

A Finite-State Parser for Use in Speech Recognition

Kenneth W. Church

NE43-307

Massachusetts Institute of Technology

Cambridge, MA. 02139

This paper is divided into two parts.¹ The first section motivates the application of finite-state parsing techniques at the phonetic level in order to exploit certain classes of contextual constraints. In the second section, the parsing framework is extended in order to account for 'feature spreading' (e.g., agreement and co-articulation) in a natural way.

1. Parsing at the Phonetic Level

It is well known that phonemes have different acoustic/phonetic realizations depending on the context. For example, the phoneme /ʊ/ is typically realized with a different allophone (phonetic variant) in syllable initial position than in syllable final position. In syllable initial position (e.g., *Tom*), /ʊ/ is almost always released (with a strong burst of energy) and aspirated (with h-like noise), whereas in syllable final position (e.g., *cat*), /ʊ/ is often unreleased and unaspirated. It is common practice in speech research to distinguish acoustic/phonetic properties that vary a great deal with context (e.g., release and aspiration) from those that are relatively invariant to context (e.g., place, manner and voicing).² In the past, the emphasis has been on invariants; allophonic variation is traditionally seen as problematic for recognition.

(1) "In most systems for sentence recognition, such modifications must be viewed as a kind of 'noise' that makes it more difficult to hypothesize lexical candidates given an input phonetic transcription. To see that this must be the case, we note that each phonological rule [in an example to be presented below]

results in irreversible ambiguity – the phonological rule does not have a unique inverse that could be used to recover the underlying phonemic representation for a lexical item. For example, ... schwa vowels could be the first vowel in a word like 'about' or the surface realization of almost any English vowel appearing in a sufficiently destressed word. The tongue flap [ɾ] could have come from a /t/ or a /d/." Klatt (MIT) [21, pp. 548-549]

This view of allophonic variation is representative of much of the speech recognition literature, especially during the ARPA speech project. One can find similar statements by Cole and Jakimik (CMU) [5] and by Jelinek (IBM) [17].

I prefer to think of variation as *useful*. It is well known that allophonic contrasts can be distinctive, as illustrated by the following famous minimal pairs where the crucial distinctions seem to lie in the allophonic realization of the /ʊ/:

- | | | |
|------|--|---------------------------------|
| (2a) | a <i>ʒ</i> ease / a <i>ʒ</i> ease | <i>aspirated / flapped</i> |
| (2b) | nigh <i>ʒ</i> rate / ni <i>ʒ</i> rate | <i>unreleased / retroflexed</i> |
| (2c) | gr <i>ea</i> ʒ wine / gr <i>ay</i> ʒwine | <i>unreleased / rounded</i> |

This evidence suggests that allophonic variation provides a rich source of constraints on syllable structure and word stress. The recognizer to be discussed here (and partly implemented in Church [4]) is designed to exploit allophonic and phonotactic cues by parsing the input utterance into syllables and other suprasegmental constituents using phrase-structure parsing techniques.

1.1 An Example of Lexical Retrieval

It might be helpful to work out an example in order to illustrate how parsing can play a role in lexical retrieval. Consider the phonetic transcription, mentioned above in the citation from Klatt [20, p. 1346] [21, pp. 548-549]:

1. This research was supported (in part) by the National Institutes of Health Grant No. 1 P01 GM 03374-01 and 03374-02 from the National Library of Medicine.

2. *Place* refers to the location of the constriction in the vocal tract. Examples include: labial (at the lips) /p, b, f, v, m/, velar /k, g, ŋ/, dental (at the teeth) /s, z, t, d, l, n/ and palatal /ʃ, ʒ, tʃ/. *Manner* distinguishes among vowels, liquids and glides (e.g., /l, r, y, w/), fricatives (e.g., /s, z, f, v/), nasals (e.g., /n, m, ŋ/) and stops (e.g., /p, t, k, b, d, g/). *Voicing* (periodic vibration of the vocal folds) distinguishes sounds like /b, d, g/ from sounds like /p, t, k/.

(3) [dʃðhɪtʌtəm]

It is desired to decode (3) into the string of words:

(4) Did you hit it to Tom?

In practice, the lexical retrieval problem is complicated by errors in the front end. However, even with an ideal error-free front-end, it is difficult to decode (3) because, among other things, there are extensive rule-governed changes affecting the way that words are pronounced in different sentence contexts, as Klatt's example illustrates:

- (5a) Palatalization of /d/ before /y/ in *did you*
- (5b) Reduction of unstressed /u/ to schwa in *you*
- (5c) Flapping of intervocalic /t/ in *hit it*
- (5d) Reduction of schwa and devoicing of /u/ in *to*
- (5e) Reduction of geminate /t/ in *it to*

These allophonic processes often appear to *neutralize* phonemic distinctions. For example, the voicing contrast between /t/ and /d/, which is usually distinctive, is almost completely lost in *writer/riders*, where both /t/ and /d/ are realized in American English with a tongue flap [ɾ].

1.2 An Optimistic View of Neutralization

Fortunately, there are many fewer cases of true neutralization than it might seem. Even in *writer/riders*, the voicing contrast is not completely lost. The vowel in *riders* tends to be longer than the vowel in *writer* due to a general process that lengthens vowels before voiced consonants (e.g., /d/) and shortens them before unvoiced consonants (e.g., /t/).

A similar lengthening argument can be used to separate /n/ and /nd/ (at least in some cases). It might be suggested that /n/ is merged with /nd/ by a /d/ deletion rule that applies in words like *mend*, *wind* (noun), *wind* (verb), and *find*. (Admittedly there is little if any direct acoustic evidence for a /d/ segment in this environment.) However, I suspect that these words can often be distinguished from *men*, *win*, *wine*, and *fine* mostly on the basis of the duration of the nasal murmur which is lengthened in the precedence of a voiced obstruent like /d/. Thus, this /d/-deletion process is probably not a true case of neutralization.

Recent studies in acoustic/phonetics seem to indicate that more and more cases of apparent neutralization can be separated as the field progresses. For instance, it has been said that /s/ merges with /ʃ/ in a

context like *gas shortage* [12]. However, a recent experiment [27] suggests that the /sʃ/ sequence can be distinguished from /ʃʃ/ (as in *fish shortage*) on the basis of a spectral tilt: the /sʃ/ spectrum is more /s/-like in the beginning and more /ʃ/-like at the end, whereas the /ʃʃ/ spectrum is relatively constant throughout. A similar spectral tilt argument can be used to separate other cases of apparent gemination (e.g., /zð/ in *is the*).

As a final example of apparent neutralization, consider the portion of the spectrogram in Figure 1 between 0.85 and 1.1 seconds. This corresponds to the two adjacent /t/s in *Did you hit it to Tom?* Klatt analyzed this region with a single geminated /t/. However, upon further investigation of the spectrum, I believe that there are acoustic cues for two segments. Note especially the total energy, which displays two peaks at 0.95 and 1.02 seconds. On the basis of this evidence, I will replace Klatt's transcription (6a) with (6b):

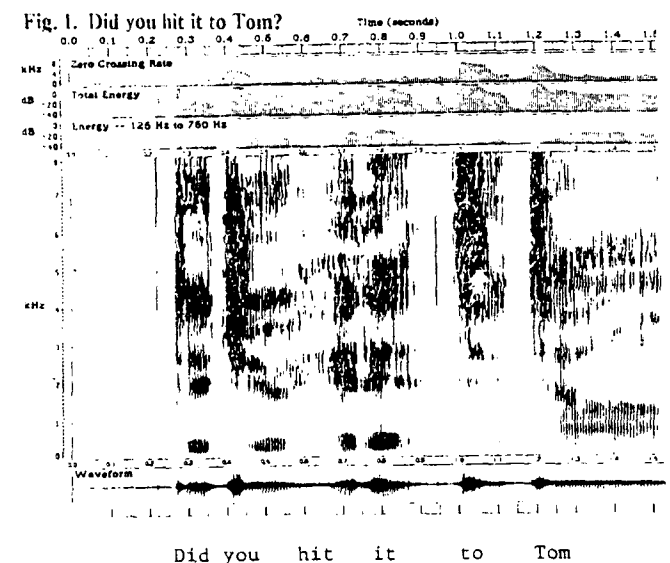
(6a) [dʃðhɪtʌtəm]

(6b) [dʃðhɪtʌtʌtəm]

↑↑

1.3 Parsing and Matching

Even though I might be able to re-interpret many cases of apparent neutralization, it remains extremely difficult to "undo" the allophonic rules by inverse transformational parsing techniques. Let me suggest an alternative proposal. I will treat syllable structure as an intermediate level of representation between the input segment lattice and the output word lattice. In so doing, I have replaced the lexical retrieval problem with two (hopefully simpler) problems: (a) parse the segment lattice into syllable structure, and (b) match the resulting constituents against the lexicon. I will illustrate the approach with



Klatt's example (enhanced with allophonic diacritics to show aspiration and glottalization):

(7) [dʃəhɪlɪʔtʰɪtʰam]
 ↑ ↑ ↑

Using phonotactic and allophonic constraints on syllable structure such as:³

- | | | |
|------|-----------------------------------|--------------------|
| (8a) | /h/ is always syllable initial. | <i>phonotactic</i> |
| (8b) | [ɪ] is always syllable final. | <i>allophonic</i> |
| (8c) | [ʔ] is always syllable final, and | <i>allophonic</i> |
| (8d) | [tʰ] is always syllable initial. | <i>allophonic</i> |

the parser can insert the following syllable boundaries:

(9) [dʃ]ə # hɪɪ # ɪʔ # tʰɪ # tʰam

It is now it is relatively easy to decode the utterance with lexical matching routines similar to those in Smith's Noah program at CMU [24].

| <u>parsed transcription</u> | | <u>decoding</u> |
|-----------------------------|---|-----------------|
| dʃə | → | did you |
| hɪɪ | → | hit |
| ɪʔ | → | it |
| tʰɪ | → | to |
| tʰam | → | Tom |

In summary, I believe that the lexical retrieval device will be in a superior position to hypothesize word candidates if it exploits allophonic and phonotactic constraints on syllable structure.

1.4 Exploiting Redundancy

In many cases, allophonic and phonotactic constraints are redundant. Even if the parser should miss a few of the cues for syllable structure, it will often be able to find the correct structure by taking advantage of some other redundant cue. For example, suppose that the front end failed to notice the glottalized /ɪ/ in the word *it*.

(10) dʃə # hɪɪ # ɪ # tʰɪ # tʰam
 ↑

The parser could deduce that the input transcription (10) is internally inconsistent, because of a phonotactic constraint on the lax vowel /ɪ/.

Lax vowels are restricted to closed syllables (syllables ending in a consonant) [1]. However, in this case, /ɪ/ cannot meet the closed syllable restriction because the following consonant is aspirated (and therefore syllable initial). Thus the transcription is internally inconsistent. The parser should probably reject the transcription and hope that the front end can fix the problem. Alternatively, the parser might attempt to correct the error by hypothesizing a second /ɪ/.⁴

There are many other examples like (10) where phonotactic constraints and allophonic constraints overlap. Consider the pairs found in figure 2, where there are multiple arguments for assigning the crucial syllable boundary. In *de-prive vs. dep-rivation*, for instance, the difference is revealed by the vowel argument above⁵ and by the aspiration rule.⁶ In addition, the stress contrast will probably be correlated with a number of so-called 'suprasegmental' cues, e.g., duration, fundamental frequency, and intensity [8].

In general, there seem to be a large number of multiple low level cues for syllable structure. This observation, if correct, could be viewed as a form of a 'constituency hypothesis'. Just as syntacticians have argued for the constituent-hood of noun phrases, verb phrases and sentences on the grounds that these constituents seem to capture crucial linguistic generalizations (e.g., question formation, wh-movement), so too, I might argue (along with certain phonologists such as Kahn [13]) that syllables, onsets, and rhymes are constituents because they also capture important generalizations such as aspiration, tensing and laxing. If this constituency hypothesis for phonology is correct (and I believe

Fig. 2. Some Structural Contrasts

| | <u>ɪ</u> | <u>ɪ</u> | <u>w</u> |
|----------|----------------------------|----------------------------|--------------------------|
| <u>d</u> | de-prive dep-rivation | di-plomacy dip-lomatic | |
| <u>t</u> | a-ttribute att-ribute | | |
| <u>k</u> | de-crease dec-riment | de-cline dec-lation | a-cquire acq-uisition |
| <u>b</u> | cele-bration celeb-rity | o-bligatory ob-ligation | |
| <u>d</u> | a-ddress add-ress | | |
| <u>g</u> | de-grade deg-radation | | |

4. Personally, I favor the first alternative: after years of witnessing Victor Zue read spectrograms, I have become most impressed with the richness of low level phonetic cues.

5. The syllable *de-* is open because the vowel is tense (diphthongized); *dep-* is closed because the vowel is lax.

6. The /p/ in *-prive* is syllable initial because it is aspirated whereas the /p/ in *dep-* is syllable final because it is unaspirated.

3. This formulation of the constraints is oversimplified for expository convenience; see [10, 13, 15] and references therein for discussion of the more subtle issues.

The agreement problem also arises in phonology. Consider the example of homorganic nasal clusters (e.g., *camp*, *can't*, *sank*), where the nasal agrees with the following obstruent in place of articulation. That is, the labial nasal /m/ is found before the labial stop /p/, the coronal nasal /n/ before the coronal stop /t/, and the velar nasal /ŋ/ before the velar stop /k/. This constraint, like subject-verb agreement, poses a problem for pure unaugmented context-free rules; it seems to be necessary to expand out each of the three cases:

- (13a) homorganic-nasal-cluster → labial-nasal labial-obstruent
- (13b) homorganic-nasal-cluster → coronal-nasal coronal-obstruent
- (13c) homorganic-nasal-cluster → velar-nasal velar-obstruent

In an effort to alleviate this expansion problem, many researchers have proposed augmentations of various sorts (e.g., ATN registers [26], LFG constraint equations [16], GPSG meta-rules [11], local constraints [18], bit vectors [6, 22]). My own solution will be suggested after I have had a chance to describe the parser in further detail.

2.2 A Parser Based on Matrix Operations

This section will show how the grammar can be implemented in terms of operations on binary matrices. Suppose that the chart is decomposed into a sum of binary matrices:

$$(14) \text{ Chart} = \text{syl } M_{\text{syll}} + \text{onset } M_{\text{onset}} + \text{peak } M_{\text{peak}} + \dots$$

where M_{syll} is a binary matrix⁸ describing the location of syllables and M_{onset} is a binary matrix describing the location of onsets, and so forth. Each of these binary matrices has a 1 in position (i, j) if there is a constituent of the appropriate part of speech spanning from the ith position in the input sentence to the jth position.⁹ (See figure 3).

Phrase-structure rules will be implemented with simple operations on these binary matrices. For example, the homorganic rule (13) could be implemented as:

8. These matrices will sometimes be called *segmentation lattices* for historical reasons. Technically, these matrices need not conform to the restrictions of a lattice, and therefore, the weaker term *graph* is more correct.

9. In a probabilistic framework, one could replace all of the 1's and 0's with probabilities. A high probability in location (i, j) of the syllable matrix would say that there probably is a syllable from position i to position j; a low probability would say that there probably isn't a syllable between i and j. Most of the following applies to probability matrices as well as binary matrices, though the probability matrices may be less sparse and consequently less efficient.

Fig. 3. M_{syll} , M_{onset} , and M_{rhyme} for: "ŋ t s ʒ ʒ"

| | | |
|-------------|-------------|-------------|
| 0 0 1 1 0 0 | 0 1 0 0 0 0 | 0 0 0 0 0 0 |
| 0 0 1 1 0 0 | 0 0 0 0 0 0 | 0 0 1 1 0 0 |
| 0 0 0 0 1 1 | 0 0 0 1 0 0 | 0 0 0 0 0 0 |
| 0 0 0 0 1 1 | 0 0 0 0 0 1 | 0 0 0 0 1 1 |
| 0 0 0 0 0 0 | 0 0 0 0 0 0 | 0 0 0 0 0 0 |
| 0 0 0 0 0 0 | 0 0 0 0 0 0 | 0 0 0 0 0 0 |

The matrices tend to be very *sparse* (almost entirely full of 0's) because syllable grammars are highly constrained. In principle, there could be n^2 entries. However, it can be shown that e (the number of 1's) is linearly related to n because syllables have finite length. In Church [4], I sharpen this result by arguing that e tends to be bounded by 4n as a consequence of a phonotactic principle known as sonority. Many more edges will be ruled out by a number of other linguistic constraints mentioned above: voicing and place assimilation, aspiration, flapping, etc. In short, these matrices are sparse because *allophonic and phonotactic constraints are useful*.

-
- (15) (setq homorganic-nasal-lattice
 - (M + (M* (phoneme-lattice #/m) labial-lattice)
 - (M* (phoneme-lattice #/n) coronal-lattice)
 - (M* (phoneme-lattice #/G) velar-lattice)))

illustrating the use of M+ (matrix addition) to express the union of several alternatives and M* (matrix multiplication) to express the concatenation of subparts. It is well known that any finite-state grammar could be implemented in this way with just three matrix operations: M*, M+, and M** (transitive closure). If context-free power were required, Valiant's algorithm [25] could be employed. However, since there doesn't seem to be a need for additional generative capacity in speech applications, the system is restricted to handle only the simpler finite state case.¹⁰

2.3 Feature Manipulation

Although "pure" unaugmented finite state grammars may be adequate for speech applications (in the weak generative capacity sense), I may, nevertheless, wish to introduce additional mechanism in order to account for agreement facts in a natural way. As discussed above, the formulation of the homorganic rule in (15) is unattractive because it splits the rule into three cases, one for each place of articulation. It would be preferable to state the agreement constraint just once, by defining a homorganic nasal cluster to be a nasal cluster

10. I personally hold a much more controversial position, that finite state grammars are sufficient for most, if not all, natural language tasks [3].

subject to place assimilation. In my language of matrix operations, I can say just exactly that:

(16) (setq homorganic-nasal-cluster-lattice
 (M& nasal-cluster-lattice
 place-assimilation))

where M& (element-wise intersection) implements the *subject to* constraint. Nasal-cluster and place-assimilation are defined as:

(17a) (setq nasal-cluster-lattice
 (M* nasal-lattice obstruent-lattice))

(17b) (setq place-assimilation-lattice
 (M + (M** labial-lattice)
 (M** dental-lattice)
 (M** velar-lattice)))

In this way, M& seems to be an attractive solution to the agreement problem.

In addition, M& might also shed some light on co-articulation, another problem of 'feature spreading'. Co-articulation (articulation of multiple phonemes at the same time) makes it extremely difficult (perhaps impossible) to segment the speech waveform into phoneme-sized units without overlapping time intervals. In order to account for co-articulation, Fujimura suggests that place, manner and other articulatory features be thought of as asynchronous processes, which have a certain amount of freedom to overlap in time.

(18a) "Speech is commonly viewed as the result of concatenating phonetic segments. In most discussions of the temporal structure of speech, a segment in such a model is assumed to represent a phoneme-sized phonetic unit, which possesses an inherent [invariant] target value in terms of articulation or acoustic manifestation. Any deviation from such an interpretation of observed phenomena requires special attention ... [B]ased on some preliminary results of X-ray microbeam studies [which associate lip, tongue and jaw movements with phonetic events in the utterance], it will be suggested that understanding articulatory processes, which are inherently multi-dimensional [and (more or less) asynchronous], may be essential for a successful description of temporal structures of speech." [9 p. 66]

In light of Fujimura's suggestion, I might re-interpret my parser as a highly parallel feature-based asynchronous architecture. For example, the parser can process homorganic nasal clusters by processing place and manner phrases in parallel, and then synchronizing the results at the coda node with M&. That is, (17a) can be computed in parallel with

(17b), and then the results are aligned when the coda is computed with (16), as illustrated below for the word *tent*. Imagine that the front end produces the following analysis:

(19)

| | | | | |
|---------------|---|---|---|---|
| | t | e | n | t |
| dental: | · | · | · | · |
| vowel: | · | · | · | · |
| stop: | · | · | · | · |
| nasalization: | · | · | · | · |

where many of the features overlap in an asynchronous way. The parser will correctly locate the coda by intersecting the nasal cluster lattice (computed with (17a)) with the homorganic lattice (computed with (17b)).

(20)

| | | | | |
|----------------|---|---|---|---|
| | t | e | n | t |
| nasal cluster: | · | · | · | · |
| homorganic: | · | · | · | · |
| coda: | · | · | · | · |

This parser is a bold departure from a standard practice in two respects: (1) the input stream is feature-based rather than segmental, and (2) the output parse is a heterarchy of overlapping constituents (e.g., place and manner phrases) as opposed to a list of hierarchical parse-trees. I find these two modifications most exciting and worthy of further investigation.

In summary, two points have been made. First, I suggested the use of parsing techniques at the segmental/feature level in speech applications. Secondly, I introduced M& as a possible solution to the agreement/co-articulation problem.

3. Acknowledgements

I have received a considerable amount of help and support over the course of this project. Let me mention just a few of the people that I should thank: Jon Allen, Glenn Burke, Francine Chen, Scott Cyphers, Sarah Ferguson, Margaret Fleck, Dan Huttenlocher, Jay Keyser, Lori Lamel, Ramesh Patil, Janet Pierrehumbert, Dave Shipman, Pete Szolovits, Meg Withgott and Victor Zue.

References

1. Barnwell, T., *An Algorithm for Segment Durations in a Reading Machine Context*, unpublished doctoral dissertation, department of Electrical Engineering and Computer Science, MIT, 1970.
2. Chomsky, N. and Halle, M., *The Sound Pattern of English*, Harper & Row, 1968.
3. Church, K., *On Memory Limitations in Natural Language Processing*, MS Thesis, MIT, MIT/LCS/TR-245, 1980 (also available from Indiana University Linguistics Club).

4. Church, K., *Phrase-Structure Parsing: A Method for Taking Advantage of Allophonic Constraints*, unpublished doctoral dissertation, department of Electrical Engineering and Computer Science, MIT, 1983 (also to appear, LCS and RLE publications, MIT).
5. Cole, R., and Jakimik, J., *A Model of Speech Perception*, in R. Cole (ed.), *Perception and Production of Fluent Speech*, Lawrence Erlbaum, Hillsdale, N.J., 1980.
6. Dostert, B., and Thompson, F., *How Features Resolve Syntactic Ambiguity*, in Proceedings of the Symposium on Information Storage and Retrieval, Minker, J., and Rosenfeld, S. (ed.), 1971.
7. Earley, J., *An Efficient Context-Free Parsing Algorithm*, CACM, 13:2, February, 1970.
8. Fry, D., *Duration and Intensity as Physical Correlates of Linguistic Stress*, JASA 17:4, 1955, (reprinted in Lechiste (ed.), *Readings in Acoustic Phonetics*, MIT Press, 1967.)
9. Fujimura, O., *Temporal Organization of Articulatory Movements as Multidimensional Phrasal Structure*, *Phonetica*, 33: pp. 66-83, 1981.
10. Fujimura, O., and Lovins, J., *Syllables as Concatenative Phonetic Units*, Indiana University Linguistics Club, 1982.
11. Gazdar, G., *Phrase Structure Grammar*, in P. Jacobson and G. Pullum (eds.), *The Nature of Syntactic Representation*, D. Reidel, Dordrecht, in press, 1982.
12. Heffner, R., *General Phonetics*, The University of Wisconsin Press, 1960.
13. Kahn, D., *Syllable-Based Generalizations in English Phonology*, Indiana University Linguistics Club, 1976.
14. Kiparsky, P., *Remarks on the Metrical Structure of the Syllable*, in W. Dressler (ed.) *Phonologica 1980. Proceedings of the Fourth International Phonology Meeting*, 1981.
15. Kiparsky, P., *Metrical Structure Assignments in Cyclic*, *Linguistic Inquiry*, 10, pp. 421-441, 1979.
16. Kaplan, R. and Bresnan, J., *Lexical-Functional Grammar: A Formal System for Grammatical Representation*, in Bresnan (ed.), *The Mental Representation of Grammatical Relations*, MIT Press, 1982.
17. Jelinek, F., course notes, MIT, 1982.
18. Joshi, A., and Levy, L., *Phrase Structure Trees Bear More Fruit Than You Would Have Thought*, *AJCL*, 8:1, 1982.
19. Klatt, D., *Word Verification in a Speech Understanding System*, in R. Reddy (ed.), *Speech Recognition, Invited Papers Presented at the 1974 IEEE Symposium*, Academic Press, pp. 321-344, 1974.
20. Klatt, D., *Review of the ARPA Speech Understanding Project*, JASA, 62:6, December 1977.
21. Klatt, D., *Scriber and Lafs: Two New Approaches to Speech Analysis*, chapter 25 in W. Lea, *Trends in Speech Recognition*, Prentice-Hall, 1980.
22. Martin, W., Church, K., and Patil, R., *Preliminary Analysis of a Breadth-First Parsing Algorithm: Theoretical and Experimental Results*, MIT/LCS/TR-261, 1981 (also to appear in L. Bolc (ed.), *Natural Language Parsing Systems*, Macmillan, London).
23. Reddy, R., *Speech Recognition by Machine: A Review*, Proceedings of the IEEE, pp. 501-531, April 1976.
24. Smith, A., *Word Hypothesis in the Hearsay-II Speech System*, Proc. IEEE Int. Conf. ASSP, pp. 549-552, 1976.
25. Valient, L., *General Context Free Recognition in Less Than Cubic Time*, J. Computer and System Sciences 10, pp. 308-315, 1975.
26. Woods, W., *Transition Network Grammars for Natural Language Analysis*, CACM, 13:10, 1970.
27. Zue, V., and Shattuck-Hufnagel, S., *When is a /s/ not a /s/?*, ASA, Atlanta, 1980.