there will be a middle ground—larger than syllables and smaller than phrases—where hypotheses will be relatively successful. Certainly the alternative “no-optlen” case, where the length of hypotheses grows without bound, does not seem to be cognitively very plausible. We will pursue these and other cognitive issues in more detail in Chapter 6.

Chapter 4. The O-B Algorithm: Segmentation Results

Here we present the results of the application of the O-B algorithm to many different corpora and corpus variants, in order to gather evidence about what characteristics of the language input affect segmentation, at least in this computer simulation but perhaps also for the language-learning infant. We methodically investigate the quantity of language (corpus size), the representational code, the qualitative nature of the language (written vs. spoken, spoken to children vs. spoken to adults), and the language itself (English, Spanish, or Japanese).

4.1 Quantity of text matters

4.1.1 Quantity of text for training

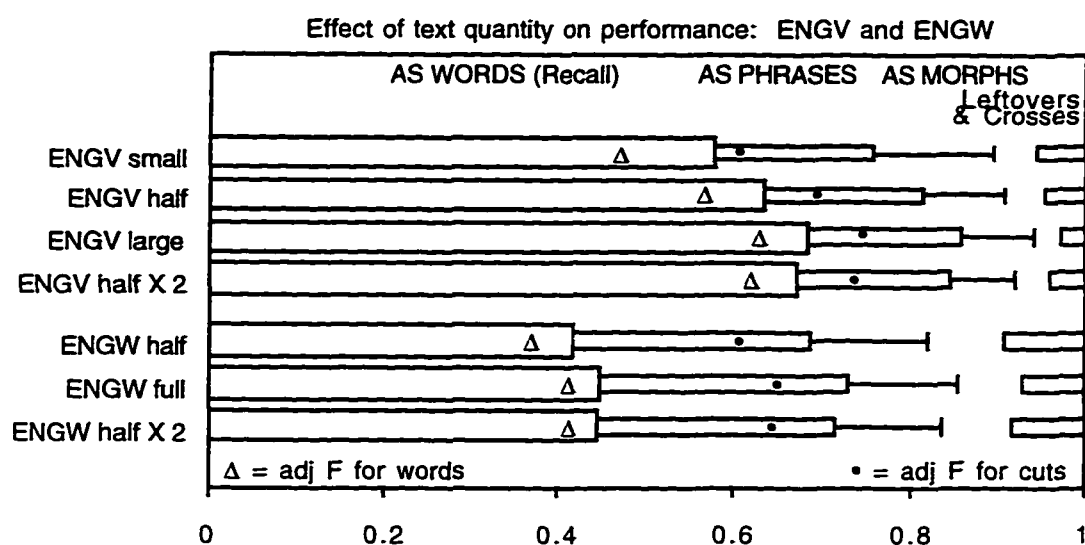
It is a truism that the quantity of text available for training affects the results of the training—more text means better results. This assumption comes from information theory, which treats text as ergodic, or homogeneous, to which the Law of Large Numbers applies, leading to the assumption that whatever inconsistencies are observed in language will be netted out in the long run. That is, statistical or quantitative patterns of language (relative frequencies of lexicon, syntax, etc.) can be observed to any desired degree of precision given only a large enough sample.

Intuitively, also, it seems reasonable that more rather than less language would be better, no matter what information source or learning method is hypothesized. In the case of our hypothesis that frequency information is useful, it

would certainly seem to be a corollary that the more experience with language, the better.

In order to demonstrate this with a computer algorithm, we need to hold constant a number of other factors which might influence performance. The table of corpus characteristics which accompanies each typology graph will enable the reader to verify that whatever can be controlled is set equal in all the corpora we compare.

We wanted to run a small and a large version of the English texts, but in order to guard against a possible lack of homogeneity, we created the "half" versions by removing every second line from the full text. Then we created a pseudo-full-corpus ("half X 2") by concatenating two copies of the half corpus. The "small" ENGV is the familiar Brent-Cartwright data, which consists of the beginning of the first session for each of the nine mother-child dyads in the full corpus:



Corpus and Version	#Tokens		#Types	Avglen(chs)		Repetition		F% baselines	
	Char	Word	Char	Line	Word	TTR	G	Words	Cuts
ENG V small	13,690	4,746	50	8.9	2.88	9.7	0.292	13.4	50.2
ENG V half	48,379	16,863	50	9.9	2.87	15.8	0.306	12.5	48.6
ENG V large	95,809	33,399	50	9.8	2.87	25.3	0.332	12.6	48.7
ENG V half X 2	96,758	33,726	50	9.9	2.87	31.7	0.355	12.5	48.6
ENG W half	156,821	37,475	26	42.2	4.18	7.6	0.206	4.2	29.8
ENG W full	313,321	74,951	26	42.1	4.18	10.6	0.224	4.2	29.8
ENG W half X 2	313,642	74,950	26	42.2	4.18	15.2	0.258	4.2	29.8

The plot shows that for both corpora the larger original version performed better on all measures than the half-version, but the doubled version did not perform quite as well as its same-size original. Thus we see that sheer size is not the only factor, but by moving away from the natural sequencing of utterances we have also somehow slightly reduced its “learnability.” In addition, the ENG W corpus performed much less well than the ENG V full corpus, though the table of corpus characteristics makes clear that ENG W is at least twice as large as ENG V, by any measure.

In these comparisons, what is “doing the work”? Is it the mere fact of more

text (more word or character tokens) or is it the higher amount of repetitiveness (TTR or G) that accounts for higher performance? This is another way of saying, can we attribute the observed effect of more text solely, or primarily, to the repetition of words and word combinations, or are there other factors that come along with text quantity that we are not aware of? Wouldn't a measure of repetitiveness, such as TTR or G, be a better predictor of performance than size? Since a minimum amount of repetitiveness is necessary for learning to occur, if the additional training text does not serve to increase the TTR significantly (by adding more tokens of the types already seen), it will not be "more text" in any useful sense. Weischedel, Meteer, Schwartz, Ramshaw, & Palmucci (1993:363f.) cite their rule of thumb that training of trigrams of a small symbol set (47 tags, in their case) needs only enough data to give an average of ten instances of each trigram type. If we extend this heuristic to the word TTR in our corpora, that would seem to be a good estimate.

The complexity of the interaction between corpus and algorithm are apparent in this comparison. First of all, TTR cannot be said to account for all of this difference, since the smallest ENGV corpus is both smaller and less repetitive (by TTR measure) than either of the large ENGW texts, yet it does better on most measures. The G measure correlates better with the performance here, and may thus be a better guide to the repetitiveness.¹ (Remember that G is less affected by absolute corpus size than TTR.)

¹ Pearson correlation coefficient for G and word recall, $r(5) = 0.958$, $p < .01$; for TTR and word recall, $r(5) = 0.77$, $p < .05$.

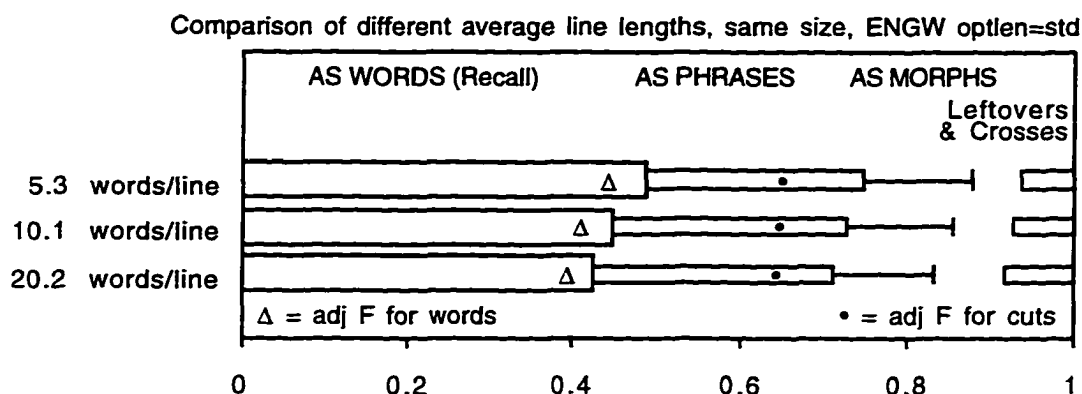
The exception to the rule here is the adjusted F cut measure (black dots in the figure above), which shows much less difference between ENG V and ENG W than do the other measures. This is largely due to the much lower random baseline for this measure for ENG W which, as discussed in Chapter 3, is in turn a side-effect of ENG W's greater line length. That is, the random baseline for the cut-F measure is greatly affected by the number of line-ends (which count as "automatic" word boundaries) relative to the number of words. The ENG V corpora have a much higher baseline than the ENG W and thus do not perform as well relative to its baseline as does ENG W, thus accounting for the discrepancy here.

When text is limited, could we reuse the same text for training purposes, thus increasing both the number of word tokens and the repetition factor? Of course, this is not a cognitive issue, but a question of computational techniques. As seen above, "doubling up" gives somewhat better results than a smaller natural corpus, but less good than a natural corpus of the same size. The difference between the doubled and natural corpus is very small, but in the direction of the natural one, even though the natural one is slightly smaller and has lower repetitiveness. This is one occasion, and we will see more, when statistical indicators may not tell the whole story.

4.1.2 Length of line as parse span

We noticed in Section 3.3.4 that a shorter line was associated with somewhat better performance. This is relevant because the line in these tests defines the amount of text which is parsed at a time. The results of each parse are used to update the lexicon, and thus may affect the overall results of the algorithm. The

following figure shows that a much shorter line length (parse length) does translate to better performance, at least in a situation where all other factors are equal.



<u>Corpus and Version</u>	<u>#Tokens</u>		<u>#Types</u>	<u>Avglen(chs)</u>		<u>Repetition</u>		<u>F% baselines</u>	
	<u>Char</u>	<u>Word</u>	<u>Char</u>	<u>Line</u>	<u>Word</u>	<u>TTR</u>	<u>G</u>	<u>Words</u>	<u>Cuts</u>
ENGW short line	313,321	74,951	26	22.2	4.18	10.6	0.224	5.5	35.4
ENGW full	313,321	74,951	26	42.1	4.18	10.6	0.224	4.2	29.8
ENGW long line	313,321	74,951	26	84.3	4.18	10.6	0.224	3.5	26.8

This comparison involves the ENGW corpus, plus two modifications² of the corpus—one where lines were cut in half and one where they were doubled, so that the same text is spread across more or fewer parses. The differences are small relative to the difference in line length, but nevertheless present.

From the table we can also see clearly the effect of line length on the random baseline measures, where shorter lines lead to higher baselines, especially for the cut-F measure, which in turn lowers the adjusted scores. The effect on these matched corpora, however, is to equalize the adjusted cut-F percentage across all three corpora, as shown by the black dots in the graph above.

² The longer-line version was created by joining each pair of lines in the original. The shorter-line version was made by dividing after the fifth word every line which contained more than five words; however, even if the remainder of the line had more than five words, it was not divided again.

4.2 Symbol set sometimes matters

There is another difference between the ENG V and ENG W corpora, a difference that we have not yet mentioned. This is the symbol set in which the language is represented. ENG W, the novel by W. Somerset Maugham, is in standard orthography—or what remains of it after capital letters are lower-cased and all non-alphabetic characters are removed. ENG V, however, is in a phonemic representation. Does the particular coding system make a difference to the O-B algorithm? To learn more about the effects of coding, we will focus in this section on corpora which are identical (or nearly so) in word tokens and lines, and differ only in their coding, at the same time looking at Spanish and Japanese texts and their orthographies.

Although as computational linguists we process text corpora in order to study language, of course the “language” we look at is only a particular coded representation. In many respects, spoken and written language differ, with the spoken form being much richer in information, and our texts are even poorer than written language, which has a number of visual cues which are missing here—not only capital letters and punctuation, but the physical arrangement of text on a page with paragraphing, indentations, italics, etc. In texts used for computer demonstrations like these, all information—linguistic or otherwise—must be contained in linear sequences of symbols (unigrams). So the nature of these symbols should be of great interest to us.

From an information theoretic perspective, there are only two attributes of a code: the number of symbols (types) used, and their distribution. It is clear, for instance, that if we exchange each letter of a text for some other (unused) letter

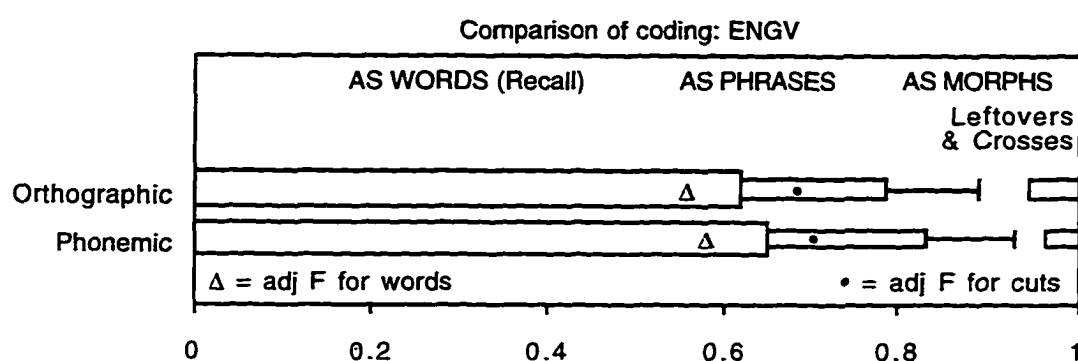
and we do this consistently throughout the text, then we have not changed it one iota from the point of view of information theory. On the other hand, if we combine two symbols into one, such as using the English letter 'k' to represent both the orthographic standard 'k' and 'q', then we have changed both the number of symbols and their distribution. Most coding differences also affect the number and distribution of n-grams (strings of contiguous symbols). For example, a change which affects bigrams is replacing the English combination 'qu' with 'k' , or replacing 'th' with 'T'.

4.2.1 English: standard orthography vs. phonemes

The ENG V corpora that we have been discussing so far are all in a phonemic transcription made by Timothy Cartwright and used in his work with Michael Brent. By matching this corpus with the original version from CHILDES (Bernstein-Ratner 1996), which is in standard English orthography, we constructed a matched pair of corpora (see Appendix B for further explanation and samples):

orthographic: look theres a boy with his hat
 phonetic: lUk D*z 6 b7 wIT hIz h&t

The phonemic version does somewhat better by all measures:



<u>Corpus and Version</u>	<u>#Tokens</u>		<u>#Types</u>	<u>Avglen(chs)</u>		<u>Repetition</u>		<u>F% baselines</u>	
	<u>Char</u>	<u>Word</u>	<u>Char</u>	<u>Line</u>	<u>Word</u>	<u>TTR</u>	<u>G</u>	<u>Words</u>	<u>Cuts</u>
ENGv orthographic	53,439	14,188	26	12.4	3.77	15.5	0.309	9.6	44.4
ENGv phonemic	40,792	14,184	50	9.4	2.88	16.0	0.313	12.9	49.3

The table shows two characteristics which might help account for this difference in performance: the greater number of symbols used and the slightly higher word repetition rate ($G = 0.309$ vs. 0.313) in the ENGV phonemic version. (The phonemicization collapsed some of the orthographic word types: For instance, { to two too } became a single phonemic word type.) The phonemic version also has a shorter average word length and thus a shorter average line, since line and word boundaries were unchanged across all three variants. We have already seen that both repetition and short lines have been associated with better performance.

A similar result is found with ENGW, the English novel. Starting from the standard orthographic version, we created a phonetic version by applying a standard text-to-phone computer program (details in Appendix B). My intent

was not to get the most accurate phonemic representation, but simply to abstract away from the inconsistencies of orthography and see what happened. Unlike the Cartwright phonemes, where each of 50 characters represented one phoneme, this scheme used 35 character symbols in combination to express diphthongs, vowel length, etc.:

original: i confess that when first i made acquaintance with charles
 phonetic: AY kAAanfEHs DHAEt WHEHn fERst AY mEYd AEkkwEYntAEEns wIHTH CHAArAXlz

Since this conversion to phonemes had the side effect of creating a longer word (increased from 4.18 to 5.1 characters), we wanted to test if that circumstance in itself would have an effect, in the absence of any phonetic significance. So, we created a third version of the corpus which had a longer word unassociated with phonetic value by expanding each orthographic vowel into an arbitrary (but consistent) two-letter combination throughout the text: a → az; e → eh; i → ij; o → ow; u → ur:

original i confess that when first i made acquaintance with charles
 expanded ij cownfehss thazt whehn fijrst ij mazdeh azcqurazijntaznceh wijth
 chazrlehs

This third version has the same character set as the orthographic version, but an average word length longer than either the orthographic or phonetic version.

Here, too, the phonemic version performed a little better than the other two:

<u>Corpus and Version</u>	<u>#Tokens</u>		<u>#Types</u>	<u>Avglen(chs)</u>		<u>Repetition</u>		<u>F% baselines</u>	
	<u>Char</u>	<u>Word</u>	<u>Char</u>	<u>Line</u>	<u>Word</u>	<u>TTR</u>	<u>G</u>	<u>Words</u>	<u>Cuts</u>
ENGW orthog.	313,321	74,951	26	42.1	4.18	10.6	0.224	4.2	29.8
ENGW phonemic	382,328	74,951	35	51.4	5.10	10.8	0.226	3.0	18.0
ENGW expanded	432,940	74,951	26	58.2	5.78	10.6	0.224	2.5	15.8

Here we have a slightly better performance from the phonemic version, and

almost identical results from the two others. This would argue for an effect of symbol set size (larger is better) and for the slightly higher repetition rate of the phonemic version. Shorter line length doesn't seem to be felt here, and word length also shows no influence. The slight differences in adjusted ratios between the orthographic and expanded versions favor the expanded one, and may be attributed to the lower random baseline for that version, which in turn is largely due to the longer word length.

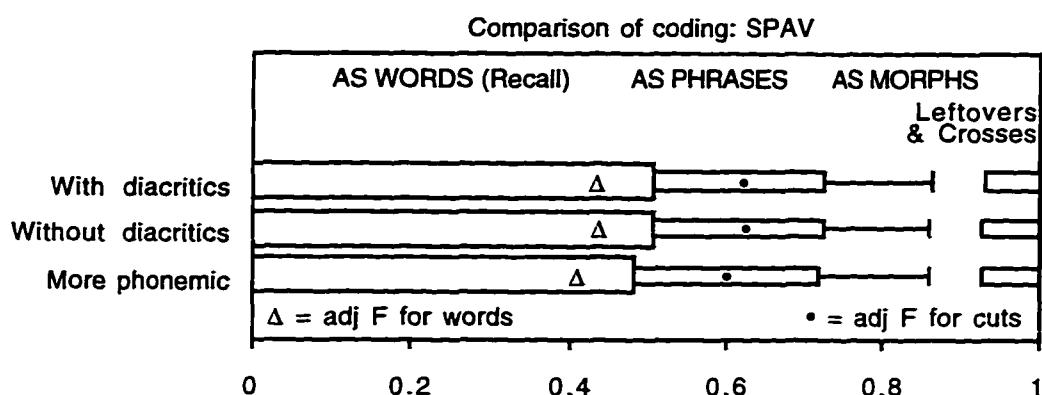
So, in both ENG V and ENG W, we see a somewhat greater learning rate when the representation is closer to the speech signal, though the effect might also be attributable to other differences, such as number of symbols or repetition rate.

4.2.2 Spanish: more or less phonetic representations

The Spanish language uses an orthography that is much closer to the speech signal than is the English system. For instance, there are five pure vowels, each with its own orthographic symbol, unlike the English language with its many different vowels and diphthongs represented by identical spellings. However, Spanish orthography varies more from environment to environment than does English, since it uses some diacritics that are left out of email, etc., and which, as a result, seem not to be strictly necessary for comprehension. One main diacritic is the tilde over *n* (*ñ*), which is a different sound and a different letter of the Spanish alphabet from plain *n* (compare *cano* 'white-haired' and *caño* 'pipe'). The other diacritic used widely is the accent mark over the vowel of an irregularly stressed syllable (*cambio* 'I change' vs. *cambió* 'he changed') or to disambiguate a few high-frequency homonyms (*ese* 'that' [adj.] vs. *ése* 'that one' [pronoun]).

The CHILDES corpus SPaV uses diacritics to distinguish accented vowels from unaccented and the *ñ* and *n* letters. This results in the separation of some word types which would be combined if the accent were not considered: *que* vs. *qué*, *esta* vs. *está*. In the CHILDES texts, also, there are occasional errors in the use of these variants, evidence that they are a perceptual luxury, not a necessity. (By comparison, spelling errors that involve the use of wrong letters rather than wrong or absent diacritics are rare in the CHILDES corpus.) To test how much information the diacritics transmit, we ran the O-B algorithm on SPaV with and without diacritics. That is, we converted these six special symbols to the unadorned lower-case letter, removing vowel accents and combining *n* and *ñ*.

The next revision was to make the corpus even more faithful to the speech signal by retaining the accented-vowel and *ñ* symbols, and then redistributing the other characters more phonetically. For instance, the sequences *gue* and *gui* are pronounced /ge/ and /gi/, while *gua* is pronounced /gwa/. To more clearly represent the pronunciation, therefore, *u* was removed in the first two sequences and retained in the third. The full list of changes is given in Appendix B. The results of running these three versions of SPaV are shown below:



Corpus and Version	#Tokens		#Types	Avglen(chs)		Repetition		F% baselines	
	Char	Word	Char	Line	Word	TTR	G	Words	Cuts
SPAV w/ diacritics	103,892	26,552	32	15.6	3.91	10.3	0.246	8.3	40.4
SPAV no diacritics	103,892	26,552	26	15.6	3.91	10.7	0.249	8.3	40.4
SPAV phonemic	99,288	26,552	32	14.9	3.74	10.4	0.247	8.8	41.2

The performances of these three corpora in the O-B algorithm are very similar to each other but, contrary to the results for the English corpora, the more phonemic version seems to perform slightly less well than the traditional orthographies. Further, the presence or absence of diacritics in the orthographic version did not seem to make much difference, despite the fact that the two versions differ significantly in the number of unigram types, repetition, etc. We can conclude that the diacritics neither help nor hinder the algorithm in segmentation.

For these three corpora, the number of symbol types have no effect on the outcome: The two runs with the most similar performance have the least similar symbol counts and repetition measures, though they are identical in other respects. The measures which might account for the poorer performance of the pseudo-phonemic version are a smaller size (by characters), but the shorter average line and word would argue for the contrary result.

So there is no consensus between the English and Spanish tests as to the effect on performance of these characteristics. We can only conclude that neither

faithfulness to the speech stream, nor any of the related information-theoretic characteristics, can be clearly and necessarily shown to predict performance in the O-B algorithm—at least not when the differences are small, as they are in these tests. It may be that a combination of influences is at work, but that remains for further research.

4.2.3 Japanese: hiragana vs. roman characters

Although we have not been able to discover very much about the influence of character sets in English and Spanish, it may nevertheless be interesting to look at a non-alphabetic language. Japanese orthography is famously complicated, using a variety of character sets, including the very large set of kanji, or Chinese characters. We will stop short of such complexity, using only the hiragana syllabary, consisting of 78 characters (more details about the character set are in Appendix B). We would like to know if the results from Japanese hiragana corpora are similar to those from a transliteration into alphabetic characters (roman characters).

Each hiragana character represents a vowel, a CV combination, or one of two mora consonants, the “mora nasal” N and the “mora obstruent” Q (Vance 1987), written in hiragana ん /n/ and っ /tsu/. More problematic for our purposes are the “small” versions of some characters, which are combined with standard characters to make palatalized sounds, like きょ /kyo/; this hiragana example is treated as a bigram by the computer, though it does not differ in phonetic length from a unigram like こ /ko/.

We have two corpora in Japanese, one from the CHILDES speech corpus

(JAPV) and one from a set of children's science books (JAPW). The first comes to us transcribed into roman characters, while the second is transcribed in hiragana, as the books appeared. In order to compare the performance in the O-B algorithm of both hiragana and roman characters, we will have to convert each corpus into the other character set.

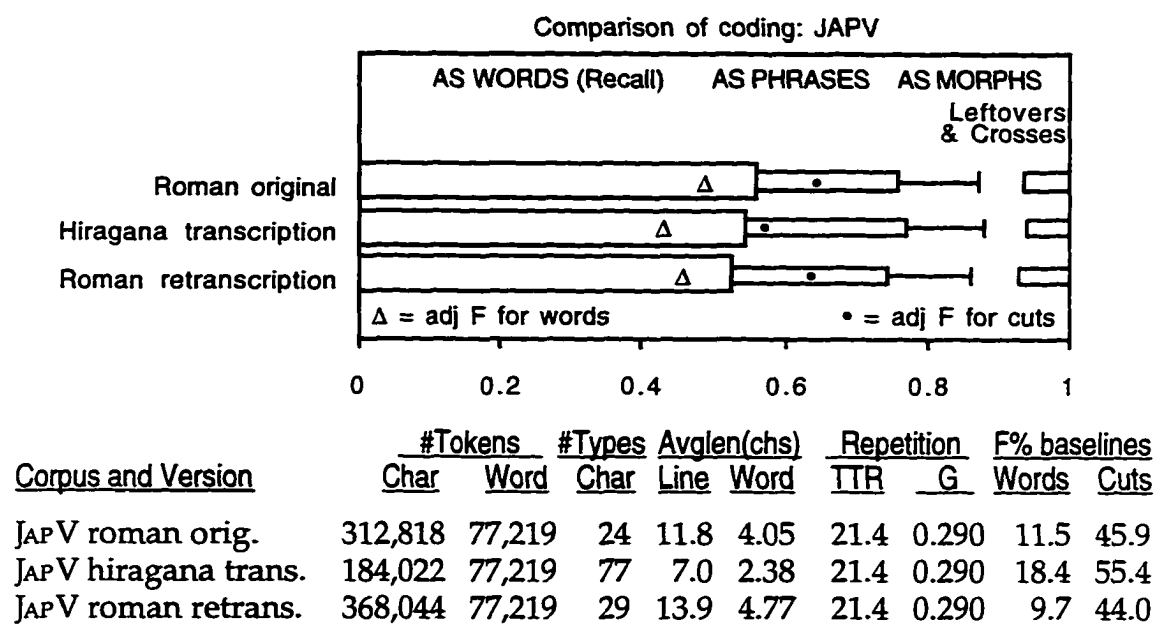
The CHILDES corpus of child-directed speech in Japanese is written in the Roman alphabet by a Japanese-speaking transcriber according to a standard system (the Hepburn system). We converted this alphabetic version into hiragana, using a computer program we wrote. The program takes into account a good deal of local context, but because it is a mechanical process, it is not entirely correct (that is, not exactly as a Japanese person would write it). As a further control, we converted this computer-generated hiragana back into roman characters, using an even more mechanical translation method which does not use any context, but equates each kana individually to a two-character transliteration. This makes in all three corpora that are matched by word and line:

Roman original:	aki chan hashigoshi ga dekiru yo
Hiragana transcription:	あき ちゃん はしごしゃ が できる よ
Roman retranscription:	A-KI TIyaN- HASIGOSIya GA DEKIRU YO

As explained in more detail in Appendix B, the hiragana transcription is fairly close to standard hiragana practice (in this case, it is exactly the same), but the mechanical conversion to roman characters is not at all standard, as can be seen by comparing it to the original, even allowing for the introduction of capital letters (to distinguish large and small kana) and hyphens (to make all targets two characters long). In this process, for instance, きよ becomes KIyo instead of kyo

and なって is NATuTE instead of natte.

Only the alphabet-to-hiragana conversion has the possibility of ambiguity, and the number of word types changed very slightly at this point. The performance results and statistics for these versions are:



Unlike the results for Spanish and English, here we have a divergence of results among the measures, with three different patterns. The identification of words themselves (recall) is best for the roman original and worst for the roman retranscription. However, words+phrases and crossings are better for the hiragana than for either of the alphabetic versions, while both adjusted F measures are worst for the hiragana. The only consistency here is that the roman original is always better than the roman retranscription, if only slightly. This should not be surprising, since linguistically the original is a human product, while the retranscription is a mechanized imitation. The differences in the adjusted figures come from the shorter words and lines for the hiragana version, which in turn create a much higher baseline.

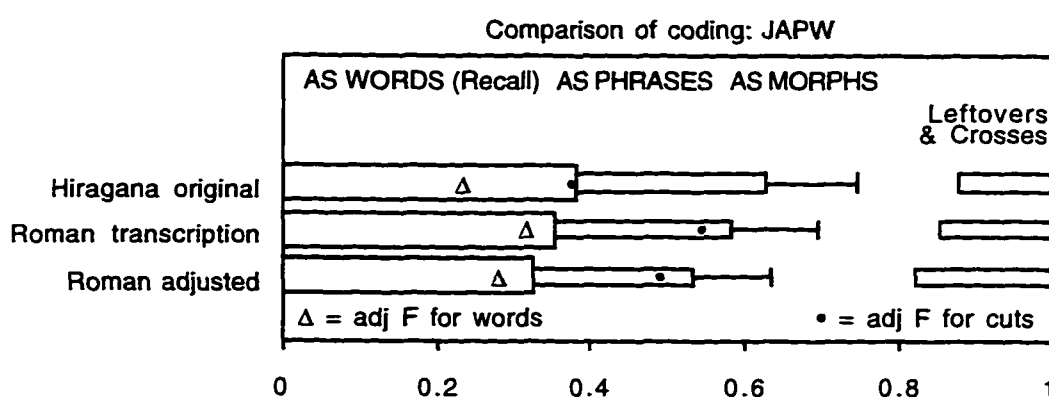
The second Japanese corpus, written all in hiragana and representing a more literary form, was transcribed into roman characters by the same mechanical method used in the JAPV “retranscription.” An adjusted version of the transcription was made by removing the hyphens used to pad out single-character transcriptions and shortening constructions using *ly* to just *y*:

Hiragana original: この きゃんぷ の ため に かんがえ られ た の が

Roman transcription: KONO KIyaN-PU NO TAME NI KAN-GAE- RARE TA NO GA

Roman adjusted: KONO KyaNPU NO TAME NI KANGAE RARE TA NO GA

The performance and characteristics are:



Corpus and Version	#Tokens		#Types	Avglen(chs)		Repetition		F% baselines	
	Char	Word	Char	Line	Word	TTR	G	Words	Cuts
JAPW hiragana orig.	37,060	19,255	78	17.7	1.92	11.0	0.262	18.7	54.6
JAPW Roman trans.	74,120	19,255	30	35.5	3.85	11.0	0.262	4.8	32.1
JAPW Roman adjust.	68,228	19,255	29	32.7	3.54	11.0	0.262	5.5	34.0

Here the discrepancy between the raw performance and the ratios adjusted for the baseline is the greatest of all, with the hiragana version showing the best raw performance on all measures, but the poorest adjusted performance. This echoes the results for JAPV, where hiragana also showed a shorter line and word, poorer performance, and a larger influence for the baseline adjustment. The larger effect in the adjusted measures for JAPW than JAPV is probably due to the

even shorter word length—at fewer than 2 hiragana or 3.8 roman characters, the shortest of all the corpora presented in this study.

4.2.4 Summary and comment

What we know about the effect of symbol set on algorithm performance is less than we would like to know. We have not found any single condition that can be said to account for performance. Neither traditional orthography, smaller (or larger) codesets, nor coding systems more faithful to what the child hears is a guarantee of better performance. Not only the codeset but also the manner in which it is applied seems to make a difference. More “lifelike” encodings—those actually used by humans—sometimes have the edge over ingenious but mechanical systems.

Further, a number of information-theoretic characteristics of the corpora, which might plausibly help account for performance differences, have not been shown to directly explain them. It is possible that combinations of these factors are responsible, but such a degree of complexity will require further study to unravel.

What relation does coding structure have to the cognitive problem of speech segmentation? Although we don’t know exactly how children represent language in their minds, it certainly is not in conventional orthographic symbols. So, how far removed are our experiments when we use convenient orthography? Or, to put it another way, would a code which is closer to the sound of language perform better? While computer simulations cannot tell us what humans do, they can give evidence as to whether there would be an advantage for one

representation over another. Particularly relevant to the present question is whether one or another representation available to infants would affect their ability to make use of frequency information. We have not found evidence that phonemes or orthography have any advantage in the segmentation process.

Of course, this research has not looked at feature clusters or other so-called sub-symbolic representations, as in Aslin et al. (1996), among others. In the present study, each symbol is a primitive, with no claim that learning benefits from any grouping other than cooccurrence frequency. For instance, when the sounds 'pe' and 'pa' are represented as bigrams, the two frequencies are assumed to be separate pieces of information, with no recognition of any similarity between the two, such as CV (consonant+vowel) or pV (p+vowel). The possibility that there would be an advantage in organizing speech information in some other way, such as a feature matrix, is not tested here.

For another example, it may be that competent readers of Spanish treat the accent mark as a "feature," adding its properties to the vowel identity, whereas to the computer *a* and *á* are two separate primitive symbols with nothing in common. However, the comparison between hiragana and roman characters gives some information about the relative merits of lower- or higher-order organizations. Hiragana has five separate primitive symbols for the combination of each consonant and a vowel (e.g. か き く け こ = ka ki ku ke ko), while the roman-character version uses five primitive symbols for the vowels and then one additional primitive for each consonant. Thus, like a feature system, the alphabetic version treats the sound /ka/ as a combination of two primitives, while hiragana treats it as a single primitive unit. As we have seen,

the hiragana system did not show a consistent advantage over the more analytic alphabetic structure for the same text, despite its shorter line length. So we can conclude that this is lack of support for the advantage of a feature-based system in the O-B algorithm.

These comparisons have thus not given any evidence that one representation or another would in itself afford a learning advantage. Even a moraic or a syllabic representation does not necessarily help, as shown by the failure of the hiragana version to consistently outperform its alphabetic transliteration. Here, we agree with Brent & Cartwright (in press:23):

The formalization of distributional regularity presented here applies to any input represented as a string of symbols, whether the symbols stand for broad phonetic classes, narrow phonetic classes, indivisible syllabic units, or something else. ...[O]ur best guess is that the distributional properties of different input codings are not fundamentally different.

It seems that the infant is free to use any idiosyncratic representation that comes to hand; at this early stage, no particular structure is required to bootstrap into segmentation.

4.3 Register does matter

We have touched indirectly on the question of different kinds of language, and in this section we will address it more directly. We use the rather controversial term “register” only for lack of a better word. The contrasts of interest here are those of child-directed vs. adult-directed spoken language, and spoken vs. written language. These are computational issues, certainly, since much computational linguistic research is performed on written rather than spoken texts, written work being more readily available in computer-readable form than

spoken language. However, there are also cognitive issues at stake. The effect of “motherese” has been much discussed as a facilitator (or prerequisite) of language learning. We would like to know in more detail what differences there are, and how they might influence learning.

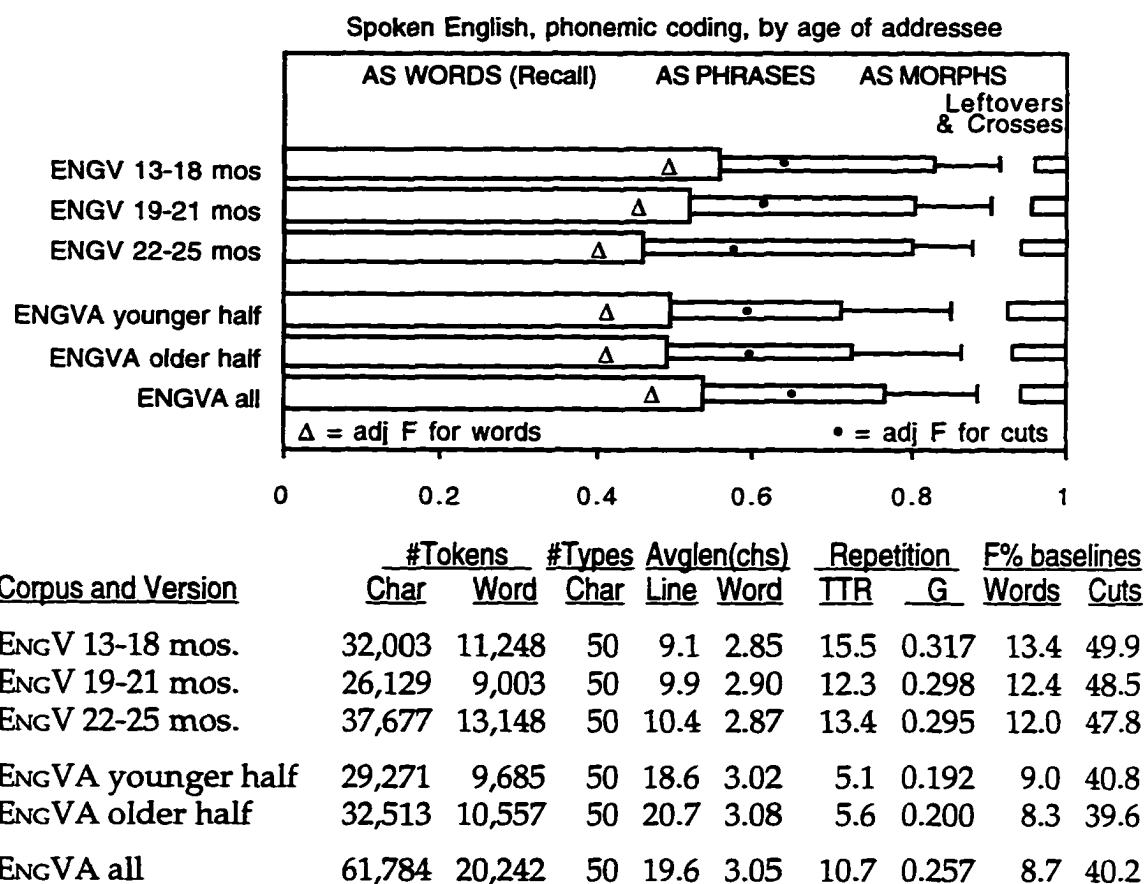
To study this, we will compare both characteristics and performance of corpora from these various sources in each language. In order to capture only differences due to register, we will try to control the samples so that corpus size, lexical repetition, line length, and codeset will all be as similar as possible. These are the factors that we have found can affect performance.

4.3.1 English

First we will look at the ENG_V corpus, which we recall is the speech of adults to children in a phonemic coding. We have a companion corpus of adults speaking to adults from the same CHILDES dataset (Bernstein-Ratner 1987), that was also converted by Cartwright to the same phonemic coding, so we can compare the two. This comparison is thus between speech to children and speech to adults, by the same individuals (mothers were interviewed by the experimenters) and transcribed according to the same conventions. We will label the adult-directed corpus ENG_{VA}, to distinguish it from the child-directed ENG_V that we have been looking at so far.

In order to further examine the relation of age of addressee to corpus characteristics, we have divided the ENG_V corpus into three parts by age of the child. Then, as controls matched by size, we divided the ENG_{VA} corpus into two parts

similarly graded, by arranging the interview transcripts in the same order as the matching mother-child transcripts:



For ENG V, the performance measures graphed above show that the adult speech to the youngest children produces more successful results in the O-B algorithm than does speech to older children, and this relation is true for all three of the age-related groupings and for all the measures, though the word+phrase measure is very close for the two older groups. The table above shows that the repetition rates are greater, and the line length shorter, for younger children across all three ENG V groups, which may help account for their better performance.

Looking at ENGVA in the same figure and table, we can make the same age comparison for the adult-directed speech and see that there is little or no perfor-

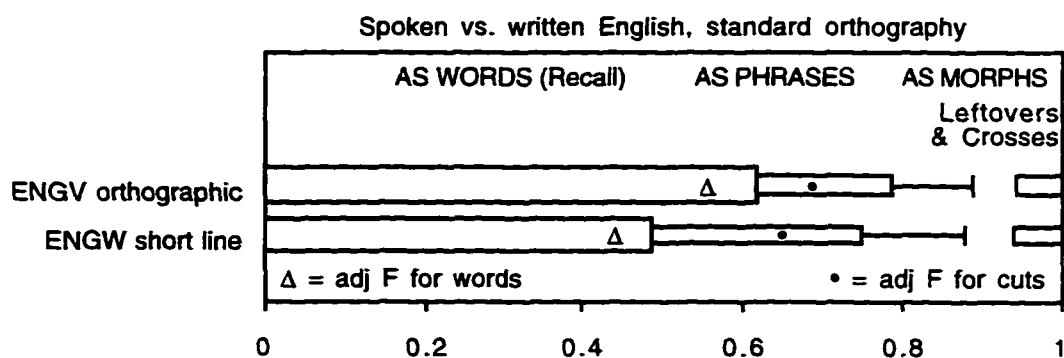
mance difference between the younger and older halves. This is evidence that the fairly large differences observed for the child-directed speech are in fact related to the child's age and not some other factor which would be the same for the same individuals recorded on the same dates in a different speech situation.³

We can also see that EngVA performed less successfully than EngV in the segmentation algorithm, when similar quantities (in characters) are compared. When the full adult corpus is used, which is twice the amount of each of the child subcorpora, the performance difference is narrowed but still not eliminated. This may be due to the lower repetition ratio, which even in the full adult corpus is lower than the oldest child corpus, and to utterances (lines) which are twice as long in the adult corpora as in the child.

To compare spoken and written English, we will compare EngW with the standard orthographic version of EngV. To match the short lines of EngV, we will use the modified version of EngW in which the lines have been cut in half. Though they are still longer than those of EngV, the short-line version of EngW performs better than the original version. Even so, we see that it does much less well on some measures than does EngV, even allowing for its much larger size.

The shorter line and greater word repetition of EngV may account for this.

³ These observations were borne out by a statistical analysis of word recall, precision, and F-measure for five corpora, with each measure calculated on the parsed version of successive 3000-character sections of each corpus. (The algorithm was exercised on each corpus as a whole, and each corpus was then partitioned into sections of equal length for scoring purposes.) A repeated-measures analysis of variance indicated a significant difference among the corpora. Recall: $F(4,28) = 20.62$, $p < .001$; precision: $F(4,28) = 31.40$, $p < .001$; F-measure [blend of recall and precision]: $F(4,28) = 28.85$, $p < .001$. To determine which individual differences between corpora were significant, post-hoc Tukey HSD tests were performed. The picture was clearest for the F-measure: the two younger child-directed corpora had significantly higher values than did the other three corpora (EngV 22-25 mos., EngVA younger, and EngVA older), which did not differ significantly among themselves.



<u>Corpus and Version</u>	<u>#Tokens</u>		<u>#Types</u>	<u>Avglen(chs)</u>		<u>Repetition</u>		<u>F% baselines</u>	
	<u>Char</u>	<u>Word</u>	<u>Char</u>	<u>Line</u>	<u>Word</u>	<u>TTR</u>	<u>G</u>	<u>Words</u>	<u>Cuts</u>
ENG _V orthographic	53,439	14,188	26	12.4	3.77	15.5	0.309	9.6	44.4
ENG _W short line	313,321	74,951	26	22.2	4.18	10.6	0.224	5.5	35.4

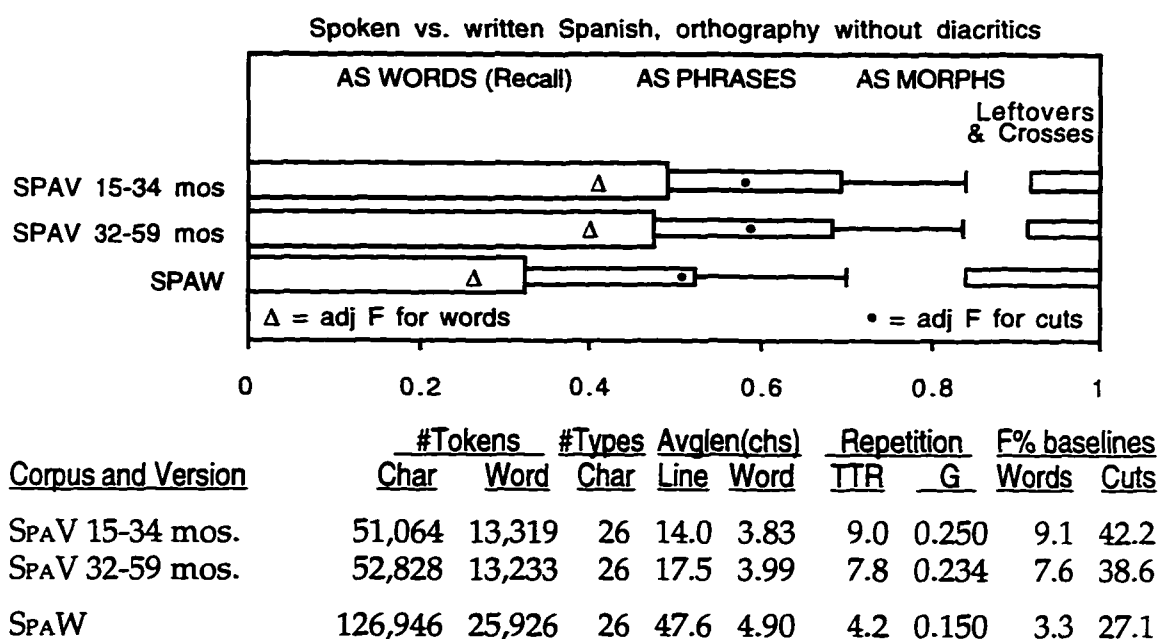
4.3.2 Spanish

The Spanish speech corpus is less diverse than the English corpus, and separating it by age of the child yields less satisfying results. While ENGW comes from a single study of nine children and their mothers, SPW is a combination of two different corpora, one of one child and one of three children, all with different age ranges but also with different individual abilities. While the ENGW children also had different abilities, the fact that so many of them were included has evened out the effect. In the Spanish case, it may be that less “even-ing” has taken place.

We created two age-related subcorpora as follows:

<u>Child</u>	<u>Age range</u>	<u>#Tokens</u>		<u>Avglen(chs)</u>		<u>Repet.</u>	<u>F% baselines</u>	
		<u>Char</u>	<u>Word</u>	<u>Line</u>	<u>Word</u>	<u>G</u>	<u>Words</u>	<u>Cuts</u>
<u>Younger subcorpus</u>								
Manuela	1;3-2;6 (15-30 mos.)	25,187	6,679	12.1	3.77	0.248	10.7	44.8
Alf	2;3-2;10 (27-34 mos.)	25,877	6,640	16.4	3.90	0.238	7.6	39.6
<u>Older subcorpus</u>								
Ida	2;8-4;7 (32-55 mos.)	24,301	6,339	14.3	3.83	0.224	9.0	41.8
Raf	4;4-4;11 (52-59 mos.)	28,527	6,894	21.6	4.14	0.228	6.3	35.7

As can be seen in the figure and table below, the age contrast within the SPAV data is not so great as we saw in ENG V, but there is a small effect in the expected direction.



The corpus SPAW is here presented for the first time. It is larger than SPAV, but has a very low amount of lexical repetition, since it is a collection of short stories that have a minimum of common vocabulary. (Recall that ENG W and JAP W are both single works.) Although the SPAW stories are all by the same author, the corpus represents the least learnable corpus in this study. Its G ratio of 0.150 stands against a range of 0.224-0.332 for the other corpora (and their variations) discussed here.

At any rate, the poor performance of SPAW relative to SPAV, despite being double the size, does not hurt our argument that spoken language is more learnable than written language, though in this case there are more plausible explanations.

4.3.3 Japanese

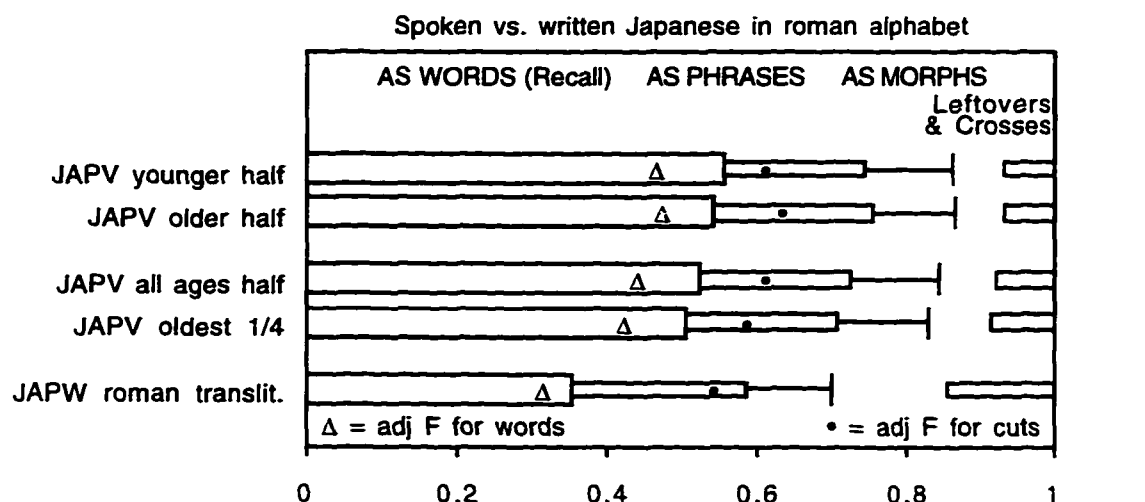
In attempting to compare spoken and written Japanese fairly, our first problem is that the JAPW corpus is much smaller than the JAPV corpus. Normal written Japanese, which is available lately in quantity, contains up to four different character sets (kanji, hiragana, katakana, and roman characters). In trying to avoid this complexity, we were limited to a particular, specialized, and small hiragana-only corpus. So first we must select a smaller portion of the JAPV corpus for comparison.

The second consideration is that we are always comparing apples and oranges, so to speak. JAPV was received in roman characters and converted by us to a less-than-perfect hiragana version. Conversely, JAPW arrived as hiragana, and was then converted to rather awkward roman characters. So each character set comparison consists of one "human" corpus and one more or less deficient "machine" version.

A third factor is that Japanese hiragana is represented as two-byte characters in the computer. This requires that special versions of the computer programs and procedures used here be written to accommodate this code structure, which was done. Additionally, it suggests that perhaps there is twice as much "information" in each hiragana character as in an alphabetic character. However, since we are using only 77 hiragana characters, versus 26 or so alphabetic characters, the ratio is actually 1.33, not 2.0. The two-byte representation is simply circumstantial, due to the normal use of kanji in conjunction with hiragana to write Japanese. It is the kanji which are so numerous (and therefore "information-laden") as to require two bytes: The standard codes for Japanese

provide upwards of 6000 kanji characters, in addition to kana, numerals, and various other graphs (Lunde 1993).

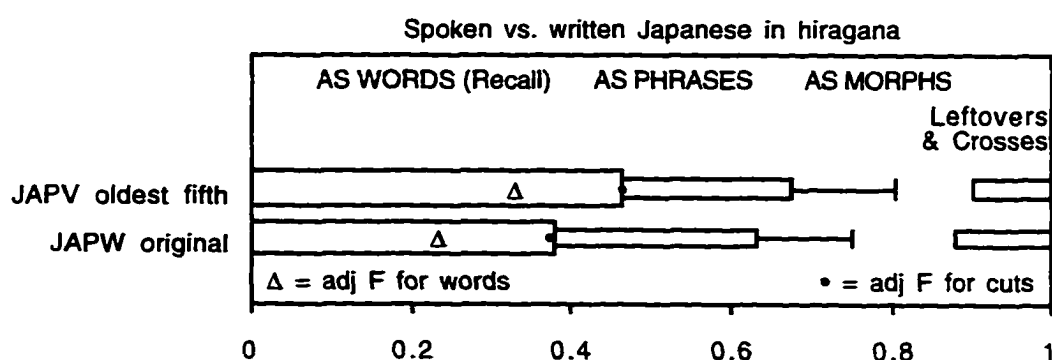
Stepping gingerly around these problems, we can still hope to present results of comparisons like those for English and Spanish. Beginning with a comparison of alphabetic corpora, we want to create age-graded subcorpora from JAPV. Because the JAPV corpus is of only one child, and contains material from 56 different dates from age 1;5 to 3;0, it should suffice to simply cut the corpus at any point and obtain age-related subcorpora. As shown in the figure and table below, this procedure yields halves that do not seem very different from one another. In fact, another "half" obtained by taking every other line across the whole corpus (labeled "all ages half" below) shows performance which is more different from the younger and older halves than these halves are from each other. It is interesting that some of its characteristics fall between the older and younger, such as line length and baselines, but others are below either, notably repetition ratios.



Corpus and Version	#Tokens		#Types	Avglen(chs)		Repetition		F% baselines	
	Char	Word	Char	Line	Word	TTR	G	Words	Cuts
JAPV younger half	144,218	35,535	24	10.9	4.06	15.9	0.283	12.9	47.9
JAPV older half	168,600	41,684	24	12.8	4.04	15.2	0.274	10.3	44.3
JAPV all ages half	156,909	38,669	24	11.9	4.06	14.0	0.268	11.5	45.8
JAPV oldest quarter	83,857	20,739	24	12.7	4.04	10.8	0.257	10.4	44.3
JAPW Rom. translit.	74,120	19,255	30	35.5	3.85	11.0	0.262	4.8	32.1

The JAPW transliteration does much less well than the spoken subcorpus of about the same size, continuing the familiar trend, even though it has a slightly higher repetition.

In order to compare speech and writing in hiragana, we next took a portion of the JAPV corpus about equal to the length (in characters) of JAPW, and we selected it from the end so that it would be the worst-performing part. Indeed, this “oldest fifth” does perform a little worse than any of the halves, although this may be due partly to its smaller size. But it still outperforms JAPW. The only notably different characteristic is the much longer line of JAPW; its repetition rate is even slightly higher than the “oldest fifth” subcorpus of JAPV.



Corpus and Version	#Tokens		#Types		Avglen(chs)		Repetition		F% baselines	
	Char	Word	Char	Line	Word		TTR	G	Words	Cuts
JAPV oldest fifth	39,294	16,738	74	7.4	2.35		9.7	0.252	17.7	54.6
JAPW hiragana orig.	37,060	19,255	78	17.7	1.92		11.0	0.262	18.7	54.6

Perhaps here is a good place to point out that there is no orthographic standard for the “word” in Japanese as there is in both English and Spanish. No word divisions appear in written Japanese, like spaces in an alphabetic language. To some extent, the shift from one character set to another serves a visual “chunking” function, but it is neither necessary nor sufficient as a guide to word boundaries. The two corpora we have been discussing, JAPV and JAPW, each came with their own word standard. JAPV was received in alphabetic transcription (roman characters), necessitating some sort of word boundaries using spaces. JAPW had been manually divided into morphemes (by a linguistics graduate student) before we received it. Thus, the “words” in each corpus are not necessarily identical. For instance, the verbal complex *いき ました* ‘went’ is treated as three “words” in JAPW but as a single “word” in JAPV: *ikimashita*. Similar constructions are very common in both corpora and are consistently analyzed into morphemes in JAPW (*iki* ‘go’ + *mashi* [polite] + *ta* [past]) but kept as conventional clusters in JAPV.

4.3.4 Conclusions

We have seen that, by a number of measures, spoken language is easier for the O-B algorithm to segment than written language. Consistent differences between the two are lexical repetition rate and line length. These characteristic differences, and the concomitant performance differences, were observed in every comparison we made.

Further, where it was possible to compare different kinds of speech—speech by adults to infants vs. speech to young children, speech to children vs. speech to other adults—the same differences in performance and characteristics were often observed there, though with somewhat less salience and reliability.

The three speech corpora all come from the CHILDES database, and thus can be assumed to share some similarities. They all are centered on young children and their worlds, and a common set of protocols was available for the use of coders. However, the corpora we extracted for testing also have several points of difference. They all consist of the utterances of adult speakers, but some are aimed at very young children, some at somewhat older children, and some are from conversation with other adults. (The child-oriented datasets of course contain some utterances of all three types, since many were recorded in a home situation where various family members and the researcher were present.)

Since all the CHILDES corpora were prepared by making each utterance a line, we can make the analogy from line length to utterance there. The written corpora, however, do not allow that equation. This is primarily because the sentence in these corpora was very long, longer than a line. While a unit smaller than a sentence could be equated to an utterance, such as a clause, this is a

complicated matter. It is not clear what utterance-like unit the child might use as an equivalent to “parsing” or segmenting speech, and since we do not know this even for speech, we know even less how to make a literary equivalent.

There is less repetition of words in written than spoken language, and in language to adults than to children, partly because the vocabulary is larger, with a wider range of topics. Speech is generally more restricted in vocabulary than writing, and speech to young children still more restricted, using primarily common and concrete words.

These differences are tendencies only, so it is of course perfectly possible for a given written and a given spoken text to present a comparison in the reverse directions. Coherence and unity has an effect. The CHILDES data, for instance, is more unified with respect to form, since the collection has been guided by a common protocol, and also more diverse since it often involves language from a variety of situations. Our SPaV corpus is from two different researchers and four different children. So, a particular spoken text might have a very large vocabulary and a particular written one an unusually low one.

4.4 Language doesn't matter

Finally, we can look at whether the particular language being learned—English, Spanish, or Japanese—makes a difference. Logically, of course, it seems very unlikely that any human language is more or less learnable than any other, by either a computer or a human. Yet so much of computational and experimental linguistics is done in English that we often overlook the similarities and differences of different languages. Here we will compare the results of our three langu-

age corpora on the same algorithm, trying to hold steady whatever differences we can control and account for the influence of those we cannot. In order to do that, we will have occasion to refer to what we have learned and reported so far.

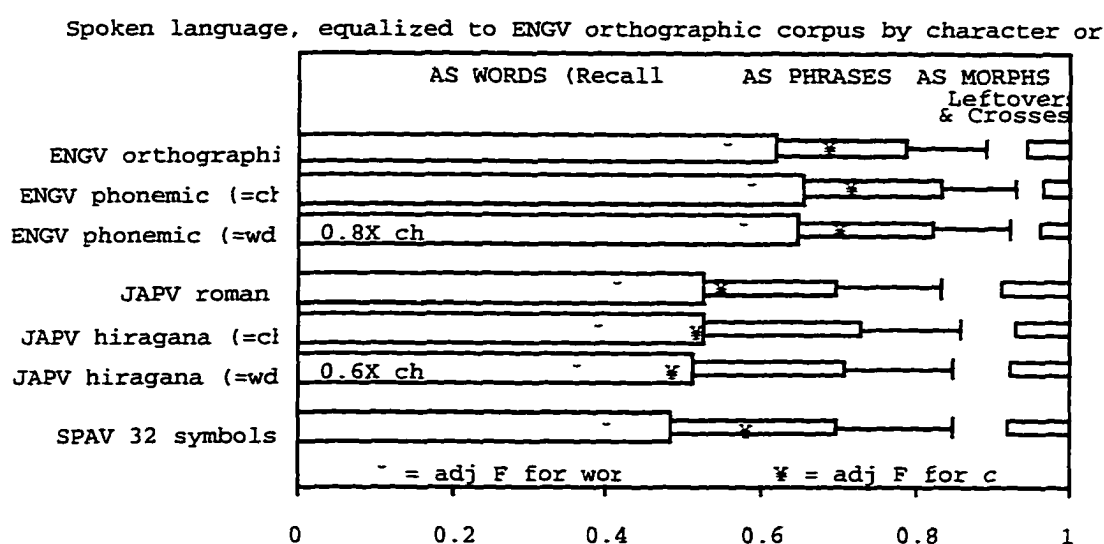
We saw that **quantity of text** is one of the most important determinants of performance, so we will control the quantity of text used for these comparisons. The results in Section 4.1.1 were indeterminate as to the unit of comparison—characters (unigrams), words, or some amount of word repetition. Here, we will set equal both characters and words, using alternate corpora where necessary, and comment upon any differences in repetitiveness. We will choose sizes equal to the smallest corpus, and then create subcorpora from all the others to match these sizes.

Second, Section 4.3 showed that **register** seems to make a big difference in results on the O-B algorithm, so we will compare spoken texts and written texts only with each other across languages, and will limit ourselves to the child-directed spoken texts.

Line length was also seen to have an influence on performance, but here we will not manipulate that, but only comment upon it. **Symbol set** was found not to make a big difference, so we will compare texts across symbol types (alphabet, phonemic, and hiragana) as well as language.

First we compare matched corpora from the CHILDES series. This comparison takes the ENG V orthographic corpus as the size referent, and then matches the others either by character count ("=ch" in the figure and table below) or word count ("=wd"), or both (no "=" label). The corpora that match both character and word count do so because their average word length in characters is close to

that of the referent. Here, JAPV roman and SPAV-32 both have a word length only a little longer than that for ENGV orthographic, the size referent, while ENGV phonemic has a much shorter word length, and thus a correspondingly smaller character size when matched by word count; the same condition is even more true for JAPV hiragana. These corpora are marked with the relative size in characters when matched by word count: e.g., JAPV hiragana is only 0.6 the size of ENGV orthographic.



	Age range in mos.	#Tokens		#Types	Avglen(chs)		Repetition		F% baselines	
		Char	Word	Char	Line	Word	TTR	G	Words	Cuts
ENG V orthog.	13-23	53,439	14,188	26	12.4	3.77	15.5	0.309	9.6	44.4
ENG V phon. (=ch)	13-21	53,354	18,582	50	9.4	2.87	19.5	0.325	13.1	49.4
ENG V phon. (=wd)	13-20	40,189	14,018	50	9.2	2.87	17.1	0.320	13.3	49.8
JAP V roman	17-26	52,938	13,124	24	10.3	4.03	10.8	0.271	13.5	49.4
JAP V hira. (=ch)	17-27	52,902	21,892	77	6.3	2.42	13.4	0.279	19.6	57.0
JAP V hira. (=wd)	17-26	33,783	13,988	77	6.2	2.42	11.0	0.270	19.8	57.4
SPA V 32 symbols	15-34	51,064	13,319	32	14.0	3.83	8.7	0.245	9.1	42.2

We see that there is not as much difference in results within each language group as there is between languages. Where there are two sizes for a given corpus, the larger one understandably performs slightly better. But matching by

size across language and codeset has not resulted in equal levels of performance. One non-statistical reason may be the difference in age range: The English texts are directed at children younger than those for the Japanese and Spanish.

Here the English texts perform the best, by all measures, and again we see that the phonemic coding does better than the orthographic, whether matched by character or word count. The Japanese versions perform moderately well, and both roman (the original here) and hiragana (the transcription) perform similarly when matched for quantity; hiragana is better than roman on some measures and worse on others. The Spanish transcript (here, with diacritics) diverges more depending on the measure. It has the lowest recall, but the other typology measures are similar to those for Japanese, and the adjusted cut-F percentage is higher than any Japanese corpus. Referring to the table, we see that SPaV has the lowest cut-F baseline, which would tend to raise the adjusted figure.

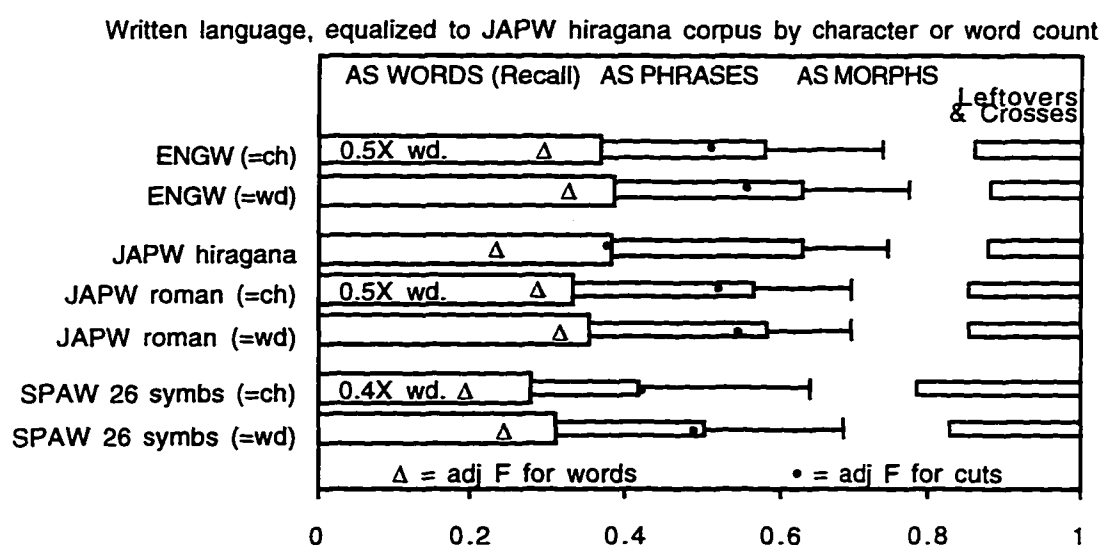
The repetition G measures here correlate well with most of the performance measures, suggesting that repetition accounts for a great deal of the differences seen here.⁴ If we assume that repetition is to large extent a voluntary characteristic—that is, a speaker can choose to be more or less repetitious, and some situations condone or evoke repetition more than others—then we can say that it is not language-specific. English is not inherently more repetitious than Spanish, though it may be the case that English mothers speak more repetitively to their infants, or that these particular English mothers may have been especially repetitious, perhaps due to some influence in the experimental environment. It is

⁴ For instance, Pearson correlation coefficient for G and word recall is $r(5) = 0.984$, $p < .01$.

probably significant that the Spanish recordings were made in the various homes of the children, while the English mothers were recorded in a common laboratory environment:

These transcripts were derived from reel-to-reel audiotaped interactions carried out in a sound-proof playroom at the Massachusetts Institute of Technology Research Laboratory of Electronics during 1979-80.... Children and their mothers played for approximately 45 minutes at each session with the same selection of toys (blocks, stuffed animals, puppets, books, etc.) (Bernstein corpus readme file)

Turning to the written corpora, we use smallest written corpus as the size referent—the JAPW hiragana corpus, with an average word length of fewer than two characters. Each of the other three corpora has a much longer word, 2 to 2.5 times longer, and the corpora matched by character count are thus correspondingly smaller, as shown by the legend “0.5X wd.” or “0.4X wd.” in the bars:



<u>Corpus and Version</u>	<u>#Tokens</u>		<u>#Types</u>	<u>Avglen(chs)</u>		<u>Repetition</u>		<u>F% baselines</u>	
	<u>Char</u>	<u>Word</u>	<u>Char</u>	<u>Line</u>	<u>Word</u>	<u>TTR</u>	<u>G</u>	<u>Words</u>	<u>Cuts</u>
ENGW (=ch)	37,018	9,038	26	41.8	4.10	4.7	0.184	4.3	30.2
ENGW (=wd)	79,150	18,832	26	42.6	4.20	5.8	0.193	4.2	29.7
JAPW hiragana orig.	37,060	19,255	78	17.7	1.92	11.0	0.262	18.7	54.6
JAPW roman (=ch)	36,948	9,598	30	35.4	3.85	8.3	0.250	4.8	32.1
JAPW roman (=wd)	74,120	19,255	30	35.5	3.85	11.0	0.262	4.8	32.1
SPAW 26 syms (=ch)	37,065	7,512	26	45.8	4.93	3.0	0.133	3.4	27.3
SPAW 26 syms (=wd)	94,445	19,256	26	47.8	4.90	3.8	0.146	3.3	27.0

The general magnitude of the written corpora results is much less than that of the spoken. Here, for instance, the recall range is 28-38%, while for the spoken corpora it is 48-66%. While the sizes of the two groups are not strictly comparable, with the written subcorpora sometimes smaller than the spoken ones, they are certainly not far enough apart to account for such a large difference. We have here another demonstration of the superiority of spoken language on this task. Some reasons for this can be found in the table, and are the same ones we have discussed before: The written corpora suffer from lower repetition and much longer lines. However, their adjusted measures benefit from lower baseline measures. Except for JAPW hiragana, all the written subcorpora here have lower baselines than all the spoken ones. In addition, the correlation of repetition with performance is not so good as in the spoken corpora. In fact, word length is slightly better as a predictor of O-B performance, though neither reaches significance.⁵

Among the written language corpora, the cross-linguistic differences are smaller than for the spoken corpora, and in general smaller than differences within their language groupings. Particularly between the English and Japanese

⁵ Pearson correlation coefficients: for G and word recall, $r(5) = 0.611$, $p > .05$; for word length and word recall, $r(5) = -0.664$, $p > .05$.

groups, there is a good deal of overlap in performance. For instance, the adjusted word-F percentages for Japanese include one higher than any English and one lower than any English. *SpAW*, however, is worse than either *JAPW* or *ENGW* on the typology measures, and lower than most other corpora on the adjusted F percentages.

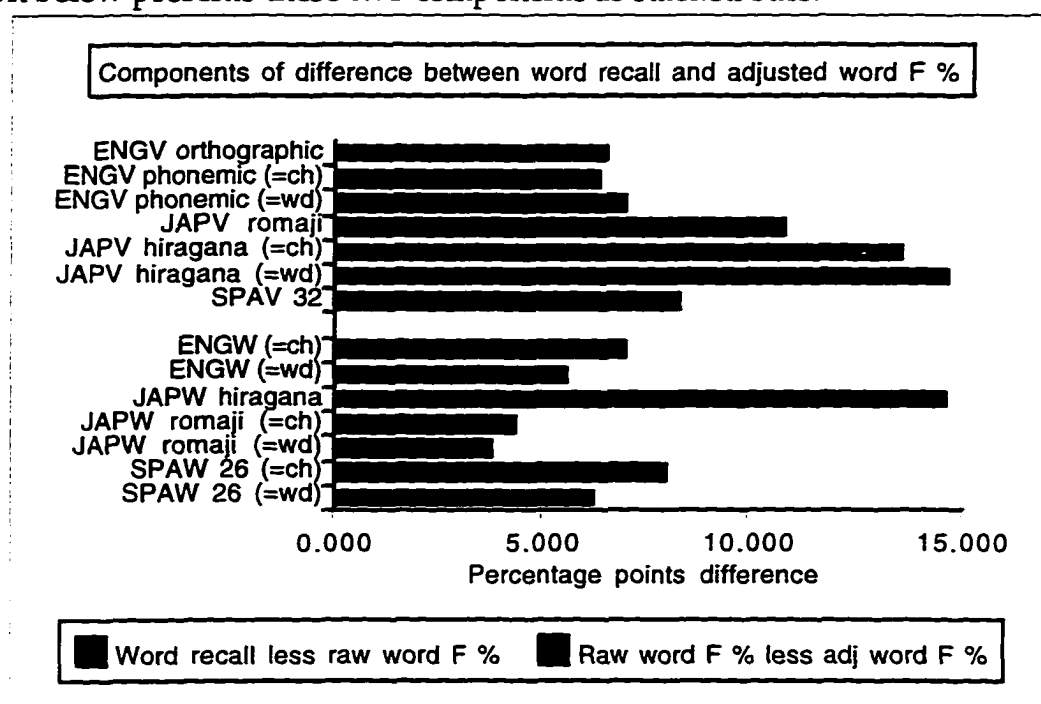
Differences within language groups are more substantial for the written than for the spoken comparisons. This is partly due to the much larger size differences within groups. While the two *ENGV* and *JAPV* corpora only differed in size at most by a factor of 1:1.6, here we have 1:2 to 1:2.5. The difference between *JAPW* hiragana (the original) and roman characters, it must be remembered, may not be a fair comparison, as one is a mechanical transcription of the other. It is of more interest that this alphabetic transcription performed very like the alphabetic English.

As with the spoken corpora, we see that the G-repetition measures cluster by language: *JAPW* has the highest measure, then *ENGW*, and lowest is *SpAW*. Here we must remember that the texts used are not strictly comparable. The English is a single work, for adults; the Spanish is a collection of stories for adults; and the Japanese is a single work aimed at children. We don't know exactly how the audience and genre facts affect the statistics, but we should not slide too far into the "a corpus is a corpus is a corpus" mirage.

Considering how low the repetition is in *SpAW*, it is surprising that it is learnable at all, and we think that this alone can account for its poor performance. However, the large repetition-measure discrepancy between the English and Japanese corpora would lead one to expect that, if anything, Japanese would

do better than English, which it does not. Not only does the repetition factor seem to favor the Japanese over the English here, but line length also favors the Japanese. Baselines are similar for the JAPW roman and ENGW, yet the two JAPW roman corpora do equal to or worse than ENGW. Only JAPW hiragana matches ENGW in typology performance, despite its multiple statistical advantages of higher baseline, shorter line, and higher repetition.

In both the written and spoken corpora, the adjusted F percentages bear a closer relationship to the typology measures for some corpora than for others. To explain the distance between word recall and adjusted word-F percentage that can be seen in the figures above, we divide the gap into two components: a) the difference between the word recall and the raw F measure, which is due to a discrepancy between recall and precision; and b) the difference between the raw and adjusted F measure, which varies according to the baseline score. The bar graph below presents these two components as stacked bars:



Those with wider total gaps are generally due to a higher baseline, which in turn is highly correlated to a shorter word and line length.⁶ The higher baseline lowers the adjusted score, increasing the gap between the recall and adjusted score.⁷

The chart above shows that the hiragana corpora, with relatively shorter lines and words (as measured in characters), have the widest gaps between raw and adjusted scores since, in most cases, the greater influence on the gap between raw and adjusted F comes from the baseline component, but the written corpora ENGW and SPAW also have a significant gap due to the distance between recall and precision. This second kind of gap is directly due to the algorithm's overdividing in these cases, producing too many parse words relative to the standard division, which in turn seems to be attributable to a very low repetition factor that results in a rigidity during parsing. With less repetition, the dictionary does not "grow" very rapidly, resulting in many low-frequency longer words and a few very short combinations that have high frequency. When optlen penalizes very long choices, the algorithm has no option but to keep repeating the too-short clusters it has found, thus overdividing.

The written corpora do not, like the spoken corpora, come from a highly standardized data base. The CHILDES datasets are unified by purpose, a broad research tradition and a more specialized social and professional network.

⁶ Pearson correlation coefficients across all of the corpora in this chart were all $p < .01$: line length vs. word F-measure, $r(12) = -0.873$; word length vs. word F-measure, $r(12) = -0.897$; line length vs. cut F-measure, $r(12) = -0.945$; word length vs. cut F-measure, $r(12) = -0.88$.

⁷ For these optlen runs, precision is nearly always lower than recall, since the number of parsed tokens is generally greater than the number of standard tokens. Therefore, F-measure—which is a kind of average of precision and recall—is never greater than recall.

Detailed and explicit procedures for collection and transcription are promulgated, and previously collected datasets are available for inspection and comparison. It is not surprising that, even across languages, they show some similarity. The written corpora, however, are a motley lot. They were plucked from different sources and are each a different literary genre. Considering these facts, it is perhaps surprising that they show even as much consistency as they do. We conclude that there are clear effects of register and text size, and less direct influences of codeset, line length, and repetition.

4.5 Conclusions

4.5.1 Computational implications

There is considerable variation within a text—any text. Whatever statistical tendencies we discover about language, we begin and end with the caution that any particular bit of language is never completely predictable. However, we must not err on the other side, since language cannot be completely novel either. Indeed, at times language seems to be a phenomenon that is perfectly balanced between the poles of repetition and innovation, constancy and creativity, rote and variation. If we can understand and characterize within-text variation, then we will be better equipped to compare different texts, perhaps even across languages.

We have shown that a simple algorithm which records the frequency of observations of various clumps of language can be surprisingly effective in discovering the standard word boundaries, despite the occasionally variable and arbitrary nature of these standards. The O-B algorithm is effective on a wide vari-

ety of texts and encodings, and it gives results equal to or better than previous computational studies on identical or similar texts, as we will show in Chapter 6.

We have introduced a set of measures of segmentation performance which we think are both compact and highly descriptive of the several dimensions that affect evaluation. Some, like recall and precision and crossing measures, are related to previous research, while others—such as the typology bar graph—are common-sense inventions of our own.

We have also seen that some quantitative characteristics of language texts are highly correlated with the performance on this segmentation task, particularly quantity of text, repetitiveness, and parse length. Some of these are seen to be related to the real-world source and nature of the text. Spoken texts tend to have higher repetitiveness and shorter utterance lengths than do written texts, and speech to very young children exceeds adult-to-adult speech in both of these tendencies.

However, quantitative considerations do not completely explain the results we found, and more research is necessary to explore several areas. Coding systems often produce differences in the algorithm's segmentation, but this is not clearly an effect of the size of the codeset, or indeed of any measures we tried. There seem to be additional attributes of "goodness" of encoding that we have not been able to describe, but that may be related to the way any given codeset is implemented.

In addition, effects on performance of text characteristics, such as we have shown, suggests that corpus studies of various sorts should report more information about the corpora they use, enabling the reader better to evaluate the

results. Further, we have demonstrated the value of running an algorithm on a wide variety of texts, not just one or two that do not greatly differ, as seems to be the current practice in computational linguistics.

4.5.2 Cognitive and linguistic implications

The results here show that the basic semantic and syntactic units of language—words or common phrases—are at least partially discoverable by a simple quantitative analysis. Using nothing but frequency observations, a significant amount of information regarding meaning-based structure in the speech stream can be learned. In Chapter 5, we will look at evidence that children have such a frequency-observing ability at their disposal, but here we can at least say that such an ability would, if available, greatly assist the child in bootstrapping into the language code.

One conclusion of this research is that the amount of language available has an effect on the amount of learning—the more language, the more learning. This is not surprising, since it has long been known that a prime requirement of language learning is exposure to goodly amounts of language. In particular, if we hypothesize that the learning proceeds from experience, it seems reasonable that more experience leads to better learning.

The amount of repetition in the language is important. Again, this seems obvious, since adults talking to small children (or foreigners learning the language) intuitively repeat and rephrase utterances, using a small vocabulary. Talking to children about the immediate environment not only involves using concrete words that help highlight salient concepts, but also limits the vocab-

ulary and promotes repetition.

We can further interpret these results cognitively by saying that “amount” of learning in this context is equivalent to speed of learning by the child. Since all children exposed to language learn it, but perhaps at different rates, it is possible that the higher-performing runs in the present research represent early-learning children and the lower-performing runs represent not children who learn poorly, but those that take a little longer to discover the keys to language. It could be that children who are exposed to language which is like our best-performing texts—lots of short utterances with a high rate of lexical repetition—are early learners, while those who have less optimal linguistic input may not speak and understand words until they are a little older.

Of course, other factors are also thought to affect learning rate, such as intonation, immediate and concrete subject matter, clear enunciation, etc., and our results in no way contradict these hypotheses. Rather, we believe that the child uses a variety of information sources to gain entry to the linguistic system. We expect that language indeed has a number of redundant and reinforcing cues to its structure, so that a learner who is deprived of one or another can make do with those remaining and still succeed.

The units in which language is recorded seem to be relatively unimportant in making this segmentation. We have seen that whether we represent the speech stream in rather arbitrary orthographies, phoneme-like representations, or syllable-like hiragana, segmentation is still possible. This suggests that babies are free to seize upon any idiosyncratic representational system that presents itself for this early effort, and then later they can build a more appropriate—and

possibly even inevitable—structure as they learn more about their language. The issue of which comes first, the word or the phonotactics, would then resolve itself to the word, with the consequence that the infant could observe word-conditioned phonotactics and use them to develop a true phonological system.

Critically, the O-B algorithm depends upon using some optimal word length to guide discovery. In these experiments, we have supplied the average length of the standard words—clearly a godlike interpolation of information the infant could not possess. We commented in Section 3.3.6 that the infant must begin with some rough idea of a good length, perhaps a range of reasonability between “too-short” and “too-long.” This could come initially from one or more of several sources: the shortest meaningful utterances (other than *mmm* and *uh*) that the child hears; the length of rhythmic units, as from one stress to the next⁸; or perhaps just from the upper bound of the infant’s memory capacity at that age.

The next chapter will look at experimental evidence to confirm these cognitive claims.

⁸ Mehler et al. (1996:102) hypothesize that infants become sensitive very early to the “rhythmic structures that are related to ‘timing units’” of the ambient language(s). Additionally, Saporta (1966) suggests a language universal: “The mean length of morphs will be inversely related to the number of phonemes in the inventory.” If so, this relation could be an effect of the infant’s information-processing capacity, and such a constraint on information load could help account for a word length limit.

Chapter 5. Experimental Evidence

Having looked at computer simulations of infant cognition in some depth, let us now return to the living organisms. We have shown that there is a significant amount of statistical information in the language stream, as evidenced by the ability of the O-B computer algorithm to extract sufficient patterning to make accurate segmentation decisions. The questions now are: Are humans also able to extract and make use of such information? In what circumstances do they actually do so? Is there any evidence for the specific hypothesis that infants actually use statistical information for speech segmentation? In this chapter, we will consider evidence from laboratory experiments with adults and children that bears on these empirical issues. We will also briefly review the cognitive literature on the issue of sensitivity to frequency and distributional information in cognition, and see how this logically applies to segmentation of language.

We will look at four experiments performed in the laboratory, two with adults as subjects and two with infants, which demonstrate that frequency information is used for speech segmentation. These experiments used artificially prepared streams of language-like stimuli so that various characteristics of the input could be controlled and so that the subjects' knowledge of English would not confound the results. We will also review two related experiments in phrasal segmentation of word-based artificial "toy languages" by adults; we will claim that frequency information can help account for the results obtained.

In reviewing this experimental evidence, we will ask precisely which characteristics of the stimuli were used as information sources by the subjects.

We will propose that some statistical indicators are better predictors of cue reliability and cue salience than others. To support this argument, we will reinterpret some of the experimental results and also look at comparable measures from corpora.

5.1 Experiments in speech segmentation

5.1.1 Hayes & Clark (1970)

The first experiments aimed at the problem of segmenting the speech stream were reported in Hayes & Clark (1970) and are still widely cited. Drawing a parallel with music, which has no “semantics” and yet is segmented, they wanted to test the existence of an unalloyed “clustering mechanism”:

Clustering mechanisms are able to detect the recurrent patterns that constitute words, even without the aid of pauses or of meanings. These abstract mechanisms are able to measure crude correlations and to differentiate between the strong inter-phoneme correlations found within words and the weaker correlations across word boundaries. They identify as words the strongly correlated clusters of adjacent phonemes in the speech stream. (p. 223)

They therefore created a “meaningless but structured artificial language,” as one commentary termed it (Wakefield et al. 1974:261), using as “phonemes” synthetic sounds produced by a computer: “complex sounds in three or four voices which lasted between 0.2 and 0.3 second” (Hayes & Clark 1970:226). Using 12 “consonants” and 7 “vowels,” they constructed four “words,” three of 8 phonemes and one of 6, in a pattern of alternating “consonant” and “vowel.” Many copies of these four words were arranged in random order (using sampling with replacement) with no boundaries between them. “What the [human] subject actually heard was a continuous stream of words, one right after the other without

pause in between words, for a duration of about 45 minutes" (p. 226).

Then the subjects were tested on 40 presentations of two contrasting stimuli, one being the four words in a random order with pauses now inserted at the word boundaries, the second being the same string of phonemes, but with each pause moved two phonemes to the left and the last two phonemes shifted to the beginning. We present here two imaginary examples (the authors do not give examples, and the stimuli were not letters):

consonants:	B	G	J	K	M	N	P	R	S	T	W	Z	vowels:	A	E	I	O	U	Y	ø												
test item A, canonical:	BIGUM				EWA				KORI				WUZø				GAJY				RESO				NøPY				TI			
test item A, shifted:	TIBIG				UME				WAKORI				WU				ZøGAJY				RE				SONøPY							
test item B, canonical:	GAJY				RESO				BIGUM				EWA				NøPY				TI				KORI				WUZø			
test item B, shifted:	ZøGAJY				RE				SOBIG				UME				WANøPY				TI				KORI				WU			

Subjects were asked which set of words sounded "most similar to the language heard in the listening phase." The result was "a small but measurable tendency to recognize the appropriate segmentation" (p. 227), with subjects scoring from 19 to 34 correct out of 40 (chance = 20 correct). A fascinating description of the subjective experience is included:

By serving as pilot subjects, the experimenters were able to obtain a subject's-eye-view of the segmentation process... At first, the sound stream seems quite amorphous and featureless. After perhaps a minute of listening, an event—perhaps a phoneme or part of a phoneme—stands out from the stream. When the event has recurred several times, the listener may notice that it is typically preceded or followed by another event. The combination of events can in turn be related to events that happen in its neighborhood. Recognition of a word, then, seems to proceed from perceptually distinctive foci outwards in both directions toward the word boundaries. Presumably, the process would tend to stop at the word boundaries because the correlations across the boundaries are weak. (p. 228)

This study has been criticized for not comparing an untrained subject on the task as a control, rather than using only chance performance as a benchmark:

“As the authors pointed out, a subject could have differed from chance, either above or below expectation, simply by adopting a strategy for responding to the sentences and following it consistently” (Wakefield et al. 1974:262). It is not clear, however, what sort of “strategy” might have been available to the subjects.

Hayes and Clark (1970) also investigated what aspects of their “language” contributed to success in the use of frequency information. In a second experiment, they varied both the number of phonemes (from 19 to 16 and 4) and the length of the “words” (from 6-and-8 phonemes to 6-only and 12-only). Here are the results of both the first experiment described above and the three versions used in the second experiment:

<u>Phonemes (C+V)</u>	<u>Word lengths</u>	<u>Correct in Test</u>		
		<u>Mean</u>	<u>Median</u>	<u>Range</u>
I. 19 (12+7)	6,8,8,8	63%	60%	48-85%
IIa. 16 (12+4)	6,6,6,6	63	67	42-88%
IIb. 4 (2+2)	6,6,6,6	53	56	25-71%
IIc. 4 (2+2)	12,12,12,12	51	50	25-83%

Here is an example of Version IIb, to compare with Version I presented above:

consonants: B G	vowels: A E
test item, canonical:	BAGABE GAGEBA GEBEGA BABEGE
test item, shifted:	GEBAGA BEGAGE BAGEBE GABABE

The 16-phoneme language performed about the same¹ as the original 19-
phoneme one, but the 4-phoneme versions performed much worse, barely above

¹ Because these were two separate experiments, it is difficult to compare them directly. The raw scores reported are higher for the 16-phoneme language, but the authors cite confidence levels that are higher for the 19-phoneme language. They also state that only the contrast between IIa and IIb is significant.

chance; there was no significant difference between the two 4-phoneme versions. The authors concluded that more phonemes were better than less, and that word length had no effect (although no word-length contrast was tried for the 16-phoneme language).

The authors observed that there was no effect of “information content,” which they measured as the number of bits of information in each word (each set had four words). Although the longer words in the 4-phoneme language had more “information” than the shorter ones in that language, yet there was no significant difference in performance. Hayes & Clark (1970) sum up:

These two experiments demonstrate that humans do, in fact, have a clustering mechanism able to segment artificial speech. It is a mechanism which (a) can segment completely unutterable sounds, (b) works on “speech” that has no semantic and no significant syntactic structure, and (c) requires relatively little time—about three-quarters of an hour in our experiments—to come to at least some parsing of the speech. It seems to us that these are important properties of a mechanism that would be useful to a child first trying to sort out the sounds he hears around him. (p. 230)

5.1.2 Goodsitt, Morgan & Kuhl (1993)

Goodsitt, Morgan & Kuhl (1993) investigated whether infants use “the aggregating effects of clustering [to] complement the segregating effects of bracketing” (p. 250) in making sense of the speech they hear. They comment that:

A clustering strategy is inherently more difficult to deploy than a bracketing strategy: whereas the latter depends upon the perception of (prosodic) cues that are available in individual speech strings, the former depends upon an analysis of multiple strings tantamount to a rudimentary distributional analysis. (p. 243)

In a first session, 7- and 8-month-old infants were trained to recognize a “target syllable” (*ti* or *de*) presented in isolation. Then, in each of two succeeding

sessions, the same syllable was presented embedded in a string with two syllables different from the target syllable called “context syllables” (*ko* and *ga*). The experiment determined to what extent the infant recognized (by head-turn) the familiar syllable and whether the recognition rate improved from the first to the second test session.

There were three conditions in the experiment, designed to explore the kind of mental representation that the infants formed. The test trisyllables (boldface below) were embedded in a “background” of repeating trisyllables, like the following:²

<u>Condition</u>	<u>Stimulus stream (sample)</u>
Invariant:	kogade dekoga kogade dekoga tikoga kogade dekoga
Variable:	kogade degako kogade degako tigako kogade degako
Redundant:	kokode dekoko kokode dekoko tikoko kokode dekoko

The “invariant” condition presented the context syllables in an unvarying sequence (*koga*). In the “variable” condition the order varied (*koga* and *gako*). The “redundant” condition used two identical context syllables (*koko*). Reasoning that “task difficulty reflects perceptual complexity” (p. 232), the authors hypothesized that if the infants were able to reduce the test stimulus from three elements to two (*ti + ko + ga* vs. *ti + koga*), they would score a higher recognition rate because they had managed to simplify the recognition task.

...[S]tretches of speech... that have coherent internal structure (allowing identification via a clustering strategy) ought to attain especially strong status as early perceptual units. (Goodsitt, Morgan & Kuhl 1993:250f.)

² Each of these trisyllables was in fact repeated three times before proceeding to the next in sequence, and there were numerous counterbalanced subconditions, including placing the target syllable sometimes before the context pair and sometimes after them.

That is, if the infants recognized a reliable repetition of two syllables always falling together, they could form a single mental representation of the two; if no such repetition was recognized, the mental construct would be more complex.

The infants in all conditions performed about the same in the first test session, but in the second session those in the invariant condition showed improved performance while infants in the other two conditions either stayed the same or declined slightly.³ The authors said that this was because only in the “invariant” condition were the **transitional probabilities**⁴ between syllables such that the infant was encouraged to form a cohesive unit of the two invariant syllables as the stimuli grew more and more familiar. In other words, while the “variable” group had to cope with two alternating orders (*koga* and *gako*), the predictability of *koga* in the “invariant” group enabled those children to treat it as a unit: “...[E]ither the perceived complexity decreased or the predictability increased across sessions for these infants” (p. 242). This was evidence that infants use:

...a clustering strategy that exploits asymmetries in transitional probabilities between successive elements, aggregating elements with high transitional probabilities and identifying points of low transitional probabilities as boundaries between units. (p. 229)

Goodsitt et al. (1993) purposely used artificial stimuli, with as many of the natural variations of speech removed as possible, so that they could see the naked effects of distributional characteristics. The use of artificial stimuli also explained why infants in the “redundant group” performed poorly despite the

³ The experiment was run twice, the second time to eliminate a possible confounding effect of prosody. The absolute scores were somewhat different in the two versions, but the trends were substantially the same.

⁴ Defined as the probability of the bigram conditional on the occurrence of the first character, or the ratio $p(ab)/p(a)$.

apparently simpler task. This group, they said, was hampered by an unnatural similarity when two of the same digitized syllables are placed together: “When reduplicated syllables occur [in natural language]..., it is never the case that they are acoustically identical” (p. 249). In addition, the artificial stimuli were all slightly separated from each other—syllables were of about 450 ms. duration and were separated by 50 ms. of silence, with 1000 ms. between trisyllabic words—but in natural language “there are no perceptible silences between reduplicated syllables” (p. 249).

5.1.3 Saffran, Newport & Aslin (1996)

Saffran, Newport & Aslin (1996) investigated whether adults use statistical information for segmentation. Stimuli were eleven computer-generated CV syllables, composed of English phonemes (four consonants and three vowels),⁵ and combined to make six different words of three syllables each. The training stream consisted of 300 tokens of each of the six words randomly arranged (but no word appeared twice in row) and synthesized in a continuous monotone voice so that no boundaries were apparent. Subjects listened to this for 21 minutes, and then were asked to choose between one of the six “words” and one of six novel trisyllables (nonwords for one group of subjects and partial words for another) for 36 test questions. Training and test stimuli looked like this:

⁵ Saffran (p.c.) acknowledges what is clear from the list of stimuli on p. 612—that only eleven syllables were used, not twelve as stated there and as would seem logically possible by combining segments.

Words:

ba bu pu / bu pa da / du ta ba / pa tu bi / pi da bu / tu ti bu

Training stream sample:

ba bu pu pa tu bi bu pa da tu ti bu du ta ba pi da bu

Test stimulus: Nonword⁶

tu pi ti

Test stimulus: Partial word, beginning pair good

pi da ta

Test stimulus: Partial word, ending pair good

bi ta ba

The differences between words and nonwords were only in the cooccurrence relationships of their component syllables in the training stream. Words were trisyllables that had always been contiguous in the training signal. Nonwords were formed of three syllables which had never been heard in contiguous positions. Each partial word had two syllables which were part of a word and thus had always been contiguous in the training signal, plus a third syllable which had never been next to the other two.

The students chose words over nonwords an average of 76% of the time, and chose words over partial words 65% of the time (chance = 50%). These scores are significantly above chance, and indicate some ability of the subjects to notice and make use of purely statistical information. The authors of the paper claimed that the transitional probabilities between syllables were the source of segmentation information in this experiment. They explained that neither syllable frequency nor bi-syllable frequency would account for the results. Words were deliberately constructed to make these frequencies unequal, and the words which were best learned were those with high transitional probabilities, not those with most frequent syllables or bigrams. They also showed that the partial

⁶ The test stimuli were not reported in full: only three of the six partial-word stimuli were given verbatim, and none of the six nonwords. We invented this nonword example in conformance with the conditions given in the paper.

words with the correct endings were more successful (more often chosen over real words) than those with correct beginnings.

Unfortunately, our ability to draw conclusions about this research is somewhat hampered by inconsistencies in the reported data and the lack of sufficient information to make revisions:

- The authors claimed that transitional probabilities for syllables spanning a word boundary ranged from 0.1 to 0.2, while those for word-internal syllables were 0.31 to 1.0, two non-overlapping ranges. However, this proves to be inaccurate. In a syllable stream created according to their specifications, the relevant ranges were 0.05-0.60 and 0.25-1.0, respectively. Thus, while their figures showed a clear separation of the boundary-spanning and word-internal pairs by transitional probability, in fact there was considerable overlap of the stimuli in this respect. In all, 36 syllable pairs were used, 12 inside words and 26 across word boundaries; two syllables (*babu* and *bupa*) occurred both inside and outside words. It is these two which caused the range overlap (transitional probability of *babu* = 0.60, *bupa*=0.35).

- The authors' stated aim was to have all word-internal syllable pairs occur an equal number of times, but two did not. As the authors said, most of them occurred 300 times each (once in each set of six words), but *bupa* occurred 420 times, since it also occurred across word boundaries. They went on to say that its higher frequency did not seem to improve its score, since "the part-word *bupabi* triggered the fewest false alarms of all" (p. 614). However, they omitted to report that *babu* also was repeated, occurring 360 times, and they therefore did not report whether this repetition had a similar negative effect.

- The paper reported that, when the six training words were ranked with respect to their transitional probabilities,⁷ the three high-TP words scored an average 79% performance vs. the three lower items' average of 72%. It is tempting to assume (by Gricean principles) that they reported these figures because they could not truthfully claim the more convincing statistic—that a ranking by performance scores was the same as a ranking by average transitional probability. It would be interesting to know the actual performance ranking of the six words, and even their individual performance percentages, so we could compare the influence of transitional probability with that of other statistical characteristics.

Although these discrepancies and omissions tend to make the reader a little wary, still the study undeniably showed a significant effect of statistical influences on the subjects' ability to segment the syllable stream.

5.1.4 Saffran, Aslin & Newport (1996a,b)

A demonstration similar to the adult experiment just described, but with somewhat simpler stimuli, was made with 8-month-old infants in Saffran, Aslin & Newport (1996a, b). Twelve synthesized syllables were arranged into four three-syllable words, which were then concatenated into a two-minute training stream with no breaks or prosodic differences. In this experiment, each syllable occurred exactly the same number of times, so there was a clear distinction between the statistics of within-word syllable pairs and pairs across boundaries: all within-word pairs had transitional probability of 1.0 and between-word pairs, 0.33.

⁷ Word transitional probability = average of the TPs of the word's two bisyllables.

Words:

tu pi ro / go la bu / bi da ku / pa do ti

Training stream sample:

tu pi ro pa do ti go la bu bi da ku pa do ti tu pi ..

Test stimulus: Word: go la bu go la bu go la bu go la bu ...

Test stimulus: Nonword: da pi ku da pi ku da pi ku da pi ku ...

The training stimuli consisted of the four words randomly concatenated so that each word (and thus each syllable) appeared the same number of times. After hearing the two-minute training stream, infants were exposed to repetitions of either one of the words from the test stream or to a novel combination of three of the same syllables. The nonwords were such that no syllable pair had been heard together in training. The infants looked longer at the novel stimulus, thus indicating that they could discriminate words from nonwords, and must therefore have formed some mental representation of the “words” they had heard.

In a second experiment, Saffran, Aslin & Newport (1996b) trained the infants similarly and then tested them with “partial words” made by joining the final syllable of a word to the first two syllables of another word. In this case, the test stimulus was not completely novel, since one of the syllable pairs (within-word) had been heard very frequently and the second pair (between-word) had been heard with moderate frequency. Still, a demonstrable novelty effect was observed, suggesting that the infants “had learned the words with sufficient specificity and completeness that sequences crossing a word boundary were relatively unfamiliar” (Saffran, Aslin & Newport 1996b:1927).

5.2 Frequency and cognition

These four sets of experiments are perhaps more a curiosity than an overwhelming demonstration. They offer some evidence that statistical information about a speech-like stream may be recorded and put to cognitive use, by both adults and infants. We will offer an argument here that this idea is not a very radical one, but is part of the implicit foundation upon which psychology and psycholinguistics is built.

It seems uncontroversial that human beings notice and remember facts about repetition and frequency of events in the environment, and that this information is an indivisible part of much of what we “know” and forms an important basis for decision making in many areas. Hasher & Zacks (1984), in an often-cited review article, gave numerous examples of evidence showing that people are sensitive to frequency, such as “experienced differences in the natural occurrence rates of single letters, pairs of letter, syllables, words, surnames, professions, and sources of mortality and morbidity” (p. 1373).

People are surprisingly accurate at answering such questions without being aware of how they acquired the relevant information in the first place... [R]ecent research in human memory [suggests] the possibility that the answer... stems from a basic operating characteristic of the information processing system. That is, the system seems to support the inevitable encoding into memory of certain fundamental aspects of experience. (Hasher & Zacks 1984:1372)

Beyond the simple recording of frequency of occurrence of events, any creature will benefit from knowing which events tend to occur together. Alloy & Tabachnik (1984) presented a thorough overview of the research on human and animal perception of covariation. Kelly & Martin (1994) presented arguments

and evidence that multiple information sources are often combined in cognition and that many of the available sources are probabilistic in nature. They argued that humans, as well as animals, tend to use a variety of probabilistic cues in combination, even when more reliable but deterministic cues are available. These tendencies and abilities seem to be widespread across species and cognitive domains.

5.2.1 Implicit learning

Evidence for the cognitive use of frequency and cooccurrence information is found from so many domains, and it is seen as so fundamental to human mental functioning, that there have been proposals that it amounts to a separate kind of cognition. The literature on this issue is quite rich, and we can do no more here than sketch some of the currently interesting theories.

Petri and Mishkin (1994) presented an hypothesis that the brain has “two fundamentally different learning and retention systems [which] use different circuitry within the brain, store different aspects of experience and follow different rules of storage” (p. 36). One system stores memories and another system develops habits; these are also termed “declarative vs. nondeclarative,” or sometimes “declarative vs. procedural.” They suggested that:

The question of what is stored in a habit system is important. It is not the cortical representation of stimulus events and their associations, as it would have to be in a cognitive memory system. Instead it is the probability that a stimulus will elicit a response, a probability that is determined by previously reinforced pairings... The product of habit learning is assumed to be a stimulus-response bond not accessible to conscious experience; it is only a tendency to respond in a particular way in a particular situation. The characteristics of a stimulus are capable of triggering a response even though those characteristics need not be recognized—that is to say, no awareness is

required. Further, it appears that response probabilities develop gradually as a result of repeated exposures, with each exposure incrementally changing the stored response probability. (p. 36)

As this quote indicates, a principal distinction to be made about mental processes is the extent to which they are conscious. It turns out to be difficult to define "consciousness" empirically, since it is not a discrete entity, but varies in distinctness from the point of view both of the subject and of the experimental observer.

Berry & Dienes (1993:2f) sketched a typology of "unconscious" learning:

<u>Subject is aware of</u>	<u>...perceiving stimuli?</u>	<u>...links between stimuli?</u>
<u>Types of learning:</u>		
"perception without awareness"	no	no
"implicit memory"	yes, then forget	no
"implicit learning"	yes	no
"automatic processing"	yes	yes, then forget

An example of "perception without awareness" is what has been called "subliminal messages." "Implicit memory" ("tacit knowledge") is at work when word recognition is faster for a recently seen word than for an initial presentation.

"Implicit learning" is when performance improves on patterned stimuli with greater repetition, even though the subject is unaware of perceiving any pattern.

"Automatic processing" ("automaticity" or "overlearning") enables us to drive a car with a minimum of conscious attention while discussing philosophy or thinking about a tennis game. The problem of a precise behavioral threshold for determination of conscious awareness causes disagreement about how to interpret results of various experimental paradigms.

Hasher & Zacks (1984:1375) showed that the encoding of frequency data is an implicit, "automatic" process, that it "occurs without intention... and neither

does it benefit from practice or from feedback,” and that it is uninfluenced by individual differences, including age:

Beyond the age of four or five, sensitivity to frequency may not change... Furthermore, the ability of newborns and infants to discriminate numerosities may indicate that some capacity for encoding frequency is present from birth... The age invariance in frequency processing supports the notion that the encoding of frequency is a fundamental cognitive process for which we are attuned early in life. This operation may belong to the category identified ... as “basic” memory processes. Such processes are thought to make up the hardwiring of the memory system. For these, developmental trends are completed at an early age and are largely the product of maturation. (Hasher & Zacks 1984:1377f.)

A cognitive process that is fundamental to the organism and which operates with great robustness seems like a logical one to be available for bootstrapping in infancy. So, it makes sense that implicit learning should be a good candidate for investigation. While it is not possible to review here the voluminous literature on implicit learning and memory, we can note that the experiments we discussed in Section 5.1 above employ stimuli with characteristics similar in several respects to those used in experiments in implicit learning (Reber 1993:27f.):

- Novel, not already familiar to the subjects.
- Made of synthetic and arbitrary materials, and possessing no intrinsic meaning or emotional content.
- Possessing a pattern which is too “rich and complex” to be consciously perceived by the subjects.

Experimental procedures can also be compared to those in the implicit learning paradigm, for instance, artificial grammar-learning studies (Reber 1993:35ff.):

- Subjects are purposely not informed that they are working with patterned stimuli.
- Subjects are repeatedly exposed to exemplars of the patterned stimuli in some context that forces them to pay close attention; for instance, subjects might be asked to study each exemplar until they can reproduce it exactly.
- Subjects are tested by discriminating between novel exemplars that show the same pattern and those that do not.

Because the stimuli in these segmentation experiments use stimuli with much less complex patterns, subjects are just asked to attend to them and are not required to study or memorize them explicitly.

These hypotheses about cooccurrence frequency and consciousness are not new to cognitive psychology, but they are not widely acknowledged or well developed in linguistics and its subfields. We would like to know what kinds of linguistic events are recorded, and how they influence linguistic knowledge. Are the authors of the experiments cited above correct in their hypotheses about which statistical relationships account for their results? Are there other possible statistical cues, or might different cues be used in different linguistic processes?

5.2.2 What is counted? Salience and redundancy

Whether frequency recording is conscious or unconscious, we may reasonably ask what is being counted. The organism is immersed in the real world of an enormous number of events. Even if this inundation is limited by the organism's sensory abilities, there must be additional mechanisms for focusing attention on certain events, or certain kinds of events, for recording. We can call this

focusing process "salience," a term which in itself does not tell us what is attended but emphasizes that not everything can be.

One candidate for salience might be the importance to the organism: A large rock falling on one's foot will be more "salient" by any measure than the same rock falling into an empty puddle some 100 meters away. Such personal events are relatively few, however, so we can look for other criteria to discriminate among the numerous non-personal events which occur.

Many features that might lend salience to events can't be quantified. As we saw above, the ends of sequences tend to be more salient than the beginnings for humans, at least, and the middles least memorable of all. Analogy of an event with another event may lend it salience idiosyncratically: A particular smell may become highly memorable because of an important event that it reminds us of. But perhaps for infants there are fewer such possibilities.

But one candidate for salience is certainly frequency. Events which recur tend to be noticed in more or less direct proportion to their recurrence, but with a proviso: Very highly frequent events take the opposite path and recede from notice, turning into background "noise." There is a paradox here, since presumably the background events must at one time have been noticed, counted, and judged to be so frequent that they were relegated to the background of awareness. Should this background noise suddenly cease, however, its very absence would be sufficient cause for becoming salient once more. In this case, it seems that infrequency would be a criterion for salience! How can we make sense of this?

We will hypothesize that event classes which exhibit a contrast in frequency are more noticeable than those in which all variants are about equally likely.

With no event much more or less likely than any other, the class is soon perceived as background noise, or texture. One's attention cannot "grab on" to any part and the tendency is to perceive the whole mix as an undifferentiated mush. This distinction between a skewed distribution and an even one is related to **entropy**, or to the amount of **variance** (Miller 1956; Garner & McGill 1956). We can describe the texture-like or monotonous end of the spectrum as having **zero redundancy** and the more noticeably contrasting mix as having a higher redundancy.

[Entropy is] ...not how loud some demon is shouting but rather the relative loudness of that demon in the crowd of relevant demons... one cue has its greatest effect when a second cue is at its most ambiguous level. Thus, the most informative cue has the greatest impact on the judgment. (Massaro 1992:326)

Somewhat counterintuitively, when the frequencies become skewed over a wide range, producing a very high redundancy, the cognitive effect is similar to the zero-redundancy case and, by hypothesis, learnability is once again decreased. Consider the following invented series, each using four letter types in varying quantities, and the unigram redundancies:

<u>Unigram Redundancy</u> ⁸	<u>Series of four types</u>	<u>Effect</u>
0%	A B C D C D B A A C B D	random
10%	D A D B C D B D C D A D	interesting
17%	D A D B D C D B D C D A D D	salient
32%	D D A D D B D D C D	boring
45%	D D D A D D D B D D D C D	monotonous

In the low-redundancy series, each stimulus is used an equal amount; as redun-

⁸ Redundancy = $100 \cdot (1 - H / H_{\max})$, where H is the entropy of the distribution of unigram (character) types, and $H_{\max}$ is $\log_2$ of the number of unigram types.

dancy increases, there is more contrast in the frequency distribution, with some types used rarely and others very frequently.

Thus, there is hypothesized to be some middle range of contrast, between randomness and monotony, where stimuli are most easily apprehended. We can see from a sampling of the corpora seen in Chapter 4 that transcribed natural languages have redundancies⁹ in the range of 10-20%:

<u>Corpus</u>	<u>Unigrams</u>			<u>Bigrams</u>		
	<u>Types</u>	<u>Entropy</u>	<u>Redund.</u>	<u>Types</u>	<u>Entropy</u>	<u>Redund.</u>
ENG V (phonemic)	50	5.11	9.4%	1327	8.67	16.4%
ENG W (orthog.)	26	4.17	11.4%	562	7.73	15.4%
SPA V	32	4.26	14.8%	593	7.62	17.3%
SPA W	26	4.01	14.6%	426	7.39	15.4%
JAP V (roman chars.)	24	3.88	14.4%	234	6.78	13.8%
JAP W (hiragana)	78	5.59	11.1%	2509	9.98	11.7%

As one very simple test of this hypothesis, we can take an experiment we have not yet discussed. Goodsitt, Morse, VerHoeve & Cowan (1984) tested whether infants would recognize a familiar syllable more readily when it was embedded in a background context of "uniformity or reduced complexity" (p. 907) versus a more varied context. In this experiment, infants 6.5 months old were acclimated to a single syllable ([du] or [ba]), and then tested on a trisyllabic sequence which included the familiar syllable and either one other syllable repeated twice, or two other syllables different from each other and from the famil-

⁹ As explained by Garner (1962:245), redundancy of n -grams for $n > 1$ can be interpreted in two ways. "Sequential redundancy" is the loss in possibilities due to some combinations not seen. For this purpose, the comparison is between the observed entropy and the entropy for an even distribution of all possible combinations of the unigram symbols. For ENG V, for instance, this would be with respect to 2500 possible bigrams (50x50), maximum entropy = $\log(2500) = 11.29$, or a redundancy of 23.2%. Instead, we are using here the "distributional redundancy" or the loss in possibilities due only to the skew in the distribution of those combinations observed, maximum entropy = $\log(1327) = 10.37$.

iar syllable. The position of the familiar syllable was varied in the repeated test stimuli, and the trisyllable groups were separated by a pause of 2 seconds. Thus, infants who had been acclimated to syllable X heard either Y Y X # Y X Y ("redundant" condition) or X Y Z # Z X Y ("mixed" condition).

By our hypothesis, the X Y Y pattern, as opposed to the X Y Z pattern, should serve to highlight the target syllable by creating a frequency contrast. The "redundant" condition has 8.2% redundancy, a fairly good value for processing, while the "mixed" context—three symbols with exactly equal frequency—has zero redundancy and should be very difficult to process. Results were that infants in the "redundant" condition recognized the familiar syllable more readily, so this experiment is evidence that discrimination is indeed easier when the signal has a nonzero redundancy, even for very young infants.

5.2.3 What is useful? Cue reliability

Our second question is: What is cognitively useful? Presumably, the organism devotes attention and energy to cognition that serves its goals. Useful observations are those with a good chance of predicting the future—those events which are highly correlated with phenomena of importance. That is to say, cognition feeds not on sheer frequency, but on correlated frequencies, on cooccurrences. For instance, knowledge that tigers frequently attack people is not cognitively useful, generating only pervasive fear. However, noticing that this event often happens near certain places or at certain times of the day, can lead to life-preserving cautions. In other words, two events which are correlated will get more notice than two more frequent ones which are not correlated. By thus building

up correlations, an organism can increase its grasp of the surrounding reality.

We would like to know in what terms humans encode this correlation. Specifically, what statistical relationships in language are most predictive, and thus likely to be an index of cognitive usefulness? We will compare three plausible statistical cues, using the discovery of word boundaries as our goal:

- **Raw frequency:** Since word-internal bigrams are fixed and bigrams across boundaries vary with the numerous possible word combinations, it seems likely that word-internal bigrams occur more frequently on average.
- **Transitional probability (TP):** This relation, which is the preferred explanation of Goodsitt et al. (1993) and Saffran and her colleagues, compares the frequency of a pair with the frequency of the left component. It draws support from the obvious fact that speech is linear, proceeding from “left” to “right,” so to speak. High transitional probability is hypothesized to be more typical of word-internal pairs than of pairs across a word boundary.
- **Mutual information (MI)** compares the frequency of a pair with the frequencies of each of its two components. Intuitively, it seems likely that the human mind keeps enough information in short-term memory to be able to record both the forward and backward context of salient elements. Therefore, it would be possible to take into account the frequency of both members of a pair. Another cognitive consideration is that many languages place function words after the related content word not, like English, in front of it, and surely the relative frequency of both halves of *inu wa* ‘dog-TOPIC’ in Japanese should be taken into account just as are both halves of *the*

dog in English. Like TP, the higher value of MI is expected to indicate a cohesive pair.

Local entropy has sometimes also been proposed as a segmentation cue, based on the intuition that entropy is a measure of choice, and there is greater choice (higher entropy) at boundaries than within cohesive units. Juola, Hall & Boggs (1994) used this idea to segment words into morphemes, with moderate success on a very small trial. We also developed an algorithm which tested local entropy (both absolute and relative) as a metric to make segmentation decisions, but we found that this technique was consistently much less successful than the approach used in the O-B algorithm, so we reluctantly abandoned it.¹⁰

We can begin investigating the reliability of these measures for word segmentation by seeing how well they are correlated with word boundaries in our corpora.

5.2.4 Cue reliability in corpus experiments

It is repeatedly assumed in the literature that one or more of these statistics correlates well with word boundaries (Saffran, Newport & Aslin 1996), but none of these claims are documented. Before we conclude this discussion, it may be useful to determine, for at least one corpus, the amount of information in the signal which can be extracted by using these bigram statistics.

¹⁰ Recently, another illusory claim for local entropy was made (Clark, Gleitman & Kroch 1997:1179): "Statistical learners can analyze a text based on a small set of words by first computing the conditional entropy (uncertainty) of points around each syllable and then associating word boundaries with points of high conditional entropy." When challenged, Robin Clark responded (p.c.): "[The procedure] produces units that are candidate words and it really works only on a very small sample. On a large sample, the syllables should distribute more evenly, so it shouldn't work (neither should transition probabilities). This was clearer in the original letter, but was edited down to an adjective in the printed version."

Using one of the corpora discussed in Chapter 4 (“EngV large,” a phonemic transcription of adult speech to infants), we segmented the “collapsed” text into words according to three different bigram criteria: raw frequency, transitional probability, and mutual information. Each of these is predicted to be higher for bigrams inside a word than for those across a word boundary. For each trial, we listed all the bigrams in order of their likelihood to fall across a word boundary as indicated by the appropriate measure, and then simply placed a space (word boundary) between each bigram pair that fell below some cutoff value. For each measure a cutoff value was chosen such that the segmented result would contain roughly the same number of words as the standard. (Note that this was not a true parsing procedure, which looks for the best combination of cuts according to some global criteria, but rather a deterministic segmenter that looks only at character bigrams.)

We will show below that the three bigram measures do not produce very different results when evaluated quantitatively, but the reader may wish to first try a qualitative evaluation. Here are the first few lines of the corpus in the standard segmentation and the segmentation of the same lines from each bigram technique:

Std: yu want tu si D6 bUk ["You want to see the book?"]

Freq: yuwant tusi D6bUk

TP: yuwant tusi D6bUk

MI: yuwan t tusi D6bUk

Std: 1Uk D*z 6 b7 wIT hIz h&t ["Look, there's a boy with his hat!"]

Freq: 1Uk D*z 6b7 wIT hIzh&t

TP: 1Uk D*z6b7wIT hIzh&t

MI: 1Uk D*z 6b7wIT hIzh&t

Std: &nd 6 dOgi ["And a doggie."]

Freq: &nd 6dOgi

TP: &nd 6dOgi

MI: &nd 6dOgi

Std: yu want tu lUk &t DIIs ["You want to look at this?"]

Freq: yuwant tu lUk&tDIIs

TP: yuwant tu lUk&tDIIs

MI: yuwan t tu lUk&t DIIs

Std: h&v 6 drINK ["Have a drink."]

Freq: h&v 6d rINK

TP: h&v6d rINK

MI: h&v 6drINK

Std: oke nQ ["Okay now."]

Freq: oke nQ

TP: oke nQ

MI: oke nQ

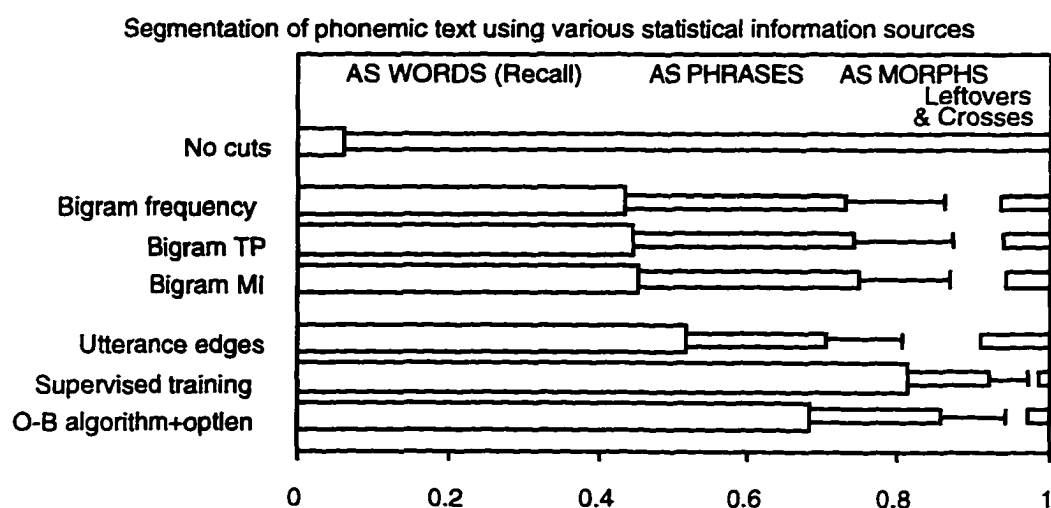
Std: WAts DIIs ["What's this?"]

Freq: WAtsDIIs

TP: WAtsDIIs

MI: WAtsDIIs

The chart below shows results for the three measures, along with several other versions for comparison, which are explained just following.



- **No cuts:** The tiny recall figure is simply the proportion of words that are found in one-word utterances (6%). The proportion of utterances that contain only one word is 21%; this is the same as word precision for this measure, but precision is not shown in the chart since “No cuts” is the only measure for which there is a noticeable difference between precision and recall.
- **Utterance-edge bigrams:** Phoneme pairs at the beginnings and ends of utterances were collected and tallied; note that these bigrams did not lie across boundaries, but occurred just following and preceding an utterance boundary. The bigram types that most often occurred in those positions were used as a key to segmenting. That is, for every bigram type which frequently began utterances, a word boundary was placed prior to every instance of it in the corpus; for every type which frequently fell at the end of an utterance, a word boundary was placed just after every instance of it.
- **Supervised:** In a supervised training procedure (unavailable to pre-lexical infants), we collected statistics for each bigram about how often it occurred across a word boundary in the standard segmentation and how often word-internally. We then ranked the bigram types in order of their propensity to occur word-internally and chose a cutoff value for this propensity. Then, starting again with the “collapsed” version of the text, we placed a word boundary between every instance of any bigram whose “propensity” value fell below this cutoff (an all-or-nothing deterministic procedure). This procedure shows the extent of actual correlation between word boundaries and bigram types.

- O-B algorithm: Finally, we repeat here the comparable results for the O-B algorithm (optlen version) as reported in Chapter 4.

The information from the utterance edges, which is probably available to the infant, is somewhat more informative than the raw bigram information. The word-boundary bigram occurrences are most informative. Although this is supervised training and thus not realistic for the infant's cognitive problem, it provides an upper bound for this kind of information (correlation of word boundaries and phonetic cooccurrence information). The chart also shows that the O-B algorithm does better than any of the bigram information sources, though not as well as the supervised training.

Which of these three statistics does the O-B algorithm use? First of all, the O-B algorithm is an "on-line" procedure: It interleaves the gathering of statistics with parsing, and records in the lexicon only instances of those "words" that it parses. In this way, it is continuously both observing and acting on its observations. Secondly, it does not use fixed-length n-grams, but strings of varying lengths; indeed, it favors the longer of two equally frequent strings. For both of these reasons, it is a more complex mechanism than rules based on n-gram statistics. Comparing it to the frequency metric: Although it does use frequency information, it does not use raw frequencies from the entire corpus, but frequency based on its prior actions. Nor does it give special priority to strings with the highest absolute frequency but looks at the relative frequencies of competing units in the section to be parsed. Comparing it to mutual information and transitional probability: Its operation is more like MI than TP, since it is not one-directional but gives equal weight to all the components of a combination.

To see how well these statistics explain experimental results, we will take a second look at each of the four experiments in speech segmentation discussed in Section 5.1, and also apply the statistics to two experiments in syntactic segmentation.

5.3 Segmentation experiments redux

Returning now to the experiments presented above, which showed that adults and infants use distributional information to segment speech-like artificial stimuli, we can ask what sort of statistical cues did they use? Do the results of these studies support our hypotheses about cue salience and reliability?

5.3.1 Cue reliability in speech segmentation experiments

In Goodsitt et al. (1993), the stimuli were trisyllabic “words” separated by silence. The research question was whether infants could make a more compact representation of one stimulus stream than another, so the comparison is between conditions. The stream in each condition was made up of two “words,” and the table below shows the distribution across the internal bigrams (in this case, bisyllables) for each of the three measures in each condition. For instance, internal bigrams in the invariant condition are *koga* (which occurs twice as often as the others), *gade*, and *deko*; the bisyllable *dede* is not internal to a trisyllabic word. Frequency here is apportioned among the bisyllables (sum to 1.0), and TP and MI are calculated to also include the boundary pause as a possible second element.

<u>Condition</u>	<u>Stimuli</u>	<u>#Bisyllable</u> <u>Types</u>	<u>Freq %</u>	<u>TP %</u>	<u>Mutual Info.</u>
Invariant	kogade dekoga	3	25, 25, 50	50, 50, 100	1.17, 1.17, 2.17
Variable	kogade degako	4	25	50	1.17
Redundant	kokode dekokoko	3	25, 25, 50	25, 50, 50	0.17

For all measures, the variable condition is clearly inferior, since there is no statistical distinction between the bigrams: All have the same characteristics. In the other two conditions, there is a frequency differential in favor of one bigram type (koga and koko, respectively). However, TP and MI measures do not pick out one bigram over the others in the redundant condition, but they do in the invariant condition. Thus, raw frequency cannot distinguish between the invariant and redundant conditions, but both TP and MI measures equally well explain the fact that in the Goodsitt et al. (1993) experiments the invariant condition performed better than the other two conditions.

In the other experiments, the subjects listened to a continuous stream of sound and were asked to locate implicit word boundaries based on the greater cohesion of some bigrams than of others. To analyze these results, we compare the statistics of the word-internal bigrams to bigrams lying across a word boundary, as the subject heard them in the training stream. If there is significantly stronger cohesion inside the word than outside, we can hypothesize that the experimental subjects could locate word division points.

In the **Hayes & Clark (1970)** experiments, both four-phoneme languages showed no difference between inside and outside bigrams for any of the three measures. This is the zero-redundancy case: All bigram types had exactly the same frequency of occurrence. With four phonemes, there were only eight

possible CV or VC bigram types and all bigram types were found both inside and outside a word, with no way for the subject to distinguish a word boundary. The values for the four-phoneme languages are shown in the table below for comparison with the more relevant ones for the larger languages:¹¹

<u>Condition</u>	<u>#Bigram</u>		<u>Frequency %</u>		<u>Trans. prob.%</u>		<u>Mutual info.</u>	
	<u>Tokens</u>	<u>Types</u>	<u>Outside</u>	<u>Inside</u>	<u>Outside</u>	<u>Inside</u>	<u>Outside</u>	<u>Inside</u>
19-phoneme	30	30	1.1	3.3	11–17	33–100	0.7–2.3	2.3–3.9
16-phoneme	24	24	1.4	4.2	11	33–100	1.4	3.0
4-phoneme	24,48	8	12.5	12.5	50	50	1.0	1.0

The “outside” values should be significantly less than the “inside” values for maximum learnability. In the language with 16 phonemes, boundaries were clearly discernible because each consonant was used only once, and there is thus a clear difference between inside and outside bigrams in all three measures: Internal bigram types were three times as frequent as outside ones, and both TP and MI showed separable ranges with no overlap. The 19-phoneme language was almost as clearly distinguishable, with only a slight overlap in MI values. In all these cases, however, there is not much difference between the three measures.

In the infant experiments of **Saffran, Aslin & Newport (1996a,b)**, the statistical structure of the training sets for both experiments was identical. Only the construction of test stimuli differed, with words and nonwords contrasted in Experiment 1, and words and part-words in Experiment 2. Here we show the statistics for inside and outside bisyllables in the training stream, and then

¹¹ The actual stimuli are not reported in the original article, only some statistics. For instance, it is stated that for the first dialect there were 30 tokens and 19 types and that “many phonemes were used in more than one word” (Hayes & Clark 1970:226). So we can guess that for consonants there were 15 tokens and 12 types, and 15/7 for vowels. The statistics in the following table are calculated from stimuli constructed according to their descriptions of this sort.

compute statistics for the test stimuli as the average of the training statistics for the two bisyllables in each test trisyllable. Thus, for instance, the test stimulus *tupiro* (a word in the training stream) contains the two bisyllables *tupi* and *piro*, both of which are word-internal and have a frequency in the training stream of 8.3% each, so their average is also 8.3%.

<u>Bisyllable statistics:</u>	<u>Frequency%</u>	<u>Trans. prob.%</u>	<u>Mutual info.</u>
Training [Outside / Inside]	[2.8 / 8.3]	[33 / 100]	[2.0 / 3.6]
Test stimuli:			
Word average	8.3	100	3.6
Part-word average	5.5	67	3.0
Nonword average	0.0	0	0.0

We can see that there was a clear statistical distinction by all measures between inside and outside bisyllables, with the outside ones only one-third the magnitude of inside. There is thus a good separation of test stimuli, even between words and part-words, by any of the three measures.

The adult experiment of **Saffran, Newport & Aslin (1996)** is a little difficult to evaluate, as discussed in Section 5.1.3 above. This experiment was similar to the infant experiment just discussed. After training on a continuous stream of six words, subjects were asked to choose between a word and either a part-word or a nonword. As just shown for the infant experiment, we would like to compare the statistical attributes for each of the test stimuli, based on their cooccurrences in the training stream. However, this adult experiment had a more complex training set, and only a few of the test stimuli used were given in the paper. Here we show the statistics (our version) for each of the known stimuli and provide a min/max range for the missing ones. For each stimulus and measure, we give

the value for each of the two bisyllables in the stimulus (first+second), and their average:

<u>Test stimuli</u>	<u>Frequency %</u>	<u>Transitional prob.%</u>	<u>Mutual info.</u>
Words:			
du ta ba	$5.6+5.6 = 5.6$	$100+100 = 100$	$4.17+3.17 = 3.67$
pa tu bi	$5.6+5.6 = 5.6$	$50+ 50 = 50$	$2.17+3.17 = 2.67$
tu ti bu	$5.6+5.6 = 5.6$	$50+100 = 75$	$3.17+2.17 = 2.67$
pi da bu	$5.6+5.6 = 5.6$	$100+ 50 = 75$	$3.17+1.17 = 2.17$
bu pa da	$7.8+5.6 = 6.7$	$35+ 50 = 43$	$1.66+2.17 = 1.91$
ba bu pu	$6.7+5.6 = 6.2$	$60+ 25 = 43$	$1.43+2.17 = 1.80$
Part-words:			
pi da ta	$5.6+0.0 = 2.8$	$100+ 0 = 50$	$3.17+0.0 = 1.59$
bi ta ba	$0.0+5.6 = 2.8$	$0+100 = 50$	$0.0+3.17 = 1.59$
bu pa bi	$7.8+0.0 = 3.9$	$35+ 0 = 18$	$1.66+0.0 = 0.83$
Part-word max. ¹²	$7.8+0.0 = 3.9$	$100+ 0 = 50$	$4.17+0.0 = 2.09$
Part-word min.	$5.6+0.0 = 2.8$	$25+ 0 = 13$	$1.43+0.0 = 0.72$
Nonword	$0.0+0.0 = 0.0$	$0+ 0 = 0$	$0.0+0.0 = 0.00$

Here, unlike the previous three analyses, we see differences among the three sets of statistics. Frequency often creates a different ordering than TP and MI; *bupada* has the highest frequency, for instance, but is at or near the low end for TP and MI. TP and MI are roughly parallel, with the same high and low candidates, but with three in the middle range patterning slightly differently.

Because we do not know all of the test stimuli, nor any of the actual scores of any of them, we cannot draw firm conclusions.¹³ But in the future, this kind of

¹² Only three of the six part-word test stimuli were given in the paper, so we have indicated the possible value range for the missing three. It appears from the three part-words given that the non-word bigram in each part-word was one that never occurred in the test stream (all statistics here=0.0), but this is not explicitly stated. Here, we assume that such is the case for all part-words.

¹³ As a side note, Saffran has said (p.c., April 1996): “‘babupu’ is the worst word in every experiment I’ve ever run with this language. And that is interesting because if you buy the idea that the ends of words are learned first... the end of that word would be really hard to find – .25 tr/pr between bu.pu, and then .2 between pu and the beginning of all the other words.” In other words, there is not enough contrast between the inside and the outside of that word. Indeed,

experiment could be used to test the usefulness of TP vs. MI. It is possible that some part-word stimuli had TP or MI values which were higher than those of the word with which they were contrasted. In these cases, the performance results (how often each was judged to be word-like) could be matched with these values and any differences in correlation noted. We can look forward to future experiments which will disentangle some of these possible explanations and give us more information about the human statistical processor.

5.3.2 Experiments in syntactic segmentation

Another kind of language segmentation is the segmentation of words into syntactic phrases. One hypothesis as to how hearers could recover the syntactic structure of a sentence, was that “in all languages there will be a small set of words or morphemes, each of which is associated with relatively few syntactic constructions and can therefore act as a marker to signal their occurrence” (Green 1979:482). Another characteristic of markers is that they are generally much more frequent than the content words they mark. In fact, frequency is thought to be one characteristic (in addition to prosody, etc.) which helps identify them as markers rather than as “content words” or referential elements.

It has proved difficult to get evidence for the “marker hypothesis” by experimenting with people processing their own languages for a number of reasons, not least of which is that, if the theory is correct then natural languages will be well supplied with markers and it will be difficult to find sufficient contrasts to find out how people perform without them. Therefore, experimenters subjects chose part-words with correct endings significantly more often than they chose those with correct beginnings (Saffran, Newport & Aslin 1996). But it is also suggestive that “the worst word” *babupu* also has the lowest MI of all six, while it is only tied with *bupada* for the lowest TP.

have invented artificial languages that can be structured both with and without markers.

One can... see how easy it is to learn the language, on the basis that if the language is hard to parse it will also be hard to learn. As a measure of parsing difficulty, learning is obviously indirect, but it ought nevertheless to be valid since it has frequently been observed that it is easier to learn material with a clear structure than material that is apparently random. (Green 1979:485)

Let us see whether our hypotheses about salience and reliability will be borne out in these experiments. We review two of them in some depth.

5.3.2.1 Green (1979)

Green (1979), the first to test this hypothesis, developed an artificial language called "Jabberish": "an artificial language devoid of meaning but possessing rules of grammar" (p. 486), as subjects were told. Subjects sat at a terminal and interacted with a computer which could either display a new Jabberish sentence, or could receive a sentence typed by the subject and report whether it was "good" or not. After 20 or so sentences, the subject was tested on 16 multiple-choice sentence-completion questions.

The language consisted of a lexicon of 18 content words, in four classes, plus 4 marker words, one to be associated with each content word class. The four content word classes (represented by capital letters) and their members are shown on the left, and the four marker words (represented by lower-case letters) on the right

A →	LOTHOG		STINY		TYAN		PANK		FLEAM		LEAM	a →	ERD
B →	BATIM		GORT		SKOP							b →	IRP
C →	FENGLE		FURD		STOPE		PUSER		RARK		CUMO	c →	ALT
D →	RELO		DOOB		PLACK							d →	ETH

A small probabilistic phrase-structure grammar controlled sentence generation:

S →	P d D	(0.2)	P →	a A c C	(0.33)
or	P	(0.7)	or	c C a A	(0.44)
or	d D	(0.1)	or	a A b B c C	(0.22)

Here are a few sample sentences:

S → P → a A c C	→	ERD LOTHOG ALT FURD.
S → P → a A b B c C	→	ERD LOTHOG IRP GORT ALT FURD.
S → P → a A b B c C	→	ERD TYAN IRP SKOP ALT PUSER.

There were three “dialects” of the language in this experiment, with five subjects studying each dialect. In the first dialect, each content word was preceded by a marker word associated exclusively with that content word class. In the second dialect no markers were used and sentences consisted only of content words. In the third dialect, marker words were assigned to marker positions randomly without regard to the class of the following content word. For illustration, corresponding sentences from each of the three dialects:

1: “effective markers”	ERD LOTHOG IRP GORT ALT FURD.
2: “no markers”	LOTHOG GORT FURD.
3: “useless markers”	ERD LOTHOG ALT GORT ERD FURD.

Dialect 1 was the best-learned (>80% accuracy), while the other two were about equally less well learned (<50% accuracy).

Let us see if statistical characteristics will help explain this discrepancy. Since the exact stimuli were generated “on the fly” by the computer during the experiment, we computed statistics on a set consisting of all possible sentences, generated in the proportions specified by the grammar and replicated so that each lexical variant was included.

	<u>Word unigrams</u>		<u>Frequency % per type</u>	
	<u>Types</u>	<u>Redund.</u>	<u>Marker (4 types)</u>	<u>Content (18 types)</u>
1: "effective markers"	22	11.7%	4.5, 6.5, 19.5(2)	1.5(3), 2.16(3), 3.25(12)
2: "no markers"	18	1.2	—	3.0(3), 4.32(3), 6.50(12)
3: "useless markers"	22	9.0	12.5 (4)	1.5(3), 2.16(3), 3.25(12)

As the table shows, the "no markers" dialect has a very low unigram redundancy, while the other two have more reasonable ones. If, as we hypothesized, a contrast in word type frequency is necessary to marker effectiveness, then Dialect 1 would be likely to do better than the others, as evidenced by the greater unigram redundancy, with Dialect 3 a close second.

In Dialects 1 and 3, markers and content words alternate, so each group accounts for half the word tokens. Content words are always fairly evenly distributed, but marker frequencies are skewed in Dialect 1 and completely even in Dialect 3, as shown by the frequency ranges in the table above. Because markers are scattered everywhere in Dialect 3, there are many more bigram types, which should lead to a loss of cue reliability. Each marker is associated with a limited (and exclusive) set of content words in Dialect 1, but not in Dialect 3. In the table below, we can see that the number of marker+content bigram types is much lower for Dialect 1 than for Dialect 3, and that both TP and MI range are absolutely higher for 1 than for 3.

	<u>TP % range (# types)</u>		<u>MI range</u>	
	<u>m+C</u>	<u>C+m¹⁴</u>	<u>m+C</u>	<u>C+m</u>
1: "effective"	16.7–33.3 (18)	12.8–84.7 (33)	3.35 – 5.47	2.19 – 3.80
3: "useless"	3.0– 6.5 (72)	25.0 (60)	2.00	-0.75 – 2.44

¹⁴ The TP and MI calculations for the C+m exclude the trailing token, which is always a content word.

Further, the patterned relationship in Dialect 1 is between a marker and the content word which follows it; content-plus-marker word pairs have no significance. We would intuitively expect, therefore, that markers would be more useful information in Dialect 1. However, when we compare marker+content bigrams and content+marker ones, neither Dialect 1 nor 3 shows a very useful statistical relationship. Most of the ranges overlap significantly, making it difficult to distinguish between marker+content—the “meaningful” pair—and content+marker, which simply fall together accidentally. Where there is no overlap (Dialect 3 TP), the statistics favor the “meaningless” pair.

This is not a good argument for any of these measures being influential in this learning experiment. Perhaps the fact that the marker words were always shorter and fewer (4 vs. 18) than the content words (and always began with a vowel, while none of the content words did) helped the learners to distinguish between them, so that the only relevant contrast is the magnitude of the values for m+C between Dialects 1 and 3. In that case, the smaller number of bigram types and the consequent higher values for the bigram relationships may be enough to account for the better performance of Dialect 1.

Green (1979) then performed a second experiment, using a more complex grammar in which markers not only signaled lexical classes but also introduced embedded phrases (analogous to complementizers). There were six dialects, with varying patterns of association between the markers, content words, and phrases. For our purposes, this experiment doesn't add much to the discussion above, since the six dialects didn't differ very much in statistical characteristics and therefore cannot readily be contrasted. So, whatever learning took place

here is not explained by these statistical indicators.

However, one of the dialects in this experiment was the occasion for a subjective comment that is noteworthy:

Subjects in Group VI disliked this experiment and claimed that their task could bear no sensible relation to natural language use. In other groups subjects enjoyed themselves and were interested in the study and its import. The difference in emotional response was surprisingly strong and would be suggestive even on its own. (p. 492)

This is impressive evidence for a rigidity that permits of no expressiveness. In this dialect, a single marker word accounted for one-half of all tokens, with the 24 content words more or less equally dividing the other half among them.

Dialect 6 had a redundancy of 37%, compared to Dialect 1 with redundancy of 9%; here are some samples for a more intuitive grasp of the contrast:

Dialect 1:	NAL GRILE EAG KON RINTH UST LAIK
Dialect 6:	NAL GRILE NAL NAL RINTH NAL LAIK

One can see why the subjects found it completely un-language-like.

5.3.2.2 Valian & Coulson (1988)

Valian & Coulson (1988) asked, as did Hayes & Clark (1970), what characteristics of the language itself might facilitate learning. They began with the assumption that “distributional analysis” is key to learning syntactic structure:

Children do make use of distributional data; distributional analysis is a form of pattern learning; pattern learning is a domain-general process. Thus it is reasonable to study it artificially as well as naturally, with simple as well as complex languages, and in adults as well as in children. (Valian & Coulson 1988:83)

They used a language similar to that of Green (1979) but varied the frequency of markers relative to the frequency of the content words. “Sentences” looked

like the following:

```
ALT LOGOTH ERD CUMO
ALT LOGOTH ERD HIFT
ERD HIFT ALT VABE
```

Subjects studied cards with sentence strings written on them, and were then tested on novel strings. In another condition, “referents” for the content words were added, in the form of colored pictures which showed commonalities within each content word group.

Results were that the dialect which had relatively high marker frequency was learned more readily than the other one. The subjects learning the high-frequency dialect showed improvement from trial to trial, while the other group did not improve much over their first attempt.

<u>Dialect</u>	<u>Word unigrams</u>		<u>% correct at last test</u>		<u>Token/ type ratio</u>
	<u>Types</u>	<u>Redundancy</u>	<u>no ref</u>	<u>w/ref</u>	
1: “high-frequency”	14	13.6%	91%	96%	6.9
2: “low-frequency”	10	0.9	78	87	9.6

We expect that there should be a difference in cue reliability in both dialects, with markers cuing content words much more reliably than content words cue markers in Dialect 1, but less contrast between the two in Dialect 2. In Dialect 1, there was a 1:6 relationship between markers and content words; that is, each marker signaled that one of six possible content words would follow. In Dialect 2, it was 2:3, meaning that each marker signaled one of three content words, but each content word was signaled by two different markers. We show the ratios:

<u>Dialect</u>	<u>TP % range</u>		<u>MI range</u>	
	<u>m+C</u>	<u>C+m</u>	<u>m+C</u>	<u>C+m</u>
1: "high-frequency"	16.7	50.0	2.41	1.41
2: "low-frequency"	33.3	25.0 – 50.0	2.41	0.41– 2.00

There was not much difference in transitional probabilities in Dialect 2, and the contrast in Dialect 1 is in the wrong direction. Mutual information ranges, however, indicate a reliably greater contrast in cohesiveness between marker+content vs. content+marker word pairs in both dialects, but more salient in Dialect 1 than in Dialect 2.

In our terms, this experiment directly measured the amount of contrast, or skewness, in the distribution of types. Note that in one sense of the term "frequency," Dialect 2 was higher—it had a higher ratio of tokens to types. We might better term the dialects "different-frequency" and "even-frequency" to reflect the relevant difference in entropy (due to even vs. skewed distributions) which would seem to account for the difference in learnability. The authors conclude:

The results strongly support the hypothesis that learners will use high-frequency items as anchor points for distributional analyses of a referentially empty language (p. 77). ...[T]he addition of a reference field allowed the low-frequency dialect to be learned as well as the high-frequency dialect was learned *without* a reference field (p. 82). ...Even in a simple language like the one used here, a grouping cue must not only be reliable, but frequent (p. 85).

5.4 Summary

We have shown experimental evidence that frequency information is actually used by adults in experiments that imitate the speech segmentation problem, and by infants in laboratory simulations of word-finding tasks. These results

have been located in the theoretical framework of cognitive science, which assumes that a large component of knowledge is conditioned by frequency and is acquired without conscious attention or awareness. Experiments in syntactic phrase-finding simulation were also cited, with an interpretation to suggest that frequency differences in stimulus categories can contribute to an account of the subjects' performance. Thus, it seems highly plausible that the pattern extraction demonstrated by the O-B algorithm is performed in a comparable fashion by human beings.

We also looked closely at specific statistical relationships which might characterize salient and reliable cues. We hypothesized that a particular range of distribution skewness might contribute to cue salience, and that an entropy or redundancy measure gives a good indication of this. We suggested that the mutual information relationship, which takes into account the relative probability of both members of a pair, is a better predictor of cue reliability than either raw frequency or transitional probability.

We can hope that these few experiments will be the inspiration for further research. The relationships between quantitative characteristics of language and qualitative linguistic knowledge that we have shown using corpora could be used to design further cognitive experiments. Only experimental work can tell us whether information that might be useful is actually being used. If care is taken in designing stimuli, such experiments can also give us information about the statistics which the mind recognizes and makes use of.

Chapter 6. Cognitive Models

We have presented the O-B computer algorithm and shown how it segments various computer corpora, and we have seen a number of other computational approaches that also tackle the segmentation problem. Our attention to this point has been largely focused on the language input—its statistical characteristics and how much information can be extracted from it by these algorithms.

In this chapter, we will take a more speculative stance. We will ask to what extent we can consider these computer programs as models of the child's learning process—as cognitive models. We wish to examine the assumptions and methods of the O-B algorithm, and five others which have similar goals, and compare them in terms of how faithful they are to the child's behavior and how plausibly they imitate the child's mental processes. In so doing, we have several questions in mind:

- How can computer models help us understand the infant's segmentation process?
- What makes a computer model a good "cognitive model"?
- How does the O-B algorithm compare as a cognitive model of the infant segmentation process?

In beginning this discussion, we are conscious of stepping out onto a very broad expanse. There is by now a massive literature on cognitive models in general and each specific paradigm in particular. We cannot review this literature here nor, in one sense, expect to do justice to the topic. What's more, these

single-purpose computer programs are so far from being a true model of human cognition, even for this limited task, that we must stretch our imaginations to even tolerate the analogy that the term “cognitive model” proposes. However, we hope that by closely comparing several specific projects that share this common goal, we can induce a heightened awareness of some important aspects—both cognitive and computational—of the child segmentation problem.

The first section of this chapter will discuss several aspects of the child’s learning process, in order to develop some key concepts that will be relevant to the project of modeling that process in a computer. The second section will evaluate the six computer models, comparing them with each other and with the child, and will refer to and further elaborate the ideas of the first section.

6.1 The infant’s cognitive development

What do we know about infant cognition? In order to determine if computer algorithms are “cognitively realistic” we must ask what constitutes the cognitive reality. In the case of infant speech perception, this means knowing: What mental representations does the infant form? What characterizes the processes that the infant uses to arrive at these representations?

6.1.1 Frequency and clustering

We expect that the infant is able to hold in mind more than just a single perceptual or cognitive unit at a time—two or perhaps three. However, just as with adults, the range of mental grasp becomes much wider with clustering. The classic article on “The Magical Number Seven” by George A. Miller (1956) showed that mental constructs are built up by “recoding”—combining smaller

units into larger “chunks,” which can then be combined into still larger ones, and so forth. We can envision the infant starting with a set of categorical sound primitives and then building these up into the first “words.”

How is it determined which sound combinations become candidates for “chunking”? Perceptual contiguity is important, but this is not very enlightening since nearly all sounds the baby hears are next to some other sound. As discussed in Chapter 1, other aspects of the speech signal might contribute to making particular sounds more salient and encourage the baby to focus on them and use them as the center of a larger chunk. Stress and other forms of prosodic highlighting might bring some sounds into the foreground. Sounds at the beginning or end of an utterance or bracketed group probably get extra attention. And bits which recur frequently might for that reason be more likely to be noticed. Then, as we discussed in Chapter 5, there are a number of possible statistical sources of cue salience and cue reliability, including distributional skewness (redundancy), transitional probability, and mutual information.

We saw in Chapter 1 that it is not known whether the infant’s first representations of words are composed of smaller elements or are holistic and unanalyzed. It seems unarguable that the clustering process is a continuous one in at least the first year of development, with more and more complex representations of sound being created until some level is reached that can be successfully matched with adult units of meaning—words. This pairing of sound and meaning are the basis of all linguistic structure, but Atkinson (1985) argues that the infant at the one-word stage of language development is not yet “linguistic,” defined as possessing

...duality of structure or double articulation, and for a system to manifest this feature it must be the case that it exhibits two levels of patterning, the relevant patterns being governed by independent principles. (Atkinson 1985:296)

If this is so, perhaps for infants each new cluster *replaces* its components in memory, rather than *coexisting* with them, as is assumed for adult linguistic knowledge. Perhaps we can interpret the theoretical disagreement as whether or not there is a *simultaneous* existence of multiple levels of mental representation.

A replacement scenario might work like this: The child might begin with a primitive phonetic representation of a vowel, then make “copies” of it and modify each copy by adding different and new details, to produce mental representations of a set of CV syllables (or moras). Primitive versions would be mentally “erased,” or subsumed into newer, more detailed versions. Thus, for the child, each V or CV unit would be a distinct representation, unrelated to any earlier versions that might survive. We would not characterize this process as “chunking” but as expansion or elaboration. We thus cannot be sure that the infant combines units into more complex representations à la Miller (1956), at least not at first. Perhaps, as has been suggested by many, the eventual arrival at a strategy of word-making by clustering smaller units is what enables the burst of word-finding which begins in the middle of the child’s second year.

Important evidence about this matter is found in the experiments of Goodstitt, Morgan & Kuhl (1993) and Morgan & Saffran (1995), reviewed in Chapter 5. Their hypotheses hinged upon the compactness and simplicity which are assumed to result from the combining of two mental units into one:

The rationale for [the discrimination maintenance] technique is that representational units are *compact*: when infants group multiple syllables into unitary percepts, their representations of the input strings should be more dense and hence require less storage capacity, than would otherwise be the case.... Conceptually, [the] noise detection technique rests on the notion that representational units are *cohesive*, the property that lies at the root of the tendency of such units to resist interruption. (p. 913)... Because cohesiveness and compactness are fundamental properties of representational units, our results are compatible with the possibility that 9-month-olds... succeeded in representing a pair of individual stimulus syllables as a single, higher-order unit. (p. 931) (Morgan & Saffran 1995)

If their interpretation of the results is correct, it constitutes evidence for either the analytic representation of words or a robust ability to form dynamically the kind of elaborated clusters we have discussed above.

Perhaps we can attempt to synthesize these conflicting points of view. Let us assume that the infant's mental representations are very fluid, forming quickly under the influence of changing stimuli and dissolving almost as quickly. Such an hypothesis would mesh with observations that babies adapt very quickly to changing environments, including linguistic ones, picking up a new language in a matter of weeks and losing most of their acquired skill in the former one. In this case, it may be that more complex mental representations are built up by chunking, but that the components of the chunks, if they are not themselves reinforced, quickly disappear. When the child achieves a clearer notion of "wordhood" and with it a first glimmering of the nature of language as occupying more than one level, there is for the first time reinforcement for maintaining truly composite mental representations.

6.1.2 Starting small

Although infants are born into a complex world, their perception and mem-

ory are limited in capacity. This was construed as a learning advantage, and even a necessity, in a recent influential contribution to child development theory by Newport (1990). In investigating the reasons that language learning abilities gradually decline toward adulthood, she proposed the “less is more” hypothesis. As we have said, one of the most critical tasks of early language learning is to locate components of meaning such as morphemes and words. Newport hypothesized that the child’s limited perception and storage would profitably restrict the number of possible analyses:

...[P]aradoxically, the more limited abilities of children may provide an advantage for tasks (like language learning) which involve componential analysis. If children perceive and store only component parts of the complex linguistic stimuli to which they are exposed, while adults more readily perceive and remember the whole complex stimulus, children may be in a better position to locate the components. (Newport 1990:24)

She went on to say that even the danger of possibly losing data would be outweighed by the “reduction and focus in the number of computations” (p. 25).

This idea was tested in a neural network by Elman (1993). He found that learning complex syntactic relationships by a simple recurrent network was facilitated by using either of two techniques. First, the input could be staged, with the syntactically simpler sentences being received first, and only after these basic concepts were learned by the network presenting the more complex data. The second technique was to use both simple and complex data together, but in the beginning to restrict the window of “short-term memory” to only three or four words, gradually expanding it as learning proceeded (“incremental memory”). This latter plan imitated the child’s expanding capacity.

If the domain is of sufficient complexity, and if there are abundant false solutions, then the opportunities for failure are great. What is required is some way to artificially constrain the solution space to just that region which contains the true solution. The initial memory limitations fill this role; they act as a filter on the input, and focus learning on just that subset of facts which lay the foundation for future success. (Elman 1993:84f.)

The network using the second technique learned very slowly in the beginning, possibly because it had to work its way through a lot of “noise”—the beginnings of all the too-long utterances. The child, however, might ignore all but appropriately simple inputs by using a filter, such as one of the “bracketing” techniques we discussed in the preceding section. Infants would be likely to acquire skill subdividing the speech stream into shorter stretches, learning to recognize prosody and stress patterns characteristic of the ambient language and to use them to structure their perception of speech into peaks and valleys, so to speak, rather than a featureless plain.

Elman worried that a small sample size might lead to incorrect generalizations, but Kareev (1995:268) saw this as “another potential benefit of limited capacity.” He pointed out that smaller samples have more skewed sampling distributions with respect to measures of correlation, so that with limited data “people are more likely to encounter a sample indicating a more extreme relationship than the actual relationship in the population” (pp. 265f.).

....[T]he fact that working memory capacity is limited increases the chances of detecting a correlation. When a correlation exists, it is more likely to draw the organism’s attention... when the world is looked at through a narrow window..., though some of the advantage is offset, of course, by the somewhat higher likelihood of false alarms. (Kareev 1995:267)

For our purposes, there is another dimension which “starts small”—time.

The developmental period that we are discussing here is a very short one. Build-

ing on the argument of the preceding section, the child gradually forms more and more complex clusters until a critical mass of meaning units is achieved. This process is self-limiting, since hypothesized words will be reinforced as their length moves into the appropriate range for the ambient language, and when the bootstrap lexicon is achieved this first tentative system is replaced by more sophisticated word-learning procedures. The small size of working memory serves, under this scenario, to slow the pace of increasing complexity to permit the levels appropriate to "words" to get reinforcement and recognition.

6.1.3 Trial and error

One more assumption about early cognitive development is that even the tiniest infants are active participants in the project of learning about their world:

What babies do know, I believe, is how to find out about their environment, and how to organize the information they obtain so it can help them obtain more... Infants carry out many kinds of information gathering activity... They look toward sounds, follow objects with their eyes, and reach out toward things that they see... Even a newborn baby, only a few minutes old, often moves his eyes toward a source of sound and rarely moves them in the wrong direction. (Neisser 1976:63f.)

At birth, infants can see, hear, taste, feel, and touch; in this sense, they are well prepared for the new world that surrounds them. An important general aspect of sensation is that even from the very beginning of life babies actively seek out information in their environment; moreover, babies show distinct preferences for some kinds of information over others. (Bornstein 1984:114)

This exploratory character must be an inborn trait:

At the very least, the child must be engaged in an active analysis of the frequency of occurrence of speech sounds at various points along phonetic continua long before the sounds are associated with arbitrary linguistic meanings. That the child should or could engage in such an analysis is surpris-

ing, as there seems to be no immediate instrumental payoff for the effort.
(Eilers 1980:28)

Learning will proceed by trying first this and then that—by hypothesis testing at a baby level. The baby observes, then analyzes the observations, and then takes action: a search for more information of a specific kind or in a specific direction. The baby gurgles a tentative *ba* and looks to see if there is an answering smile or coo. Those hypotheses which are successful ultimately lead to learning at several levels: The object of that specific hypothesis will be confirmed and will move further along the path from a guess to “knowledge,” but also the steps that led to the hypothesis will be rewarded by success, and that path, that kind of evidence, will be slightly more likely to be preferred in the future. In statistical terms, this is “cue reliability.” Those observations (“cues”) which lead to correct hypotheses are judged more “reliable,” more dependable in predicting external reality, than those which lead to errors. Billman (1989) calls this “focused sampling”:

Focused sampling is a learning procedure which capitalises on systematicity to guide learning. As any predictive regularities are discovered, the learner allocates more attention to the predictive features [=cues]. Hypothesis sampling is biased to focus on features with a history of predictive success. It guides learning based on developing knowledge of the system being learned. (Billman 1989:130)

There is plenty of leeway in this taxonomy for individual variation. Some babies simply observe and analyze for a long time, testing hypotheses by further silent observation, and only risk an action when they are pretty sure. Others are more ebullient, trying almost everything and not minding a high “false alarm rate.”

We see the end product of word segmentation as just such a cyclic and ongoing process of hypothesis generation and testing, and then incorporating the results by further analysis. In this sense, the segmentation process does not produce results suddenly and across the board, but incrementally.

6.1.4 Frequency and rules

The issue of cue reliability returns us to a question raised in Chapter 2: When we say that children are “learning to segment language,” do we mean that they are discovering meaning units one by one and entering them in the first mental lexicon? Or do we mean that they are learning “rules,” or regularities, for segmenting their native language into words?

There is a nice distinction here that is intimately related to the child’s role as active learner. As we noted in the previous section, we can differentiate between the goal object of learning and the path by which it is reached. Babies are not trying to learn rules or regularities, but to communicate with those around them, so they use an artful combination of imitation and analysis to enter the linguistic system. Instinctively trying to minimize effort and maximize results, they learn to pay attention to the most productive regularities, the most reliable cues—in a sense, they learn “rules.”

“Rote vs. rule” is a primary tension in the design of human language, reflecting itself in lexicon vs. syntax, frozen vs. productive forms, and formulaic and creative patterns at every linguistic level. We can easily see that rote usage is based in experience and thus is affected by frequency, but it is less clear that rules might also have a quantitative basis. If we interpret “rules” as “regular-

ities" rather than as "regulations," the dichotomy is somewhat weakened:

Others have tried to distinguish between behavior which is *characterized* by rules, and behavior which is *governed* by rules. Presumably, in the first case, the behavior only accidentally conforms to a rule, whereas in the latter case the rule has causal effect. (Elman et al. 1996:102)

Regularities, or behavior which is "characterized by rules," can certainly be described in quantitative terms. Those parts of language which are regular ("rule-governed" or "productive") are also those which are more type-frequent—that is, there are more different variations which share the same characteristics—though rote elements may be more token-frequent. So, for instance, usages of the form "go to N" (*go to school, go to work*) are more idiomatic and less regular than the similar "go to the N" (*go to the library, go to the store, etc.*). The second form will occur with many more different nouns than the first, so "go to the N" might be called a "rule" and the various versions of "go to N" listed as "exceptions." (The idiomatic *go home* is even more restricted.)

So "rules," in this sense of probabilistic regularities, are being learned by the infant as one product of successful hypotheses. That is, as more and more words are learned, more and more regularities that led to their hypothesis can be confirmed. This is the opposite of the usual sense of "rule" in linguistics as something which generates linguistic objects, rather than something which is induced from them. Perhaps we can say that the words themselves are consciously learned, while the "rules" or cues that lead to successful learning are the kind of implicit learning that Reber (1993) discusses.

Therefore, undoubtedly infants are engaged in both "learning what" and "learning how" simultaneously, but observational evidence indicates that the

first project bears fruit much earlier than the second. The project of learning mature phonotactics, while certainly ongoing throughout this period, is unlikely to lead to the first lexicon. As discussed in Chapter 1, the child's linguistic knowledge is greatly increased as a result of discovering the first group of words, and it is probable that the first primitive methods of word discovery are soon discarded, or at least vastly reorganized, to point toward the adult system.

For a bootstrapping task, a goal of learning words seems more plausible than one of learning how to find words. In the next section we will see how this distinction affects the design of cognitive models.

6.2 Computational algorithms as cognitive models

In Chapter 2 we reviewed a number of computational models of the infant's segmentation process. We will now look in depth at their similarities and differences, both as computer algorithms and as cognitive models. They fall into three major groupings:

- Three of them were connectionist networks: the University of Rochester group (Aslin et al. 1996), the University of Edinburgh group (Cairns et al. 1997), and the University of Southern California group (Christiansen, Allen & Seidenberg in press); refer to Section 2.3.4 for a summary chart.
- Two algorithms, Brent & Cartwright (1996) and de Marcken (1996b), used the **Minimum Description Length (MDL)** principle.
- Two were statistical clustering algorithms, Wolff (1977) and Olivier (1968). This group will be represented by the O-B algorithm, discussed at length in Chapter 3, as the inheritor of their techniques.

None of these is really very “like” a human child, but we will see that we can sharpen our understanding of the issues involved in the child’s learning by pushing as far as we can the analogy between the infant mind and a computer program.

6.2.1 Criteria for cognitive models

What is a model? In one sense, any theory implies a model of reality (Broadbent 1980). In a somewhat narrower sense, a verbal description is called a “model” when it describes in detail some complex process, providing a fuller context for empirical observations whose relationship is not transparent. A computer program which simulates the inputs and outputs of some behavior, and possibly the internal processing as well, is also termed a “model.” These two kinds of models have different strengths and weaknesses. The computer model has greater precision and enforces internal consistency, but the narrative description can often be drawn closer to reality and include a wider range of relevant phenomena.

Models help increase our understanding by forcing us to look at more aspects of the problem, or to look at them more closely, than we might otherwise do, and they can generate new and testable hypotheses.

Cognitive scientists who build computational models to embody psychological theories do so because they believe that such models enable them to achieve a degree of explicitness that is virtually impossible to attain when working only with pencil and paper. Such models encourage a comparison of the results of the theory to actual data, which aids in refining hypotheses and evaluating them. (Hill 1987:443)

But all models are at bottom a collection of analogies, and both their power and their limitations stem from this fact. Even a precise physical model, such as a model railroad, is limited in its reality by being much smaller than the original. We need to make explicit the various analogies involved, remembering that each

one is a human invention and all are imperfect.

Cognitive models should be evaluated at least in part on how well they simulate the cognitive process—in this case, whether the resources and processes they use are plausible ones for infants, and whether the timing of events is similar to that which is observed for children. In addition, learning models have special features that cognitive process models do not.

To be realistic, computational learning models should be **unsupervised** models—having the computer “learn” from examples by induction. Supervised models present correctly analyzed instances to be learned, and are useful to discover new ways of organizing the data so as to achieve the output (Daelemans, Berck & Gillis 1996). Unsupervised models present unanalyzed instances and require that the system discover the correct analysis. As we mentioned in Section 2.2, models which perform unsupervised learning are also known as “self-organizing” and the resulting structure is called “emergent.” All the models we have looked at are mainly unsupervised, though some have supervised portions.

Models that simulate changes in representation or behavior should be **incremental**, with the results of learning felt in small stages as successive inputs are received, rather than only all at once after all inputs have been seen. An incremental algorithm “operates with limited memory and modifies its grammar on the basis of the current input, the currently hypothesized explanation, and a limited memory window into previous input” (Powers & Daelemans 1992:140, citing Winston 1984).

Models of childhood learning should be **developmental**, with critical para-

meter(s) of the learning process being altered as the learning proceeds, imitating the maturation process in young children (Ling 1996). For example, in a decision-tree algorithm, the maximum depth of decision trees can be increased gradually, or the error limit be decreased gradually. The connectionist network of Elman (1993) that starts with a small “working memory” window is an example of this. In the present case, we hypothesize that no significant maturation takes place during the few months of this bootstrapping process, so we will not evaluate the algorithms on this characteristic.

We will now undertake a close comparison of these models as a good way to bring out some features that we would perhaps have overlooked otherwise. Broadbent (1980) insists that the best way to advance science is always to compare two models, rather than look at only one:

...[O]ne ought to do experiments on the least likely predictions from one's model, so as to disprove it. But it appears to be a general rule of human nature that people do not do this... By a suitable combination of assumptions almost any data can be explained... [C]omplex and detailed theories do not get tested at the really crucial points; and if data are found that contradict them it is too easy to add yet a further complexity and to go on believing in the theory... It is quite wrong and thoroughly dangerous to approach psychological problems with one existing model in mind. (Broadbent 1980:116f.)

All of the models discussed here use the cooccurrence of phoneme combinations to “learn word segmentation.” But there are major differences in goals and methods.

6.2.2 Strategy and goals: dividing or clustering?

With respect to overall strategy, the algorithms pattern into two groups: The connectionist nets work by dividing, while the MDL and O-B algorithms

proceed by clustering. The **clustering approach** looks for cohesive and frequent groups and treats these as the objects of interest; boundaries between words are a side-effect of this process. The **dividing approach** is also concerned with cohesion, but it looks for points of unusually low cohesion and treats these as divisions, with words arising between these boundaries as a side-effect of the process.

Our approach is to discover segmentation boundaries directly, by searching for *infrequent* sequences. (Cairns et al. 1997:120)

The contrast is even more clearly seen in terms of the goal or endpoint of the process, which is the same contrast we discussed in Section 6.1.4 above between two kinds of knowledge—rote and rule, between “knowing that” and “knowing how.” O-B and MDL are modeling the discovery of particular words, which are then recognized as words and used again. The nets are modeling the discovery of how to discover words, generally phonotactic relationships between phonemes and/or features, but in some cases also using stress information. The end product of the dividing process is a “knowledge” of the statistical regularities of word boundaries¹ encoded in the hidden unit weights of the networks.

As one indication of this difference, all three networks “train” on one body of data and then “test” using a held-out portion. This reflects the designers’ view of the process as learning a skill, which is then demonstrated on data distinct from that which was seen in training. O-B and MDL, on the other hand, are engaged in a self-organization of the phonemic stream, creating a lexicon and a

¹ Or maybe syllable boundaries. Cairns et al. (1997:136) said: “...[T]he network’s modest performance in finding lexical boundaries does not reflect a distinction between word and syllable boundaries; detecting syllable boundaries is sufficient, given the relationship between syllables and word in spoken English.”

parsed version of the input. Their goal is not the ability to find new words in new input, but the knowledge store which directly results from the “learning” process.

As we pointed out in Section 6.1.4, there are cognitive arguments for characterizing the infant bootstrapping process as one of finding words to form a first lexicon, rather than one of learning word-finding rules. Thus, the clustering algorithms have an edge in casting the problem in a more lifelike form, and the bigram and connectionist procedures are in this respect less cognitively realistic.

From the point of view of evaluating the results, also, there are advantages to setting the task as discovering words rather than “rules.” The words themselves must agree with the adult versions sufficiently to serve as entry to the local language community, and they can thus be tested against that standard. But there may be multiple paths to discovering the first words, and we do not know whether there is a “correct” one. Therefore it is both more practical and more realistic to evaluate the learning by the fact of knowing particular words than of knowing particular phonotactic rules or segmentation cues.

Is it possible that the connectionist nets, too, are learning actual words? How do we know that they are making phonotactic discriminations, not lexical ones? On the one hand, they create no lexicon, and their use of featural representations increases our perception of them as discriminating on phonotactics. But they evaluate the word boundaries as output and not the rules which presumably find them. Since “connectionist networks are notoriously hard to analyze” (Cleeremans 1993:205) and we know of no empirical test which can make the necessary distinction, we will accept the judgments of their designers:

The main empirical claim behind our approach is that subregularities within a domain can be, and are, exploited to the extent that they make useful predictions. In our case the subregularity is phonotactics, the sublexical distributional regularities of phonology. (Cairns et al. 1997:142)

The most important outcome, however, was the presence of significantly higher activation for word boundaries than for within-word phoneme triplets. This indicates that, in addition to learning phoneme triplets that preceded a phrase/utterance boundary, the model also learned phoneme triplets that preceded a word boundary. (Aslin et al. 1996:129)

6.2.3 Timing

Even algorithms with similar goals and strategies can pursue them in very different ways. One measure of the cognitive reality of a model is how closely it matches the timing of the original. Children receive input continuously and, after a period of no apparent results, begin to produce and understand words. We assume that the learning process also is taking place incrementally and continuously, with new inputs modifying the results of learning so far and affecting the outputs fairly directly. How do the computer algorithms measure up as incremental and continuous processors?

- The MDL algorithm receives the worst score as an incremental process.

Both of these models accepted all input as a single event and then processed it a number of times, at the end producing a lexicon and a segmented version of the input. Processing consisted of placing word boundaries in the text and evaluating the result, focusing on the word units which result and their repetition, and then trying a new set of boundaries, and so forth. Though the computation proceeds by stages, neither the input nor the output is incremental:

Indeed, our current implementation reads in an entire input corpus before outputting a single lexicon, whereas children add to their lexicons incrementally as new input becomes available. (Brent & Cartwright in press:22)

In terms of implementation of a model, it has been argued that, since the final result is the same, the inefficiencies involved in incremental or staged outputs are not justified (Ling 1996). In this case, however, we would like to be reassured that, however inefficient, an incremental version is at least possible. Brent has described an incremental version (Brent 1997), but so far has not presented a working incremental model.

The nature of the algorithm makes this unlikely. MDL sets itself to find the best representation of a particular and finite set of data. If the set changes slightly, then a new computation must be performed. MDL was not designed to cope with the human problem of keeping up with a continuous stream of data, and it seems to be in principle impossible to do this incrementally.

- The connectionist networks received a mixed score on being incremental. The processing itself was obviously incremental, with the weights changing very gradually throughout the training process. The input was presented phoneme by phoneme and, although some nets reprocessed their input, this was presumably in imitation of “more of the same” and was not in principle necessary to the learning process.

The outputs did not seem to be incremental, since the results took the form of a test on new data after training is complete. However, since each cycle produced outputs throughout the training process, we assume that they could have been reported at any point simply by reading the current values of the

output units. Whether the outputs are incremental depends on the pattern that would emerge as the outputs changed from the beginning to the end of training.

The major problem with any bracketing approach to the segmentation task is that it cannot produce incremental outputs that are cognitively usable, and the case of the connectionist nets is a good example. Recall that the form of the output differed among the three connectionist projects, but all had some output that was interpreted as the probability of a boundary following each phoneme.² There seem to be two logical possibilities for the change in outputs as training proceeds: 1) The likelihood of many specific boundaries increases gradually and in synchrony, so that at some point in the course of training a large number of boundaries become recognizable at the same time. 2) A few boundary points are clearly identified early in the process, with more and more added as training proceeds. The first case is not incremental by any interpretation. The second case, though discovering boundaries incrementally, produces no usable "words" until far along in the training process. The first boundaries to be identified are few and widely spaced, yielding unmanageable stretches of language in between that are too complex to be accessible to the infant as objects. Counterintuitively, the "words" get smaller and smaller as experience and knowledge increase.

- The O-B algorithm, in its "training" mode, processes each utterance ("parse section") as it is encountered and produces a segmented version as it

² The Rochester and USC nets both used the relative activation level of an output unit reserved for boundary information, trained by matching it to utterance boundary information on input. The Edinburgh net used a function of the error value in a phoneme prediction task, as had Elman (1990).

goes along. The lexicon is modified for each utterance, adding possible words and continuing to gather evidence about their likelihood. This is a cognitively realistic model, with both input and output occurring incrementally and synchronously. It resembles the child's hypothesis testing, as discussed in Section 6.1.3, since words are not discovered abruptly but gradually become more and more likely as evidence for their existence accumulates—or they may fall from consideration if not much additional evidence is encountered, or if they consistently fail to win in competition with other candidates.

O-B resembles the child in the gradual growth of the words themselves. Although all word-spaces are removed from the text before feeding it to the O-B algorithm, nevertheless the default segmentation is a series of single characters. At the very beginning of the learning process, the output will be one-character "words." After a few utterances have been seen, two-character words will appear, and so on, with words getting longer and longer as learning proceeds, just as the child's process was described in Section 1.1.1. Thus, even at the beginning of the process, a few small "words" can be extracted.

6.2.4 Constraints

Each of the algorithms also has a constraint mechanism which guides and limits the learning process. They are all assessing the relative cohesion between various groups of sounds, but these values form a continuum with no clear point at which to decide that a word or a word boundary exists. How does each algorithm make that decision, and how does each compare with the child's process? What constrains or guides the learning process in each case?

Two algorithms (Wolff 1977; de Marcken 1996b) handle the cohesion continuum by producing a nested segmentation, with the most cohesive units on the inside and larger and less cohesive groups containing them. The algorithms make no decisions as to which groups represent morphs, which words, and which phrases. In one sense, this may be more natural than rigid word boundaries. Certainly adults, and probably even older children, maintain these more complex units with several hierarchical and coexisting levels:

Unfortunately, there is no single useful definition of a word... A solution to this conundrum is to avoid distinguishing between different levels of representation altogether: in other words, to try to capture as many linguistically important generalizations as possible without labeling what particular branch of linguistics they fall into. (de Marcken 1995:2)

The graphic presentations we used in Chapter 4 were a similar attempt to incorporate this insight into the evaluation of results. However, as suggested by Atkinson (1985), it is likely that infants at this age are truly “prelinguistic,” and thus could not have this much detail in their representations.

For the O-B algorithm, we have described the “optlen” parameter, which penalizes the use of too many long words so that their use is discouraged when there are shorter candidates. This constraint can be interpreted as imitating the tendency of the child to find short rather than long words, even when frequency statistics might indicate a longer recurring group of sounds. It can be thought of as the analogue of a developmental limitation, the combination of a small working memory and a short time frame, as discussed in Section 6.1.2.

The particular value of the “optlen” parameter that we used in our experiments was the one calculated to produce about the same number of word candi-

dates as existed in the standard, which simplified scoring by bringing precision and recall closer together. For the child there would be no such value, of course, but by our hypothesis in Section 6.1.2 a gradual working up from very small to larger. In fact, the demonstration by the O-B algorithm that such an incremental clustering process needs some constraint can be seen as a result that confirms the “less is more” hypothesis of Newport (1990).

In the connectionist networks, it is not clear what constrained the learning process. As we mentioned in the previous section, we do not know how the outputs changed over the training process, and the authors did not report how they decided to terminate training. In two cases, since the precision and recall percentages were fairly close together, we can see that the final state of the network produced roughly the same number of word divisions as in the standard (Rochester 86% of the standard, USC 108%). The Edinburgh results, however, showed that the number of division made by the network was only 35% of the standard number of divisions, a severe shortfall, and only 60% of these were correct (21% of the standard).

The MDL algorithm is controlled by the Minimum Description Length principle, in which the constraint is the tension between the length of the potential word and its frequency. The algorithm tries to minimize the combined length of the coded representation and the lexicon, resulting in a constraint on extremes in both directions: a long word which occurs rarely is penalized, as is a very short word which occurs too frequently. In the former case the lexicon entry is too long to result in net savings for only a few occurrences. In the latter case, each of the many occurrences will require an index reference that is almost

as long as the original word would have been.

As we discussed in Chapter 2, MDL is derived from information theory and defines the most efficient encoding of a text:

The MDL Principle is a well-motivated and theoretically sound principle for data compression and estimation from information theory and statistics. As a strategy of statistical estimation, MDL is guaranteed to be near optimal.
Li & Abe (1996:1)

Brent & Cartwright extend this to the cognitive sphere:

...[T]he notion that the best segmentation of the input is the one with the shortest representation can be interpreted as a formalization of Occam's Razor—the notion that the best explanation of a set of observations (e.g., linguistic inputs) is the simplest. (Brent & Cartwright in press:10f.)

It is not clear, however, that infant minds—or even adult minds—work according to such idealized principles, and neither de Marcken nor Brent do more than assert this. For the human brain, where computation is slow and storage is plentiful, there seems no justification for a scheme which does a lot of work in order to save storage. We agree with Christiansen et al. (in press:29) in characterizing the MDL-derived DR model as “an abstract strategy for the calculation of an optimized description of the input” rather than as “a psychological mechanism.”

To carry the argument a bit further, we can argue that MDL and other compression schemes are an attempt to render natural language—which is “natural” to the human mind—more manageable by man-made computers. The most efficient structure as gauged by information theory and compression criteria is the least redundant one, the one which has been compressed by removing as much redundancy as possible. This is not a natural characteristic of human language, which typically has a fair amount of redundancy, as we showed in Chapter 4.

6.2.5 Processing resources: mind vs. machine

We want a computer model to demonstrate plausible similarities between mental representations and processes and their computational analogues, but it is difficult to evaluate the similarities when the nature of the host is so different.

Urging caution in using models, Broadbent says:

I would firmly believe that in the long run any adequate account of human beings will have to be capable of computer implementation. The problem at the moment is rather that such models are not adequate empirically; none of them behaves quite like a person behaves, and this tends to be obscured when one discusses this without looking at concrete data. (Broadbent 1980:114)

Some years later, we cannot yet claim that computers “behave quite like a person.” It is easy to say, as did Olivier, that children can’t be expected to behave like computers:

The segmentation device, in a manner of speaking, takes linguistic experience as it comes and learns what it can from it, rather than looking for evidence to answer particular linguistic questions... [avoiding] searching and sorting operations which require large amounts of computer time and are very implausible as models for a human learner. (Olivier 1968:97f.)

On the other hand, the mind might accomplish “searching and sorting” using very different methods. The issue here is the interaction of processing and storage. We know that the brain has relatively more storage and slower processing than does a computer, but because of the codependency between representation and computation, it is possible that the brain and the computer are able to perform identical functions, each in their particular way. Certainly the computer has been made to imitate many tasks formerly thought “plausible” only for human beings, and just as certainly there are many more mental processes which

are currently unimaginable for computers to perform. Where it will end nobody knows.

With respect to these algorithms, all of them depend on resources that are most un-human-like and the hardest to analogize. The O-B algorithm has a computation-intensive parsing procedure and it stores in memory a long list of word candidates. The connectionist nets also require a good deal of computation to update the many neural weights required, and these weights occupy a good deal of computer storage to retain the high degree of precision necessary.

Only the MDL procedure, however, requires the repeated reprocessing of the original input "from scratch." Computers can separate representation and computation, storing the data as input in one place, then processing it and storing the data as transformed by computation in another place. As was said about another computer algorithm which depends on heavy computation:

[The forward-backward algorithm] is not incremental in nature, and requires that an entire data set from day 1 be saved and reprocessed offline. (Klatt 1986:315n.)

For the mind, however, there is no "offline." There is no representation without computation. Remembering data for the human brain means classifying it, which requires computation and results in the transformation of the data. Even if data can be successfully recalled, it will no longer be exactly the same "raw" data. So "reprocessing" means repeatedly reorganizing mental storage, and each time the outcome will subtly depend on the classification which preceded it.

6.2.6 Representation of language

Just as the infant develops mental representations, so a computer model of

language must have some representation of language. In Chapter 4, we tried several systems for transcribing language into some consistent set of discrete symbols and saw that the O-B algorithm handled letters, phonemes, and mora/syllables with no predictable differences among them. In particular, there was no obvious relationship between the degree of fidelity to the speech signal (as judged intuitively) and performance in the segmentation algorithm. We concluded there that it appeared that the child was free to arrive at any idiosyncratic system of its own, subject only to what is workable for the particular language. This makes the learning hypothesis very robust: Any path that gets the child to the point of scanning a categorical stream will be productive of segmentation.

Since we cannot simply copy the child's representation, is there any evidence that one representation is more "cognitive" than another? Of the other algorithms, MDL and bigram techniques used segmental representations, generally phonemes, while the three connectionist approaches all used binary feature systems. Edinburgh claimed that features are more realistic:

Two criticisms can be made of [phonemic] input. First, since we intend our model to predict the very first stages in the process of extraction of lexical chunks by an infant, if we use categorial n -grams, then this necessarily post-dates the beginning of categorial phonemic perception in an infant. Second, despite the descriptive convenience afforded by the category "phoneme"..., accounts have been advanced more recently... where phonological features are mapped directly into words, and any perception of segments may be orthogonal to word recognition. An account of speech segmentation based on an input composed of phonological features may be both more parsimonious and more psychologically realistic. (Cairns et al. 1997:130)

They also saw the input representation as determined by the nature of the model:

Different statistical approaches are appropriate to subsegmental and segmental levels of representation, so we employ below a connectionist model-

ing approach to the former and more conventional statistical approaches to the latter. (Cairns et al. 1997:112)

The other two connectionist groups focused on technical aspects of features vs. segments, though they seemed to see opposite effects:

Our strategy was to code the input as a set of articulatory features. This strategy was based on our working assumption that a phonemic input scheme would not allow the model to generalize to other phonemes that were not reinforced by the model.... This was done to allow the model to learn not just which specific phoneme ended an utterance, but also to generalize this end-of-utterance information to other phonemes in English that shared one or more features with this phoneme. (Aslin et al. 1996:126f.)

Our failure to replicate Elman's [1990] results [in generalizing end-of-utterance information] may be a result of using both a much larger number of words... and the fact that the input is coded in terms of features rather than phonemes (the contributions of the weights from a single feature are less constraining of the following phoneme than the set of weights from a particular phoneme because each feature participates in many phonemes). (Christiansen et al. in press:20f.)

We claim that binary features have no more "reality" than do phones or phonemes, and no less. The feature systems used in these nets were notational variants of phonemes derived by simply converting phonemes into features using mechanical rules. At first, Cairns et al. (1997) seemed to be an exception:

...Government Phonology elements [features] can be automatically extracted from the speech stream, so we have reason to believe that coding in this way represents a step further toward ecological validity.³ (Cairns et al. 1997:132)

But, after all, they did not code their features by any automatic extraction process, but generated them from phonemes by rule.

³ "The concept of ecological validity has become familiar to psychologists. It reminds them that the artificial situation created for an experiment may differ from the everyday world in crucial ways. When this is so, the results may be irrelevant to the phenomena that one would really like to explain" (Neisser 1976:33).

Feature systems of this sort do not encode the acoustic or articulatory details of actual sounds, but those attributed to idealized sound categories to define their position in the linguistic system.⁴ For instance, if a child's mother speaks with a lisp, the child will hear a different distribution of acoustic features, but will presumably still learn the same sound/meaning distinctions as will any other child learning the same language. In fact, the three models here each used a different set of features, based on three different abstract structural analysis of phonemes.

The conviction that features are more informative than segmental representations is sometimes even used as an argument against their use. One commentator insisted that phonemes be coded as atomic elements ("random coding" is the binary representation of numbers, as phoneme #1, phoneme #2, etc.) rather than broken down into features:

This random encoding is economic and avoids weakening the results by explicitly encoding linguistic knowledge into the patterns. (Bosch & Daelemans 1992:186)

Since language is composed of a stream of abstract categorical units that are the product of the human mind, there seems no way to represent language without using one or another of these abstractions. But it is not obvious whether more information is contributed by features than segments toward the linguistic problem of segmentation. For instance, in Spanish words end only in a vowel or one of the consonant phonemes (d l n r s z) that do not appear to follow any natural

⁴ This is not to deny that features might be part of a child's representational system. As examples in Chapter 1 showed, we can certainly hypothesize a categorical system based on features. However, the feature systems used in the connectionist systems discussed in Chapter 2 were not novel creations, but were counterparts to phonemic systems, as argued here.

class in terms of binary features. Another example: In English, the phoneme /ng/ cannot begin a word (or syllable). This phoneme has features +velar and +nasal, but none of the other nasal or velar phonemes are similarly restricted. The difference between the representations, if any, needs to be demonstrated, not simply asserted, whether it is seen as an advantage or a liability for a simulation.

The technical effect of features on the net's operation is beyond our competence to evaluate. It may be that a network using features can "learn" more rapidly, since its inputs are already partially analyzed and it may thus have less work to do. Or, it may prove the opposite, that the network has to work harder to "digest" the greater number of correlations which are possible with features. This, too, needs to be tested.

In summary, just as there is no evidence that one or the other representational unit is closer to the child's mental unit, neither is there any evidence that one or the other is more "realistic" or closer to the physical speech signal, nor that either has a clear advantage in a simulation.

6.2.7 Quantitative comparison of algorithms

Since in this research we have a cognitive goal, not an engineering one, in one sense it is not relevant to compare algorithms quantitatively. Our goal here is to show that frequency information can be mobilized in a cognitively realistic fashion, and to explore the effect of characteristics of the language input on segmentation results, not to win a "numbers game" with other algorithms. The algorithm that finds the most correct words is not necessarily the "best" or the most interesting approach. As discussed at length above, cognitive analogies are

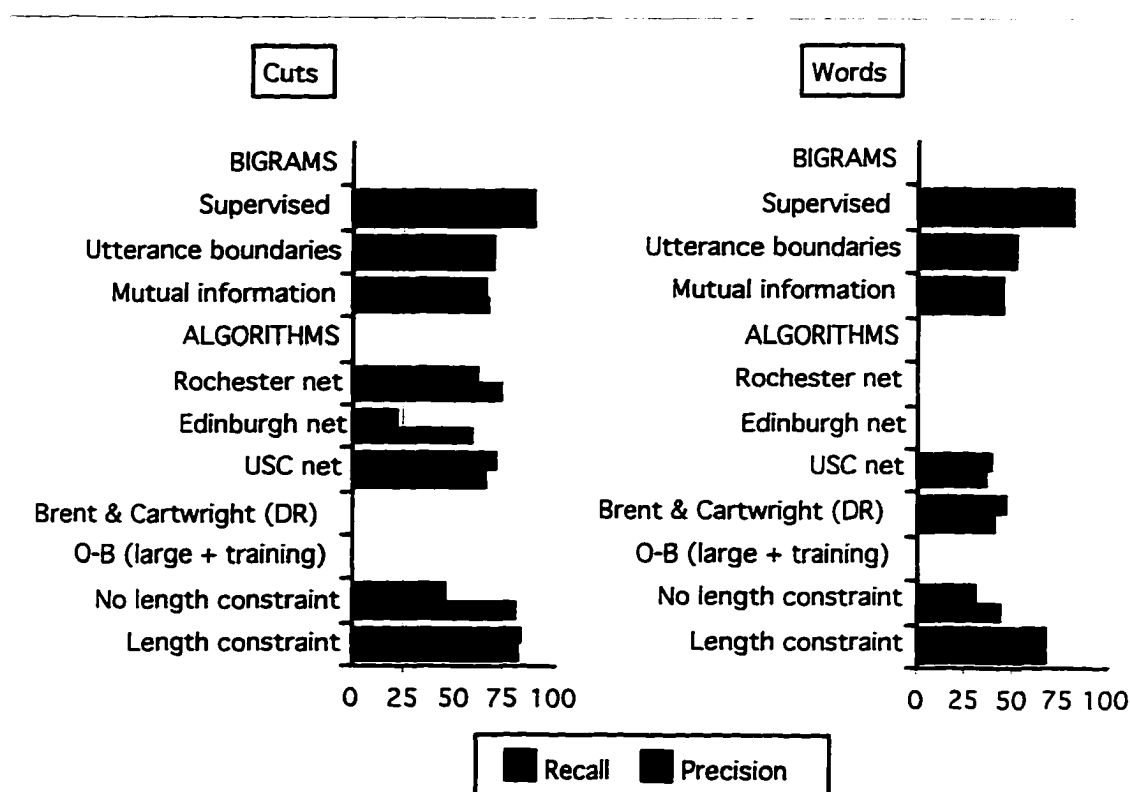
more important than percentage points in the results.

On the other hand, one criterion of a successful cognitive strategy is its effectiveness. An important conclusion of this study is that there is a good deal of quantitative information in the speech signal that could help the infant locate initial word boundaries. If this is true, it ought to be the case that there are various ways of harnessing this information and putting it to use. Just as we concluded that the infant could adopt any of a number of primitive mental representations of the speech signal and successfully use them to find word boundaries, so it would not be surprising to learn that various techniques can be equally successful in using the inherent statistics of language.

When two computationally powerful systems are given the same set of input data, they both extract every bit of data regularity from that input. (MacWhinney 1993:295)

Below we present a table of recall and precision in terms of cuts and of words (some algorithms only report one way or the other) and, following the table, a summary of what each algorithm assumed. As we have argued in various contexts above, inflexible scores like these are not necessarily the best measure of cognitive effectiveness, but they have the practical value of being widely available and easy to compare. In all cases, to enable us to best compare across different approaches, we report only the results due to distributional and utterance boundary information, omitting those from combined cues (frequency statistics plus prosody information or phonotactic rules). De Marcken (1996b) is also omitted since, as we saw in Chapter 2, he segmented hierarchically, at several levels, rather than creating a single series of words. He cited recall rates of

over 90%, but he did not give precision rates to balance against them because the number of units created (at various levels) was many times the number of words in the standard. Of the other algorithms, only the USC net reports results in terms of both divisions and words, so there are gaps where the missing scores would be.



	<u>Cuts</u>		<u>Words</u>	
	<u>Recall</u>	<u>Precision</u>	<u>Recall</u>	<u>Precision</u>
<u>Bigrams (with length constraint):</u>				
Supervised	90	89	81	81
Utterance boundaries	69	70	52	52
Mutual information	66	67	45	45
<u>Connectionist nets:</u>				
Rochester	62	74	–	–
Edinburgh	21	59	–	–
USC	71	66	40	37
<u>Brent & Cartwright (DR)</u>	–	–	47	41
<u>O-B (no length constraint):</u>				
Small, test mode	47	80	35	48
Large, test mode	36	90	25	43
Large, training mode	45	80	31	44
<u>O-B (with length constraint):</u>				
Small, test mode	81	80	66	66
Large, test mode	85	86	72	73
Large, training mode	83	81	68	67

- **Bigrams:** As we reported in Chapter 5, simple bigram statistics (with a constraint to find the standard number of words) could discover about 45% of the words in our large (33,000 words) CHILDES phonemic corpus of English. All three statistics were very similar; here we show mutual information, which was a shade higher than frequency and transitional probability. Utterance boundary bigram information counts bigrams preceding and following utterance boundaries, and supervised learning used standard word boundary information.

- **Connectionist nets:** All the nets used feature representations. Aslin et al. (1996) of the University of Rochester used a small corpus of speech to infants and tested whether a connectionist architecture using moving triplets could generalize from utterance boundary (pause) information to word boundaries. Cairns et

al. (1997) of the University of Edinburgh ran a very large spoken English corpus using no boundary information, and did not get as good results as the other two. Christiansen et al. (in press) of the University of Southern California (USC) used utterance boundary information similar to the Rochester group but with a larger corpus of child-directed speech and a simple recurrent network (SRN), with similar results.

- **Brent & Cartwright** (1996) ran a MDL algorithm on very small corpora of speech of mothers to children, transcribed in a phonemic code.

- The **O-B algorithm** is designed to work incrementally as data arrives (training mode), and with a length constraint, and most of the scores presented in Chapter 4 were for that configuration. In general, with a length constraint, more data will perform better than less, and the test mode will perform somewhat better than the training mode. Without a length constraint, the accuracy gets worse as training proceeds, since words grow too long, so the training mode will generally be better than the test mode in these cases.

Although it will not show the O-B algorithm at its best, in the interests of full disclosure we offer the most direct comparison possible with Brent & Cartwright's DR algorithm. We copied their procedure of running each of the ninety sets separately and averaging the results. Means and standard deviations for each category are shown, and the O-B results are for no length constraint, test mode:

	<u>O-B</u>	<u>DR</u>
Word recall	37.6 (6.3)	47.3 (9.1)
Word precision	34.9 (7.4)	41.3 (12.7)

The table shows that O-B has a substantially lower mean value than DR, with however somewhat less variation. Of course, the tiny amounts of data made necessary by the DR algorithm's computational requirements are distinctly to the disadvantage of O-B.

All of these algorithms (except the bigram supervised) are largely unsupervised. As we have discussed above, they make different assumptions about the learning process. O-B (and the bigram procedures) need some goal for average length. Brent & Cartwright and de Marcken use the computation-intensive MDL principle to guide their choices. The two better-performing networks receive supervised training for utterance boundaries, but then generalize this to word boundaries without supervision.

6.3 Summary

In this chapter we have emphasized the following aspects of the child's cognitive processes during early word segmentation, and showed how they can be paralleled in a cognitive computer model to a greater or lesser degree:

- The child's task is more aptly characterized as a search for particular words than as a search for word-finding techniques.
- The discovery of words is an incremental process. Language input is received continuously, and words are discovered one by one, with a few primitive words being followed by more words and more complex words.
- Word discovery is likely to issue most directly from a clustering process rather than a dividing one, since only by clustering can words proceed from small to large.

- The infant's learning is constrained by small working memory and by the interaction of an incremental process with a short time frame.

Therefore, we have evaluated the computer algorithms on the following criteria:

	<u>O-B</u>	<u>MDL</u>		<u>Nets</u>
		<u>Brent</u>	<u>de Marcken</u>	
Find words or rules?	words	words	words	rules
Continuous input?	yes	no	no	yes
Continuous output?	yes	no	no	no
Incremental word complexity?	yes	no	no	yes?
Clustering or dividing strategy?	cluster	cluster	cluster	divide
Constraint	word length	MDL	MDL/hier	?

The real gains of this examination, however, have been two. We have a clearer understanding of the nature of the infant segmentation process and the logical relationships of its various parts. In addition, by virtue of our small case study of six particular computer simulations of infant segmentation, we have an appreciation for what a cognitive computer algorithm can show us about this problem and what it cannot.

Chapter 7. Conclusions about Language

What have we shown about language, and about children learning it? Our hypothesis was that infants trying to segment the speech stream into meaningful units might use frequency information—observations about the cooccurrence of certain combinations of sounds. But the child is only half of the process of learning language. Cognitive psychology argues that an account of cognition may owe much to the nature of the input. In one such argument, Neisser (1976) discusses Gibsonian theory:

The organism is not thought of as buffeted about by stimuli, but rather as attuned to properties of its environment that are objectively present, accurately specified, and veridically perceived... The most important thrust of the theory is to suggest that students of perception should develop new and richer descriptions of the stimulus information, rather than ever-subtler hypotheses about mental mechanisms. (Neisser 1976:19)

Language itself is tightly molded to the human mind, so that learning it is virtually unavoidable for every baby.

We turn to computer models because they can extract certain kinds of structure in language and in so doing document the existence of such structure. We are thus interested to know the source of the statistical regularities that these learning models exploit. Are they in all language, and all texts? Or only in language to children? What measures can be used to demonstrate this? These are the questions we set out to answer and in this final chapter we will summarize the answers we found.

Since this dissertation has ranged over a fairly wide area, it may not be immediately apparent how the various chapters and their different background

disciplines are related. We will try to show here that it has all been about the nature of language.

7.1 The effect of input on O-B segmentation

It is indisputable that perfect segmentation into morphemes or “words” in a manner that will match human intuitions is not possible by purely statistical methods, or, as de Marcken (1996b:335) puts it: “unsupervised induction will never converge to ideal grammars and lexicons.” Natural language, for some reason, uses ambiguity on every level, so that even such an apparently simple matter as word division requires the interaction of several linguistic levels to resolve completely. We have asked, however, how far statistical information will go.

Because the O-B algorithm is not very computation-intensive, it has been possible to run a large number of experiments using different parameters, corpora, and input representations. In Chapter 4, we reported results for a wide variety of such comparisons showing that, across three languages and two registers, purely distributional information enables a significant amount of segmentation.

The most robust finding was that our spoken texts performed much better on the segmentation task than did our written texts, and that speech directed to children performed best of all. In fact, to take a representative indicator, there was no overlap in word recall rates between texts of comparable size: The worst-performing spoken text did better than the best-performing written text, even though one written corpus (JAPW) was aimed at young children. Of course, this result may be dependent on the particular texts we used. But it accords with other findings concerning the impact of quantitative characteristics.

Texts which performed better on the O-B segmentation task tended to be those with:

- Shorter words (shorter average word token length)
- Shorter utterances (shorter average parse sections)
- Higher word type repetition rate (greater G value and, less revealingly, higher token/type ratio)

These characteristics are also those which characterize spoken texts more often than written texts, and more formal or “higher” speech as opposed to less formal or “lower” speech. In particular, our studies showed clearly that adult-to-adult speech was less likely to have these characteristics than was the speech of adults to children, and speech to older children less than speech to very young children. Thus, our quantitative measures confirm general intuitions:

...[W]e suspected that speech to 12-month-olds, who are just learning their first words, is different in important respects from speech to 2- and 3-year-olds, who are producing multiword utterances. (Aslin et al. 1996:127)

...[S]peech addressed to preverbal infants is likely to be different with respect to type-token frequency and utterance length compared with speech addressed to children who have already begun to utter multi-word utterances. (Christiansen et al. in press:14)

Another clear finding was that larger texts were learned better than smaller ones, when the texts were comparable in other ways. However, this effect was much less potent than the differences listed above. For instance, even small spoken texts nearly always did better than large written texts.

None of these results was particularly unexpected, but it was noteworthy that they could be demonstrated on a wide variety of corpora. In contrast, it was

also noteworthy what could not be demonstrated. As discussed above, the representational code of the language input didn't seem to reliably affect performance. Also, the equality of various natural languages could not be demonstrated—we tried English, Spanish, and Japanese, and found that corpora controlled for register and size still showed substantial differences in performance across languages and similarities within language. We feel this must be a product of the variety inherent in different texts even within one language and the problems of comparing unlike coding structures, rather than any actual inequality of natural languages.

An implicit finding from the experiments was a respect for the current limits of the quantitative approach and a wariness of treating natural language as an information-theoretic object. We saw that even small texts have a lot of variation in them, as measured by performance on the O-B segmentation task, and we cannot always put our finger on the exact nature of the variation. For instance, it seemed that “naturalness” promoted learning: When we manipulated texts, their segmentation performance often worsened. Artificially created coding systems did less well than codes used by humans for reading and writing. Deleting every other line created a text that did less well than one the same size which was made up of contiguous lines. Perhaps these effects should not surprise us. Natural language is like a complex fabric, with many secret connections and hidden ligatures. But such behavior is quite out of keeping with the information-theoretic concept of ergodicity. If language were ergodic, or homogeneous, we would be able to rearrange sentences and paragraphs with impunity.

While our conclusions are not very surprising with respect to standard

assumptions about child learning, they do imply a revision of some current practices in modeling language. We observed many large and small differences (and similarities) in performance across our range of corpora that could not be correlated with any of the myriad of quantitative corpus characteristics that we tested. However, if we only looked at the results for two or three corpora, there were always several characteristics which appeared to account for the performances. On the basis of these adventures, we argue that conclusions about “language” which are based on a comparison of two or three similar corpora are likely to be misleading. As the number of publicly available corpora continues to increase, we need to extend our experiments to more of them, and to more different ones.

7.2 Modeling the input

What do these experiments tell us about the enterprise of language modeling? The use of computer-readable corpora to represent “language” in computer experiments is now widespread, and equally widespread are certain principles concerning the similarity of these strings of symbols to the complex reality of human language. Here is a typical statement about “realism” in language input:

[It is necessary to use] real data that are representative of the input that confronts human listeners and learners... [W]e now present what we believe to be much more adequate input for any model of segmentation: a large corpus of real speech... Use of this data will allow us to make much more substantive claims than is possible using small-scale, unrepresentative input... Linked to this issue is the tendency to use training data that are idealized and noise-free. One of the goals of this paper has been to show that real, noisy data can be used in psycholinguistic modeling and that such data allow our models to claim increased ecological validity. (Cairns et al. 1997:123f.,142)

As this quote illustrates, three factors are thought to increase the chance of cap-

turing the essence of language.

- Get a lot of it.
- Include as much variety as possible.
- Use spoken language, and transcribe it faithfully.

We will compare these with our results and show that these guidelines are not so obvious or uncomplicated as they may appear.

Received wisdom is that the larger the language sample, the more likely that variation will “average out.” This reasoning has a respectable lineage in the Law of Large Numbers, and satisfies the information-theoretic requirement for an ergodic source, or homogeneous stream. The Brown Corpus, the first and still widely used large text corpus, contains 500 text samples, each of about 2000 words, “representing a wide range of styles and authors, from press reporting through belles lettres, from learned and scientific writing through love stories” (Bell, Cleary & Witten 1990:79).

However, it is long since acknowledged that language is not ergodic; computational linguists simply treat it as if it were:

In real life, as in dealing with English, there is no such perfect test suite. Nevertheless, it is not hard to make approximations... with the understanding that the results of the model testing might be off because of biases in the suite. (Charniak 1993:35f.)

Language is so variable and multifarious that we may question whether a “representative sample” is even a possibility. One corpus-builder speaks from experience:

In building a natural language corpus one would like ideally to adhere to the theoretical principles of statistic sampling and inference. Unfortunately, the standard approaches to statistical sampling are hardly applicable to building a language corpus. First, it is very difficult (often impossible) to delimit the total population in any rigorous way... Secondly, even if the population could be delimited, because of the sheer size of the population and given current and foreseeable resources, it will always be possible to demonstrate that some feature of the population is not adequately represented in the sample. Thirdly, there is no obvious unit of language (words? sentences? texts?) which is to be sampled and which can be used to define the population. (Atkins, Clear & Ostler 1992:4)

Another desideratum is that speech, as opposed to written texts, is “real” language—generated by human beings extemporaneously, not edited and processed “offline” to conform to some prescriptive ideal. Our experiments lend support to the belief that speech and writing are different, demonstrating not only different quantitative characteristics but also robust differences in segmentation performance. However, we do not know all the dimensions of variation in language, and there are many known dimensions that our experiments did not try to quantify, such as the following:

...[S]pontaneous conversational language differs from most genres of text. More than half of the lexical items in conversational English are closed-class words, like “and”, “is”, “him” and “by”. This proportion is greater than in many genres of text, and covers a different range.... Furthermore, a corpus of real speech allows us to study a psychologically realistic level of inflected forms. (Shillcock et al. in press:4)

In our experiments reported in Chapter 4, we found that more text and spoken text did better at segmentation, all other things being equal—but they seldom were. For instance, simply reusing the same text again was better than using it only once, but was not as good as an equivalent amount of new text (Section 4.1.1). On the other hand, a large written corpus did much worse than a spoken corpus a fraction of the size (75K vs. 5K words). Similarly, the Edinburgh

network experiment did considerably worse than the two other nets, despite the fact that it used a much larger corpus, a spoken one, and went to considerable trouble to transcribe it with as much detail as possible. Of course, there were other differences as well among those experiments, so we cannot be sure that the greater amount of language harmed the results, but at least we can say that much greater size does not seem to have overcome other disadvantages.

Results like this suggest that it is difficult to quantify or otherwise characterize a sample of language. The characteristics that we can quantify can often indicate gross differences between texts, but these do not always correlate with functional differences. While intuition and the Law of Large Numbers argue that quantity and a variety of sources is one way to reduce the amount of overall variation, the prescription to seek variety in order to achieve homogeneity is a paradox. Whether more quantity and more variety are “better” depends on the purpose to which the language is being put. Many engineering applications thrive on enormous amounts of text, and are probably more robust to the extent that they are trained and tested on a wide variety of language. For lexical sampling, a grab-bag of texts like the Brown Corpus may well result in a more even or “representative” distribution. For other purposes, such as modeling acquisition, such a corpus would be undesirably discontinuous to serve as the analog of naturally connected language input.

We intuitively feel that children need to hear a moderately large amount of speech to learn the language, but it is harder to be sure that learning would improve with a greater variety of speech. It may be the case that each particular child hears a fairly restricted range of language, not only limited by being an

infant, but also by their situation in a particular family, language community, and culture. Is it important that infants hear a range of language? Or important that they not? We don't know.

As we saw in Chapter 4, some kinds of variation improve segmentation performance, while others worsen it. In Chapter 5 we showed that some sorts of variation may be an important component of salience. Building "representative" corpora is not possible until we know more about the characteristics and effects of different sources of variation. While there is so much that we don't know about variability in language and its effects on cognition, we should be cautious about expecting easy answers from corpora and models.

7.3 The future of statistics in language

It has long been acknowledged that natural language is statistically redundant, or predictable. Clearly, there is a great deal of structure in any written text, which can be described by simple statistics such as letter frequencies, probabilities of letter and word sequences, etc. Shannon (1948) explicitly showed this by generating samples of pseudo-text which looked more and more like English as the probabilities used involved longer and longer strings, and thus took into account more and more of the contextual dependencies embedded in language. From this exercise comes the hypothesis that the limit of such a process would yield the "true" amount of structure in the language, expressed as entropy. In practice, the amount of text needed to derive the statistics necessary for such a measure is an unreasonably large amount. According to Lucky (1989:113), all the text in the Library of Congress contains "only enough data to characterize the

[letter] statistics out to about ninth order—barely more than one word!”

In a classic experiment, Miller & Selfridge (1950) tried reading lists of English words to adults and asking them afterwards to write down all the words they remembered, regardless of order. The lists were of four different lengths (10-50 words), and were of eight degrees of structure, being drawn with respect to n-gram statistics of various lengths. For instance, the first list was a random list, the second was order-1, or with respect to the frequency of the individual words, the third respected pair-order frequencies, and so on, with the eighth list being taken directly from connected English prose. Two factors affected the results: Subjects remembered a higher percentage of the words from shorter passages and from passages of higher statistical “order.” The authors concluded:

The results indicate that meaningful material is easy to learn, not because it is meaningful *per se*, but because it preserves the short range associations that are familiar to the subjects. Nonsense materials that retain these short range associations are also easy to learn. (Miller & Selfridge 1950:183).

Many studies have been made of the statistics of language (e.g. Zipf 1935; Miller 1951; Pierce 1961; Kucera & Francis 1967; Lucky 1989; Bell, Cleary & Witten 1990), but little of it has been put to use in cognitive accounts of language.

This dissertation has shown the effectiveness of the statistical structure of language in one small task of first language acquisition. We have proposed that infants accomplish their first segmentation by using a variety of information sources and cognitive skills, and that one important source of information might be the cooccurrence frequency of events in the speech stream. In addition to demonstrating effectiveness, using a computational simulation, we have cited experimental evidence that adults and infants make use of such cues for segmen-

tation tasks, and we have shown that statistical information about many other kinds of events are involved in human learning over the life span.

Another contribution of this dissertation is the synthesis of evidence from three disparate fields. The literature in each of the fields is not particularly controversial or remarkable, but there are powerful implications for linguistic theory in their confluence. Despite a long tradition of recognizing the “oddly consistent” statistics of language,⁵ linguistic theory continues to step neatly over any chance to incorporate these facts. Psycholinguistics goes some little way toward acknowledging the influence of lexical frequency and parsing preferences (essentially statistical patterns of use). In syntax and morphology, the terms “regularity” and “productivity” are used rather loosely, both of which ultimately come down to quantitative measures. In semantics and syntax, the notions of “compositionality” and “idiom” refer to the two poles of a continuum, a graded distinction that cannot be applied without an implicitly quantitative scale. But these matters are often subordinated to near invisibility in favor of “principles and parameters” which involve categorical distinctions, not graded or quantifiable ones.

Recently there are a few winds of change. The increase in computational linguistics research in the past decade has focused attention on the possibility of using quantitative distinctions in natural language processing. Abney (1996) details eloquently and at length several areas of linguistics which require graded or probabilistic distinctions, including topics in language acquisition, language change, language variation, language generation, and language comprehension. The experiments cited in Chapter 5 are a further indication that interest in statisti-

⁵ I can't recall who said this, but some distinguished scholar did.

cal influences in language processing is growing. We urge more of such experiments, along with increased careful attention to controlling and reporting the statistics of the stimuli for readers to evaluate.

A recent exchange of letters in *Science* was revealing of the current conflict between inquiries into statistical influences on language and prevailing theoretical positions that discourages such exploration. The occasion was a report of infants' ability to use statistical information to segment speech (Saffran, Aslin & Newport 1996b) and an accompanying enthusiastic commentary (Bates & Elman 1996). Two rejoinders were: "Learning words and learning grammar are... different computational problems" (Pinker 1997:1178); and that the statistical procedure used to discover word boundaries must then be considered innate knowledge and "the learner must be built from the beginning to use such a procedure to discover linguistics units" (Clark et al. 1997:1179). Bates and Elman (1996; Elman et al. 1996) vigorously challenged innatist tenets, but Saffran et al. responded more moderately:

Whether this mechanism is particular to language, or is instead applied to many segmentation problems, is yet to be determined... In either case, this type of rapid and rather complex learning, while it may or may not be tied specifically to learning languages, might well be the kind of remarkable skill that permits language learning in humans to occur... Our findings... offer the possibility of a mechanism that could turn out to be suitably powerful, but also biased and constrained, so as to perform a piece of the task. (Saffran, Aslin & Newport 1997:1181,1276)

If we add these newly appreciated sources of information and new cognitive skills to the traditional repertoire of language knowledge and abilities, we may be able better to account for the many levels of language, and language acquisition, that cohere to one organic whole.

Appendix A. Parsing Algorithm

The parsing algorithm takes in a single section of text with no word spacing, usually one line or one utterance, and returns a version of this text segmented into words. A word boundary is assumed preceding and following the text section. The lexicon does not change during the course of this parse, but is altered only as a result of the parse outcome.

This appendix contains a detailed demonstration of the parsing algorithm, exercised on the sentence fragment "youthisawhim." As in the full system, the text is parsed first from the left end, and then from the right end backwards. Each column in the following table will be referred to by a letter: (a), (b), etc. The explanation here is intended to supplement the more general description in Chapter 3.

The parse begins at cycle 0 (a) with a single letter (e) – the leftmost or the rightmost. Since the lexicon is initialized with an instance of each character in the symbol set, cycle 0 always has one and only one possible parse. The number of occurrences (f) of this letter as recorded in the lexicon is the basis for calculating its probability as occurrences divided by sum of the occurrences of all items in the lexicon. Adding the log of this probability (g) to the score from the previous cycle (d)—in this case 0.00—gives the score for this cycle (h). In the table, the possible parse with the highest score in each cycle is starred (i).

In succeeding cycles, several possible parses are considered – one for each preceding cycle plus one for the current cycle. Thus, for cycle 1, two possible parses are considered, for cycle 2, three, etc. Each new parse consists of a prev-

ious parse (column c) plus one new “word” (column e). In cycle 1, the first parse considered is the entire string as a single word: *y_o*. The second parse is the winning parse from cycle 0 (*y*) plus the most recent character as a one-letter word: *y+o*. In cycle 2, first the entire string (now 3 letters) as a single word is considered: *you*. Then, the parse from cycle 1 plus the most recent letter: *y+o+u*. Then, the parse from cycle 0 plus the letters seen since that cycle as a single word: *y+ou*.

The first step in evaluating a possible parse is to determine if the new “word” is in the lexicon. If not (if column f = 0), the parse is abandoned, since only words previously entered in the lexicon may be used in parsing. If yes, its probability is computed and the log of the frequency (g) is added to the previous score for this parse (d) to get the current score (h). (Note that adding logs of probabilities is the same as taking a log of the product of the probabilities.) When all possible parses for this cycle have been scored, the highest score (least negative number) is marked the winner (stars in column i) for that cycle and becomes the basis for continuing parses in succeeding cycles. When the new score and the previous best score are the same, the previous score is retained as the new winner.

These processes continue until all characters (cycles) have been completed. The winner of the last cycle becomes the final parse. The second table below demonstrates the same process starting at the right end of a string and working to the left. This second parsing is done simply to reduce the possibility of missing a good parse by this optimized method. If both forward and backward parses have equal scores, the backward version is used, purely because of circum-

stances in the historical development of the computer program.

The same process as constrained by a length calculation is shown on the right side of the table. For each parse, the number of segments ("words") (column l) is used to compute the average segment length (m) in characters. This average is compared with an optimum segment length which is an input parameter for the run; here it is 4.18. The difference between these two is calculated (column n), and this value plus 1.00 is multiplied times the basic score (h) to get a compound score (o). The highest such score (least negative) is the winner for the cycle (marked "++" in column p).

For the most part, the winner is the same without optlen and with it. Occasionally, however, the basic scores for two candidates are close enough and/or the average lengths distant enough that the optlen calculation creates a new winner (see cycle 7 in the forward parse and cycle 6 in the backward parse). The table uses columns (j) and (k) to present the previous parse in the optlen process when it is different from that in the unconstrained process. (For a given cycle, the new segment to be added to the previous parse is always the same, so columns (e), (f), and (g) are not repeated for the optlen calculation here.)

In both the forward and backward parses, the winning parse is different for the unconstrained and constrained processes, but the forward and backward versions of each agree. Neither of these conditions is universally true. All four processes can be the same and all four can be different in a particular case.

A.1 Building the Lexicon

There are three steps in creating the lexicon:

- 1) Initialize the lexicon with all unigrams, and set frequency = 2.¹
- 2) After each parse, increment the count for every word used in the parse (value A below).
- 3) After each parse, identify all word pairs in the parse as “word candidates” for future parses. For each such candidate, add it to the lexicon if not already present, with an initial value (B in table below); if it is already in the lexicon, increment its count (C in table).

Three possible variations in this procedure exist, varying in how frequency counts are incremented. In one version, only pair candidates are tallied; in another, only words used in a parse are tallied; in a third (P+W), both parse words and pair candidates are tallied. These differences are summarized:

<u>Value in O-B algorithm:</u>	<u>Tally procedure</u>		
	<u>Pair</u>	<u>P+W</u>	<u>Word</u>
A. Increment for parse word (per word token)	0	+1	+1
B. Initial frequency for word-pair candidates	0	0	=1
C. Increment for pair of parse words (word candidate)	+1	+1	0

To reiterate, in the “word-tally” procedure, which is the principal one used in this study, pairs are added to the dictionary with a count of 1, and not incremented unless they are used in a parse (thus becoming “words”). The first time such a “potential” word is used in a parse, its frequency is incremented from 1 to 2; the next time from 2 to 3; and so on. Thus, the “word-tally” lexicon potentially

¹ Unigrams are initialized at 2 instead of 1 to prevent their being purged even if never used. The original Olivier algorithm did “garbage collection” on the lexicon, removing unused entries, and the O-B procedure allows for reducing the size of the lexicon offline before “dictin” runs. In both cases, entries with a tally of 0 or 1 are deleted.

contains three kinds of entries: a) single letters, which are initialized at 2 and then incremented whenever they are used in a parse; b) "potential" words, with frequency of 1, formed by contiguous pairs of words in a parse but never (or not yet) used in a parse; and c) those words (longer than one letter) which began as pairs and were then subsequently used in at least one parse, with frequencies greater than 1.

A.2 Comparing the Olivier and O-B Algorithms

<u>Olivier</u>	<u>O-B</u>
Parse sections contain exactly 480 characters.	Parse sections contain varying numbers of characters, depending on utterance or line length (but less than 150 characters).
Parse sections may end mid-word.	Parse sections never break a word.
Dictionary is purged of unused candidates when it gets too full.	Dictionary is allowed to expand (now we have larger computers).
Lexicon tallies are updated for each word in the parse, and also for each word-pair candidate.	Lexicon tallies are updated for each word in the parse, but word-pair candidates are not tallied until used in a parse.
No constraint; words are allowed to get longer and longer.	The parsing algorithm is constrained by an ideal average word length (optlen) and parses are penalized if they exceed it.

A.3 Forward parse of: youthisawhim

Cycle & Subcycle	Previous parse		Candidate for next segment				Total		Optlen-constrained parse						
	Segments	Score	Segment	Freq	Log(p)	Score	Score		Previous parse Segments	Score	Segs Avglen	Diff Score	Total		
(a)(b)	-----	(c)-----	(d)	-----	(e)-----	(f)	(g)	(h) (i)	-----	(j)-----	(k)	(l)	(m)	(n)	(o) (p)
0	-	-	0.00	y	185	-8.89	-8.89	**				1	1.00	3.18	-37.18 ++
1	-	-	0.00	yo	0										
	0	y	-8.89	o	90	-9.93	-18.83	**				2	1.00	3.18	-78.72 ++
2	-	-	0.00	you	546	-7.33	-7.33	**				1	3.00	1.18	-15.99 ++
	1	y o	-18.83	u	68	-10.34	-29.17					3	1.00	3.18	-121.94
	0	y	-8.89	ou	28	-11.62	-20.51					2	1.50	2.68	-75.50
3	-	-	0.00	yout	0										
	2	you	-7.33	t	268	-8.36	-15.69	**				2	2.00	2.18	-49.91 ++
	1	y o	-18.83	ut	21	-12.03	-30.86					3	1.33	2.84	-118.74
	0	y	-8.89	out	118	-9.54	-18.44					2	2.00	2.18	-58.64
4	-	-	0.00	youth	12	-12.84	-12.84	**				1	5.00	0.82	-23.37 ++
	3	you t	-15.69	h	92	-9.90	-25.60					3	1.66	2.51	-89.94
	2	you	-7.33	th	120	-9.52	-16.85					2	2.50	1.68	-45.17
	1	y o	-18.83	uth	0										
	0	y	-8.89	outh	3	-14.84	-23.74					2	2.50	1.68	-63.62
5	-	-	0.00	youthi	0										
	4	youth	-12.84	i	871	-6.66	-19.50	**				2	3.00	1.18	-42.52 ++
	3	you t	-15.69	hi	25	-11.78	-27.48					3	2.00	2.18	-87.39
	2	you	-7.33	thi	2	-15.42	-22.76					2	3.00	1.18	-49.62
	1	y o	-18.83	uthi	0										
	0	y	-8.89	outhi	0										

Cycle & Subcycle	Previous parse		Candidate for next segment			Total		Optlen-constrained parse						
	Segments	Score	Segment	Freq	Log(p)	Score		Previous parse Segments	Score	Segs Avglen	Diff Score	Total		
(a)(b)	-----(c)-----	(d)	-----(e)-----	(f)	(g)	(h)	(i)	----- (j) -----	(k)	(l)	(m)	(n)	(o)	(p)
6	- -	0.00	youthis	0										
5	youth i	-19.50	s	1195	-6.20	-25.71				3	2.33	1.84	-73.18	
4	youth	-12.84	is	226	-8.60	-21.45				2	3.50	0.68	-36.03	
3	you t	-15.69	his	481	-7.51	-23.21				3	2.33	1.84	-66.08	
2	you	-7.33	this	109	-9.66	-16.99 **				2	3.50	0.68	-28.55 ++	
1	y o	-18.83	uthis	0										
0	y	-8.89	outhis	0										
7	- -	0.00	youthisa	0										
6	you this	-16.99	a	686	-7.00	-24.00 **				3	2.66	1.51	-60.32	
5	youth i	-19.50	sa	30	-11.52	-31.02				3	2.66	1.51	-77.97	
4	youth	-12.84	isa	10	-13.10	-25.94				2	4.00	0.18	-30.62 ++	
3	you t	-15.69	hisa	3	-14.84	-30.54				3	2.66	1.51	-76.75	
2	you	-7.33	thisa	0										
1	y o	-18.83	uthisa	0										
0	y	-8.89	outhisa	0										
8	- -	0.00	youthisaw	0										
7	you this a	-24.00	w	99	-9.79	-33.80		youth isa	-25.94	3	3.00	1.18	-77.93	
6	you this	-16.99	aw	7	-13.62	-30.61				3	3.00	1.18	-66.74	
5	youth i	-19.50	saw	26	-11.72	-31.23				3	3.00	1.18	-68.08	
4	youth	-12.84	isaw	41	-11.07	-23.91 **				2	4.50	0.32	-31.56 ++	
3	you t	-15.69	hisaw	0										
2	you	-7.33	thisaw	0										
1	y o	-18.83	uthisaw	0										
0	y	-8.89	outhisaw	0										

Cycle & Subcycle	Previous parse		Candidate for next segment			Optlen-constrained parse								Total			
	Segments	Score	Segment	Freq	Log(p)	Score		Previous parse Segments	Score	Segs Avglen	Diff Score						
(a)(b)	-----	(c)-----	(d)	-----	(e)-----	(f)	(g)	(h)	(i)	-----	(j)-----	(k)	(l)	(m)	(n)	(o)	(p)
9	-	-	0.00	youthisawh	0												
8	youth	isaw	-23.91	h	92	-9.90	-33.81			3	3.33	0.84	-62.45				
7	you	this a	-24.00	wh	10	-13.10	-37.10	youth isa	-25.94	3	3.33	0.84	-72.12				
6	you	this	-16.99	awh	2	-15.42	-32.42 **			3	3.33	0.84	-59.87	++			
5	youth	i	-19.50	sawh	0												
4	youth		-12.84	isawh	0												
3	you	t	-15.69	hisawh	0												
2	you		-7.33	thisawh	0												
1	y	o	-18.83	uthisawh	0												
0	y		-8.89	outhisawh	0												
10	-	-	0.00	youthisawhi	0												
9	you	this awh	-32.42	i	871	-6.66	-39.08			4	2.75	1.43	-94.97				
8	youth	isaw	-23.91	hi	25	-11.78	-35.69 **			3	3.66	0.51	-54.02	++			
7	you	this a	-24.00	whi	3	-14.84	-38.84	youth isa	-25.94	3	3.66	0.51	-61.73				
6	you	this	-16.99	awhi	0												
5	youth	i	-19.50	sawhi	0												
4	youth		-12.84	isawhi	0												
3	you	t	-15.69	hisawhi	0												
2	you		-7.33	thisawhi	0												
1	y	o	-18.83	uthisawhi	0												
0	y		-8.89	outhisawhi	0												

Cycle & Subcycle		Previous parse		Candidate for next segment			Total Score	Optlen-constrained parse										
		Segments	Score	Segment	Freq	Log(p)		Segments	Score	Segs Avglen	Diff Score	Total						
(a)(b)		-----	(c)-----	(d)	-----	(e)-----	(f)	(g)	(h)	(i)	-----	(j)-----	(k)	(l)	(m)	(n)	(o)	(p)
11	-	-		0.00	youthisawhim	0												
	10	youth isaw hi	-35.69	m	101	-9.77	-45.46							4	3.00	1.18	-99.12	
	9	you this awh	-32.42	im	102	-9.75	-42.17							4	3.00	1.18	-91.95	
	8	youth isaw	-23.91	him	473	-7.54	-31.45							3	4.00	0.18	-37.11	++
	7	you this a	-24.00	whim	0						youth isa	-25.94						
	6	you this	-16.99	awhim	2	-15.42	-32.42							3	4.00	0.18	-38.25	
	5	youth i	-19.50	sawhim	2	-15.42	-34.93							3	4.00	0.18	-41.22	
	4	youth	-12.84	isawhim	5	-14.10	-26.94 **							2	6.00	1.82	-75.99	
	3	you t	-15.69	hisawhim	0													
	2	you	-7.33	thisawhim	0													
	1	y o	-18.83	uthisawhim	0													
	0	y	-8.89	outhisawhim	0													
Parse:				youth isawhim						youth isaw him								

A.4 Backward parse of: youthisawhim

Cycle & Subcycle		Candidate for next segment				Optlen-constrained parse								
		Previous parse		Total		Previous parse		Segs		Diff		Total		
(a)(b)	(c)	(d)	(e)	Segment	Freq	Log(p)	Score	(j)	(k)	(l)	(m)	(n)	(o)	(p)
11	-	-	0.00	m	101	-9.77	-9.77 **			1	1.00	3.18	-40.83	++
10	-	-	0.00	im	102	-9.75	-9.75 **			1	2.00	2.18	-31.02	++
11	m	-9.77	i		871	-6.66	-16.43			2	1.00	3.18	-68.68	
9	-	-	0.00	him	473	-7.54	-7.54 **			1	3.00	1.18	-16.44	++
10	im	-9.75	h		92	-9.90	-19.66			2	1.50	2.68	-72.35	
11	m	-9.77	hi		25	-11.78	-21.55			2	1.50	2.68	-79.32	
8	-	-	0.00	whim	0									
9	him	-7.54	w		99	-9.79	-17.34 **			2	2.00	3.68	-81.15	++
10	im	-9.75	wh		10	-13.10	-22.86			2	2.00	3.68	-106.99	
11	m	-9.77	whi		3	-14.84	-24.61			2	2.00	3.68	-115.19	
7	-	-	0.00	awhim	2	-15.42	-15.42 **			1	5.00	0.82	-28.07	++
8	w him	-17.34	a		686	-7.00	-24.34			3	1.66	2.51	-85.54	
9	him	-7.54	aw		7	-13.62	-21.16			2	2.50	1.68	-56.71	
10	im	-9.75	awh		2	-15.42	-25.18			2	2.50	1.68	-67.49	
11	m	-9.77	awhi		0									
6	-	-	0.00	sawhim	2	-15.42	-15.42 **			1	6.00	1.82	-43.50	
7	awhim	-15.42	s		1195	-6.20	-21.63			2	3.00	1.18	-47.16	
8	w him	-17.34	sa		30	-11.52	-28.86			3	2.00	2.18	-91.78	
9	him	-7.54	saw		26	-11.72	-19.27			2	3.00	1.18	-42.00	++
10	im	-9.75	sawh		0									
11	m	-9.77	sawhi		0									

Cycle & Subcycle	Previous parse		Candidate for next segment			Total Score	Optlen-constrained parse										
	Segments	Score	Segment	Freq	Log(p)		Previous parse Segments	Score	Segs Avglen	Diff Score	Total						
(a)(b)	-----	(c)-----	(d)	-----	(e)-----	(f)	(g)	(h)	(i)	-----	(j)-----	(k)	(l)	(m)	(n)	(o)	(p)
5	-	-	0.00	isawhim	5	-14.10	-14.10 **					1	7.00	2.82	-53.88		
	6	sawhim	-15.42	i	871	-6.66	-22.08	saw him	-19.27	3	2.33	1.84	-73.81				
	7	awhim	-15.42	is	236	-8.54	-23.97			2	3.50	0.68	-40.27				
	8	w him	-17.34	isa	10	-13.10	-30.44			3	2.33	1.84	-86.67				
	9	him	-7.54	isaw	41	-11.07	-18.61			2	3.50	0.68	-31.27 ++				
	10	im	-9.75	isawh	0												
	11	m	-9.77	isawhi	0												
4	-	-	0.00	hisawhim	0												
	5	isawhim	-14.10	h	92	-9.90	-24.01	isaw him	-18.61	3	2.66	1.51	-71.67				
	6	sawhim	-15.42	hi	25	-11.78	-27.21	saw him	-19.27	3	2.66	1.51	-78.05				
	7	awhim	-15.42	his	481	-7.51	-22.94 **			2	4.00	0.18	-27.07 ++				
	8	w him	-17.34	hisa	3	-14.84	-32.18			3	2.66	1.51	-80.89				
	9	him	-7.54	hisaw	0												
	10	im	-9.75	hisawh	0												
	11	m	-9.77	hisawhi	0												
3	-	-	0.00	thisawhim	0												
	4	his awhim	-24.00	t	268	-8.36	-32.36			3	3.00	1.18	-70.55				
	5	isawhim	-14.10	th	120	-9.52	-23.62 **	isaw him	-18.61	3	3.00	1.18	-61.33				
	6	sawhim	-15.42	thi	2	-15.42	-30.85	saw him	-19.27	3	3.00	1.18	-75.64				
	7	awhim	-15.42	this	109	-9.66	-25.08			2	4.50	0.32	-33.11 ++				
	8	w him	-17.34	thisa	0												
	9	him	-7.54	thisaw	0												
	10	im	-9.75	thisawh	0												
	11	m	-9.77	thisawhi	0												

Cycle & Subcycle	Previous parse		Candidate for next segment			Total Score	Optlen-constrained parse										
	Segments	Score	Segment	Freq	Log(p)		Previous parse		Segs Avglen	Diff Score	Total						
							Segments	Score									
(a)(b)	-----	(c)-----	(d)	-----	(e)-----	(f)	(g)	(h)	(i)	-----	(j)-----	(k)	(l)	(m)	(n)	(o)	(p)
2	-	-	0.00	uthisawhim	0												
3	th	isawhim	-23.62	u	68	-10.34	-33.96**	this	awhim	-25.08	3	3.33	0.84	-65.42	++		
4	his	awhim	-24.00	ut	21	-12.03	-36.03				3	3.33	0.84	-66.54			
5	isawhim		-14.10	uth	0			isaw	him	-18.61							
6	sawhim		-15.42	uthi	0			saw	him	-19.27							
7	awhim		-15.42	uthis	0												
8	w him		-17.34	uthisa	0												
9	him		-7.54	uthisaw	0												
10	im		-9.75	uthisawh	0												
11	m		-9.77	uthisawhi	0												
1	-	-	0.00	outhisawhim	0												
2	u th	isawhim	-33.96	o	90	-9.93	-43.90				4	2.75	1.43	-106.68			
3	th	isawhim	-23.62	ou	28	-11.62	-35.24	this	awhim	-25.08	3	3.66	0.51	-55.55			
4	his	awhim	-24.00	out	118	-9.54	-33.54				3	3.66	0.51	-50.76			
5	isawhim		-14.10	outh	3	-14.84	-28.94**	isaw	him	-18.61	3	3.66	0.51	-50.63	++		
6	sawhim		-15.42	outhi	0			saw	him	-19.27							
7	awhim		-15.42	outhis	0												
8	w him		-17.34	outhisa	0												
9	him		-7.54	outhisaw	0												
10	im		-9.75	outhisawh	0												
11	m		-9.77	outhisawhi	0												

Cycle & Subcycle	Previous parse		Candidate for next segment			Total Score	Optlen-constrained parse										
	Segments	Score	Segment	Freq	Log(p)		Previous parse		Segs Avglen	Diff Score	Total						
							Segments	Score									
(a)(b)	-----	(c)-----	(d)	-----	(e)-----	(f)	(g)	(h)	(i)	-----	(j)-----	(k)	(l)	(m)	(n)	(o)	(p)
0	-	-	0.00	youthisawhim	0												
1	outh	isawhim	-28.94	y	185	-8.89	-37.84					3	4.00	0.18	-44.65		
2	u th	isawhim	-33.96	yo	0												
3	th	isawhim	-23.62	you	546	-7.33	-30.96	this awhim	-25.08	3	4.00	0.18	-38.25				
4	his	awhim	-24.00	yout	0												
5	isawhim	-14.10	youth	12	-12.84	-26.94**	isaw him	-18.61	3	4.00	0.18	-37.11	++				
6	sawhim	-15.42	youthi	0			saw him	-19.27									
7	awhim	-15.42	youthis	0													
8	w him	-17.34	youthisa	0													
9	him	-7.54	youthisaw	0													
10	im	-9.75	youthisawh	0													
11	m	-9.77	youthisawhi	0													
Parse:			youth isawhim					youth isaw him									

Appendix B. Corpus Descriptions

B.1.0 INTRODUCTION

This appendix gives details about how the corpora used in this study were acquired and processed, and shows samples of the text at various stages. Each corpus required some modification before it could be used in the O-B segmentation algorithm; all had particular idiosyncrasies or problems that had to be solved on a case-by-case basis. Where several different versions of the corpus were used, I describe the modifications made. None of these corpora is very large, and therefore it was possible when necessary to do a certain amount of hand-editing. I have tried to give enough information that the reader can judge the nature of each corpus.

After each corpus was standardized, and capitals, punctuation, numerals, etc. were removed, it was named and stored, with the word divisions intact, for use in scoring the results of the algorithm and parse. Just before the corpus is run through the algorithm, the word divisions (spaces) are removed, but line breaks are left in (each an implicit word division), and the task of the algorithm is to put word divisions back in. Below, I show a sample of each standardized corpus ("after preparation") used for scoring, and the corpus as output from one version of the algorithm ("after dictin parsing"). I have tried to select samples so as to show the same text in various stages.

Section 2.0 of this appendix describes certain procedures common to ENG V, SPA V, and JAP V, all from the same data base. Then the individual corpora are described, each in its own section.

B.2.0 ABOUT CHILDES

Three of the corpora described below were taken from datasets in the Child Language Data Exchange System (CHILDES) collection (MacWhinney 1995). This project consists of a fairly well standardized protocol for collecting and recording speech of young children and their caretakers, and of a computer archive where researchers may deposit data they have encoded for others to use with no charge. It is available over the Internet as well [<http://poppy.psy.cmu.edu/childes>]. The archived data is intended to be useful for research with many different focuses (phonological, semantic, pragmatic, etc.), structures for recording a number of different aspects of language are provided, and the contributing researcher may choose to use as many of them as she wishes.

Following is the beginning of one of the Spanish files, verbatim, to give a flavor of the amount of information included (most of which our study ignores).

Sample of text as received:

```
@Begin
@Participants:  MAN Manuela Target_Child,  FAT Sergio Father,
                MOT Margaret Mother
@Age of MAN:    1;3.4
@Sex of MAN:    female
@Filename:      860928s.FAT
@Coding:        CHAT 1.0
@Date:          28-SEP-1986
@Situation:     at home with Father
@Activities:    playing with bricks
@Location:      Brighton, Sussex, England
@Coder:         Angeles Clark
@Tape location: videotape 1, audiotape 4536-4863 (stops at 00:10:05;
                does not correspond to videotape)
@Language:      Spanish
*MOT:           Sunday twenty eight September in the afternoon Manuela is
                playing with a <few> [>] bricks and things.
```

```

%tim: 00:00:00
*MAN: <yyy> [<].
%pho: w ae
%act: crawling towards the fireplace
%tim: 00:00:07
*MAN: yyy.
%pho: ae, :
%act: stops crawling
%tim: 00:00:10
*FAT: Manuela.
*MOT: in a minute she's going to do the <names and things> [>] # we
      hope.
*MAN: yyy.
%pho: 6 $ 6, :
%act: turning her head to look at FAT
%tim: 00:00:16
*FAT: ven ac^a'.
%eng: come here
*MAN: yyy.
%pho: ? 6 ? 6 ? 6
%par: laughing
%tim: 00:00:19
*FAT: ven ac^a'.
%eng: come here
*MAN: yyy.
%pho: 6
%act: picking up an egg shaped rattle and trying to stand it on the
      floor
%tim: 00:00:23
*MAN: yyy.
%pho: ? 6
%act: looking at the rattle standing up on the floor
%tim: 00:00:26
*FAT: qu^e' es esto Manuela?

```

Corpus preparation: First, I selected just the speech of adults, ignoring all child speech and descriptive lines. Then I separated each speaker's "speech" into "utterances" by beginning a new line at each occurrence of a final punctuation { ; . ? ! } or pause mark (# or ## are used for pauses of 2 seconds or more). Then I removed all but the standard orthographic representation of utterances, leaving behind words which were annotated as baby talk, onomatopoeia, undecipherable (xxx), etc., as well as words that seemed hard to standardize

(*mm, ha, ooh*, etc.). In the Spanish and Japanese cases, I removed the occasional English sentence which crept in.

- Because CHILDES provides a very rich annotation system and the coders seem to put enormous effort into recording as many details as possible, it would be possible to spend much more time than I did to prepare these corpora. However, as it was, I did a lot of hand-and-eyeball processing of the data, and I think the result is more than adequate for the tasks here.
- In the short sample above, twelve utterances appear, of which six are the child Manuela's and six are adults', and two of the adult speech are explanations in English, which we removed by hand. This leaves four speeches in Spanish by the father, totaling nine words.

Sample after preparation [same as text above]:

manuela
 ven aCÁ
 ven aCÁ
 quE es esto manuela

B.3.1 ENG V

Sources This corpus was received from two different sources. First, the Bernstein corpus from CHILDES was downloaded. Then, a phoneticized version of the same data was received from Michael Brent and Timothy Cartwright, who kindly gave me permission to work with their data. So I used the phonetic version as a first source, and then matched the standard orthographic version to it.

Source of orthographic texts The data were collected by Nan Bernstein-Ratner (Bernstein Ratner 1996) from nine mothers, each speaking to her daughter aged 1;1 to 1;11. There are three samples from each mother. This dataset is unusual in that the children are very young. Most of the sessions were followed by a interview of the mother by the researcher, and the mother's speech on these occasions was transcribed into a separate series of files.

The purpose of the interview was to gather a sample of the mothers' speech for comparison against the mother-child condition. In order to obtain exemplars of mother-adult words to match to the mother-child condition, sometimes the questioning wandered. My apologies to the bored readers among you. In addition, to save some transcription time, in some cases the mothers' responses to my questions are provided, but my own conversation is not. [Bernstein-Ratner corpus, readme file]

Format Similar to other CHILDES datasets; see "About CHILDES" above.

Sample of text as received (child-directed):

```
@Participants:    CHI Alice Child, MOT Mother
@Age of CHI:      1;1.0
@Sex of CHI:      female
@Situation:       first session, mother/child play session
*MOT:    She's really into books right now.
%add:    NAN
```

*MOT: You want to see the book?
 *MOT: Oh, look there's a boy with his hat.
 *MOT: And a doggie.
 *MOT: Oh, you want to look at this?
 *MOT: W' look at this?
 *MOT: Have a drink.
 *MOT: OK now.
 *MOT: Oh what's this?
 *MOT: What's that?
 *MOT: What is it, huh?
 *MOT: Oh.
 *MOT: Look can you take it out?
 *MOT: Take it out.
 *MOT: You want it in?
 *MOT: xxx.
 *MOT: Put that on.
 *MOT: Can you xxx it?
 *CHI: g&.
 *MOT: That.
 *MOT: yes.
 *MOT: OK.
 *MOT: Oh.
 *MOT: Open it up. [=! falsetto]
 *MOT: Take the doggie out.
 *MOT: I think it will come out.
 *MOT: Let's see.
 *MOT: Yeah.
 *MOT: Pull it out.

Sample of text as received (adult-directed):

@Participants: MOT Mother, NAN Nan Investigator
 @Age of CHI: 1;1.0
 @Situation: first session, Maternal interview
 *NAN: the blocks.
 *MOT: yeah.
 *MOT: Well, (be)cause that really got her # she was the most enthusiastic playing with that.
 *NAN: Does she have any at home?
 *MOT: Not like this.
 *MOT: She (ha)s got # um!
 *MOT: # actually, I was gon (t)a ask you where you got (th)em -- she (ha)s got bigger and smaller ones -- she (ha)s got small wooden ones and great big cardboard ones.
 *MOT: Um -: but these are more fun to knock down.
 *NAN: um hum.
 *NAN: Do you have any routines you play a lot with her.
 *NAN: She seems to play a lot of hiding routines # !

- *MOT: yeah - on the changing table, I do put a diaper over her face and peek a boo with her.
- *NAN: D(o) you do it with blocks also?
- *MOT: I (ha)ve never done it with blocks before.
- *MOT: I (ha)ve never really been # right in front of her and had blocks.
- *MOT: Usually, , y(ou) know, if we (a)re in this situation we (a)re eating # And there (i)s no fooling around!
- *NAN: um -: when you took the doll and the dog with her, you tried to get her to do things you do at home +...
- *MOT: yeah - like combing the dog and stuff?

Corpus preparation: In this case, preparation consisted of selecting from the standard orthographic text those lines which matched the phonetic text, and then editing the standard lines so that the word and line boundaries matched the phonetic version as nearly as possible – in other words, trying to duplicate the preparation that Brent and Cartwright had done.

Sample after preparation (child-directed):

you want to see the book
 look theres a boy with his hat
 and a doggie
 you want to look at this
 look at this
 have a drink
 okay now
 whats this
 whats that
 what is it
 look can you take it out
 take it out
 you want it in
 put that on
 that
 yes
 okay
 open it up
 take the doggie out
 i think it will come out
 lets see
 yeah
 pull it out

Source of Phonetic Texts Downloaded by permission of Timothy Cartwright.

[<ftp://cog.jhu.edu/pub/Cartwright/old-inputs/>] The files match the CHILDES files, both adult- and child-directed speech, and have been transcribed into a phonetic notation of Cartwright's devising. These files were further divided into "short" and "input" versions, with the "short" version containing only the first 500 or so words.

Phonetic character set 50 ASCII characters are used, as detailed by Cartwright in a "readme" file [annotated "Timothy Andrew Cartwright, 1993.6.11, revised 1994.6.15"]:

<u>Vowels</u>		<u>Vowel+r comb.</u>	<u>Consonants</u>				
bit	I	bird	3	pan	p	zap	z
bet	E	butter	R (unstressed)	ban	b	ship	S
bat	&	arm	#	man	m	azure	Z
but	A	horn	%	tan	t	hand	h
hot	a	air	*	dan	d	chip	c
law	O	ear	(	nan	n	judge	G
put	U	lure	)	can	k	lap	l
the	6			gone	g	rap	r
beet	i			sing	N	bottle	L
bait	e			fan	f	button	~
buy	9			van	v	rhythm	M
about	Q			thin	T	yet	y
boot	u			than	D	wet	w
boat	o			sand	s	when	W
boy	7						

Format The phonetic texts were already fully edited, one utterance to a line, as I had done for other CHILDES texts. I did no further preparation here except to align the orthographic texts to the phonetic ones for the child-directed data. I did not use matching orthographic texts for the adult-directed data, as it was used only as a contrast to the child-directed speech.

Sample of phonetic text as received:

yu want tu si D6 buk
 lUk D*z 6 b7 wIT hIz h&t
 &nd 6 dOgi
 yu want tu lUk &t DIIs
 lUk &t DIIs
 h&v 6 drINK
 oke nQ
 WAts DIIs
 WAts D&t
 WAt Iz It
 lUk k&n yu tek It Qt
 tek It Qt
 yu want It In
 pUt D&t an
 D&t
 yEs
 oke
 op~ It Ap
 tek D6 dOgi Qt
 9 TINK It wIl kAm Qt
 lEts si
 y&
 pUl It Qt

B.3.2 ENG W

Source The novel *Of Moon and Sixpence* by W. Somerset Maugham, downloaded in April 1995 from the Gutenberg EText archive of ASCII texts. [<ftp://mrcnext.cso.uiuc.edu> or <http://www.nettap.com/~drc/>]

Format Standard English orthography, including digits and brackets, etc. Text was divided into lines of roughly 60 characters each, ending with line-end (hex'0d', or CR control character), as if ready for display at a computer terminal or printing on narrow paper. Also found paragraphing, occasional blank lines for visual spacing, and four footnotes with bibliographic data.

Sample of text as received:

The Moon and Sixpence

by W. Somerset Maugham

Author of "Of Human Bondage"

Chapter I

I confess that when first I made acquaintance with Charles Strickland I never for a moment discerned that there was in him anything out of the ordinary. Yet now few will be found to deny his greatness. I do not speak of that greatness which is achieved by the fortunate politician or the successful soldier; that is a quality which belongs to the place he occupies rather than to the man; and a change of circumstances reduces it to very discreet proportions. The Prime Minister out of office is seen, too often, to have been but a pompous rhetorician, and the General without an army is but the tame hero of a market town. The greatness of Charles Strickland was authentic. It may be that you do not like his art, but at all events you can hardly refuse it the tribute of your interest. He disturbs and arrests. The time has passed when

Corpus preparation: Lines were left as found, but blank lines were removed. All punctuation and numbers were removed, and capital letters were lower-cased. Spaces were left in the scoring version and removed from the parse input. Note that this procedure results in treating as a single word characters separated only by an apostrophe or a hyphen, such as:

one's	ones
his father's life	his fathers life
there'd be a quarrel	thered be a quarrel
palm-trees	palmtrees
well-meaning	wellmeaning
ill-cut bread-and-butter	illcut breadandbutter

Sample after preparation:

the moon and sixpence
by w somerset maugham
author of of human bondage
chapter i
i confess that when first i made acquaintance with charles
strickland i never for a moment discerned that there was in
him anything out of the ordinary yet now few will be found
to deny his greatness i do not speak of that greatness which
is achieved by the fortunate politician or the successful
soldier that is a quality which belongs to the place he
occupies rather than to the man and a change of circumstances
reduces it to very discreet proportions the prime minister
out of office is seen too often to have been but a pompous
rhetorician and the general without an army is but the tame
hero of a market town the greatness of charles strickland
was authentic it may be that you do not like his art but at
all events you can hardly refuse it the tribute of your
interest he disturbs and arrests the time has passed when

Character set: Received 79 characters: 26 lower-case, 26 capitals, space, 10 digits, and 16 punctuation marks { - . , ; ? ! ' " ` () [] < > }. All but 26 lower-case letters were removed.

Sample after dictin parsing:

them onand sixp ence
 by w some rse tm augh am
 author of of human bondage
 chapter i
 i confess that when first i made acquaint ance with charl es
 stric kland i never for amoment disce rned that there was in
 him any thing outof the ordinary yet now few willbe found
 tode ny his great ness idonot speak of that great ness which
 is achiev edby the fortun ate poli tici an orthe success ful
 s oldier thatis aqu ality which be long stothe place he
 occupi es rather than to themanand a change of circum stance s
 red uces itto very discreet pro port ions the prime min ister
 outof office isseen too often to havebeen but ap omp ous
 rhe to rician andthe gener al without an army is butthe tame
 hero ofa market town the great nessof char les stric kland
 was authent ic itmaybe that you donot like his artbut at
 all events you can hardly refuse it the tribut eof your
 interest he disturb sand ar rests the time has passed when

Translation to phonetic symbols: In order to compare the results of processing

the corpus in two different orthographies, I looked for a "standard" phonetici-
 zation (or phonemicization) routine. From the Internet [[ftp://svr-ftp.eng.cam.
 ac.uk/pub/comp.speech/synthesis/english2phoneme.shar](ftp://svr-ftp.eng.cam.ac.uk/pub/comp.speech/synthesis/english2phoneme.shar)], I downloaded a
 program which worked (several didn't compile or didn't run), and which
 seems to be widely distributed around the Internet under slightly different
 names. The version I got has a "readme" file dated 1985 and signed by John
 A. Wasser of Digital Equipment Corp. in Littleton, Mass., which says:

Here it is one last time. I have fixed all of the bugs I heard about and added
 a new feature or two.... I think that this version is good enough for most
 purposes.... As before, this is all public domain and I make no copyright
 claims on it. The part derived from the Naval Research Lab should be
 public anyway.

Phonetic character set The version resulting from submitting the corpus to the program **phoneme** uses 35 characters—17 capitals and 18 lower-case:

{ ACDEGHI NORSTUWXYZbdfghjklmnpqrstvwyz }.

Sample after phoneticization:

DHAX mUWn AEnd sIHkspEHns
 bAY dAHblyUWw sAHmrSEht mAOAEm
 AOTHER AXv AXv hyUWmAEh bAHndIHj
 DHAX mUWn AEnd sIHkspEHns
 DHAX mUWn AEnd sIHkspEHns
 CHAEptER AY
 AY kAAfEHs DHAet WHEhN fERst AY mEYd AEkkwEYntAEhs wIHTH CHAArAXlz
 strIHkLAEnd AY nEHvER fAOr EY mOWmEHnt dIHsERnd DHAet DHEhr wAAz IHn
 hIHm EHnIYTHIHNG AWt AXv DHAX AOrdIHnAArIY yEHt nAW fyUW wIHl bIY fAWnd
 tUW dEHnIY hIHZ grEYtnEHs AY dUW nAAt spIYk AXv DHAet grEYtnEHs WHIICH
 IHZ AECHYIYvd bAY DHAX fAOrtUWnEYt pAAlIHtIHSHAEn AOr DHAX sAHkseHsfUhl

B.4.1 SPaV

Source Recorded speech from the CHILDES datasets “deuchar” and “marrero,” retrieved in March 1996 and combined into a single corpus for use here. The “deuchar” set accounts for about 25% of the SPaV corpus, measured by characters and by words.

- **Deuchar** (Deuchar & Clark 1992) records the language development of Manuela, “an infant acquiring English and Spanish simultaneously... recordings were made... of spontaneous interactions between Manuela and her Spanish-speaking father...” [readme file with the corpus]. I used the adult utterances from the eight Spanish transcripts, made from age 1;3 to 2;6, six with the Spanish-speaking father, one with the mother, for whom Spanish is a second language, and one with both parents. (A sample of the Deuchar text is in Section 2.0 of this Appendix.)
- **Marrero** [contributed by Victoria Marrero and Maria Jose Alba in Madrid, Spain]. Recordings of interactions with three Spanish children: Alf, age 2;3-2;10; Ida, age 2;8-4;7; and Raf, age 4;4-4;11.

Format See above, “About CHILDES.”

Sample of text as received (end of Marrero/Raf):

*MJO: no, Alfon, no te enfades con Rafa; es que a él no le gusta que le des una patada a sus cositas; venga, no te disgustes, no; venga, vamos a jugar ahora a otra cosa; a qué jugamos?
 *RAF: pero no sé.
 *MJO: jugamos con la pizarra, por ejemplo?
 *RAF: no.
 *MJO: no -? , con qué?
 *RAF: con otra cosa.

*MJO: con cuál?
 *RAF: Yiye, sabes qué quería que vosotros me comprareis?
 %err: <8> comprareis = comprarais
 *MJO: qué?
 *RAF: esto, esto quería que me comprareis.
 %err: <6> comprareis = comprarais
 %act: mostrando una foto de otro juego de plastilina.
 *MJO: a ver, no lo veo bien desde aquí, enséñamelo.
 *RAF: mira.
 *MJO: espera, que me voy a sentar aquí con Alfon porque pesa mucho.
 *RAF: es muy diver.
 %err: <3> diver = divertido
 *MJO: a ver cómo es -? , es para hacer un arco iris?
 *RAF: mira.
 *MJO: ah, es para hacer formas.
 *RAF: claro, es para haser formas, y se pueden haser éstas.
 %err: <4> haser = hacer
 <9> haser = hacer
 %act: señalando las que vienen dibujadas en la caja.
 *MJO: ah, claro, puedes hacer # mira, un triángulo, un cuadrado, un círculo, una cruz, y luego puedes hacer también un arco iris, una trenza, verdad?
 *RAF: o estos filiquitosos@.
 %act: señalándolo en el dibujo.

Character set As with all CHILDES files, there are a very large number of special characters used to encode linguistic and extralinguistic information.

The special problem with the Spanish datasets is the occurrence of accented characters. These were preserved as capital letters: *á* or *^a* = *A*, *é* or *^e* = *E*, *í* or *^i* = *I*, *ó* or *^o* = *O*, *ú* or *^u* = *U*, *ñ* or *^n* = *N*. The resulting character set, therefore, was 32 symbols—26 lower-case letters and 6 capital letters.

Corpus preparation See above, "About CHILDES."

Sample after preparation (long lines are shown with indented runovers):

no alfon no te enfades con rafa
 es que El no le gusta que le des una patada sus cositas
 venga no te disgustes no
 venga vamos jugar ahora otra cosa
 quE jugamos
 jugamos con la pizarra por ejemplo

no con quE
 con cuAl
 quE
 ver no lo veo bien desde aquí ensENamelo
 espera que me voy sentar aquí con alfon porque pesa mucho
 ver cÓmo es es para hacer un arco iris
 es para hacer formas
 claro puedes hacer
 mira un triAngulo un cuadrado un cÍrculo una cruz y luego puedes hacer
 tambiEn un arco iris una trenza verdad

Sample after dictin parsing:

no alfon no te enfades con rafa
 esque El no legusta que ledes una pata da sus cositas
 venga note dis gu ste sno
 venga vamos juga r ahora otracos a
 quE jugamos
 jugamos con lapi za rra porejemp lo
 no conquE
 con cuAl
 quE
 ver no loveo bien desde aquí ensENa melo
 espera que me voy s entaraquí con alfon por que pesa mucho
 ver cÓmo es es para hacer una rcoir is
 espara hacer form as
 claro puede s hacer
 mira un tri An gulo uncua drado unc Irculo una c ru z y luego puedes
 hacer tambiEn una rcoir is una tren za verdad

Phonetic translation: The following changes were made to convert the standard orthography (with accents) into a more phonetic notation (Langenscheidt 1985). Note that interactions between the two sets of characters are avoided. For instance, orthographic z becomes phonetic [c], while orthographic s in some contexts becomes phonetic [z], but it is never the case that s will be changed to z which will then be changed to c.

- Changes in orthography, from two-character to one-character symbol:

<u>from:</u>	ch	<u>to:</u>	C
	ll		L
	rr		R
	qu		k
	gu (before e, i)		g

- Change in symbol to reflect true phonetic value, in some cases differing by context:

<u>from:</u>	v	<u>to:</u>	b
	z		c
	h		deleted (= silent)
	j (word-final)		deleted (= silent)
	s (before voiced consonant)		z
	s (other)		s
	g (before e, i)		j
	g (other)		g
	x (before vowel)		gs
	x (other)		s
	c (before e, i, E, I)		c
	c (other)		k

Note: [c] is pronounced as *th* in peninsular Spanish and like *s* in South American Spanish, so I left it alone.

Phonetic character set: Nine capital letters { ACETLNORU } and 22 lower-case letters, omitting { hqv x }.

Sample after phonetic translation:

no alfon no te enfades kon rafa
 es ke El no le gusta ke le des una patada sus kositas
 bengal no te dizgustes no
 bengal bamos jugar aora otra kosa
 kE jugamos
 jugamos kon la pikaRa por ejemplo
 no kon kE
 kon kuAl
 kE

ber no lo beo bien dezde akI ensENamelo
espera ke me boy sentar akI kon alfon porke pesa muCo
ber kOmo es es para acer un arko iris
es para acer formas
klaro puedes acer
mira un triAngulo un kuadrado un cIrkulo una kruc y luego puedes acer
tambiEn un arko iris una trenca berdad

B.4.2 SPAW

Source: A series of short stories by Alejandro Dolina, downloaded from the Internet in November 1995. [<http://arcadia.informatik.uni-muenchen.de/rec/argentina/pub/local/rec/argentina/aportes/dolina>]

Format: Received as copies of a number of email messages, each containing a portion of the text in addition to the usual email headers, etc. The text was divided into lines of about 70 characters each.

Sample of text as received:

From mandel Thu Sep 5 16:16:53 1991
 Subject: BALADA DE LA PRIMERA NOVIA
 To: mandel (Luis Mandel)
 Date: Thu, 5 Sep 91 16:16:53 MET DST
 Status: OR

BALADA DE LA PRIMERA NOVIA (Por Alejandro Dolina)

El poeta Jorge Allen tuvo su primera novia a la edad de doce a-nos. Guarden las personas mayores sus sonrisas condescendientes. Porque en la vida de un hombre hay pocas cosas mas serias que su amor inaugural.

Por cierto, los mercaderes, los Refutadores de Leyendas y los aplicadores de inyecciones parecen opinar en forma diferente y resaltan en sus discursos la importancia del automovil, la higiene, las tarjetas de credito y las comunicaciones instantaneas. El pensamiento de estas gentes no debe preocuparnos. Despues de todo han venido al mundo con propositos tan diferentes de los nuestros, que casi es imposible que nos molesten.

Corpus preparation: Headers and title lines were removed, plus a few lines which seemed to be duplicates, and some extraneous text at the end. As usual, capital letters were changed to lower case and punctuation was removed. Some hand editing was done to determine whether punctuation should be replaced with a space or not.

Sample after preparation:

el poeta jorge allen tuvo su primera novia a la edad de doce años
 guarden las personas mayores sus sonrisas condescendientes porque en la
 vida de un hombre hay pocas cosas mas serias que su amor inaugural
 por cierto los mercaderes los refutadores de leyendas y los aplicadores
 de inyecciones parecen opinar en forma diferente y resaltan en sus discursos
 la importancia del automovil la higiene las tarjetas de credito y
 las comunicaciones instantaneas el pensamiento de estas gentes no debe
 preocuparnos despues de todo han venido al mundo con propositos tan
 diferentes de los nuestros que casi es imposible que nos molesten

Character set: Received as upper- and lower-case Roman alphabet, 10 digits,

space, and 13 punctuation characters { . , : ; ! ? () ' " - * ~ }. No accent marks,

but ñ appears as ~n, which I changed to simple n. Reduced to 26 lower-case
 letters.

Sample after dictin parsing:

el poeta jorgeallen tuvo su primeranovia ala edad de do ceanos
 guar den las personas mayores sus s onrisa scon des cen dientes porque enla
 vida de unhombre hay pocas cosas mas seria squesu amor in au gur al
 por cierto los mer caderes losrefuta dores de leyendas ylos aplica dores
 dein ye cciones parece n opinar en forma diferente yres al tan ensus discurs os
 la importa ncia dela utomovi lla hi giene las tar je tasde cre dito y
 lasco muni cacion esin stan tan eas el pensa miento deesta sgentes no debe
 preocup a mos despuesde todo han ven ido al mundo con propositos tan
 difer entes delos nuestros que casi esim posible quenos mole sten

B.5.1 JApV

Source: Recorded speech from the CHILDES "Aki" corpus, collected by Susanne Miyata (1995). She recorded this Japanese child more or less weekly in his home, from age 1;5 to 3;0, resulting in 56 transcripts. I used the adult speech, which included Aki's mother, father, two grandmothers, a grandmother-aged baby-sitter, and the researcher.

Format: See above, "About CHILDES."

Sample of text as received (some non-speech lines omitted for brevity):

```
@Participants: AKI Akifumi Target_child, AMO Okaasan Mother,
                SUZ Suze Investigator
@Age of AKI:    1;5.7
@Situation:     Aki and AMO looking at picture book
*AMO:  issho !
%act:  comparing the picture of a bear with Aki's toy bear
*AKI:  &6 !
%gpx:  points video camera
*AKI:  &Nga: !
%gpx:  points video camera
*AMO:  pikapika@o .
*AKI:  &wa: .
%gpx:  points book
*AMO:  issho ne .
*AMO:  konnichiwa tte .
*AKI:  &godte [: kore] .
%act:  Aki touches bear's glove, which is fixed with a band
*AMO:  un .
*AMO:  kuma+san tebukuro shiyoo ka ?
%act:  puts glove on bear
*AMO:  samui+samui .
*AMO:  hora .
*AMO:  konnichiwa tte .
*AKI:  &a .
%act:  bows
*AMO:  un .
*AMO:  konnichiwa shita .
*AKI:  &wadto .
%gpx:  points letter B in the book
```

*AMO: un .
 *AMO: bi@l .
 *AKI: &we: [: bi@l] .
 *AMO: un .
 *AMO: ato nani ga suki kanaa ?
 %act: turning pages
 *AKI: oowa@o .
 %act: touching his big toy lion
 *AMO: raion dete+kuru tokoro doko datta kke ?
 %act: turns pages while Aki peers into the book
 *AMO: hora !
 %gpx: points picture of lion
 *AKI: aa !
 %gpx: points toy lion
 %act: Aki takes his toy bear
 *AMO: kuma-san dai#suki shite !
 *AMO: ii+ko+ii+ko shite !
 %gpx: shows how to pat the bear

Character set Received similar to other CHILDES files, with Japanese utterances

transliterated into the Roman alphabet, and explanatory statements in English.

The Japanese transliteration uses 24 characters—the lower-case letters, omitting { l q v x }, plus Q and the apostrophe.

- Q is commonly used to indicate geminated consonants, but here it indicates a glottal stop, as *un chikuQ to sasareru to itai ne* or *areQ* (the latter 104 times).
Geminations here are indicated by double consonants (*kaette; konnichiwa*).
- The apostrophe is used occasionally, as is standard, to distinguish initial from syllabic *n*, as *man'in*. Other instances, rather idiosyncratically distinguishing monomorphemic from bimorphemic combinations, have been eliminated and single words created: *oo'otoko* -> *oootoko*.
- The colons used occasionally to indicate lengthening were not retained, as they seemed to indicate emphatic longations (*yo:ku mite; nan:imo inai*) rather than lexical ones, which were spelled using doubled vowels (*kaa-san; jidoosha*).

Corpus preparation: The transcripts have spaces between “words,” even though Japanese does not have an orthographic standard for “words,” as do English and Spanish. I assume that the (Japanese) coders used their own intuitions in order to place these spaces, and I did not alter them, but scored the parsing accordingly.

- I did have to make a decision about how to treat morphemes which were explicitly indicated, using notation provided by the CHILDES scheme (MacWhinney 1995:37f.) to distinguish between suffix (-), prefix (#), compound (+), and clitic (~). There was much more morpheme coding here than in the CHILDES English and Spanish files (though the same notation was available for them). The “readme” file says that the Morikawa system was used for “morphemecization.” However, I was unable to make a principled distinction between these various levels in Japanese, so I broke up all such combinations into two or more words:

Aki-chan	aki chan
O#kaa-san ni kiite	o kaa san ni kiite
o#futon ?	o futon
yoku mite+goran !	yoku mite goran
itai+itai+itai .	itai itai itai
kuma-san dai#suki shite !	kuma san dai suki shite
ame+ame fure+fure kaa-san	ame ame fure fure kaa san

- It should be noted that the characteristically Japanese onomatopoeia (“giseigo/gitaigo”) are not broken into morphemes here, though they are sometimes annotated as onomatopoeia by a trailing @o (which I remove):

pyonpyon@o .
 dondon ookiku narimashita .
 otoko ga dondondondon@o o uchi e kaette kita .

- A further note of interest is that there is only one instance of the object marker を *wo* in this corpus. Matsuda (1992) reported that in spoken Japanese it is not unusual to find deletion of 70-90% instances of を, and he discussed some possible reasons for this.

Sample after preparation:

issho
 pikapika
 issho ne
 konnichiwa tte
 un
 kuma san tebukuro shiyoo ka
 samui samui
 hora
 konnichiwa tte
 un
 konnichiwa shita
 un
 un
 ato nani ga suki kanaa
 raion dete kuru tokoro doko datta kke
 hora
 kuma san dai suki shite
 ii ko ii ko shite

Sample after dictin parsing:

issho
 pika pika
 issho ne
 konni chiwa tte
 un
 kuma san te bukuro shi yooka
 samui samui
 hora
 konni chiwa tte
 un
 konni chi wa shita
 un
 un
 ato nani gasuki kanaa
 raion dete kuru tokoro doko da ttakke
 hora
 kuma san dai suki shite
 iikoi ko shite

Translation to hiragana: The corpus as edited was then translated into hiragana so that the two coding systems (roman alphabet and hiragana) could be compared. This was not a trivial task, it turned out, and required a special program and a lot of hand fiddling. Because I wanted the two versions to be very similar, I went so far as to remove the odd line from the alphabetic corpus if it was clear that it would be difficult or impossible to convert into kana. Some of these violate Japanese phonology (*krkrkr*), and others are just unusual (*fyufyufyu*) or are in a different romanization system (e.g., *tyan* = chan; *crayon* is English), or are misspelled (e.g., *miseter* for *miseteiru* 'showing'; *kuukoolainaa* for *kuukoorainaa* 'airliner'):

- The resulting hiragana translation is a mechanical one, and therefore departs in several respects from the usual Japanese writing. A notable example is that the topic marker is spelled as *wa* and thus is converted as わ *wa* instead of the usual は *ha*; this is the result of an idiosyncratic feature of the Japanese writing customs. However, reducing the instance of orthographic conventions that are not related to speech understanding should not affect the learnability of the corpus; it might even enhance it slightly.
- Conversion to hiragana leads to a few new homonyms, resulting in slightly different numbers of word types in the two versions. For instance, the following pair of utterances with identical meaning ('[the lion is] saying grr') illustrate both orthographic and lexical variants in the corpus (I show line number in the corpus and reverse romanization from the hiragana):

1389	gao tte yutteru
20674	gaou tte itteru yo

Hiragana character set: Of the 85 standard hiragana characters, we use 77,

shown in the table below in the standard Japanese sequence and with their roman equivalents.

A	あ	I	い	U	う	E	え	O	お
KA	か	KI	き	KU	く	KE	け	KO	こ
GA	が	GI	ぎ	GU	ぐ	GE	げ	GO	ご
SA	さ	SI	し	SU	す	SE	せ	SO	そ
ZA	ざ	JI	じ	ZU	ず	ZE	ぜ	ZO	ぞ
TA	た	TI	ち	TU	つ	TE	て	TO	と
DA	だ	DI	ぢ	DU	づ	DE	で	DO	ど
NA	な	NI	に	NU	ぬ	NE	ね	NO	の
HA	は	HI	ひ	FU	ふ	HE	へ	HO	ほ
BA	ば	BI	び	BU	ぶ	BE	べ	BO	ぼ
PA	ぱ	PI	ぴ	PU	ぷ	PE	ぺ	PO	ぽ
MA	ま	MI	み	MU	む	ME	め	MO	も
YA	や			YU	ゆ			YO	よ
RA	ら	RI	り	RU	る	RE	れ	RO	ろ
WA	わ							WO	を
N	ん	i	い	tu	っ	e	え	o	お
ya	や			yu	ゅ			yo	よ

Sample after hiragana translation:

いっしょ
 ぴかぴか
 いっしょ ね
 こんにちわ って
 うん
 くま さん てぶくろ しょう か
 さむい さむい
 ほら
 こんにちわ って
 うん
 こんにちわ した
 うん
 うん
 あと なに が すき かなあ
 らいおん でて くる ところ どこ だった っけ
 ほら
 くま さん だい すき して
 いい こ いい こ して

B.5.2 JAPW

Source: The computer transcription of a series of 18 children's picture books about science were made available by Prof. Syun Tutiya of Chiba University in Japan. Since virtually all Japanese texts are written with a mixture of phonetic symbols (kana) and meaning-based characters (kanji), it is unusual to find one like this that is all in hiragana. There are fewer than 100 characters in the kana set, while there are thousands of kanji. I wanted to compare the alphabet with kana orthographies without getting involved with the complexities of kanji.

- Although these books are intended for children in kindergarten or early elementary school, the vocabulary is fairly sophisticated. The orthography is unusual not only in that there are no kanji, but there are also no katakana, the kana symbol set which mirrors hiragana and is used for foreign words. All words normally written in kanji or katakana are here transcribed in hiragana.

Format: This corpus consists of eighteen datasets, each in two files—one file of text and one containing part-of-speech tags. The text was transcribed to the file exactly as it appeared in the book, including space between characters on a line and breaks at the end of each line. The tag files consist of the same text, but separated into morphemes and each morpheme assigned a part of speech from one of 30 tags from 13 major groups. I did not use the tags themselves, but only the division into morphemes which is implicit in the tag files.

Sample from text file as received:

けものは、からだじゅうに けが
はえて いる どうぶつのことです。
きつねも けものの なかまです。

けものは あかちゃん で うまれて、
おちちを のんで そだちます。

さむい ところに すむ けものには
けが いっぱい はえています。
けは おーぼーのように あたたかです。

Sample from tag file as received (read down the columns to match the text

above):

0100010 DB- - 	0100180 ZN- - きつね	0100360 HP- - を
0100020 ZN- - けもの	0100190 HP- - も	0100370 ZV2- - のん
0100030 HP- - は	0100200 ZN- - けもの	0100380 HP- - で
0100040 DTT- - 、	0100210 HP- - の	0100390 ZV2- - そだち
0100050 ZN- - からだ	0100220 ZN- - なかま	0100400 HA3- - ます
0100060 JB- - じゅう	0100230 HA3- - です	0100410 DTK- - 。
0100070 HP- - に	0100240 DTK- - 。	0100420 DB- -
0100080 ZN- - け	0100250 DB- - 	0100430 ZJ4- - さむい
0100090 HP- - が	0100260 ZN- - けもの	0100440 ZN- - ところ
0100100 ZV2- - はえ	0100270 HP- - は	0100450 HP- - に
0100110 HP- - て	0100280 ZN- - あかちゃん	0100460 ZV4- - すむ
0100120 ZV4- - いる	0100290 HP- - で	0100470 ZN- - けもの
0100130 ZN- - どうぶつ	0100300 ZV2- - うまれ	0100480 HP- - に
0100140 HP- - の	0100320 HP- - て	0100490 HP- - は
0100150 ZN- - こと	0100330 DTT- - 、	0100500 ZN- - け
0100160 HA3- - です	0100340 JT- - お	0100510 HP- - が
0100170 DTK- - 。	0100350 ZN- - ちち	0100520 ZN- - いっぱい

Character set This corpus uses the same 77 hiragana characters shown in the previous section. In addition, a dash character is used to express long vowels for those few words that are normally written in katakana; other words use the standard hiragana methods of indicating vowel lengthening. A few instances of four other punctuation marks were deleted. Thus, the total character set is 78 characters.

Corpus preparation: The JAPW corpus was prepared from the tag files, placing a blank after each morpheme indicated there to create a word-separated standard to check parsing against. I broke lines at all periods, and at commas unless the resulting line would be less than 10 characters long. All punctuation and digits were then removed.

Sample after preparation (word-spacing has been added):

けものは からだ じゅう に けが はえ て いる どうぶつ の ことです
 きつね も けもの の なかま です
 けものは あかちゃん で うまれ て
 お ちち を のん で そだち ます
 さむい ところ に すむ けもの には けが いっぱい はえ て い ます
 けは おーぼー の よう に あた たか です

Sample after dictin parsing:

けものは からだ じゅう に けが はえて い る どうぶつ の ことです
 きつね も けもの の なか ま です
 けものは あか ちゃん で うまれ て
 おちち を の んで そだち ます
 さむ い ところ に すむ けもの には けが い っぱい はえ てい ます
 けは おー ぼー の ように あた たか です

Transliteration to roman characters: In order to compare the results of processing the corpus in two different orthographies, the hiragana was translated to roman characters according to the table above. It is a completely mechanical process, but it results in a version which no Japanese would recognize as normal. For instance, the palatalized きょ is usually romanized as *kyo* but here becomes *KIyo*, which is simply a literal conversion of the kana characters.

Roman character set: As shown in the table above, 21 capital letters and 7 lower-case letters are used. In addition, a hyphen is added to each 1-letter symbol to force two ASCII characters for each Japanese character. The Japanese elonga-

tion dash is represented here by two equal marks: "==" . So, in total, 30 ASCII characters are used.

Sample after roman transliteration:

KEMONO HA KARADA JIyuU- NI KE GA HAE- TE I-RU DOU-BUTU NO KOTO DESU
 KITUNE MO KEMONO NO NAKAMA DESU
 KEMONO HA A-KATiyan- DE U-MARE TE
 O- TITI WO NON- DE SODATI MASU
 SAMUI- TOKORO NI SUMU KEMONO NI HA KE GA I-tuPAI- HAE- TE I- MASU
 KE HA O==BA== NO YOU- NI A-TATAKA DESU

Sample after roman transliteration and dictin parsing:

KEMONO HA KARA DA JIyuU- NI KEGA HAE-TEI- RU DO U-BUTU NO KOTO DESU
 KI TUNE MO KEMONO NO NAKAMA DESU
 KEMONO HA A-KAT Iyan- DE U-MARE TE
 O-TITI WO NO N-DE SODATI MASU
 SAMU I- TOKORO NI SUMU KEMONO NIHA KEGA I- tuPAI- HA E-TEI- MASU
 KEHA O- == BA == NOYOU-NI A-TATAKA DESU

Appendix C: Summary Table of Corpus Characteristics

	# Tokens		#Types	Avglen(chs)		Repetition		F% baseline		Dataset	Text sections
	Char	Word		Line	Word	TTR	G	Words	Cuts		
ENG V small/"B&C"	13,690	4,746	50	8.9	2.88	9.7	0.292	13.4	50.2	brecarsh	3.2.1-3, 3.3.3-4, 4.1.1
ENG V med./phonemic	40,792	14,184	50	9.4	2.88	16.0	0.313	12.9	49.3	brecar	3.3.3, 4.2.1, 5.2.4
ENG V large	95,809	33,399	50	9.8	2.87	25.3	0.332	12.6	48.7	brechall	3.3.3, 4.1.1
ENG V longer lines	13,690	4,746	50	11.0	2.88	9.7	0.292	10.2	46.9	brecarsc	3.3.4
ENG V bounds. moved	13,688	4,745	50	8.9	2.88	9.7	0.292	13.8	50.2	brecarshut	3.3.4
ENG V half of large	48,379	16,863	50	9.9	2.87	15.8	0.306	12.5	48.6	brechalf	4.1.1
ENG V half X 2	96,758	33,726	50	9.9	2.87	31.7	0.355	12.5	48.6	brechalf2	4.1.1
ENG V orthographic	53,439	14,188	26	12.4	3.77	15.5	0.309	9.6	44.4	breconew	4.2.1, 4.3.1, 4.4
ENG V 13-18 mos	32,003	11,248	50	9.1	2.85	15.5	0.317	13.4	49.9	brekid1	4.3.1
ENG V 19-21 mos	26,129	9,003	50	9.9	2.90	12.3	0.298	12.4	48.5	brekid2	4.3.1
ENG V 22-25 mos	37,677	13,148	50	10.4	2.87	13.4	0.295	12.0	47.8	brekid3	4.3.1
ENG V phon. (=ch) ¹	53,354	18,582	50	9.4	2.87	19.5	0.325	13.1	49.4	brekidk	4.4
ENG V phon. (=wd)	40,189	14,018	50	9.2	2.87	17.1	0.320	13.3	49.8	brekidw	4.4
ENG VA younger	29,271	9,685	50	18.6	3.02	5.1	0.192	9.0	40.8	bread1	4.3.1
ENG VA older	32,513	10,557	50	20.7	3.08	5.6	0.200	8.3	39.6	bread2	4.3.1
ENG VA all	61,784	20,242	50	19.6	3.05	10.7	0.257	8.7	40.2	bread	4.3.1

¹ Note: (=ch) means matched for size by character count; (=wd) means matched for size by word count.

	# Tokens		#Types	Avglen(chs)		Repetition		F% baseline		Dataset	Text sections
	Char	Word		Line	Word	TTR	G	Words	Cuts		
ENGW full/orthog.	313,321	74,951	26	42.1	4.18	10.6	0.224	4.2	29.8	maughamt	4.1.1, 3.2, 4.2.1
ENGW half	156,821	37,475	26	42.2	4.18	7.6	0.206	4.2	29.8	mauhalf	4.1.1
ENGW half X 2	313,642	74,950	26	42.2	4.18	15.2	0.258	4.2	29.8	mauhalf2	4.1.1
ENGW short line	313,321	74,951	26	22.2	4.18	10.6	0.224	5.5	35.4	maushort	4.1.2, 4.3.1
ENGW long line	313,321	74,951	26	84.3	4.18	10.6	0.224	3.5	26.8	maulong	4.1.2
ENGW phonemic	382,328	74,951	35	51.4	5.10	10.8	0.226	3.0	18.0	maughamph	4.2.1
ENGW expanded	432,940	74,951	26	58.2	5.78	10.6	0.224	2.5	15.8	mauword	4.2.1
ENGW (=ch)	79,150	18,832	26	42.6	4.20	5.8	0.193	4.2	29.7	mau34	4.4
ENGW (=wd)	37,018	9,038	26	41.8	4.10	4.7	0.184	4.3	30.2	mau34k	4.4
SPAV w/ diacritics	103,892	26,552	32	15.6	3.91	10.3	0.246	8.3	40.4	spanall	4.2.2
SPAV no diacritics	103,892	26,552	26	15.6	3.91	10.7	0.249	8.3	40.4	spano	4.2.2
SPAV phonemic	99,288	26,552	32	14.9	3.74	10.4	0.247	8.8	41.2	spanph	4.2.2
SPAV Manuela (15-30)	25,187	6,679	32	12.1	3.77	7.5	0.248	10.7	44.8	deualle	4.3.2
SPAV Alf (27-34 mos)	25,877	6,640	31	16.4	3.90	6.9	0.238	7.6	39.6	alfc	4.3.2
SPAV Ida (32-55 mos)	24,301	6,339	32	14.3	3.83	6.1	0.224	9.0	41.8	idac	4.3.2
SPAV Raf (52-59 mos)	28,527	6,894	32	21.6	4.14	6.4	0.228	6.3	35.7	rafc	4.3.2
SPAV 15-34 mos	51,064	13,319	26	14.0	3.83	9.0	0.250	9.1	42.2	spanok1	4.3.2
SPAV 32-59 mos	52,828	13,233	26	17.5	3.99	7.8	0.234	7.6	38.6	spanok2	4.3.2
SPAV 32 symbols	51,064	13,319	32	14.0	3.83	8.7	0.245	9.1	42.2	spank1	4.4
SPAW 26 symb	126,946	25,926	26	47.6	4.90	4.2	0.150	3.3	27.1	doledt	4.3.2
SPAW 26 symb (=ch)	37,065	7,512	26	45.8	4.93	3.0	0.133	3.4	27.3	doledtk	4.4
SPAW 26 symb (=wd)	94,445	19,256	26	47.8	4.90	3.8	0.146	3.3	27.0	doledtw	4.4

	# Tokens		#Types	Avglen(chs)		Repetition		F% baseline		Dataset	Text sections
	Char	Word		Line	Word	TTR	G	Words	Cuts		
JAP V roman orig.	312,818	77,219	24	11.8	4.05	21.4	0.290	11.5	45.9	akirom	4.2.3
JAP V hiragana trans.	184,022	77,219	77	7.0	2.38	21.4	0.290	18.4	55.4	akih	4.2.3
JAP V roman retrans.	368,044	77,219	29	13.9	4.77	21.4	0.290	9.7	44.0	akihr	4.2.3
JAP V younger half	144,218	35,535	24	10.9	4.06	15.9	0.283	12.9	47.9	akiroma	4.3.3
JAP V older half	168,600	41,684	24	12.8	4.04	15.2	0.274	10.3	44.3	akiromb	4.3.3
JAP V all ages half	156,909	38,669	24	11.9	4.06	14.0	0.268	11.5	45.8	akihalf	4.3.3
JAP V oldest quarter	83,857	20,739	24	12.7	4.04	10.8	0.257	10.4	44.3	akirombb	4.3.3
JAP V oldest fifth (hir.)	39,294	16,738	74	7.4	2.35	9.7	0.252	17.7	54.6	akih5	4.3.3
JAP V roman subset	52,938	13,124	24	10.3	4.03	10.8	0.271	13.5	49.4	akiromk	4.4
JAP V hiragana (=ch)	52,902	21,892	77	6.3	2.42	13.4	0.279	19.6	57.0	akihk	4.4
JAP V hiragana (=wd)	33,783	13,988	77	6.2	2.42	11.0	0.270	19.8	57.4	akihw	4.4
JAP W hiragana orig.	37,060	19,255	78	17.7	1.92	11.0	0.262	18.7	54.6	eucnew	4.2.3, 4.3.3, 4.4
JAP W roman trans.	74,120	19,255	30	35.5	3.85	11.0	0.262	4.8	32.1	eucnewr	4.2.3, 4.3.3, 4.4
JAP W roman adj.	68,228	19,255	29	32.7	3.54	11.0	0.262	5.5	34.0	eucnewrx	4.2.3
JAP W roman (=wd)	36,948	9,598	30	35.4	3.85	8.3	0.250	4.8	32.1	eucnewrk	4.4

References

- Abney, Steven (1996), "Statistical methods and linguistics," in Klavans & Resnik, eds. [<http://www.sfs.nphil.uni-tuebingen.de/~abney/>]
- Alloy, Lauren B., and Naomi Tabachnik (1984), "Assessment of covariation by humans and animals: The joint influence of prior expectations and current situational information," *Psychological Review* 91:112-149.
- Anderson, Stephen R. (1994), "Parsing morphology: 'Factoring' words," pp. 167-183 in Eric Sven Ristad, ed., *Language computations: DIMACS workshop on human language*, March 20-22, 1992, Providence, R.I.: American Mathematical Society (DIMACS Series in Discrete Mathematics and Theoretical Computer Science, Volume 17).
- Aslin, Richard N., Julide Z. Woodward, Nicholas P. LaMendola, and Thomas G. Bever (1996), "Models of word segmentation in fluent maternal speech to infants," pp. 117-134 in Morgan & Demuth, eds.
- Atkins, Sue, Jeremy Clear, and Nicholas Ostler (1992), "Corpus design criteria," *Literary and Linguistic Computing*, 7(1):1-16.
- Atkinson, Martin (1985), "How linguistic is the one-word stage?" pp. 289-312 in Martyn D. Barrett, ed., *Children's single-word speech*, New York: Wiley.
- Bates, Elizabeth, and Jeffrey Elman (1996), "Perspective: Learning rediscovered," *Science* 274:1849.
- Bell, Timothy C., John G. Cleary, and Ian H. Witten (1990), *Text compression*, Englewood Cliffs, N.J.: Prentice-Hall.
- Benedict, Helen (1979), "Early lexical development: comprehension and production," *Journal of Child Language* 6:183-200.
- Bernstein Ratner, Nan (1996), "From 'signal to syntax': But what is the nature of the signal?" pp. 135-150 in Morgan & Demuth, eds.
- Berry, Dianne C., and Zoltan Dienes (1993), *Implicit learning: Theoretical and empirical issues*, Hillsdale, N.J.: Erlbaum.
- Bertoncini, Josiane, and Jacques Mehler (1981), "Syllables as units in infant speech perception," *Infant Behavior and Development* 4:247-260.

- Bever, Thomas G. (1975), "Cerebral asymmetries in humans are due to the differentiation of two incompatible processes: holistic and analytic," pp. 251-262 in Doris Aaronson and Robert W. Rieber, eds., *Developmental psycholinguistics and communication disorders*, Annals of the New York Academy of Sciences, Vol. 263.
- Billman, Dorrit (1989), "Systems of correlations in rule and category learning: Use of structured input in learning syntactic categories," *Language and Cognitive Processes* 4(2):127-155.
- Black, E., S. Abney, D. Flickinger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini, and T. Strzalkowski (1991), "A procedure for quantitatively comparing the syntactic coverage of English grammars," pp. 306-311 in *DARPA Speech and Natural Language Workshop*, San Mateo, Cal.: Morgan Kaufmann.
- Bornstein, Marc H. (1984), "Perceptual development," pp. 81-131 in Marc H. Bornstein and Michael E. Lamb, eds., *Developmental psychology: An advanced textbook*, Hillsdale, N.J.: Erlbaum.
- Bosch, Antal van den, and Walter Daelemans (1992), "Linguistic pattern matching capabilities of connectionist networks," pp. 183-196 in Walter Daelemans and David Powers, eds., *Background and experiments in machine learning of natural language: Proceedings first SHOE workshop*, Institute for Language Technology and AI, Tilburg University, The Netherlands.
- Brent, Michael R. (1997), "Toward a unified model of lexical acquisition and lexical access," *Journal of Psycholinguistic Research*, 26(3):363-375.
- Brent, Michael R., and Timothy A. Cartwright (1996), "Distributional regularity and phonotactic constraints are useful for segmentation," *Cognition* 61:93-125.
- Brent, Michael R., and Timothy A. Cartwright (in press), "The informational structure of word learning," in P. Broeder and J. M. J. Murre, eds., *Cognitive models of language acquisition*, Cambridge, Mass.: MIT Press.
- Briscoe, Ted (1989), "Lexical access in connected speech recognition," pp. 84-90 in *Proceedings of the 27th Meeting of the Association for Computational Linguistics..*
- Broadbent, D. E. (1980), "The minimization of models," pp. 113-128 in Antony J. Chapman and Dylan M. Jones, eds., *Models of man*, Leicester, England: The British Psychological Society.
- Brown, Roger (1958), *Words and things*, Glencoe, Ill.: Free Press.

- Brown, Roger (1973), *A first language*, Cambridge, Mass.: Harvard University Press.
- Cairns, Paul, Richard Shillcock, Nick Chater, and Joe Levy (1994), "Lexical segmentation: The role of sequential statistics in supervised and unsupervised models," *Proceedings of the 16th Annual Conference of the Cognitive Science Society*, Hillsdale, N.J.: Erlbaum.
- Cairns, Paul, Richard Shillcock, Nick Chater, and Joe Levy (1997), "Bootstrapping word boundaries: A bottom-up corpus-based approach to speech segmentation," *Cognitive Psychology* 33:111-153.
- Cartwright, Timothy A., and Michael R. Brent (1994), "Segmenting speech without a lexicon: Evidence for a bootstrapping model of lexical acquisition," pp. 148-152 in *Proceedings of the 16th Annual Conference of the Cognitive Science Society*, Hillsdale, N.J.: Erlbaum.
- Chang, Chao-Huang, and Gilbert K. Krulee (1991), "Resolution of ambiguity in Chinese and its application to machine translation," *Machine Translation* 6:4:279-315.
- Charniak, Eugene (1993), *Statistical language learning*, Cambridge, Mass.: MIT Press.
- Chen, Lilly Lee (1993), "Chinese: A not-so-analytic language," 20th LACUS Forum.
- Chiat, Shulamuth (1979), "The role of the word in phonological development," *Linguistics* 17:591-610.
- Christiansen, Morten H., Joseph Allen, and Mark S. Seidenberg (in press), "Learning to segment speech using multiple cues: A connectionist model," *Language and Cognitive Processes*.
- Christophe, Anne, Emmanuel Dupoux, Josiane Bertoncini, and Jacques Mehler (1994), "Do infants perceive word boundaries? An empirical study of the bootstrapping of lexical acquisition," *Journal of the Acoustical Society of America* 95(3):1570-80.
- Church, Kenneth Ward (1983), *Phrase-structure parsing: A method for taking advantage of allophonic constraints*, Indiana University Linguistics Club.
- Church, Kenneth, William Gale, Patrick Hanks, and Donald Hindle (1991), "Using statistics in lexical analysis," pp. 111-160 in Uri Zernik, ed., *Lexical acquisition: Exploiting on-line resources to build a lexicon.*, Hillsdale, N.J.: Erlbaum.

- Church, Kenneth W., and Robert L. Mercer (1993), "Introduction to the special issue on computational linguistics using large corpora," *Computational Linguistics* 19(1):1-24.
- Clark, Robin, Lila Gleitman, and Anthony Kroch (1997), "Letter," *Science*, 276: 1179.
- Cleeremans, Axel (1993), *Mechanisms of implicit learning: Connectionist models of sequence processing*, Cambridge, Mass.: MIT Press.
- Cover, Thomas M., and Joy A. Thomas (1991), *Elements of information theory*, New York: Wiley.
- Crystal, David (1987), *The Cambridge encyclopedia of language*, Cambridge: Cambridge University Press.
- Curtiss, Susan (1977), *Genie: A psycholinguistic study of a modern-day 'wild child'*, New York: Academic Press.
- Curtiss, Susan (1982), "Developmental dissociations of language and cognition," pp. 285-312 in Loraine K. Obler and Lise Menn, eds., *Exceptional language and linguistics*, New York: Academic Press.
- Cutler, Anne (1994), "Segmentation problems, rhythmic solutions," *Lingua* 92:81-104.
- Cutler, Anne, and David M. Carter (1987), "The predominance of strong initial syllables in the English vocabulary," *Computer Speech and Language* 2:133-142.
- Cutler, Anne, and D. G. Norris (1988), "The role of strong syllables in segmentation for lexical access," *Journal of Experimental Psychology: Human Perception and Performance* 14:113-121.
- Daelemans, Walter, Peter Berck, and Steven Gillis (1996), "Unsupervised discovery of phonological categories through supervised learning of morphological rules," in *Proceedings of COLING*. [<http://xxx.lanl.gov/cmp-lg/9607013>]
- DeCasper, A. J., and W. P. Fifer (1980), "Of human bonding: Newborns prefer their mothers' voices," *Science* 208:1174-1176.
- DeCasper, A. J., and M. J. Spence (1986), "Prenatal maternal speech influences newborns' perception of speech sounds," *Infant Behavior and Development* 9:133-150.

- de Marcken, Carl (1995), "The unsupervised acquisition of a lexicon from continuous speech," A.I. Memo No. 1558, MIT Artificial Intelligence Lab. [<http://xxx.lanl.gov/cmp-lg/9512002>]
- de Marcken, Carl (1996a), "Linguistic structure as composition and perturbation," pp. 335-341 in *Proceedings of the 34th Meeting of the Association for Computational Linguistics*.
- de Marcken, Carl G. (1996b), *Unsupervised language acquisition*, Ph.D. dissertation, MIT.
- de Marcken, Carl (undated), "Language acquisition by recursive text compression," ms., MIT Artificial Intelligence Laboratory.
- Deuchar, Margaret, and Angeles Clark (1992), "Bilingual acquisition of the voicing contrast in word-initial stop consonants in English and Spanish," *Cognitive Science Research Reports* No. 213, University of Sussex.
- Echols, Catharine H. (1996), "A role for stress in early speech segmentation," pp. 151-170 in Morgan & Demuth, eds.
- Echols, Catharine H., and Elissa L. Newport (1992), "The role of stress and position in determining first words," *Language Acquisition* 2(3):189-220.
- Echols, Catharine H., Megan J. Crowhurst, and Jane B. Childers (1997), "The perception of rhythmic units in speech by infants and adults," *Journal of Memory and Language* 36:202-225.
- Eilers, Rebecca E. (1980), "Infant speech perception: History and mystery," pp. 23-39 in Yeni-Komshian et al., eds.
- Ejiri, Koichi, and Adolph E. Smith (1993), "Proposal of a new 'constraint measure' for text," pp. 195-211 in R. Kohler and B. B. Rieger, eds., *Contributions to quantitative linguistics*, Boston: Kluwer Academic Publishers.
- Elbers, Loekie, and Frank Wijnen (1992), "Effort, production skill, and language learning," pp. 337-368 in Ferguson, Menn & Stoel-Gammon, eds.
- Elman, Jeffrey (1986), "Comment" on Klatt's paper, p. 323 in Perkell & Klatt, eds.
- Elman, Jeffrey L. (1990), "Finding structure in time," *Cognitive Science* 14:179-211.
- Elman, Jeffrey L. (1993), "Learning and development in neural networks: the importance of starting small," *Cognition* 48:71-99.

- Elman, Jeffrey L., Elizabeth A. Bates, Mark H. Johnson, Annette Karmiloff-Smith, Domenico Parisi, and Kim Plunkett (1996), *Rethinking innateness: A connectionist perspective on development*, Cambridge, Mass.: MIT Press.
- Ferguson, Charles A., and Carol B. Farwell (1975), "Words and sounds in early language acquisition," *Language* 51:419-39.
- Ferguson, Charles A., Lise Menn, and Carol Stoel-Gammon, eds. (1992), *Phonological development: Models, research, implications*, Timonium, Md.: York Press.
- Flores d'Arcais, Giovanni B. (1988), "Language perception," pp. 97-123 in Newmeyer, ed., Volume III.
- Friederici, Angela D., and Jeanine M. I. Wessels (1993), "Phonotactic knowledge of word boundaries and its use in infant speech perception," *Perception & Psychophysics* 54(3):287-295.
- Garner, Wendell R. (1962), *Uncertainty and structure as psychological concepts*, New York: Wiley.
- Garner, W. R., and William J. McGill (1956), "The relation between information and variance analyses," *Psychometrika* 21(3):219-228.
- Gerken, LouAnn, Peter W. Jusczyk, and Denise R. Mandel (1994), "When prosody fails to cue syntactic structure: Nine-month-olds' sensitivity to phonological *vs.* syntactic phrases," *Cognition* 51:237-265.
- Gleitman, Lila R., and Eric Wanner (1982), "Language acquisition: The state of the art," pp. 3-48 in Eric Wanner and Lila R. Gleitman, eds., *Language acquisition: The state of the art*, Cambridge: Cambridge University Press.
- Gleitman, Lila R., Henry Gleitman, Barbara Landau, and Eric Wanner (1988), "Where learning begins: initial representations for language learning," pp. 150-193 in Newmeyer, ed., Volume III.
- Goodman, Joshua (1996), "Parsing algorithms and metrics," pp. 177-183 in *Proceedings of the 34th Meeting of the Association for Computational Linguistics..* [<http://xxx.lanl.gov/cmp-lg/9605036>]
- Goodsitt, Jan V., Philip A. Morse, James N. VerHoeve, and Nelson Cowan (1984), "Infant speech recognition in multisyllabic contexts," *Child Development* 55:903-910.
- Goodsitt, Jan V., James L. Morgan, and Patricia K. Kuhl (1993), "Perceptual strategies in prelingual speech segmentation," *Journal of Child Language* 20:229-252.

- Green, David M., and John A. Swets (1966), *Signal detection theory and psychophysics*, New York: Wiley.
- Green, T. R. G. (1979), "The necessity of syntax markers: Two experiments with artificial languages," *Journal of Verbal Learning and Verbal Behavior* 18:481-496.
- Hallé, Pierre A., and Bénédicte de Boysson-Bardies (1994), "Emergence of an early receptive lexicon: Infants' recognition of words," *Infant Behavior and Development* 17:119-129.
- Hallé, Pierre A., and Bénédicte de Boysson-Bardies (1996), "The format of representation of recognized words in infants' early receptive lexicon," *Infant Behavior and Development* 19:463-481.
- Halvorsen, Per-Kristian (1988), "Computer applications of linguistic theory," pp. 198-219 in Newmeyer, ed., Volume II.
- Hasher, Lynn, and Rose T. Zacks (1984), "Automatic processing of fundamental information: The case of frequency of occurrence," *American Psychologist*, 39:1372-1388.
- Hatzivassiloglou, Vasileios (1996), "Do we need linguistics when we have statistics? A comparative analysis of the contributions of linguistic cues to a statistical word grouping system," pp. 67-94 in Klavans & Resnik, eds.
- Hayashi, Akiko, Yuji Tamekawa, Toshisada Deguchi, and Shigeru Kiritani (1996), "Developmental change in perception of clause boundaries by 6- and 1-month-old Japanese infants," Fourth International Conference on Spoken Language Processing (ICSLP).
- Hayes, John R., and Herbert H. Clark (1970), "Experiments on the segmentation of an artificial speech analogue," pp. 221-234 in John R. Hayes, ed., *Cognition and the development of language*, New York: Wiley.
- Hill, Archibald A. (1958), *Introduction to linguistic structures: From sound to sentence in English*, New York: Harcourt, Brace.
- Hill, Jane C. (1987), "Language acquisition," pp. 443-452 in S. C. Shapiro, ed., *Encyclopedia of artificial intelligence*, New York: Wiley.
- Hinton, Geoffrey E. (1989), "Connectionist learning procedures," *Artificial Intelligence* 40:185-234.
- Hohne, Elizabeth A., and Peter W. Jusczyk (1994), "Two-month-old infants' sensitivity to allophonic differences," *Perception and Psychophysics* 56:613-623.

- Huang, Chu-Ren, Kathleen Ahrens, and Keh-jiann Chen (1993), "A data-driven approach to psychological reality of the mental lexicon: Two studies on Chinese corpus linguistics," pp. 53-68 in *Proceedings of the International Conference on the Biologic Basis of Language*, Center of Cognitive Science, National Chung Cheng University.
- Hyman, Larry M. (1978), "Word demarcation," pp. 443-470 in Joseph H. Greenberg, ed., *Universals of human language*, Vol. 2: *Phonology*, Stanford: Stanford University Press.
- Ingram, David (1992), "Early phonological acquisition: A cross-linguistic perspective," pp. 423-435 in Ferguson, Menn & Stoel-Gammon, eds.
- Ito, Akinori, and Masaki Kohda (1996), "Language modeling by string pattern n-gram for Japanese speech recognition," Fourth International Conference on Spoken Language Processing (ICSLP).
- Juola, Patrick, Chris Hall, and Adam Boggs (1994), "Corpus-based morphological segmentation by entropy changes," ms., University of Colorado at Boulder.
- Jusczyk, Peter W. (1997), *The discovery of spoken language*, Cambridge, Mass.: MIT Press.
- Jusczyk, Peter W., and Carolyn Derrah (1987), "Representation of speech sounds by young infants," *Developmental Psychology* 23:648-654.
- Jusczyk, Peter W., Kathy Hirsh-Pasek, Deborah G. Kemler Nelson, Lori J. Kennedy, Amanda Woodward, and Julie Piwoz (1992), "Perception of acoustic correlates of major phrasal units by young infants," *Cognitive Psychology* 24:252-293.
- Jusczyk, Peter W., Anne Cutler, and Nancy J. Redanz (1993), "Infants' preference for the predominant stress patterns of English words," *Child Development* 64:675-687.
- Jusczyk, Peter W., Angela D. Friederici, Jeanine Wessels, V. Svenkerud, and A. M. Jusczyk (1993), "Infants' sensitivity to the sound structure of native language words," *Journal of Memory and Language* 32:402-420.
- Jusczyk, Peter W., P. A. Luce, and J. Charles-Luce (1994), "Infants' sensitivity to phonotactic patterns in the native language," *Journal of Memory and Language* 33:630-45.
- Jusczyk, Peter W., and Richard N. Aslin (1995), "Infants' detection of the sound patterns of words in fluent speech," *Cognitive Psychology* 29:1-23.

- Jusczyk, Peter W., Ann Marie Jusczyk, Lori J. Kennedy, Tracy Schomberg, and Nan Koenig (1995), "Young infants' retention of information about bisyllabic utterances," *Journal of Experimental Psychology: Human Perception and Performance* 21:822-36.
- Kareev, Yaakov (1995), "Through a narrow window: Working memory capacity and the detection of covariation," *Cognition* 56:263-269.
- Kelly, Michael H., and Susanne Martin (1994), "Domain-general abilities applied to domain-specific tasks: Sensitivity to probabilities in perception, cognition, and language," *Lingua* 92:105-140. Also published as Lila Gleitman and Barbara Landau, eds. (1995), *The acquisition of the lexicon*, Cambridge, Mass.: MIT.
- Kita, Kenji, Takashi Omoto, Yoneo Yano, and Yasuhiko Kato (1994), "Application of corpora in second language learning: The problem of collocational knowledge acquisition," Second Annual Workshop on Very Large Corpora, Kyoto.
- Klatt, Dennis (1986), "Problem of variability in speech recognition and in models of speech perception," pp. 300-319 in Perkell & Klatt, eds.
- Klavans, Judith, and Philip Resnik, eds. (1996), *The balancing act: Combining symbolic and statistical approaches to language*, Cambridge, Mass.: MIT Press.
- Klenk, Ursula, and Hagen Langer (1989), "Morphological segmentation without a lexicon," *Literary and Linguistic Computing* 4:247-253.
- Kucera, Henry, and William N. Francis (1967), *Computational analysis of present-day American English*, Providence: Brown University Press.
- Kuhl, Patricia K. (1980), "Perceptual constancy for speech-sound categories in early infancy," pp. 41-66 in Yeni-Komshian et al., eds.
- Kuhl, Patricia K. (1986), "Comments" on Peter W. Jusczyk's talk, pp. 19-29 in Perkell & Klatt, eds.
- Ladefoged, Peter (1982), *A course in phonetics*, second edition, San Diego: Harcourt Brace Jovanovich.
- Langenscheidt (1985), "The pronunciation of Spanish," pp. 10-12 in *Langenscheidt's Pocket Spanish Dictionary*, New York: Langenscheidt, 1985
- Lehiste, Ilse (1970), *Suprasegmentals*, Cambridge, Mass.: MIT Press.

- Leonard, Laurence B. (1992), "Models of phonological development and children with phonological disorders," pp. 495-507 in Ferguson, Menn & Stoel-Gammon, eds.
- Li, Hang, and Naoki Abe (1996), "Clustering words with the MDL principle," in *Proceedings of COLING*. [<http://xxx.lanl.gov/cmp-lg/9605014v2>]
- Lieberman, Philip (1991), *Uniquely human: The evolution of speech, thought, and selfless behavior*, Cambridge, Mass.: Harvard University Press.
- Ling, Charles X. (1996), "Can symbolic algorithms model cognitive development?" *Proceedings of the 18th Annual Conference of the Cognitive Science Society*, Hillsdale, N.J.: Erlbaum. [<http://www.csd.uwo.ca/faculty/ling/>]
- Lucky, Robert W. (1989), *Silicon dreams: Information, man, and machine*, New York: St. Martin's Press.
- Lunde, Ken (1993), *Understanding Japanese information processing*, Sebastopol, CA: O'Reilly & Associates, Inc.
- Luo, Xiaoqiang, and Salim Ruokos (1996), "An iterative algorithm to build Chinese language models," pp. 139-143 in *Proceedings of the 34th Meeting of the Association for Computational Linguistics*. [<http://xxx.lanl.gov/cmp-lg/9606021>]
- Macken, Marlys A. (1979), "Developmental reorganization of phonology: A hierarchy of basic units of acquisition," *Lingua* 49:11-49.
- MacWhinney, Brian (1995), *The CHILDES project: Tools for analyzing talk*, 2nd ed., Hillsdale, N.J.: Erlbaum.
- Makino, Hiroshi (1985), "Beta: an automatic kana-kanji translation system," *Computer*, January 1985.
- Mandel, Denise R., Peter W. Jusczyk, and David P. Pisoni (1995), "Infants' recognition of the sound patterns of their own names," *Psychological Science* 6(5):314-317.
- Mandel, Denise R., Deborah G. Kemler Nelson, and Peter W. Jusczyk (1996), "Infants remember the order of words in a spoken sentence," *Cognitive Development* 11:181-196.
- Martinet, André (1960/1964), *Elements of general linguistics*, tr. Elisabeth Palmer, Chicago: University of Chicago Press.

- Massaro, Dominic W. (1992), "Connectionist models of speech perception," pp. 321-350 in Ronan G. Reilly and Noel E. Sharkey, eds., *Connectionist approaches to natural language processing*, Hillsdale, N.J.: Erlbaum.
- Matisoff, James A. (1989), "The bulging monosyllable, or the mora the merrier: Echo-vowel adverbialization in Lahu," pp. 163-197 in Jeremy H. C. S. Davison, ed., *South-east Asian linguistics: Essays in honour of Eugénie J. A. Henderson*, School of Oriental and African Studies, University of London.
- Matsuda, Kenjiro (1992), "Accusative case marker -o deletion in Tokyo Japanese," *Annual Meeting of the Chicago Linguistic Society* 28:356-366.
- Mehler, Jacques, Peter Jusczyk, Ghislaine Lambertz, Nilofar Halsted, Josiane Bertoncini, and Claudine Amiel-Tison (1988), "A precursor of language acquisition in young infants," *Cognition* 29:143-178.
- Mehler, Jacques, and Anne Christophe (1993), "Maturation and learning of language in the first year of life," ms.
- Mehler, Jacques, Emmanuel Dupoux, Thierry Nazzi, and Ghislaine Dehaene-Lambertz (1996), "Coping with linguistic diversity: The infant's viewpoint," pp. 101-116 in Morgan & Demuth, eds.
- Menn, Lise, and Edward Matthei (1992), "The 'two-lexicon' account of child phonology: Looking back, looking ahead," pp. 211-247 in Ferguson, Menn & Stoel-Gammon, eds.
- Menyuk, Paula (1984), "Continuities and discontinuities in language development over the first two years," pp. 195-201 in Lawrence J. Raphael, Carolyn B. Raphael, and Miriam R. Valdovinos, eds., *Language and cognition: Essays in honor of Arthur J. Bronstein*, New York: Plenum Press.
- Merkel, Magnus, Bernt Nilsson and Lars Ahrenberg (1994), "A phrase-retrieval system based on recurrence," Second Annual Workshop on Very Large Corpora, Kyoto.
- Miller, George A. (1951), *Language and communication*, New York: McGraw-Hill.
- Miller, George A. (1956), "The magical number seven, plus or minus two: Some limits on our capacity for processing information," *Psychological Review* 63(2):81-96. Also pp. 14-44 in George A. Miller (1967), *The psychology of communication: Seven essays*, New York: Basic Books; and *Psychological Review* (1994) 101(2):343-352.

- Miller, George A. and Jennifer A. Selfridge (1950), "Verbal context and the recall of meaningful material," *American Journal of Psychology* 63:176-185.
- Miyata, Susanne (1995), "The Aki corpus: Longitudinal speech data of a Japanese boy aged 1.6-2.12," *Bulletin of Aichi Shukutoku Junior College* 34:183-191.
- Morgan, James L. (1994), "Converging measures of speech segmentation in preverbal infants," *Infant Behavior and Development* 17:389-403.
- Morgan, James L. (1996a), "A rhythmic bias in preverbal speech segmentation," *Journal of Memory and Language* 35(5):666-688.
- Morgan, James L. (1996b), "Prosody and the roots of parsing," *Language and Cognitive Processes* 11(1/2):69-106.
- Morgan, James L., Richard P. Meier, and Elissa L. Newport (1987), "Structural packaging in the input to language learning: Contributions of prosodic and morphological marking of phrases to the acquisition of language," *Cognitive Psychology* 19:498-550.
- Morgan, James L., and Jenny R. Saffran (1995), "Emerging integration of sequential and suprasegmental information in preverbal speech segmentation," *Child Development* 66(4):911-936.
- Morgan, James L., and Katherine Demuth (1996a), "Signal to syntax: an overview," pp. 1-22 in Morgan & Demuth, eds.
- Morgan, James L., and Katherine Demuth, eds. (1996), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*, Hillsdale, N.J.: Erlbaum.
- Morgan, James L., Rushen Shi, and Paul Allopenna (1996), "Perceptual bases of rudimentary grammatical categories: Toward a broader conceptualization of bootstrapping," pp. 263-283 in Morgan & Demuth, eds.
- Myers, James, Peter W. Jusczyk, Deborah G. Kemler Nelson, Jan Charles-Luce, Amanda L. Woodward, and Kathryn Hirsh-Pasek (1996), "Infants' sensitivity to word boundaries in fluent speech," *Journal of Child Language* 23:1-30.
- Nagao, Makoto, and Shinsuke Mori (1994), "A new method of n-gram statistics for large number of n and automatic extraction of words and phrases from large text data of Japanese," pp. 611-615 in *Proceedings of COLING*..
- Nagata, Masaaki (1994), "A stochastic Japanese morphological analyzer using a forward-DP backward-A* N-Best search algorithm," pp. 201-207 in *Proceedings of COLING*.

- Neisser, Ulric (1976), *Cognition and reality: Principles and implications of cognitive psychology*, San Francisco: W. H. Freeman & Co.
- Newmeyer, Frederick J., ed. (1988), *Linguistics: The Cambridge Survey*. Volume II: *Linguistic theory: Extensions and implications*. Volume III: *Language: Psychological and biological aspects*, Cambridge: Cambridge University Press.
- Newport, Elissa L. (1990), "Maturation constraints on language learning," *Cognitive Science* 14:11-28.
- Nobesawa, Shiho, Junya Tsutsumi, Tomoaki Nitta, Kotaro Ono, Sun Da Jiang, and Masakazu Nakanishi (1994), "Segmenting a sentence into morphemes using statistic information between words," pp. 227-232 in *Proceedings of COLING*.
- Nobesawa, Shiho, Junya Tsutsumi, Dajiang Sun, Tomohisa Sano, Kengo Sato, and Masakazu Nakanishi (1996), "Segmenting a sentences into linky strings using d-bigram statistics," in *Proceedings of COLING*.
- Olivier, Donald Cort (1968), *Stochastic grammars and language acquisition mechanisms*, Ph.D. dissertation, Harvard University.
- Oller, D. Kimbrough, and Michael P. Lynch (1992), "Infant vocalizations and innovations in infraphonology: Toward a broader theory of development and disorders," pp. 509-536 in Ferguson, Menn & Stoel-Gammon, eds.
- Pawley, Andrew, and Frances Hodgetts Syder (1983), "Two puzzles for linguistic theory: Nativelike selection and nativelike fluency," pp. 191-226 in Jack C. Richards and Richard W. Schmidt, eds., *Language and communication*, London: Longman.
- Perkell, Joseph S., and Dennis H. Klatt, eds. (1986), *Invariance and variability in speech processes*, Hillsdale, N.J.: Erlbaum.
- Peters, Ann M. (1981), "Language typology and the segmentation problem in early child language acquisition," *Proceedings of the seventh annual meeting of the Berkeley Linguistics Society*, Berkeley Linguistics Society.
- Peters, Ann M. (1983), *The units of language acquisition*, Cambridge: Cambridge University Press.
- Petri, Herbert L., and Mortimer Mishkin (1994), "Behaviorism, cognitivism and the neuropsychology of memory," *American Scientist* 82:30-37.

- Pierce, J. R. (1961), *Symbols, signals and noise: The nature and process of communication*, New York: Harper.
- Pinker, Steven (1997), "Letter," *Science*, 276:1178.
- Powers, David, and Walter Daelemans (1992), "SHOE: The extraction of hierarchical structure for machine learning of natural language (project summary)," pp. 125-159 in Walter Daelemans and David Powers, eds., *Background and experiments in machine learning of natural language: Proceedings first SHOE workshop*, Institute for Language Technology and AI, Tilburg University, The Netherlands.
- Pye, Clifton (1983), "Mayan telegraphese: Intonational determinants of inflectional development in Quiché Mayan," *Language* 59(3):583-604.
- Reber, Arthur S. (1993), *Implicit learning and tacit knowledge: An essay on the cognitive unconscious*, New York: Oxford University Press.
- Redlich, A. Norman (1993), "Redundancy reduction as a strategy for unsupervised learning," *Neural Computation* 5:289-304.
- Ristad, Eric Sven, and Robert G. Thomas (1995), "New techniques for context modeling," ms. [<http://xxx.lanl.gov/cmp-lg/9505002>]
- Rivest, Ronald L. (1987), "Learning decision lists," *Machine Learning* 2:229-246.
- Saffran, Jenny R., Elissa L. Newport, and Richard N. Aslin (1996), "Word segmentation: The role of distributional cues," *Journal of Memory and Language* 35(4):606-621.
- Saffran, Jenny R., Richard N. Aslin, and Elissa L. Newport (1996a), "Statistical cues in language acquisition: Word segmentation by infants," *Proceedings of the 18th Annual Conference of the Cognitive Science Society*, Hillsdale, N.J.: Erlbaum.
- Saffran, Jenny R., Richard N. Aslin, and Elissa L. Newport (1996b), "Statistical learning by 8-month-old infants," *Science* 274:1926-1928.
- Saffran, Jenny R., Richard N. Aslin, and Elissa L. Newport (1997), "Letter," *Science*, 276:1180-1181, 1276.
- Sakai, Shinsuke, and Michael Phillips (1993), "J-SUMMIT: Japanese spontaneous speech recognition," EUROSPEECH-93, Berlin, Germany.

- Sampson, Geoffrey (1985), *Writing systems: A linguistic introduction*, Stanford: Stanford University Press.
- Saporta, Sol (1966), "Phoneme distribution and language universals," pp. 61-72 in Joseph H. Greenberg, ed., *Universals of language*, 2nd ed., Cambridge, Mass.: MIT Press.
- Schieffelin, Bambi B. (1979), "Getting it together: An ethnographic approach to the study of the development of communicative competence," pp. 73-108 in Elinor Ochs and Bambi B. Schieffelin, eds., *Developmental pragmatics*, New York: Academic Press.
- Schwartz, Richard G. (1988), "Phonological factors in early lexical acquisition," pp. 185-222 in Michael D. Smith and John L. Locke, eds., *The emergent lexicon: The child's development of a linguistic vocabulary*, New York: Academic Press.
- Shannon, Claude E. (1948), "A mathematical theory of communication," *Bell System Technical Journal* 27:398-403; reprinted in Claude E. Shannon and Warren Weaver (1949), *The Mathematical Theory of Communication*, Urbana: University of Illinois Press.
- Shillcock, Richard, John Hicks, Paul Cairns, Joe Levy, and Nick Chater (in press), "A statistical analysis of an idealised phonological transcription of the London-Lund corpus," *Computer Speech and Language*. [<http://cogsci.ed.ac.uk/ftp/pub/rcs/corpuspaper>]
- Siegel, Sidney, and N. John Castellan, Jr. (1988), *Nonparametric statistics for the behavioral sciences*, 2nd ed., New York: McGraw-Hill.
- Slobin, D. I. (1982), "Universal and particular in the acquisition of language," in Eric Wanner and Lila R. Gleitman, eds., *Language acquisition: The state of the art*, Cambridge: Cambridge University Press.
- Slobin, Dan I. (1985), "Crosslinguistic evidence for the language-making capacity," pp. 1157-1256 in Dan I. Slobin, ed., *The crosslinguistic study of language acquisition: Volume 2: Theoretical issues*, Hillsdale, N.J.: Erlbaum.
- Smadja, Frank (1993), "Retrieving collocations from text: Xtract," *Computational Linguistics* 19:143-177.
- Sproat, Richard, Chin Shih, William Gale, and Nancy Chang (1996), "A stochastic finite-state word-segmentation algorithm for Chinese," *Computational Linguistics* 22(3):377-404.

- Stemberger, Joseph Paul (1992), "A connectionist view of child phonology: Phonological processing without phonological processes," pp. 165-189 in Ferguson, Menn & Stoel-Gammon, eds.
- Stoel-Gammon, Carol (1992a), "Research on phonological development: Recent advances," pp. 273-282 in Ferguson, Menn & Stoel-Gammon, eds.
- Stoel-Gammon, Carol (1992b), "Prelinguistic vocal development: Measurement and predictions," pp. 439-456 in Ferguson, Menn & Stoel-Gammon, eds.
- Strange, Winifred, ed. (1995), "Cross-linguistic studies of speech perception: A historical review," pp. 3-45 in Winifred Strange, ed., *Speech perception and linguistic experience: Issues in cross-language research*, Baltimore: York Press.
- Teller, Virginia, and Eleanor Olds Batchelder (1994), "A probabilistic algorithm for segmenting non-kanji Japanese strings," pp. 742-747 in *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI-94)*.
- Valian, Virginia, and Seana Coulson (1988), "Anchor points in language learning: The role of marker frequency," *Journal of Memory and Language* 27:71-86.
- Vance, Timothy J. (1987), *An introduction to Japanese phonology*, Albany: State University of New York Press.
- van de Weijer, Joost (1997), "The suprasegmental structure of infant-directed speech," poster presented at Groningen Assembly on Language Acquisition (GALA), University of Edinburgh, Scotland.
- Vihman, Marilyn May (1981), "Phonology and the development of the lexicon: evidence from children's errors," *Journal of Child Language* 8:239-64.
- Vihman, Marilyn May (1996), *Phonological development: The origins of language in the child*, Cambridge, Mass.: Blackwell.
- Vihman, Marilyn May, and Lorraine McCune (1994), "When is a word a word?" *Journal of Child Language* 21:517-542.
- Wakefield, James A., Jr., Eugene B. Doughtie, and Byong-Hee Lee Yom (1974), "The identification of structural components of an unknown language," *Journal of Psycholinguistic Research* 3(3):261-269.
- Walley, Amanda C. (1993), "The role of vocabulary development in children's spoken word recognition and segmentation ability," *Developmental Review* 13:286-350.

- Waterson, Natalie (1971), "Child phonology: a prosodic view," *Journal of Linguistics*, 7:179-211.
- Weischedel, Ralph, Marie Meteer, Richard Schwartz, Lance Ramshaw, and Jeff Palmucci (1993), "Coping with ambiguity and unknown words through probabilistic models," *Computational Linguistics* 19(2):359-382.
- Werker, Janet F., and Judith E. Pegg (1992), "Infant speech perception and phonological acquisition," pp. 285-311 in Ferguson, Menn & Stoel-Gammon, eds.
- Werker, Janet F., Valerie L. Lloyd, Judith E. Pegg, and Linda Polka (1996), "Putting the baby in the bootstraps: Toward a more complete understanding of the role of the input in infant speech processing," pp. 427-447 in Morgan & Demuth, eds.
- Winston, P. H. (1984), *Artificial intelligence*, Reading, Mass.: Addison-Wesley.
- Wode, Henning (1992), "Categorical perception and segmental coding in the ontogeny of sound systems: A universal approach," pp. 605-631 in Ferguson, Menn & Stoel-Gammon, eds.
- Wolff, J. G. (1975), "An algorithm for the segmentation of an artificial language analogue," *British Journal of Psychology* 66:79-90.
- Wolff, J. G. (1977), "The discovery of segments in natural language," *British Journal of Psychology* 68:97-106.
- Wu, Zimin, and Gwyneth Tseng (1993), "Chinese text segmentation for text retrieval: Achievements and problems," *Journal of the American Society for Information Science* 44(9):532-542.
- Yarowsky, David (1994), "Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French," *Proceedings of the 32nd Meeting of the Association for Computational Linguistics*.
- Yeni-Komshian, Grace H., James F. Kavanagh, and Charles A. Ferguson, eds. (1980), *Child phonology*, Vol. 2: *Perception*, New York: Academic Press.
- Zipf, George K. (1935/1965), *The psycho-biology of language*, Cambridge, Mass.: MIT Press.