THE USE OF LINGUISTIC HIERARCHIES IN SPEECH UNDERSTANDING[1]

Stephanie Seneff

Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139 USA
http://www.sls.lcs.mit.edu, mailto: seneff@sls.lcs.mit.edu

ABSTRACT

This paper describes two related systems which provide frameworks for encoding linguistic knowledge into formal rules within the context of a trainable probabilistic model. The first system, TINA [33], drives top-down from sentence level structure, terminating in either words or syllables. Its main purpose is to provide a meaning representation for the sentence. The other system, ANGIE [36], operates bottom-up from phonetic or orthographic units, characterizing the substructure of syllables/words. It provides a framework for both phonological rule modelling and letter-to-sound/sound-to-letter transformations. The two systems logically converge on the syllable or word layer. We have recently been successful in integrating their combined constraint into a recognizer search, achieving considerable improvement in understanding accuracy [9, 23]. In this paper, I will look both toward the past and the future, identifying and motivating the decisions that were made in the design of TINA and ANGIE and the associated rule formalisms, and contemplating various remaining open research issues.

1. INTRODUCTION

Speech is first and foremost a *communicative* signal. It is a complex encoding of linguistic messages for the purpose of conveying information among humans who share the code. Speech scientists have been studying various aspects of the speech code for many decades, and engineers have been involved in designing computer systems that attain a certain degree of competence in understanding the code.

At the core of human communication is the notion of "words" as the fundamental units. Above the word level, it is apparent that words group into phrases, and phrases group into higher level units such as clauses. Linguists have done much to describe the syntactic structure of speech [19], and have attempted to address the issues of how syntax and semantics might interact [16].

Studies of the structures of words in multiple languages have revealed a great deal of substructure [31]. The exact specification of that substructure still eludes us, however, particularly for languages such as English with a rich borrowing from other languages. We are now reasonably confident that the syllable exists as an intermediate layer between words and phonemes, although most speech recognition systems make little or no use of this syllable layer. There is also the possibility of breaking words down into *meaning* units (i.e., morphemes), which may not necessarily align precisely with syllable units based strictly on phonology and sonority. The difficulty of defining exactly how the phonemes of a word might group themselves into natural subunits has been a major hurdle to the design of systems that utilize this intermediate structure.

Over the last decade, members of the Spoken Language Systems group at the MIT Laboratory for Computer Science have been involved with building systems that attempt to *understand* conversational speech within highly restricted domains. These systems typically interact with a user in order to provide some information available in local databases or on the Web, such as flight schedules [44], weather information [45], or direction-finding in a city [14]. Throughout this time, I have been intrigued by the notion of representing linguistic knowledge in hierarchies, within a trainable probabilistic framework. It is my belief that such representations may have significant advantages over a flatter structure, in terms of being able to generalize knowledge across similar contexts. Above the word level, it seems appropriate to intermix syntax and semantics within the rules rather than to commit to one or the other operating alone, because this is a good way to realize strong semantic constraint in the grammar while preserving syntactic structure necessary for a proper meaning representation. Below the word level, it seems analogously appropriate to intermix morphology and syllabification, both of which are necessary to achieve adequate letter-to-sound rules. Within restricted domains this approach has been feasible, although it remains unclear whether it will scale to handle all of English.

The system that parses words into meaning, called TINA [33], operates top-down, and produces a meaning representation that is used by the backend of all of our systems for database lookup, response generation, etc. The system that parses phones into words, called ANGIE [36], operates bottom-up, and functions in part to represent phonological effects and letter-sound patterns probabilistically. It also produces a structural analysis, which has been used effectively by a hierarchical duration model to further improve recognition performance [8]. We believe ANGIE will be useful as well for characterizing unseen words or adding new vocabulary items incrementally through generalizations of learned structure [23]. The terminal layer in the hierarchy can be either phones or letters. The phone terminals capture phonological rules such as palatalization, devoicing, stop-deletion, glottaliza-

[1] This material is based upon work supported by DARPA under contract N66001-96-C-8526, monitored through Naval Command, Control and Ocean Surveillance Center, and by the National Science Foundation under Grant No. IRI-96187321.

tion, and gemination. The letter terminals provide a reversible letter-to-sound/sound-to-letter system [26, 27].

In the remainder of this paper, I will first reflect on the intellectual context that stimulated the initial development of the TINA and ANGIE systems. The design of the systems will then be motivated based on a set of both theoretical and pragmatic design goals. Following this, significant aspects of first the TINA and then the ANGIE systems will be developed in some detail. After a discussion on some system integration issues, some remaining open research questions will be addressed. Since the scope of this paper is very broad and the space is rather limited, each topic will be addressed at a rather superficial level. A more detailed treatment can be found in the cited literature.

2. A PERSONAL RETROSPECTIVE

It is interesting for me to look back on the intellectual setting at MIT during the '80s, an exciting and inspirational context influencing the design choices of the TINA system, and planting the seeds for the later ANGIE system. The Chomsky and Halle theory of generative phonology had long since been introduced [6], and Dan Kahn had proposed the notions of organizing phonological constraints around syllable structure [18]. A team of researchers led by Jon Allen was developing a sophisticated letter-to-sound generation system called MITalk, based on a decomposition of words into meaning units called morphs [2]. Mark Randolph, a fellow student in the speech group, was parsing words into syllables, with the aim of formally encoding a distinctive-feature formalism [30]. Victor Zue, then a researcher in Ken Stevens' Speech Group, was beginning to codify his acoustic phonetic knowledge and utilizing it in the development of speech recognition systems that made use of ordered context-sensitive phonological rules to expand the lexicon [43]. Ken Church's doctoral thesis [10] proposed applying context-*free* rules[2] to parse syllables, in order to capture phonological effects, arguing that conditions for phonological phenomena could be encoded effectively in category names.

While these activities were going on around me, I was completing a doctorate on auditory modelling for speech recognition [32]. After graduation, I joined the Speech Group as a researcher, and, as a member of a team assembled by Victor Zue, began to become interested in the notion of having the computer actually understand the sentences it was recognizing, in order to perform some useful function. Due to my prior involvement in the ARPA-SUR program, I had some knowledge of research activities in the computational linguistics community, particularly the work of Bill Woods at BBN on Augmented Transition Networks (ATN's) [42].

Meanwhile, Noam Chomsky had by this time abandoned transformational rules as applied to syntactic parsing, and had moved on to the notions of government/binding theory and the "move anything anywhere plus constraint" idea [5]. I paid a lot of attention to Chomsky's work – it was clear that movement phenomena were of paramount importance to him, and I felt that the principle-based parsing of government-binding theory was a vast improvement over transformational rules. I could however also see the computational appeal of ATN's from a pragmatic standpoint[3].

Two other complicating factors for me were the issues of a trainable probabilistic framework demanded of the recognition community, and the experience from my prior work on auditory modelling that gave me insight into what might be computationally feasible in biological systems. Long before, Fred Jelinek had talked to me about his passions for statistical language modelling [17], and by the mid to late '80s the recognition community was converging on word and class n-grams as the language model of choice [3, 29]. The notion of probabilities had not yet crept into the linguistic community, however.

The TINA system [33] was designed and implemented in the late '80s. It is based on the idea of context-free rules plus constraints, but also includes a trigram probability framework with local temporal and spatial dimensions. There is a trace mechanism to handle movement phenomena, and syntactic and semantic features are passed along for unification from node to node in the parsing process. The core design remains intact today, although a myriad little improvements have been introduced through the years, most notably the addition of a robust-parsing capability [34] and a somewhat altered probability framework to increase constraint [35]. The notion of relaxing the constraint that the entire sentence must be accounted for was inspired by the work of Wayne Ward on the Phoenix system [39].

By the early '90s, the Spoken Language Systems Group, headed by Victor Zue, had spun off from Ken Stevens' Speech Group. We invited Sheri Hunnicutt, a key member of the MITalk team, to spend a sabbatical year with us. Together with one of my PhD students, we worked out a parsing scheme that could do letter-to-sound/sound-to-letter conversion reversibly [26, 27]. The rules were context-free, but the probability framework carried a great deal of constraint. The ANGIE system later emerged out of these ideas, where phonological rules were modelled in the same way as letter-to-sound rules. Since then, I have been supervising several other students whose theses have explored different aspects of the ANGIE and/or TINA systems [7, 28, 22].

3. DESIGN CONSIDERATIONS

In developing TINA, and later ANGIE, a set of design conditions were imposed based on the premise of simultaneously providing constraint for the recognizer and a formal specification of the encoded linguistic knowledge. The design was also guided by knowledge of plausible restrictions on the processing capabilities of biological systems. A formal specification of these conditions is as follows:

- The grammar should be characterized by a set of context free rules, which would however be decomposed into nodes in a spatio-temporal field, with communications restricted to nearest neighbors.
- The system must be trainable from a set of automatically parsed data, and should yield a low perplexity[4] when prop-

[2] A context free rule is a rule that rewrites a symbol generally into a sequence of zero or more symbols. A context-*sensitive* rule attaches conditions under which the symbol is permitted to be rewritten.

[3] For an overview of the different approaches to parsing natural language see [1, 19].

[4] Roughly defined as the geometric mean of the number of choices at each terminal advance.

erly trained.

- The framework should be *causal*; in particular, the search should be able to predict the probability of the next event in time, based on both short term and long term history, but not taking into account any information about the future.
- Long distance constraints would be realized via propagation of features among the nodes. The specification of these features should be unidimensional[5].

These design goals have been met for both TINA and ANGIE. TINA parses top-down, mainly because the movement phenomena that are prevalent in wh-query domains would be difficult to implement in a bottom-up context. ANGIE operates bottom-up, which to me was clearly the right choice due to the desire to share low-level structure among similar words in a large-vocabulary recognizer. Thus, for example, "fly," "flies," "flight," "flights," and "flying" all share the first three phonemes in a common partial theory. ANGIE, unlike TINA, does not yet make use of any feature unification, although I think features marking part-of-speech and/or stress would be an interesting augmentation. Both systems can produce a probabilistic score for the next terminal advance (phone in ANGIE, word in TINA), given the preceding context, which makes integration with a recognizer relatively straightforward.

ANGIE and TINA have been developed mainly in the context of limited-domain conversational systems, and the English language has dominated over all other languages. However, TINA has been successfully used for many other languages including Japanese (in conjunction with a researcher from NEC [14]), French (in conjunction with researchers at LIMSI [4]), Spanish [45], Italian [14], Mandarin Chinese [38], and, at Lincoln Laboratory, Korean [40]. Our exposure to these other languages has given us a wider scope for evaluating the design framework, although it has not led to the point where a major design change seemed necessary. The ANGIE system, which parses words into their substructure, is much newer, and has only been applied in recognition in the context of two domains (ATIS [22, 23] and our JUPITER weather domain [9, 15]).

Perhaps the most difficult aspect of designing rules for TINA is to devise a scheme to simultaneously encode both syntax and semantics, while maintaining a conceptually manageable knowledge space. Furthermore, the desire to realize a low perplexity often conflicts with the goal of greater coverage. The situation is far less complicated below the word level, perhaps because the step of forming words from sequences of phonemes presumably occurred much earlier in our evolutionary history than the step of forming sentences from words. The most difficult aspect of developing rules to encode syllable structure is the issue of ambisyllabicity [18]. This phenomenon is a widespread problem for English, since it was derived from a mixture of Germanic and Romance languages. The former tend to have closed syllables (ending in one or more consonants) whereas the latter tend to have open syllables (ending in a vowel). We have found it feasible to adhere to a short set of guiding principles to decide where to place a syllable boundary, as elucidated more fully in Section 5. Many other languages (such as French, Spanish, and Mandarin) are completely unambiguous about their syllable boundaries. In fact, I suspect that, if English were not the dominant language for speech recognition research, a syllable-based approach to speech recognition would have been much more popular than it currently is.

<table>
<tr><td colspan="10">word</td></tr>
<tr><td colspan="3">pre</td><td colspan="3">sroot</td><td colspan="4">uroot</td></tr>
<tr><td>uonset</td><td colspan="2">nuc</td><td>onset</td><td>nuc_lax+</td><td>coda</td><td colspan="2">uonset</td><td colspan="2">nuc</td></tr>
<tr><td>k!</td><td colspan="2">em</td><td>m!</td><td>ih+</td><td>s</td><td colspan="2">sh!</td><td colspan="2">en</td></tr>
<tr><td>c</td><td>o</td><td>m</td><td>m2</td><td>i</td><td>s</td><td>s2</td><td>i</td><td>o</td><td>n</td></tr>
<tr><td colspan="3">com-</td><td colspan="3">mis+</td><td colspan="4">sion</td></tr>
</table>

Figure 1: ANGIE parse tree for the word "commission," with letters as the terminals. An aligned sequence of morphs is shown below the parse tree. *Note:* "!" denotes onset position and "+" marks stress. The second letter in a doubleton is specially tagged (m2, s2).

<table>
<tr><td colspan="10"><i>word</i></td></tr>
<tr><td colspan="4"><i>pre</i></td><td colspan="3">sroot</td><td colspan="3">uroot</td></tr>
<tr><td colspan="2"><i>uonset</i></td><td colspan="2">nuc</td><td>onset</td><td>nuc_lax+</td><td>coda</td><td>uonset</td><td colspan="2">nuc</td></tr>
<tr><td colspan="2"><b>k!</b></td><td colspan="2"><b>em</b></td><td>m!</td><td>ih+</td><td>s</td><td>sh!</td><td colspan="2">en</td></tr>
<tr><td>kcl</td><td><i>k</i></td><td><b>ax</b></td><td>m</td><td>-m</td><td>ih</td><td>sh</td><td>-sh</td><td>ax</td><td>n</td></tr>
<tr><td colspan="4">com-</td><td colspan="3">mis+</td><td colspan="3">sion</td></tr>
</table>

Figure 2: ANGIE parse tree for the word "commission," with phones as the terminals. An aligned sequence of morphs is shown below the parse tree.

I mentioned earlier that TINA and ANGIE converge at the syllable/word layer. We are exploring both of these layers as a possible convergence point for top-down and bottom-up processing, and we have not yet determined which one is more appropriate. It may be that languages that support clear syllable boundaries converge at the syllable layer, whereas languages that exhibit widespread ambisyllabicity converge at the word layer.

3.1. Hierarchical Probability Models

With a hierarchical linguistic framework based on context free rules, it is not immediately apparent how to lay down a trainable probability model that describes the resulting structures. One needs to choose context conditions that are specific enough to be highly constraining, while not so specific that sparse data problems become a critical issue. The problem of characterizing the substructure of words seems much more tractable than the problem of characterizing how words are put together to form sentences. It has been feasible to define a single fully specified matrix for subword structure as shown in the ANGIE parse tree in Figures 1 (letter terminals), and 2 (phone terminals)[6]. This parse tree has four layers below the word representing from top to bottom morphology, syllable structure, phonemes, and phonetic realizations/letters as the terminal units. With only a few alternate choices at each layer, it becomes practical to encode the entire column above the left terminal as the bigram context for the predicted phone/letter on the right. For example, in Figure 2,

[5] Specified at the category level rather than the rule level.

[6] We will revisit these figures in Section 5.

the probability of the terminal [ax][7] is conditioned on the italicized column to the left: $Prob$([ax]|word,pre,uonset,/k!/,[k]). At the present time, columns are built bottom-up based on trigram probabilities conditioned on the child and the immediate left sibling (e.g., $Prob$(/em/|k!,[ax]) highlighted in boldface in Figure 2). The process terminates when the column merges with the left sibling's column into the same parent category.

It is far less obvious how to lay out a grammar specifying syntax and semantics. While it is clear to us that syntax alone is insufficient for our needs, it has also become evident that a semantic grammar that is not laid down on a syntactic base quickly becomes unwieldy. We believe that explicit representation of major syntactic constituents, such as subject, predicate, direct object, and predicate adjective, is an appropriate strategy for the organization of clauses, with major semantic classes such as "flight_event" or "a_location" appearing in the layer just below the syntactic-level node. Prepositional phrases are generally grouped into case-frame like units such as "time_event" or "source." At any point in a parse tree, it is important to try to group possible alternatives into the highest level unit that makes sense in the context. Thus there is a general trend towards more specific categories near the leaves of the tree. Long distance constraints such as number agreement are best realized through feature unification. The parse trees that are produced do not lay out in tidy two-dimensional grids, and so it is not as clear how to organize a probability model around the structure.

In wh-query domains there is a preponderance of sentences with wh-marked constituents that are moved from their underlying position in the clause to the front of the sentence, as in "<What street> is this bank on <trace>?". A trace mechanism to restore the moved constituent to its natural position has benefit in the resulting ability to share a much larger portion of the grammar rules, the reduced perplexity due to explicit accounting of the misplaced noun phrase, and a superior semantic representation for translation and/or database access.

3.2. Current Practices in Language Models for Recognition

Most of the work in speech recognition to date has been focused on the task of correctly producing the sequence of words that were spoken. The notion of characterizing any information beyond the word sequences is usually not treated as part of the explicit goal, although some amount of phonological and semantic knowledge is generally viewed as a necessary adjunct to success. Usually, each word is represented in the lexicon as a sequence of phonemes, and in some systems a phonological rule framework permits the expansion of lexical entries to explicitly account for phonological effects like flapping or devoicing [11, 15, 13, 41]. Typically the rules are precompiled into the lexicon, yielding an expanded lexicon of alternate pronunciations.

For language models above the word level, the usual choice is class n-grams, where words are grouped into semantic classes and each instance of a class member is viewed as representing all words in the class [29]. For instance, the month names, January, February, etc., form a natural class, and every time any one occurs, it is logical to assume that the others would also be appropriate. The goal is to achieve as low a perplexity as possible, and to use the classes mainly to overcome sparse data problems.

Class n-grams have difficulty when logical members are multiword sequences. For example, Boston, San Diego, and Salt Lake City form an obvious class of city_name, but they are written as one, two, and three words, respectively. A simple solution is to introduce the concept of an "underbar word," enhancing the lexicon with such superwords.

Another problem encountered by the n-gram representation is that words can generally be associated with only one class. The English word "to" is thus at issue because as a preposition it forms an obvious class with "from" but as the infinitive marker "to go" it would be highly inappropriate to substitute "from." A part-of-speech tagger could be used to pre-label all instances of "to" before verbs in a training corpus as "to_inf", for example, allowing the other usage of "to" to merge with "from." Such "tricks" of creating underbarred superwords and semantically tagged twins are examples of a very rudimentary linguistic model.

4. LINGUISTIC HIERARCHIES ABOVE THE WORD LEVEL

A TINA grammar can be viewed as a large collection of subworlds, with each subworld defined by a set of rules that share a common left-hand side category. All of the categories appearing on the right-hand side of the rules in a given set are treated analogously to words in a traditional bigram language model, but restricted to a subworld associated with the given left-hand side category. During recognition, the actual probability of the next word is the *product* of conditional probabilities for all the classes traversed in climbing the parse tree from the terminal leaf to the point where the parse tree merges with the branch leading to the word's immediate left sibling. This design yields a causal system with an easily trainable probability base, as was our goal laid out in Section 3.

It is informative to understand TINA's relationship to n-gram language models through a couple of examples. TINA can achieve the same effects achieved by the underbarred words of a class bigram, but without requiring them to be lexicalized. Consider an example consisting of a class "Cal_Cities" containing cities in California. Three of the cities start with the word "San": San Diego, San Francisco, and San Jose. Assume these three words represented 6%, 11%, and 3% respectively of the total instances of Cal_Cities in the training set. Then TINA's grammar would expect the "subworld" Cal_Cities to start with the word "San" with a 20% probability (6 + 11 + 3). "San" would advance to one of three possibilities: "Diego" (30%), "Francisco" (55%), and "Jose" (15%). All of these paths would end (and exit the subworld) with probability 1.0. The net result is probabilistically identical to what would be produced by a class n-gram, with these three city names lexicalized via underbars. The analogy breaks down above the preterminal layer, however, since the Cal_Cities preterminal would be likely to itself occur in several different subworlds, and in each subworld it would have a unique probabilistic characterization, based on its frequency of occurrence after its specified subworld-dependent left siblings.

[7] Following established conventions, we denote phonemes with // and phones with [].

<table>
<tr><td colspan="9">sentence</td></tr>
<tr><td colspan="9">full_parse</td></tr>
<tr><td colspan="2">**q_subject:** *Generator*</td><td colspan="7">**do_question**: *Activator*</td></tr>
<tr><td>which</td><td>meal</td><td>do</td><td colspan="4">subject</td><td colspan="2">predicate</td></tr>
<tr><td></td><td></td><td></td><td colspan="4">flight_event</td><td colspan="2">vp_serve</td></tr>
<tr><td></td><td></td><td></td><td colspan="4">a_flight</td><td>serve</td><td>**meal_object**: *Absorber*</td></tr>
<tr><td></td><td></td><td></td><td>flight</td><td colspan="3">flight_number</td><td></td><td></td></tr>
<tr><td>*what*</td><td>*meals*</td><td>*does*</td><td>*flight*</td><td>*nine*</td><td>*sixty*</td><td>*three*</td><td>*serve*</td><td>*<trace>*</td></tr>
</table>

Figure 4: TINA parse tree for the sentence "<What meals> does flight nine sixty three serve <trace>." The *<trace>* terminal is linked to the initial q_subject via an explicit trace mechanism. See text for details.

<table>
<tr><td colspan="7">sentence</td></tr>
<tr><td colspan="7">full_parse</td></tr>
<tr><td colspan="7">do_question</td></tr>
<tr><td>do</td><td colspan="4">subject</td><td colspan="2">**predicate**</td></tr>
<tr><td></td><td colspan="4">**flight_event**</td><td colspan="2">**vp_serve**</td></tr>
<tr><td></td><td colspan="4">a_flight</td><td>serve</td><td>meal_object</td></tr>
<tr><td></td><td>flight</td><td colspan="3">flight_number</td><td></td><td>meal_type</td></tr>
<tr><td>*does*</td><td>*flight*</td><td>*nine*</td><td>*sixty*</td><td>*three*</td><td>*serve*</td><td>*dinner*</td></tr>
</table>

Figure 3: TINA parse tree for the sentence "Does flight nine sixty three serve dinner?" The highlighted categories are involved in a trigram probability discussed in the text.

In the grammars we have written, TINA's categories are very specific near the leaves of the parse tree, but become increasingly general at higher levels. Near the top the nodes are mostly syntactic in nature, with category labels such as "subject" and "predicate." It is important to explicitly encode syntactic as well as semantic structure, in order to impose additional regularity on the grammar leading to well formed meaning representations, and to help the grammar developer organize a systematic mental model of the structure.

4.1. Example Parse Tree

An example parse tree for TINA is shown in Figure 3 for the sentence, "Does flight nine sixty three serve dinner." TINA's bigrams within parent classes can also be interpreted as trigrams with both a temporal and a spatial component. Within phrasal groups they behave very similarly to a class bigram, but across major syntactic boundaries TINA can capture the appropriate constraint much more effectively by explicitly representing probabilities in the higher layers of the parse tree. Thus, in the example, even a trigram language model would be ineffective at predicting the word "serve" based on the two numerals preceding it. TINA's prediction of "serve" is mostly carried by the prediction of the "vp_serve" category just below the predicate layer. The probability that is measured is the likelihood of a predicate category beginning with the semantic class "vp_serve" conditioned on the left-sibling "flight_event."

4.2. Long Distance Movement

TINA is able to exploit long-distant constraints through the use of a trace mechanism to explicitly model movement. Consider for example the sentence, "<What meals> does flight nine sixty three serve <trace>?" This works via an implicit partnership among three privileged nodes in the parse tree structure, a "generator" (q_subject), an "activator" (do_question), and an "absorber" (meal_object), as shown in Figure 4. The activator passes along to its descendents the generated constituent, and if no absorber picks it up the parse is rejected. The language model predicts the trace marker after "serve" with a probability of 1.0 (having confirmed a semantic match on "food" for the proposed trace). The two example parse trees in Figures 3 and 4 can share the majority of their rules, while still disallowing the inappropriate generalizations "what meals does flight nine sixty three serve dinner?" and "Does flight nine sixty three serve?" Such rule sharing is important to reduce computation and to ameliorate sparse data problems in training.

4.3. Feature Unification

TINA also has a mechanism to enforce syntactic constraints on features such as number (singular, plural) and verb mode (finite, root, past participle, etc.). For instance, in the parse tree shown in Figures 3 and 4, the auxiliary verb "does" sets the mode to be "root." This feature is passed along passively to the main verb, and enforces the selection of "serve" rather than "serves" or "serving." These features not only provide constraint to the recognizer but can also be essential in some cases to disambiguate redundant parse solutions, where alternatives with incorrect feature values would lead to erroneous meaning representations.

4.4. Robust Parsing

In conversational speech, people often violate the strict rules of syntax. Furthermore, even for narrow domains, it is essentially impossible to write a grammar that fully covers all the ways people can ask questions. In our grammars, we generally include mechanisms to cope with parse failure that involve licensing skippable words, and piecing parsed fragments together. The perplexity is generally very high at the seams between fragments and/or skipped words, so it is a mechanism to be used conservatively, if possible. It is also sometimes difficult to infer how to combine the fragments to form a coherent meaning representation. We make use of explicit tables of appropriate noun-attribute relationships to aid in the process of constructing a coherent frame from fragments. The mechanism is viewed as a sentence-internal discourse mechanism, and utilizes procedures that are shared with the normal sentence-to-sentence history mechanism [34].

4.5. Portability Issues

At this point we have developed grammars that support conversational systems in several distinct limited domains: a city guide for Boston and vicinity [14], a flight travel planning and reservations system [44], a weather information system [45], a system for accessing classified ads for used cars [25], and a restaurant guide [37]. These are all clearly very narrow domains, and it is possible that the focus on such restricted systems has led us to solutions that would not generalize to all of English. However, for the foreseeable future, our group will continue to focus on such narrow domains, so these systems have provided examples of the degree of complexity TINA will be required to handle in domains of interest to us. While I cannot yet visualize the possibility of a fully automatic procedure for acquiring new grammars, nonetheless we are generally able to reuse large portions of prior grammars in new domains, particularly as the conceptual view of the grammar structure becomes more stabilized. For example, we can insert entire subgrammars to handle time rules, date rules, and number rules, which recur in many of our domains.

5. LINGUISTIC HIERARCHIES BELOW THE WORD LEVEL

The purpose for building hierarchical structure below the word level is multifold. One main goal is to develop a language model to predict phone sequences of the language without explicit ties to a particular vocabulary. A bottom-up parsing procedure has the important property that it supports significant structure sharing among words that begin with the same phone sequence. If words are further decomposed into syllables, which then form the basic recognition unit, even greater sharing is possible, since words such as "retention" and "contention" can share everything except their prefix in common syllable nodes.

Another important goal is to model phonological rules in a trainable probabilistic framework. The phonological phenomena are captured through simple context-*free* rules, but the probability model allows the system to learn the appropriate context conditions for the rules automatically from aligned corpora.

ANGIE's language model, while restricted to phone-to-phone transitions, is very powerful, and captures generic linguistic knowledge of English while a partial word is under construction. We have determined empirically that, within the ATIS domain, ANGIE is able to achieve a significantly lower perplexity on unseen data than a phone trigram similarly trained [21]. Once a word is completed, higher level language models can be incorporated as well (e.g., syllable/word n-grams and full parse trees).

The substructure that is captured in ANGIE's grammar rules includes morphology, stress, syllable structure, and phonological effects. As in TINA, probabilities are trained automatically from a parsed corpus. However, in the case of ANGIE, the training data are a little more difficult to obtain, since it is not nearly as straightforward to provide a phonetic transcription as it is to provide an orthography. We have used the approach of seeding on phonetic transcriptions provided by automatic alignment of training data using our SUMMIT speech recognizer [12, 15].

The shared probability model is important for generalizing phenomena over similar contexts. Rare words can benefit from observations of common words that have the same local phonetic environment. And words that are completely unknown to the recognizer can be generated with a non-zero probability by following the parse tree fragments of words with localized equivalent patterns. For example, "queen" can be decomposed into the onset of "quick" and the rhyme of "seen."

In ANGIE, we currently represent our lexicon in two tiers – words are entered as sequences of "morphs,"[8] and morphs are in turn entered as sequences of phonemes. The morphs are essentially syllabic units specially marked for spelling and positional constraints. We currently distinguish for English five different possible morph positions: prefix, stressed root, unstressed root, "dsuf" and "isuf"[9]. Context-free rules encode positional constraints for the morph units – for example, unstressed root always follows immediately after stressed root, and isuf's are always terminal.

As mentioned previously, it is often not obvious where to place syllable boundaries in English words. There are many cases of ambisyllabicity, where it is not clear whether the intermediate consonant belongs with the preceding or following syllable. Placement of the boundary can also be influenced by the underlying morphology – when there is a clear inflectional ending we do not attempt to shift the terminal consonant of the root into onset position, even though this would be in accord with a maximal-onset rule. Hence "dancing" becomes "danc ing" rather than "dan cing". Often we introduce a double consonant as a means of implementing explicit ambisyllabicity, which reduces via a gemination rule to a single phonetic realization. Hence, "connect" becomes "con- nect" with two /n/ phonemes at the phonemic layer reducing to one at the phonetic layer. This makes the boundary between the word-internal syllables behave analogously to boundaries between word sequences like "on next" or "seven nine." Such lexicalized geminations are nearly always associated with a spelling that includes a doubleton letter "nn."

5.1. Example Parse Tree

ANGIE's framework supports two sets of terminals with shared parse trees above the terminal layer. The preterminal layer contains the phonemic sequence exactly matched to the entries in the two-tiered lexicon. The terminals are either the letters of the spelling of the word or the phones of the particular spoken realization. Thus letter-to-sound and phonological rules are licensed on the preterminal-to-terminal mappings. The upper layers capture syllabification, morphology, and stress.

Example parse trees in ANGIE for the word "commission" were given in Figure 1 (letter terminals) and 2 (phone terminals). The word decomposes into a prefix (com-) a stressed root (mis+) and an unstressed root (sion). Phonemically, there are both a final /m/ for the prefix and an onset /m!/ for the root. These geminate in the phonetic realization into a single [m]. ([-m] is a code for "deleted in the context of preceding [m]"). Similarly, the "mis+" unit ends phonemically with an /s/. The /s/ is palatalized to a [sh]

[8] This follows roughly the definition given in [2], p. 24, which is a representation of morphological units such as prefix and root that is also tied to the word's spelling.

[9] "dsuf" roughly corresponds to "derivational suffix," and "isuf" to "inflectional suffix," but we are willing to violate strict conventions for pragmatic reasons.

Word Lexicon

commission	com- mis+ sion
mister	mis+ ter
mansion	man+ sion

Morph Lexicon

com-	k! em
man+	m! ae+ n
mis+	m! ih+ s
sion	sh! en
ter	t! er

Figure 5: Selected entries from a word and morph lexicon for ANGIE.

at the phonetic level, with the onset /sh!/ of the "sion" marked as deleted. Figure 5 illustrates how sharing of subword units can be achieved, using the examples "mis+" and "sion."

5.2. Lexicon Creation

ANGIE relies heavily on the availability of a specifically prepared two-tiered lexicon, in which words are represented in terms of their underlying morphs. We have already obtained, through careful hand-editing, a seed lexicon of some 10,000 words, derived from the common words of the Brown corpus [20] augmented with words from some of our conversational domains such as ATIS and JUPITER. We are in the process of converting all the words of Pronlex[10] into ANGIE's lexical format [28]. We are utilizing a semi-automatic process which first parses the letters of each word into a set of hypothesized phonemic alternatives, and then parses the phonetic units as provided by Pronlex into phonemes, constrained by the choices produced by the letter-parsing step.

We hope to use the resulting morph lexicon as a basis for a generic morph-based recognizer for general English. A phonological model would need to be trained on a large corpus such as Wall Street Journal. There would still be some possibility of unseen morphs in new material, but these would likely be covered generatively by the rule base. We also believe the lexicon would be useful for training a letter-to-sound system. Ultimately, we would like to augment it with additional information such as part-of-speech, and perhaps add feature propagation to ANGIE's framework to utilize such features. Of course the automatic procedures are not error-free, so extensive hand correction is required to perfect the lexicon. This work is ongoing.

5.3. Phonological Rule Expression

ANGIE's ability to encode and generalize phonological rules is best illustrated through an example. Consider the parse tree shown in Figure 6 for the word "introduce" pronounced casually as "innerduce." The two special phones [-n] and [-rx] are "deletion" phones, meaning that they occupy no temporal space and have no acoustic model. The deletion category is tied to the preceding phone's identity. The grammar developer would specify that /t!/ can be realized as "[-n]", meaning "/t/ in onset position can be deleted after [n]." The probability model captures the important context conditions – falling stress and following schwa. The deletion of the /ow/ is predicated on the realization of the preceding /r/ as a retroflexed schwa ([rx]).

<table>
<tr><td colspan="9">sentence</td></tr>
<tr><td colspan="9">word</td></tr>
<tr><td colspan="2">sroot</td><td colspan="3">uroot</td><td colspan="4">sroot2</td></tr>
<tr><td>nuc_lax+</td><td>coda</td><td colspan="2">uonset</td><td>nuc</td><td colspan="2">onset</td><td>lnuc+</td><td>lcoda</td></tr>
<tr><td>ih+</td><td>n</td><td>t!</td><td>r</td><td>ow</td><td colspan="2">d!</td><td>uw+</td><td>s</td></tr>
<tr><td>ih</td><td>n</td><td>-n</td><td>rx</td><td>-rx</td><td>dcl</td><td>d</td><td>uw</td><td>s</td></tr>
<tr><td colspan="2">in+</td><td colspan="3">tro</td><td colspan="4">duce+</td></tr>
</table>

Figure 6: ANGIE parse tree for the word "introduce," showing phonological rules expressed in preterminal-to-terminal mappings. The morph sequence is shown below the terminal phones.

sentence			sentence	
word		n→ t!	word	
sroot		↑	uroot	
coda		-n	uonset	
n			t!	
n	→-n		-n	→rx

Figure 7: Schematic of probability model in ANGIE, and its accounting of the context conditions for t-deletion in words such as "introduce."

Figure 7 illustrates the context conditions that are learned, with regard to this t-deletion rule. The column above the [n] encodes coda position in a stressed syllable. It predicts a deletion after [n] with no awareness of which phoneme actually follows. The trigram column-building step decides which phoneme was deleted. Other possibilities would be /t/, /d/, /d!/, and /n/. The training procedure would collapse together the /t/ deletion here with other similar environments, such as "integrate," "cantaloupe," "entertain," "Santa Clause," "hunter," and "pantyhose." The column above the [-n] would learn through training that it is rarely followed by anything other than [ax], [rx], and [ix]. The system would thus learn from examples that the right context must be a schwa, but it could be front, back or retroflexed. This "fact" was not informed by any rule, but rather discovered from observation of training data.

5.4. Duration Modelling

ANGIE's parse trees can provide access to intermediate structures within words, which can be useful for characterizing prosodic information. Thus far we have only attempted to characterize prosody through *timing* measures. However, we have found that significant improvements in both phonetic recognition and word spotting can be gained through the use of relative duration models relating parents to children at all layers of an ANGIE parse tree [7, 8]. The approach is to normalize the duration of each constituent in the parse tree with respect to its particular *children*, and then to measure the portion it occupies of its *parent's* total duration. The procedure propagates to the top of the tree to yield a word-by-word speaking rate parameter,

[10] A pronunciation lexicon for the words in the Comlex lexicon, produced and distributed by the Proteus Project at New York University, under the auspices of the Linguistic Data Consortium (see http://www.ldc.upenn.edu).

which can then be folded back into the phonemic layer to tighten the distributions on absolute phoneme duration. This too leads to improved overall recognition. We believe that this direction of research has many as yet unexplored branches, both in terms of incorporating hierarchies above the word level and in incorporating other prosodic measures such as fundamental frequency and energy. Now that we have a framework that includes both TINA and ANGIE parse trees in an integrated environment, we should be able to begin to explore this rich research area.

6. UNIFYING THE HIERARCHIES

It is at the present time not obvious to me what is the "optimum" design of a recognition system that supports integration of linguistic hierarchies into the recognizer search. There are a number of issues involved, which mostly break down into the question of prioritizing the various constraint application steps. For example, we have determined empirically that hierarchical duration modelling is far more effective when applied late rather than early, presumably because it makes assumptions about word structure that are utilized in its scoring process. Linguistic processing above the word level is computationally expensive, and therefore should probably be delayed until lower level constraints have already eliminated large portions of the search space. The ANGIE linguistic model could quite easily be converted into a finite state network, especially if restricted to syllables or morphs as the highest level recognition units. I believe a promising choice as a first step is to train up the probability space of ANGIE from a large corpus of aligned and parsed phonetic data, and precompile the resulting probability model into a right-branching network of phone sequences representing all syllables/morphs of the domain/language. Subword linguistic probabilities would be associated with each branch. This network could be incorporated into a recognizer to produce a syllable lattice, supported by a syllable n-gram (or perhaps a morph n-gram) as further linguistic constraint.

A promising approach we are exploring currently is to use a syllable recognizer to produce a short N-best list and then use this N-best list as a strong filter on a phonetic lattice [9]. The resulting highly pruned lattice can then be processed through a second stage search where ANGIE and TINA are both included in their entirety, with words as the point of conjunction between bottom-up (ANGIE) and top-down (TINA) processing. Hierarchical duration modelling and any more refined prosodic modelling that we hope to develop for future systems could also be applied at this stage. We believe this approach is feasible in a real-time system, which is a necessary constraint for the conversational systems we are developing.

7. FUTURE CHALLENGES

Through the development of our conversational systems we have become increasingly aware of the need to design recognizers that support multiple domains and flexible vocabulary within each domain. They should also be able to deal with unknown words and false starts that involve partially uttered words. We would like to design the recognizer such that there is a core engine that produces a manageable-sized high quality phonetic lattice independent of the domain and/or vocabulary. This phonetic lattice could then be processed by a suite of domain-dependent recognizers, with the final decision mediated by an informed top-level selection algorithm, that should take into account dialogue context.

We have thus far trained ANGIE's phonological modelling in only two domains: ATIS and JUPITER. As we acquire a broader base of telephone quality speech from users of our evolving conversational systems, we could train ANGIE's phonological models on material covering all of the domains combined. This would permit us to develop a core syllable/morph-based recognizer that would hopefully produce a high quality phone lattice, which could then be processed efficiently in a second stage by multiple domain-specific systems, each integrating with TINA's trained grammar as well and specializing in one of many application domains. We think such a system would allow a user to explore several topics of interest in a single phone call. Thus domain-dependencies would be introduced within a computationally tractable second stage of processing, yielding a more flexible recognition capability than exists in our current systems.

Each domain-dependent recognizer would have domain-specialized versions of a TINA grammar, along with a generic ANGIE grammar that has however been trained on a domain-dependent corpus. The probabilities on higher level nodes in the TINA grammar could be adapted to reflect dialogue context– for example enhancing the probabilities on a "price" category when the system asks for a price range. Furthermore, both the TINA and the ANGIE grammars could be adjusted to account for materials being presented to the user. When the system displays a list of restaurants it could add them to both TINA's and ANGIE's vocabularies, while at the same time adjusting upwards the probabilities associated with restaurant_name. A very preliminary exploration into some of these ideas is discussed in [23].

False starts and unknown words are very challenging aspects of spontaneous speech recognition in limited domains. False starts are fortunately usually prosodically marked; the search space becomes explosive if they are permitted to occur anywhere. ANGIE's models would support an abort part way through word substructure, given a prosodically signalled break, and it is conceivable that TINA's grammar could be used effectively to restrict the possibilities after a false start to be a restart of all partial theories currently under construction.

The ANGIE system, due to its generative model, permits the novel construction of syllables and syllable sequences unknown to its explicit lexicon. Therefore, theoretically, there is no problem with proposing an unknown word bottom-up, although its probability would likely be low. TINA can easily support unknown words in proper noun classes, and it is conceivable that the sentential context would even in some cases lead to the correct class selection. The system, having identified the class, could then query the user for more information – "I'm unaware of this restaurant; could you spell it for me?" The ANGIE system could then combine the spoken letter sequence and the phonetic sequence obtained from the original utterance into a set of plausible spellings that could be matched against restaurant names in available databases. We have not yet attempted to implement these ideas, but much of the infrastructure is currently in place to make it possible to explore them.

Many researchers believe it should be possible for a system to acquire a formal grammar automatically from a large set of example sentences. I have little hope that this is possible in the near future, particularly when there are many ways in which an inappropriate bracketing can yield a pathological meaning representation. However, I do believe it would be feasible to propagate lexical semantic classes (as in George Miller's WordNet [24]) up into a syntactic tree, in order to produce a semantic grammar semi-automatically, and this could be a powerful technique for expediting the process of constructing rich grammars for more unrestricted domains.

8. ACKNOWLEDGEMENTS

The research described in this paper is the result of collaboration with many past and current students and staff of the Spoken Langauge Systems Group at the MIT Laboratory for Computer Science, especially Victor Zue, Jim Glass, Ray Lau, Christine Pao, Joe Polifroni, Grace Chung, Helen Meng, and Aarati Parmar. I especially thank Victor Zue for many stimulating discussions and for carefully reading an early version of the manuscript.

9. REFERENCES

1. J. Allen, *Natural Language Understanding*, The Benjamin/Cummings Publishing Company, Inc., Menlo Park, California, 1987.

2. J. Allen, M. S. Hunnicuttt and D. Klatt, *From Text to Speech: The MITalk System*, Cambridge Studies in Speech and Science Communication, Cambridge Univ. Press, Cambridge, 1987.

3. L. R. Bahl, F. Jelinek, and R. L. Mercer, "A Maximum Likelihood Approach to Continuous Speech Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-5, No. 2, March, 1983.

4. H. Bonneau-Maynard, J. L. Gauvain, D. Goodine, L.F. Lamel, J. Polifroni, and S. Seneff, "A French Version of the MIT-ATIS System: Portability Issues," *Proc. EUROSPEECH '93*, pp. 2059-2062, Sept. 21-23, 1993.

5. N. Chomsky, *Some Concepts and Consequences of the Theory of Government and Binding*, The MIT Press, Cambridge, MA., 1982.

6. N. Chomsky and M. Halle, *The Sound Pattern of English*, New York, NY, Harper & Row, 1968. republished in paperback, Cambridge, MA: MIT Press, 1991.

7. G. Chung, *Hierarchical Duration Modelling for a Speech Recognition System*, S.M. Thesis, MIT Department of Electrical Engineering and Computer Science, 1997.

8. G. Chung and S. Seneff, "Hierarchical Duration Modelling for Speech Recognition using the ANGIE Framework," *Proc. EUROSPEECH '97*, pp. 1475–1478, Rhodes, Greece, September, 1997.

9. G. Chung and S. Seneff, "Improvements in Speech Understanding Accuracy through the Integration of Hierarchical Linguistic, Prosodic, and Phonological Constraints in the Jupiter Domain," *These Proceedings*.

10. K. W. Church, *Phrase-Structure Parsing: A Method for Taking Advantage of Allophonic Constraints*, Ph.D. Thesis, Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, 1983.

11. M. H. Cohen, *Phonological Structures for Speech Recognition*, Ph.D. Dissertation, U. of California, Berkeley, CA., 1989.

12. J. Glass, J. Chang, M. McCandless, "A probabilistic framework for feature-based speech recognition," *Proc. ICSLP '96*, Philadelphia, PA, pp. 2277–2280, October, 1996.

13. J.L. Gauvain, L.F. Lamel, G. Adda, and M. Adda-Decker, "Speaker Independent Continuous Speech Dictation," *Proc. EUROSPEECH '93*, pp.125–128, Berlin, Germany, Sept. 1993.

14. J. Glass, G. Flammia, D. Goodine, M. Phillips, J. Polifroni, S. Sakai, S. Seneff, and V. Zue, "Multilingual Spoken-language Understanding in the MIT VOYAGER System," *Speech Communications*, Vol. 17, No. 1-2, pp. 1–19, Aug., 1995.

15. J. R. Glass and T. J. Hazen, "Telephone-based Conversational Speech Recognition in the Jupiter Domain," *These Proceedings.*

16. R. Jackendoff, *Semantics and Cognition*, MIT Press, 1983.

17. F. Jelinek, Personal Communication.

18. D. Kahn, *Syllable-based Generalizations in English Phonology*, Ph.D. Thesis, Department of Linguistics and Philosophy, MIT, Cambridge, MA, 1976.

19. M. King, ed., *Parsing Natural Language*, Academic Press, New York, New York, 1983.

20. H. Kucera and W. Francis, *Computational Analysis of Present-Day American English*, Brown University Press, 1967.

21. R. Lau and S. Seneff, "Providing Sublexical Constraints for Word Spotting within the ANGIE Framework," *Proc. EUROSPEECH '97*, pp. 263–266, Rhodes, Greece, Sep. 22–25, 1997.

22. R. Lau, *Subword Lexical Modelling for Speech Recognition*, PhD Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, May 1998.

23. R. Lau and S. Seneff, "A Unified System for Sublexical and Linguistic Modelling Using ANGIE and TINA," These proceedings.

24. G. Miller, "Nouns in WordNet: A Lexical Inheritance System," *International Journal of Lexicography*, Vol. 3, No. 4, pp. 245-264, 1990.

25. H. Meng, S. Busayapongchai, J. Glass, D. Goddeau, L. Hetherington, E. Hurley, C. Pao. J. Polifroni, S. Seneff, and V. Zue, "WHEELS: A Conversational System in the Automobile Classifieds Domain," *Proc. ICSLP '96*, Philadelphia, PA, pp. 542-545, October, 1996.

26. H. Meng, *Phonological Parsing for Bi-directional Letter-to-Sound/Sound-to-Letter Generation*, Ph.D. Thesis, Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, June 1995.

27. H. Meng, S. Hunnicutt, S. Seneff, and V. Zue, "Reversible letter-to-sound/sound-to-letter generation based on parsing word morphology," *Speech Communication*, 18, pp. 47-63, 1996.

28. A. D. Parmar, *A Semi-Automatic System for the Syllabification and Stress Assignment of Large Lexicons*, MEng Thesis, Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, June 1997.

29. A. Nadas, "Estimation of Probabilities in the Language Model of the IBM Speech Recognition System," *IEEE Trans. ASSP*, Vol. ASSP-3, pp. 859-861, Aug., 1984.

30. M. A. Randolph, *Syllable-based Constraints on Properties of English Sounds*, Ph.D. Thesis, Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, Sept. 1989.

31. S. Scalise, *Generative Morphology*, Foris Publications, Dordrecht, Netherlands, 1986.

32. S. Seneff, "A Joint Synchrony/Mean-rate Model of Auditory Speech Processing," *Journal of Phonetics*, Vol. 16, pp. 55–76, 1988.

33. S. Seneff, "TINA : A Natural Language System for Spoken Language Applications," *Computational Linguistics*, Vol. 18, No. 1, pp. 61–86, March, 1992.

34. S. Seneff, "Robust Parsing for Spoken Language Systems," *Proc. ICASSP '92*, pp. 189-192, March, 1992.

35. S. Seneff, H. Meng, and V. Zue, "Language Modelling for Recognition and Understanding Using Layered Bigrams," *Proc. ICSLP*, pp. 317–320, Oct. 12-16, 1992.

36. S. Seneff, R. Lau, and H. Meng, "ANGIE: A new framework for speech analysis based on morpho-phonological modelling," *Proc. ICSLP '96*, Philadelphia, PA, vol. 1, pp. 110–113, Oct. 1996. URL http://www.sls.lcs.mit.edu/raylau/icslp96_angie.pdf

37. S. Seneff and J. Polifroni, "A New Restaurant Guide Conversational System: Issues in Rapid Prototyping for Specialized Domains," *Proc. ICSLP '96*, Philadelphia, PA, pp. 665-668, October, 1996.

38. C. Wang, J. Glass, H. Meng, J. Polifroni, S. Seneff, and V. Zue, "YINHE: A Mandarin Chinese Version of the GALAXY System," *Proc. EUROSPEECH '97*, pp. 351–354, Rhodes, Greece, Sep. 22–25, 1997.

39. W. Ward, "Understanding Spontaneous Speech: The Phoenix System," *Proc. ICASSP '91*, pp. 365-367, May, 1991.

40. C. J. Weinstein, Y-S Lee, S. Seneff, D. R. Tummala, B. Carlson, J. T. Lynch, J-T Hwang, and L. C. Kukolich, "Automated English-Korean Translation for Enhanced Coalition Communications," MIT Lincoln Laboratory Journal, Vol. 10, No. 1, 1997.

41. M. Weintraub, H. Murveit, M. Cohen, P. Price, J. Bernstein, G. Baldwin, and D. Bell, "Linguistic Constraints in Hidden Markov Model Based Speech Recognition," *Proc. ICASSP '89*, pp. 699–702, Glasgow, Scotland, May, 1989.

42. W. Woods, "Transition Network Grammars for Natural Language Analysis," *Commun. of the ACM* Vol. 13, pp. 591-606, 1970.

43. V. Zue, "The Use of Phonetic Rules in Automatic Speech Recognition," *Speech Communication* 2, pp. 181-186, 1983.

44. V. Zue, S. Seneff, J. Polifroni, M. Phillips, C. Pao, D. Goodine, D. Goddeau, and J. Glass, "PEGASUS: A Spoken Dialogue Interface for On-Line Air Travel Planning," *Proceedings, International Symposium on Spoken Dialogue*, Waseda University, Tokyo, Japan, Nov. 10-12, 1993.

45. V. Zue, S. Seneff, J. Glass, L. Hetherington, E. Hurley, H. Meng, C. Pao, J. Polifroni, R. Schloming, P. Schmid, "From interface to content: translingual access and delivery of on-line information," *Proc. EUROSPEECH '97*, Rhodes, Greece, pp. 100–200, Sept. 1997.