A Semi-Automatic System for the Syllabification and Stress Assignment of Large Lexicons

by

Aarati D. Parmar

Submitted to the Department of Electrical Engineering and Computer Science
in Partial Fulfillment of the Requirements for the Degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

Massachusetts Institute of Technology

June 1997

Author .....
Department of Electrical Engineering and Computer Science
May 23, 1997

Certified by .....
Dr. Stephanie Seneff
Principal Research Scientist
Thesis Supervisor

Certified by .....
Dr. Helen Meng
Research Scientist
Thesis Co-Supervisor

Accepted by .....
Arthur C. Smith
Chairman, Departmental Committee on Graduate Theses

A Semi-Automatic System for the Syllabification and Stress Assignment of Large Lexicons

by

Aarati D. Parmar

Submitted to the
Department of Electrical Engineering and Computer Science
May 23, 1997
In Partial Fulfillment of the Requirements for the Degree of
Master of Engineering in Electrical Engineering and Computer Science

ABSTRACT

This Master's Thesis concerns research in the automatic analysis of the sub-lexical structure of English words. Sub-lexical structure includes linguistic categories such as syllabification, stress, phonemic representation, phonetics, and spelling. This information could be very useful in all sorts of speech applications, including duration modeling and speech recognition.

ANGIE is a system that can parse words, given either their phonetic or orthographic representation, into a common hierarchical framework with the categories mentioned above. A new feature enforcing morphological constraints has recently been added to this paradigm. We define "morphs" to be somewhat like syllable units of a word, but each of them are tagged morphologically, and associated with both an orthographic sequence and a phonemic representation. Each word is represented as concatenations of these morphs, which then encode both the orthography and the phonemics of the word.

This thesis defines a procedure to semi-automatically derive a sub-lexical representation of new words in terms of these morphs, using ANGIE's hierarchical framework. One distinctive characteristic of this procedure is that both the phonetics and the spelling information are utilized. The procedure is developed using several corpora. When this procedure is used to derive the sub-lexical representations, some words will fail, either because the word is rejected by the hierarchical framework, or a morph needed to transcribe the word is missing. The words that successfully obtain morphological decompositions are used to evaluate the coverage and accuracy of the existing procedure. The words that fail to be represented are a valuable resource because they provide new information about the sub-lexical structure of English. This new information can be incorporated into our procedure to improve its coverage and accuracy.

Thesis Supervisor: Dr. Stephanie Seneff
Title: Principal Research Scientist

Thesis Co-Supervisor: Dr. Helen Meng
Title: Research Scientist

Acknowledgments

Completing this thesis is the culmination of five years of hard work, intellectual expansion, and pure fun which has composed my MIT career. I am much obliged to my thesis advisors, Dr. Stephanie Seneff and Dr. Helen Meng, for their patient explanations, active encouragement, and enormous amount of knowledge they have provided during my tenure at SLS. They have both been a great inspiration, and I plan to follow in their footsteps.

I would like to thank Stephanie specifically for the fun she brings to any endeavor. I am grateful to her for all excitement she instills when commencing any project. I admire her ability to discover hidden regularities in any set of words, and her immensely creative ideas.

I am grateful to Helen for all the support she has given me over the last two years. I thank her for all her sisterly advice, and the enormous amount of knowledge she has bequeathed to me. Not only is she my thesis supervisor, but also a very good friend.

Having a group leader like Dr. Victor Zue has made my tenure at SLS incredibly rewarding. I thank him for all the support he has given, both directly and indirectly.

Raymond Lau has helped in innumerable ways. I want to thank Ray for all of his patient explanations and technical support, and of course, the chocolate.

Many many thanks go to my officemates Mike and Michelle. They have made my tenure here amazingly fun and entertaining. I will miss the great talks Mike and I had at 4AM concerning computer science and biology.

I could not have traveled so far in my life without the help of all of my teachers, starting from kindergarten through grade school, and all the way up to my instructors at MIT. They have imbued me with a wealth of knowledge which is infinitely valuable.

All my friends at MIT deserve much acknowledgment for their intellectual stimulation, and of course fun, we've all had together. I especially want to thank my best friends Neeta, Sal, and Suma, for being so understanding about all the times I could not join them, but had to work instead.

I reserve deepest gratitude for my parents. I could not have even begun my journey were it not for the incredible amount of love, support and dedication they have provided. I also want to acknowledge my siblings, Anjali, Abhishek, and Amy, for all their encouragement and especially entertainment.

This thesis is dedicated to my dear Aristotle.

This research was supported by a research contract from BellSouth Intelliventures.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem Definition

As part of ongoing research in the Spoken Language Systems Group, we are attempting to establish a representation for words in terms of sub-word units. This representation captures knowledge on multiple linguistic levels including morphology, syllabification, stress, phonemics, and graphemics. This new paradigm could potentially be used to more efficiently model words in a language. Certainly the information can be utilized in a variety of different speech applications, hopefully with enhanced performance.

At least two immediate applications of this new representation exist. First, words can be composed from the set of these finite units, much like a function can be composed from a basis set. A speech recognizer could operate with these underlying sub-word units, leading to unlimited vocabulary recognition. Second, this theorized "alphabet" could also be used in letter-to-sound/sound-to-letter generation. Once the correct sequence of these units is found for a word, the phonological information could be directly inferred from these units, or vice-versa.

We propose here a knowledge representation, known as "morphs," that embodies the multiple levels of linguistic information described in the first paragraph[1]. We would like to try to extract these units from an inventory of English words to accumulate a complete set for English. This thesis explores the mechanics of finding these morphs, and tries to discover if these units can be used to represent the majority of English words.

In the next section, we describe prior research that uses portions of the linguistic hierarchy (usually only one of the levels mentioned between phones and words) to improve performance across different tasks and domains. We start with those applications that employ phonemes and syllables, and move upwards to end at morphology.

[1] They are not to be confused with morphemes, which are the smallest linguistic units that still contain meaning.

1.2 Previous Work

Sub-Lexical structure between the phone and word levels has been used in various speech tasks, with promising results in performance improvement. These sub-lexical units include phones, phonemes, syllables, and morphemes. The first two subsections describe results using syllabic information in speech recognition. Following that is a subsection concerning the use of morphemes as basic recognition units.

The above three examples relate the use of only one of the sub-word categories to improve speech recognition performance. The remaining two subsections describe frameworks that integrate all of these levels into a hierarchical structure. The first is the Speech Maker Formalism, which is used to perform text-to-speech synthesis. The second formalism, known as ANGIE, has been used in many speech-related tasks, including letter-to-sound/sound-to-letter generation, duration modeling, word spotting, as well as speech recognition. ANGIE also forms the framework for the research in this thesis.

All of these results suggest that more knowledge between the word and phone level can improve the performance of speech applications.

1.2.1 Incorporating Syllabic Constraint to Model Unknown Words

Syllabification information is used by Kemp and Jusek [5] to construct a word model for unknown words. The motivation behind using syllables is to better cover word fragments that are abundant (up to 50%) in spontaneous speech. An unknown word model consisting of a weighted phoneme graph is computed from syllables from the 359,611 word CELEX dictionary. The JANUS-2 recognizer is tested with this model on 265 utterances from the June 1995 VERBMOBIL test set. This set has 3,823 words, 122 of which are out of vocabulary words. The word accuracy increases slightly, from a baseline of 68.5%, to 68.7%, with a 10.9% unknown word detection rate. The false alarm rate for the unknown words is 13.1%.

The unknown word models are also built from syllables from a 1,987 word corpus, which results in a 68.9% accuracy, an 18.0% detection rate of unknown words, and 28.6% false alarm rate. The improvements in accuracy are considered to be statistically insignificant, but would probably improve if tested on much larger corpora of spontaneous speech. Some encouraging news is that many of the false alarms for unknown words occur where there is a recognition error, so that the model is actually applicable in those cases.

1.2.2 Incorporating Syllable Boundaries in Speech Recognition

Wu et al. [14] use syllabic boundaries to improve speech recognition performance. These boundaries are derived from two different sources. One is from the acoustic signal, and the other is from the

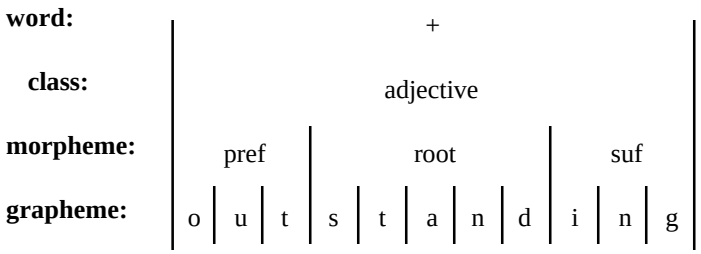

Figure 1-1: *Part of a Speechmaker grid for the word "outstanding."*

forced alignments of transcriptions.

The speech decoder uses syllables as an intermediate level between phones and words. Phones traverse a syllable graph with a bigram model instead of a word graph. The words are extracted from the syllables using a stack decoder and word bigram probabilities. The syllabic onset information is specifically encoded as probabilities into the syllable graph.

When the syllabic information is used, and derived from the transcription of words, the word error decreases by 38%, from a baseline of 10.8% to 6.7%, on a subset of the OGI Numbers corpus. If the information is instead derived from the acoustic signal, the accuracy improves, but it is not quite as significant. The word error rate here drops by 10% to 9.2%.

1.2.3 Incorporating Morphology into Large Vocabulary Speech Recognition Systems

Geutner's [3] motivation behind using morphemes in speech recognition is twofold. First, the language being recognized, German, is a highly inflected language, where new words are created simply by adding short, syllable level affixes. Since nouns can be concatenated indefinitely, there are an uncountable number of compound words. Secondly, the number of morphemes needed to represent a set of utterances is much smaller than the number of words.

Representing basic recognition units as morphemes is then an obvious avenue for exploration. A morpheme based model using the JANUS-2 recognizer has a slightly better word accuracy (65.4%) than a word model (64.7%), when unknown words are allowed in the test set. This news is encouraging, and expected, as smaller syllable-sized models can be used to cover a larger set of words, some of which are unknown. A word model on a closed vocabulary still out-performs both (66.9%), however.

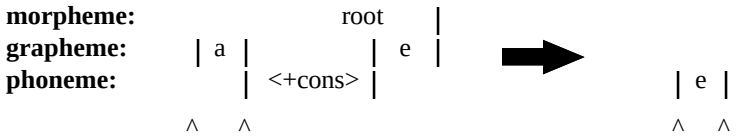

Figure 1-2: *A two-dimensional Speechmaker rule for words like "bathe" and "bake."*

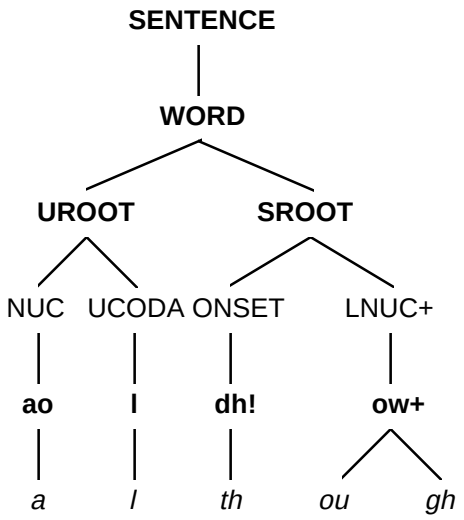

Figure 1-3: *An* ANGIE *parse tree, for the word "although," with letter terminals.*

1.2.4 Speech Maker: A Multi-Layer Formalism for Text-to-Speech Synthesis

The Speech Maker Formalism [13] is used to produce speech from text. It uses a multi-level structure known as a "grid" to capture constraints on different linguistic levels. Figure 1-1 contains a part of a grid for the word "outstanding."

As the grid is two-dimensional, so are the text-to-sound rules. An example of a rule is given in Figure 1-2, to show how the letter "a" is pronounced in "e"-terminal words such as "bathe", and "bake". Clearly the upper layers help to show that the "e" is a terminal vowel, so that the "a" is a long vowel. The carets are used to specify the context of the rule.

The motivation behind the Speech Maker Formalism is to allow linguists to write powerful, compact rules for text-to-speech synthesis. The ability to transform text to sound using information from different linguistic levels is a vast improvement, in contrast to rules that are applied to a one-dimensional string.

1.2.5 ANGIE: A Hierarchical Framework

ANGIE [12] is a system used to parse words into a tree structure with different linguistic levels. Phonemes, syllables, and words are three layers that are included in this framework. A word's spelling (or phonetics) is parsed into this two-dimensional structure, with the help of trained probabilities and some sub-word constraints. An example of an ANGIE parse tree is shown in Figure 1-3 for the word "although."

One of the advantages of ANGIE's framework is that sub-lexical patterns can be "learned," with minimal supervision, by means of trained probabilistic models, in addition to human-engineered knowledge. These models are built using guidance provided by *context-free* rules. Another strength of ANGIE is that either phones *or* letters can be parsed into this hierarchical framework. This means that the orthography and phonology share upper levels of structure. A much more thorough description of ANGIE is given in Chapter 2.

ANGIE has been used in four different speech tasks. They are letter-to-sound/sound-to-letter generation, speech recognition, duration modeling, and word spotting. The results of using ANGIE are described next.

Morpho-phonological Modeling of Letter to Sound Generation

Meng [9] relates how a pre-cursor of the ANGIE framework is used in the task of letter-to-sound and sound-to-letter generation. The phoneme accuracies on a test set is 91.7%, with a word accuracy of 69.3%. The system achieves a letter accuracy of 88.6%, and a word performance of 51.9%, when converting phones to letters. ANGIE [12] obtains slightly better results on sound-to-letter, with a 53.2% word accuracy, and an 89.2% letter accuracy.

A Hierarchical Language Model for Speech Recognition

Preliminary results are offered by Seneff et al. [12], in the context of phone recognition using the SUMMIT segment-based recognizer. Instead of using a phone bigram model, the ANGIE framework is used. A phone error rate of 36% is achieved, compared to the baseline result of 40%.

Hierarchical Duration Modeling for Speech Recognition Using ANGIE

Chung [1] uses ANGIE to determine the duration models for different sub-word segments. Interesting regularities are discovered, such as the fact that the duration of suffixes is affected more than prefixes by speaking rate. Words before pauses are spoken slower. The stressed vowel in pre-pausal words is lengthened the most, as opposed to onset consonants or unstressed vowels.

When this Chung hierarchical duration model is incorporated into a recognizer constrained by ANGIE sub-lexical constraints, phone error rate drops from a baseline of 35% to 33.4%. Adding

implicit lexical knowledge to the framework reduces the baseline error to 29.7%. If phoneme duration scores are also used, the error falls further to 28%.

The duration models are also used to discriminate between confusable city names, such as "New York" and "Newark", in a word-spotting task. As a post-processing stage, all words labeled "New York" are input to a discriminator based on the duration models. The number of confusions is reduced by 65%, from 60 to 21, from a total of 324 occurrences of "New York". Clearly the information derived from these hierarchically defined models is useful.

Using Sub-Lexical Constraints for Word-Spotting

Lau [7] builds a word-spotter on top of the ANGIE framework. Both the keywords and the filler are modeled by ANGIE. In this paper, different types of sub-lexical structures are used to model the filler. The Lau word-spotting FOMs range from 86.3% to 89.3%, with better accuracies resulting from more constrained sub-word filler models. Along with better performance comes an increase in the speed of the computations, probably because more constraint reduces the number of possibilities to explore.

When the hierarchical Chung duration model examined previously is added to the keyword phone models, the performance improves, for all conditions explored, from FOMs of 88.4% to 89.3% for varying sub-word constraint models, to 89.8% to 91.6%. The FOM for the system with the greatest linguistic constraint increases from 89.3% to 91.6% with the addition of this duration model.

1.3 Research Goals

The previous discussions show how linguistic information between the word and phone levels, such as morphology, syllables, and phone context, or an integration of them all, improves performances in many different tasks. By now it should be apparent that this sub-lexical knowledge would be useful for speech applications, including word-spotting, speech recognition, and letter/sound generation.

The primary goal of this thesis is to define a procedure to automatically extract sub-lexical information, in terms of a proposed set of morph units, from words from various corpora. We plan to achieve this by parsing words into the ANGIE framework and extracting information from the parse trees. Then this information can be utilized for the above applications, as well as others. The sub-lexical information can be piped back into the ANGIE framework, to train its probabilities. Better trained models should improve ANGIE's parse coverage and performance. We could also attempt to "homogenize" corpora. If we can consistently transcribe many different corpora using our morph units as a sort of alphabet, then we have ultimately translated many lexicons, with various phone(me) sets, into one large dictionary.

We can take advantage of this process in order to measure how well our morph knowledge

representation can cover a set of words. We define "morphs" as a particular spelling convention representing the syllables of words, which attempt to code the way syllables may be conceptually represented by a human. The set of morphs is different from the set of syllables, which only contain phonetic information. They are also not exactly morphemes, which embody the smallest unit of meaning. Morphs contain stress information and some morphology, such as whether a particular syllable is a prefix, suffix, or root. For example, the next-to-last syllable in "fundamental" and the syllable "meant" sound the same, while their morphs are completely different, not only because of the way that they are spelled ("ment" and "meant"), but also because of the different morphological structure – "-ment" is a suffix, while "meant" is a stressed root.

While the number of syllables is finite, it remains to be seen if the number of morphs can all be listed. It is very likely that prefix and suffix morphs can be wholly enumerated. However, it is unclear whether the set of stressed syllable morphs can be similarly enumerated, or if they grow as new words are encountered.

We plan to determine whether morphs make a closed set by acquiring a set of morphs from one corpus of words, and then observing how well the knowledge can cover another corpus of words. To make the comparison more than fair, we use a much larger lexicon to test the coverage of morphs. If we can show that a small set of morphs can represent many words, or that their growth, as new words appear, increases asymptotically, we have partially fulfilled our purpose of finding a new alphabet to represent words.

This opportunity can also be utilized to explore the difference in accuracies when less human intervention is used to train ANGIE's models. As described in Chapter 2, rules are used to guide the formation of ANGIE's probability models. Usually these rules are hand-written by an expert. A method for automatically inducting a subset of the rules has recently been formulated, and we would like to test its accuracy and coverage.

Along the way of satisfying these goals we do some "exploratory data analysis." We would like to examine how and why particular data sets do not fit into our hierarchical framework with morphs, and if there are smoothing solutions to counteract this problem. These parse failures may provide missing information in our knowledge base, which we can incorporate to extend coverage and accuracy. Evaluating the ANGIE framework itself on a large set of words is also a feat we would like to accomplish.

1.4 Research Plan

The primary goal of this thesis is develop a procedure to extract sub-lexical structure from a large corpus of words, using the ANGIE hierarchical framework. To do this, we parse words into the ANGIE framework, and then extract the sub-lexical information, in the form of morphs, from the parse tree.

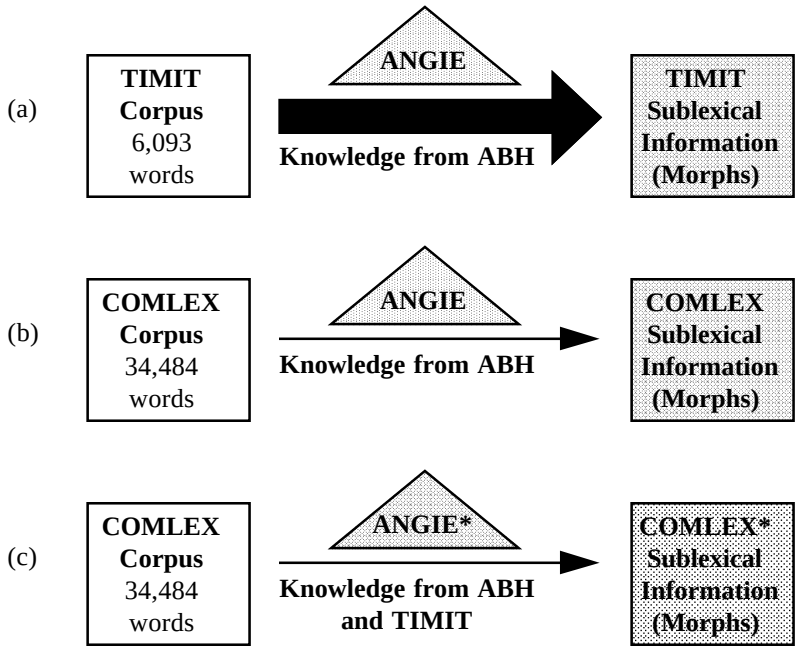

Figure 1-4: *An overview of the research plan, divided into three steps.*

It is possible that certain words will not parse into the framework or a required morph does not exist. These parse failures should provide interesting information about the structure of words in the English language.

6,093 words from the TIMIT corpus [6] with their phoneme realizations will be used to ascertain this procedure. We plan to exploit ANGIE's features so that both the spelling and phonological information can be used to extract the sub-lexical information. Once the basic algorithm is in place, we plan to apply it to a much larger corpus, the 34,484 words from the COMLEX corpus. This will allow us to examine how easily our procedure to extract morphs can be extended, using the methods developed with TIMIT.

Figure 1-4 contains a high-level block diagram of this procedure. In part (a), the TIMIT corpus is used as a pilot corpus to develop our procedure for extracting sub-lexical information. This process is symbolized by the darkened arrow. The sub-lexical information in TIMIT is obtained with the help of ANGIE. In this step, ANGIE's knowledge base is augmented with prior knowledge obtained from a 9,083 word lexicon that we call ABH. ABH is described in section 2.5.

In the next step, (b), we observe how well our procedure applies to the much larger COMLEX corpus, and how well the morph representations we extracted from a set of 9,083 words can be applied to 34,484. This step should also let us know how well the ANGIE framework can accommodate all the variations possible in a large corpus.

Finally, we can add the information acquired from TIMIT, and then try to extract sub-lexical knowledge from COMLEX again (c). The previous step should provide a baseline against which we can compare the amount of knowledge added to ANGIE by TIMIT. From this we can estimate how much new knowledge comes from TIMIT.

Along the way, we can study the data, particularly parse failures, to evaluate the feasibility of generalizing our ANGIE framework with morph constraint to many words. Some of the results should provide insights into the sub-lexical structure of English, as well as the coverage of morphs. These experiments also provide a backdrop against which the performance of automatically generated rules can be compared.

1.5 Chapter Summary

We would like to establish a sub-lexical representation for words which incorporates knowledge on multiple linguistic levels including morphology, syllabification, stress, phonemics, and graphemics. We believe that this more compact representation can improve performance in many speech applications.

"Morphs" are a proposed representation for these ideal units. These morphs can be collected from a database of words, and then analyzed to observe how well they can cover other words.

Some examples where this sub-lexical representation is beneficial are outlined. Syllables are used to construct word models for unknown words, with very slight improvements. They can also be used in speech recognition, as a distinct level between the phone and word levels. Using this information decreases word error rate. Finally, using morphemes in German speech recognition improves word accuracies, compared to words, when unknown words are included.

These previous examples only employ one of the proposed levels to improve performance. The Speech Maker Formalism and ANGIE both integrate all of these levels into one hierarchical structure. These models are aesthetically appealing, since relationships between different linguistic levels are so elegantly and compactly characterized. Rules from the Speech Maker Formalism can describe text-to-speech conversion powerfully. ANGIE encompasses a similar framework, except that transitions are governed by probabilities. ANGIE parses either the spelling or phonetic of a word into its framework, with sharing of higher layers.

The ANGIE framework is used in various tasks. In both letter-to-sound and sound-to-letter tasks, ANGIE is competitive with other models. It also reduces the word error rate for phone recognition when substituted in place of a phone-bigram model. Sub-lexical information can be used to condition phone durations. This information improves phone recognition performance, and can also be used to discriminate between confusable word-pairs. Finally, word-spotting can use the sub-word models produced by ANGIE to improve performance. The more constraining the model, the better the

performance, with higher speed being an additional bonus.

The primary goal of this thesis is to define a procedure to extract sub-lexical information, in the form of our morphs, from large lexicons, using our ANGIE framework. The process of obtaining this information should reveal how well morphs represent words.

We would also like to take advantage of the context of these experiments to do some exploratory data analysis on the words we encounter, particularly those which fail to parse. These failures can be studied for missing knowledge, which can then be incorporated into ANGIE's lexicon. This analysis should also provide an interesting evaluation of the sub-lexical properties of English, as well as any serious limitations of the ANGIE framework. Evaluating less human-engineered ANGIE training models is also a goal.

The research plan consists of three steps. The TIMIT corpus is used as a pilot to develop the procedure for extracting sub-lexical information. The accuracy of these extractions is a metric for the performance of this procedure. Then this procedure can be applied to the larger COMLEX lexicon, both with and without the new knowledge learned from TIMIT. In this way we can measure how well ANGIE, as well as the knowledge gained from TIMIT, can be extended. Data analysis and other goals can be accomplished along the way.

1.6 Thesis Outline

This introductory chapter provides the purpose of this thesis, the motivation, and some reasons why sub-lexical information is beneficial in speech recognition. An outline of the goals of this thesis is provided, along with a research plan. The next chapters each focus on a particular aspect of the plan.

Since ANGIE is used extensively in this thesis, Chapter 2 describes the operation of ANGIE in detail. This should help readers understand details concerning the sub-lexical extraction procedure, as well as the actual structure of our morphs.

Chapter 3 details the TIMIT corpus, and relates the basic procedure developed to extract sub-lexical information from this corpus. It also examines the words which are rejected by the ANGIE framework, and attempts to find some smoothing solutions for these variants. This section provides an opportunity for data analysis as well. An evaluation of the procedure is also included.

In Chapter 4, the same procedure is applied to COMLEX, and the results evaluated. Then, knowledge from TIMIT is added to ANGIE which is then re-applied to COMLEX.

Chapter 5 provides some reflections on the differences between parsing TIMIT and COMLEX. In this chapter, the results of using rules generated from automatically determining phoneme-to-phoneme mappings are compared to those based on manually written rules.

The process of transcribing morphs manually, a feature critical to this thesis, is very complex

and time-consuming. A tool that has been crafted to facilitate the process is described in Chapter 6.

Finally, we end with some conclusions about this thesis in Chapter 7. Some ideas for future study are also included.

Chapter 2

ANGIE

2.1 Motivation

ANGIE [12] is a system used to parse words into a hierarchical framework, based on either orthography or phonology. This hierarchical framework is used to statistically model the linguistic information present in a word. Categories of information include:

- Morphs/Syllables
- Phonemes, including stress information
- Phones/Letters

There are two distinctive characteristics of the ANGIE framework. The first is the framework itself. Multiple linguistic levels are combined into a unified structure. Such a formalism provides a simultaneous analysis of these multiple levels. Different levels of constraint and context are automatically included as well.

The second key idea is that the upper layers (phonemes and above) can be shared between parses of letters and phones. Fitting both the orthographic and phonetic information into the same framework makes the idea of *reversible* generation possible. (Most conventional systems are only capable of generating from letters to sounds, and a few from sounds to letters.) Parse trees generated from letters and phones can be compared, or even intersected, based on the upper layers, to constrain parses, a feature that is central to this thesis.

The combination of these two concepts makes ANGIE a powerful tool for evaluating sub-lexical structures. High level transcription conventions between phones and a spelling in a lexicon can easily be captured. One example is the transcription of the suffix "-tion". Within one lexicon, occurrences of this unit are usually transcribed with a consistent phone or phoneme sequence. An example is given in Figure 2-1, where instances of "tion" are always transcribed using the same cluster of tokens.

Summit:

connections	k ax n eh kd sh ax n z
destination	d eh s t ax n ey sh ax n
intersection	ih n td rx s eh kd sh ax n
restriction	r ax s t r ih kd sh ax n

Comlex:

additions	.xd'IS.Inz
deliberation	d.Il+Ib.xr'eS.In
inhibition	+Inh.Ib'IS.In
interrogations	.Int+Er.xg'eS.In

ABH:

corruption	k! er r! ah+ p sh! en
formulations	f! aor+ m yu l! ey+ sh! en s*pl
participation	p! er t! ih+ s ih p! ey+ sh! en
reflection	r! iy f! l eh+ k sh! en

Figure 2-1: *Words with an internally consistent transcription of "tion", from various lexicons.*

ANGIE has the hierarchical framework to guarantee this high level convention, and the reversible characteristic to determine the mapping between the phone and letter categories.

ANGIE's hierarchy is useful also in the context of letter-to-sound generation. A naive mechanism, such as a string-to-string converter, might transcribe the "sch" in "discharge" to the phonemes **/s k/**, as in "school". Since ANGIE employs higher level constraints, "dis" and "charge" can be recognized as separate morphological units, so that the correct letter-to-phoneme rules are applied. See Figure 2-2 for an ANGIE parse tree of the word "discharge"[1].

A lexical representation in terms of a hierarchical, multi-level structure turns out to be very versatile for speech and language applications. So far the structure provided by ANGIE has been used in various tasks, including letter-to-sound/sound-to-letter generation [10], phone recognition [12], duration modeling [1], and word-spotting [7], as described in the introductory chapter.

The extra linguistic information, all unified into one framework, can be applied to many other applications, including speech recognition, dynamic vocabulary extension, and phoneme-to-phone alignment.

2.2 Basic Operation

ANGIE works by parsing a set of terminals, bottom up, into a hierarchical framework. The allowed transitions between categories are defined by rules with associated probabilities. A sample parse

[1]Note how the sequence "dis" is categorized under the node PRE, or prefix, and "charge" is similarly classified as an SROOT, or stressed root. See Appendix A for a thorough explanation of these categories.

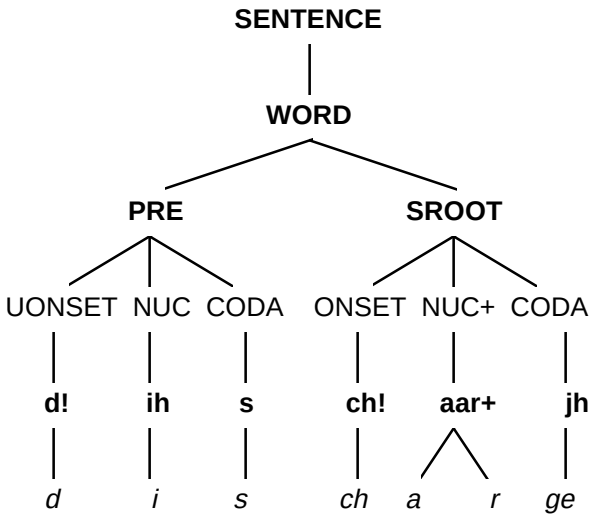

Figure 2-2: ANGIE *letter parse tree for the word "discharge."*

tree with letter terminals is shown in Figure 2-3. Figure 2-4 illustrates a parse tree for the same sentence, with phone terminals.

2.2.1 Parse Tree Structure

These parse trees have six layers. The top root node, SENTENCE, may sprout any number of nodes on the WORD layer. The following layers are, in order: morphological, sub-syllabic, phonemic, and letter/phone. The bottom letter/phone layer is referred to as the terminal layer. For a given word, the upper five layers always use the same categories, but the terminal layer may contain phones, graphemes, or phonemes encoded in the conventions of a particular corpus. A table of the categories for each of the six layers and brief descriptions may be found in Appendix A.

A few distinctions about the phoneme set (on the fifth layer) used by ANGIE should be known. Consonant phonemes are marked with an "!" (as in /**t!**/ in "interested") to refer to a phoneme that must be in onset position. Phonemes that are vowels may be marked with a "+" (as in /**ih+**/) to indicate stress. At present, only two levels of stress are used.

2.2.2 Probabilistic Parsing Algorithm

ANGIE parses either the spelling or phonetics of a word into this structure using a left-to-right, bottom-up algorithm. Allowed transitions are defined by rules and augmented by trained probabilities. A parse begins as follows. The first terminal node (either a letter or a phone) is retrieved from the tokens of the given terminal sequence (which is either a letter or phone string). In the example of Figure 2-4, the terminal node would be [*q*], a glottal stop.

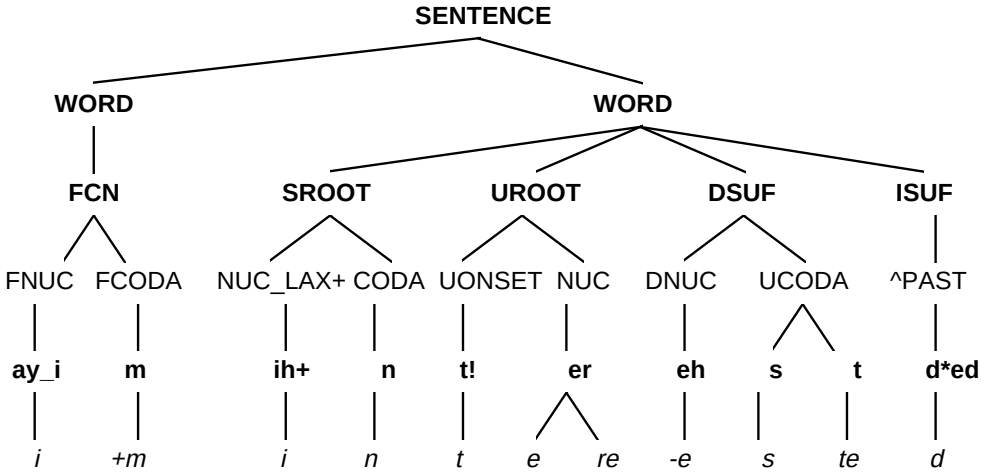

Figure 2-3: *An* ANGIE *parse tree for the the phrase "I'm interested", with letter terminals.*

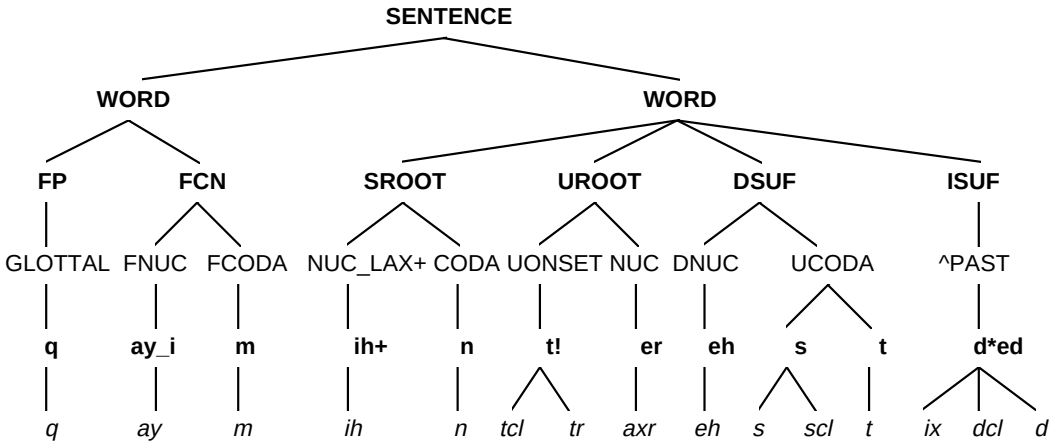

Figure 2-4: *An* ANGIE *parse tree for the phrase "I'm interested", with phone terminals.*

After retrieving this first node, parsing proceeds bottom up. For each node, the next higher node is conditioned on its child (the node below it) and its own sibling (the node immediately to its left on the same layer). Since two other categories are used, the probability is a trigram probability. This bottom up procedure continues until the entire column is built. (A column is defined as the set of nodes along the path from the root SENTENCE node down to the terminal node (in the case of Figure 2-4, [*q*]).) In the example, the next higher node would be the glottal stop phoneme /**q**/. Since it is at the beginning, the left sibling is an implicit START node.

As multiple transitions are possible, there can be multiple columns, also known as parse theories. Each one of these theories will have different transitions, and thus different probabilities. They can be ranked in order of likelihood. In this way, probabilities provide a theoretically sound mechanism for scoring different parse trees. How these probabilities are derived is discussed in subsection 2.2.3.

After the columns are built, parsing then proceeds left to right, beginning with the next terminal node in Figure 2-4, which is [*ay*]. The probability of this terminal ([*ay*]) is conditioned on the preceding column. Probabilities set to zero, or the omission of a rule, disallow terminals to follow certain columns. Then, parsing proceeds bottom up again to produce the next column, and so on, until all of the tokens in the terminal string are incorporated into the framework.

This is the basic algorithm ANGIE uses to parse trees. In order to improve parses in letter mode, ANGIE also considers doubletons of letters as possible terminals. An example is the word "shack." When beginning the parse, not only is the "s" considered as a possible terminal node, but so is "sh." Theories with these two different terminals compete against each other. Figure 2-3 contains two occurrences of doubletons, "re" and "te."

2.2.3 Derivation of Probabilities by Rules

ANGIE uses probabilities to parse a word into its hierarchical framework. These probabilities are derived by generating parse trees and counting the occurrences of the trigrams, as well as the column-to-terminal node transitions. Since no probabilities are available at this point, parse trees are initially generated in this case by using context-free, hand-written rules, that specify transitions between layers. Examples of these rules are shown in Figure 2-5. These rules specify only local constraints, spanning from one layer to the adjacent one.

The format of these rules allows an efficient representation of the licensed transitions from layer to layer. The first rule allows the category SROOT to parse to the syllable structure {ONSET NUC_LAX+ CODA}. The brackets around the category ONSET indicate optionality, so that {NUC_LAX+ CODA} is the other possible sequence. Parentheses around a group indicate an *or* operation, as in the second rule. In this rule, NUC+ may transition to either of the phonemes /**el**+/, /**oy**+/, /**aw**+/, or /**ao**+/. Finally, dashed terminals can be used to specify context. In the case of the third rule, an /**s**/ may go to the letter /$e/, but only if the /$e/ is preceded by the letter

```
sroot → [onset] nuc_lax+ coda

nuc+  → (el+ oy+ aw+ ao+)

s     → $-x $e
```

Figure 2-5: *Selected context-free rules. A "$" indicates the symbol is a terminal category, such as a phone or a letter (in this case, a letter). Brackets indicate optional tokens and parentheses enclose alternates.*

/$x/. A rule like this is appropriate for a word like "axe", which has the phonemes /**ae+ k s**/. A similar context dependency appears in Figure 2-3 for "e", where the phoneme /**eh**/ is allowed to follow, as long as the previous letter is an "e".

In training, counts are tabulated from parse trees that are generated by these rules. These counts are then normalized to become probabilities, which are stored as a trained grammar.

The upper five layers of parse trees have the same structure, regardless of whether the terminals are phones or letters, and hence the rules are also the same. The rules describing transitions for these top layers are named "high level rules," while those describing transitions from the ANGIE phonemes on the fifth layer to the terminals on the sixth are, of course, "low level rules." By separating the rules into these classifications it should be apparent that ANGIE can accommodate any new phone, letter, or phoneme set, just by composing a new set of low level rules, which specifies the allowed transitions between ANGIE phonemes and terminals.

2.3 Parsing Modes

The previous section deals with the basic operation of ANGIE. This next section attempts to explain the plethora of different parsing modes possible.

2.3.1 Letter versus Phone Mode

As mentioned earlier, a word can be parsed into the ANGIE framework based on either its spelling or phonetics. The parsing operation is the same for either mode. The only difference lies in the different low level rules files, and grammars (trained probabilities) that are used. For example, low level letter rules specify the mapping between ANGIE phonemes (on the fifth layer) and graphemes on the sixth layer. The low level phone rules denote the allowed transitions between the ANGIE phonemes and the target phone set.

2.3.2 Train versus Recognition Mode

Both Train and Recognition modes employ either probabilities or rules in order to parse words into the framework. In Train mode, the phoneme sequence of a word is given, and is used to constrain the fifth (pre-terminal) phonemic layer in the word's parse tree. The phonemes contain enough information to almost fully constrain the upper layers. (Some of this constraining information includes stress markings (+), which constrain a phoneme to be categorized under an SROOT, or the onset marking (!) which forces a phoneme to be in ONSET position.) Train mode is used for collecting counts for a trained grammar, since the desired parse trees can be so well specified.

Historically, Recognition mode comes from the fact that phonemes are not available for a word, and must be derived. Recognition mode allows any set of phonemes, provided they are licensed by the rules. Since the phoneme sequence is not required, words that are not in the word-to-phoneme lexicon can be parsed, unlike in Train mode. However, the phonemes can be further constrained, by being tracked by a given lexicon. This means that the only phoneme-to-phoneme transitions allowed are those that already exist in the provided lexicon of words.

2.3.3 Morph Mode

This thesis uses ANGIE's newly added "morph mode" extensively. We define "morphs" as syllable-sized units of a word, with both a phonemic and orthographic representation. Morphs are tagged to belong to one of the nine categories (SROOT, DSUF, PRE, etc.) that are possible on the morphological (third) layer. See Appendix A for a complete listing and description of these categories. Some example morphs are shown in Table 2.1, along with their phonemic transcriptions.

Symbols such as "-", "+" and "*" are used to denote different morphological categories. Upper case letters are used to distinguish morphs with the same letters but different pronunciations, such as **nat+** and **nAt+** in Table 2.1. A few morphs are allowed to have an alternate pronunciation, to reflect subtle differences. Appendix B relates these symbolic tags to the categories, along with an explanation of each category.

The basic motivation behind using morphs in ANGIE is to further capture and represent the structure inherent in words. From experience, humans appear to internally represent words in terms of discrete sub-word units, with a consistent spelling and phonemic transcription. Morphs are also used in ANGIE to reduce computation of parse trees by constraining search – those trees that do not agree with the morph constraints are pruned. Morphs also compactly represent the sub-lexical structure of words, which includes stress and syllabification.

Morphs are specially constructed so that the removal of their tags, and their subsequent concatenation will result in the correct spelling of the word. A simple concatenation of the morphs' phonemes establishes a phonemic representation of the word. Table 2.2 contains a list of words

Table 2.1: *Selected examples of morphs. A "+" indicates a stressed morph. A dash at the beginning signifies a suffix, while one at the end is a prefix. "*" denotes a morph belonging to a function word. A morph beginning with "=" is another type of suffix.*

Morph	Phoneme Representation	Associated Node
-al	**/el/**	DSUF
-ing	**/ing/**	DSUF
=ly	**/l! iy/**	ISUF
Ur	**/y! er/**	UROOT
ca+	**/k! ey+/**	SROOT
fasc+	**/f! ae+ s/**	SROOT
i	**/ih/**	UROOT
nAt+	**/n! ey+ t/**	SROOT
nat+	**/n! ae+ t/**	SROOT
phis+	**/f! ih+ s/**	SROOT
so	**/s! ow/**	UROOT
so-	**/s! ow/**	PRE
that+s*	**/dh! ae t s*pl/**	FCN
ti	**/t! ih/**	UROOT
tion	**/sh! en/**	UROOT

with their morphological decompositions. For example, the word "sophistication" has the morph sequence **so- phis+ ti ca+ tion**, which can be converted by direct table lookup to the phonemes **/s! ow f! ih+ s t! ih k! ey+ sh! en/**.

ANGIE parses can be constrained by morphs. As an ANGIE parse tree is built, both the morphological and phoneme layer are tracked against the list of morphs and the morphs' phonemes. Each time a morphological boundary appears on the third layer (the node changes), the phonemes belonging to that morph node are matched against morphs in a pre-defined morph-phoneme "lexicon." Not only must the morph's phonemes match those in the parse tree, but the category (SROOT, UROOT, etc.) must match as well. Parses that do not have a morph matching the two conditions are rejected. For the example in Figure 2-6, at the boundary after the node PRE, the phonemes **/s! ow/** are looked up in a table like that in Table 2.1. The morphs that match are **so** and **so-**. However, since the category is a prefix (PRE) and not an unstressed root (UROOT), only the morph **so-** is legitimate. This morph lookup continues through the entire parse, so that at the end the morph sequence **so- phis+ ti ca+ tion** is extracted from the parse tree.

Table 2.2: *Selected words from the ABH corpus with their morphological decompositions.*

Word	Morphological Decomposition
sophistication	**so- phis+ ti ca+ tion**
that+s	**that+s***
naturally	**nat+ Ur -al =ly**
fascinating	**fasc+ i nAt+ -ing**

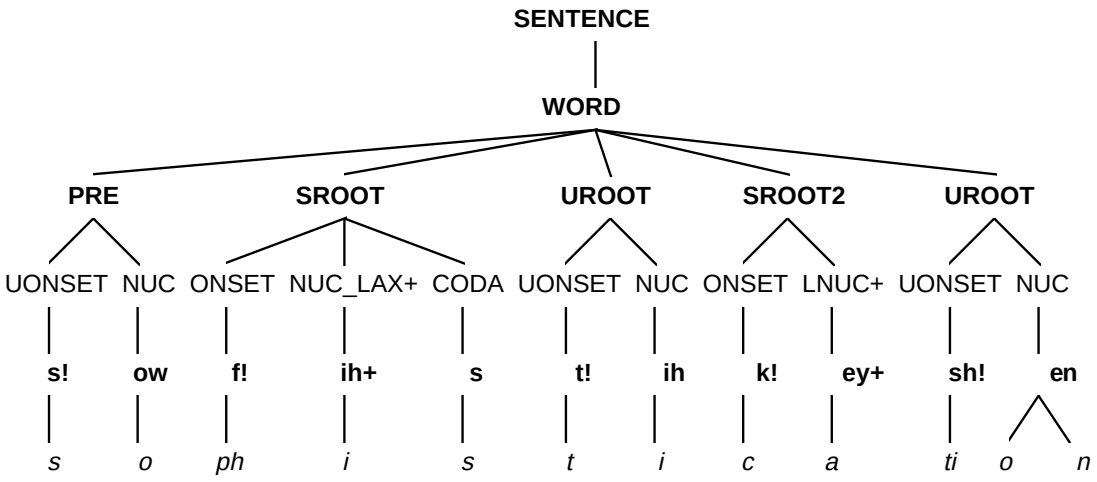

Figure 2-6: *Angie letter parse tree for the word "sophistication."*

Table 2.3: *Selected words from the ABH corpus with their phoneme decompositions.*

Word	Phoneme Decomposition
sophistication	**/s! ow f! ih+ s t! ih k! ey+ sh! en/**
that+s	**/dh! ae t s*pl/**
naturally	**/n! ae+ t y! er el l! iy/**
fascinating	**/f! ae+ s ih n! ey+ t ing/**

In order to better understand the new morph feature, examples of morph decompositions are provided in Table 2.2. Note how the morph sequences in Table 2.2 can be compressed into the spelling, if the symbols and spaces are removed. The morphological decompositions can also be combined with the morph-phoneme dictionary in Table 2.1 to produce phoneme sequences for the words. This information is available in Table 2.3.

2.4 Added Capabilities

To boost performance or reduce computation, other processing has been added to the ANGIE framework.

2.4.1 Meta Rules

Meta rules handle spelling changes in English words, and thus operate only on letter terminal sequences. One example of these spelling changes is when the ending "e" is dropped in "recognize" if the suffix "ing" is added. A meta rule adds the "e" back, to produce "recogniz_eing." This is the

Table 2.4: *Examples of preprocessing accomplished by meta rules.*

Terminal String	After Meta Rules
r e c o g n i z i n g	r e c o g n i z_e i n g
s o p h i s t i c a t i o n	s o p h i s t i c a t i o n

first example in Table 2.4. In the second example, the word "sophistication" has the letters "t" and "i" combined into the unit "ti", for the suffix "tion."

Encoding the terminals improves performance. For example, "z_e" can prevent the "i" from being associated with /**ih**/ by way of the column probabilities. If trained correctly, "z_e" would be deemed more likely to follow a long vowel (such as /**ay**/) rather than a short one (/**ih**/). The "ti" in "sophistication" can be phonemically treated as the unit /**sh!**/ in /**sh! en**/, instead of the two separate phonemes /**t!**/ and /**iy**/.

Of course these simple rules can misfire. ANGIE has an automatic backup mechanism: if the parse fails with the use of meta rules, the terminals are parsed again, this time without the rules.

2.4.2 Pruning

For a given input sequence of terminals, there can be a large number of possible parse theories. It is sometimes impossible to generate every possible theory, due to memory constraints and speed. Hence ANGIE does some pruning to remove unlikely theories.

ANGIE does two types of pruning. The first is a simple cutoff of the number of theories that are kept. As a column is built, any number of theories are allowed to form. But, after every possible column has been built, only the top n are retained for further parsing expansion. In the following experiments, up to forty theories are kept after each column iteration.

The other pruning deals with identical twins. If two partial theories have the same column, the less probable one is deleted. This is possible in the case that there are two theories which end with the same column, but have different previous columns.

2.5 ABH Corpus

The studies in this thesis are based on knowledge, in the form of a trained grammar and a list of allowed morphs, derived from the ABH corpus. The ABH corpus is a collection of 9,083 words extracted from three domains. These domains are the ATIS (flight information) domain, the 10,000 most frequent words in the Brown corpus, and the Harvard List.

The words in the ABH corpus have already been hand parsed into the ANGIE framework using hand-tailored, context-free letter rules. They also have accompanying morphological decompositions,

as exemplified in Table 2.2. A trained grammar for letters has been generated from the top parse trees for these words. A morph-to-phoneme lexicon has also been created for this set. This lexicon contains 5,168 morphs, which compose the morph pronunciations for all 9,083 words in the ABH corpus. The parses and morphological decompositions have been checked by experts and are reasonably accurate.

It should be noted that words can have many, ambiguous morphological decompositions, all of which are correct. The transcribers have attempted to maintain consistency by choosing transcriptions based on those of similar entries. We hope that ANGIE can learn this consistency, and apply it systematically to new words.

2.6 Chapter Summary

ANGIE is a system that parses words into a hierarchical framework, encompassing linguistic categories such as morphology, syllabification, stress, and phonemics. A parse tree can be built given either a sequence of letters or phones for a word. ANGIE's two distinctions are *reversible* letter/sound generation, and sharing of higher linguistic levels in a structural framework.

Words are parsed into the framework, known as a parse tree, using a left-to-right, bottom-up algorithm. Allowed transitions are defined by rules augmented by probabilities.

The probabilities that drive the parse are generated by collecting counts from parse trees and normalizing them to produce probabilities. In this training procedure, parse trees are generated solely under the direction of hand-written rules.

ANGIE parses in letter or phone mode. Additionally, it can parse in Train or Recognition mode. In Train mode, the phonemes of the word are required. Recognition mode does not demand this information but can constrain parses based on phoneme transitions in a provided word-to-phoneme lexicon.

Morphs are used to constrain the parses further. Not only must a word (either phones or letters) fit into the hierarchical framework, but it must also be compatible with a morph sequence. Words can be represented both phonemically and orthographically using these morphs.

Meta rules and pruning both serve to improve ANGIE's performance and reduce computation. Meta rules pre-process the letter sequence to assure more reliable parses. The two methods of pruning include a limit on the number of theories possible after each column advance, and the removal of identical twins.

The ABH corpus is a collection of 9,083 words from ATIS, the Brown corpus, and the Harvard List. A letter grammar has been trained from these words. A 5,168 morph-to-phoneme lexicon contains all the morphs needed to create morphological decompositions for these words. This corpus is used in our experiments to provide a baseline grammar and morph lexicon.

Chapter 3

Experiments with the TIMIT Corpus

3.1 Motivation

This thesis attempts to define a method to accurately extract linguistic information from words in a large lexicon. The TIMIT corpus is used in a pilot experiment to determine this procedure. Once we succeed in formulating a procedure capable of producing quality transcriptions for TIMIT, we can apply this method to a much larger corpus, such as COMLEX.

The TIMIT corpus is used as a development set for two reasons. First, TIMIT is a medium sized corpus, which means that it is large enough to ensure that it has good coverage of different sub-lexical structures, but small enough to be manageable by a human, especially for evaluating accuracies. Secondly, TIMIT is a good candidate for our pilot because it is a "phonetically rich" corpus. Special care has been taken [6] to ensure that it includes a wide variety of phone-to-phone transition. In this respect, TIMIT should be a demanding corpus for ANGIE.

3.2 Goals

The primary goal of these experiments is to obtain sub-lexical structure for all of the words in the TIMIT corpus. This sub-lexical structure is encoded in terms of the morphological decompositions described in section 2.3.3. We would like this sub-lexical information to adhere to our pre-defined conventions as much as possible. However, at times it is somewhat difficult to measure accuracy, since often more than one correct morphological decomposition is possible for a word. The morphs that are deemed "optimal" are those that agree with the experts' conventions, and are the most consistent with other words in the development corpus. Hence a word may have multiple cor-

rect decompositions, such as **mas+ quE rAde+**, **masqu+ er ade+**, and **masqu+ e rade+** for "masquerade."

There are some secondary goals as well. One of them is to extend the coverage of the TIMIT corpus, in terms of our morph conventions. In this process, we acquire additional sub-lexical knowledge. This knowledge can be added to our procedure, so that it is better prepared to transcribing other larger lexicons, such as COMLEX. This knowledge is encoded in two forms. One is in terms of the probabilistic framework that is used by ANGIE. We would like to better train the probabilities, and fill in sparse data gaps, from the sub-lexical information we extract from TIMIT. The other type of knowledge that needs to be acquired is new morphs. We have found 5,168 unique morphs to cover the words in ABH, but they are not sufficient to cover all words in English.

We plan to acquire this new knowledge through the parse failures. Any words that fail to parse into the framework, or do not get morphological decompositions, will highlight the gaps in our knowledge base, due to sparse data, that need to be filled. By filling in these gaps we can augment ANGIE's knowledge base, enabling it to incorporate new sub-lexical structures. In this sense, failures are actually favorable to our cause, in fulfilling our secondary goal of acquiring new knowledge.

We like to describe our other secondary goal as "exploratory data analysis." The morphological information we extract from the TIMIT corpus provides a clear window into the sub-lexical structure of English. Patterns of stress, syllabification, and phonemics should be readily apparent. It is also informative to explore how many different types of morphs are needed to cover a large set of words, and the distribution of these different types.

3.3 Corpus Description

6,093 words from the TIMIT corpus [6] are used in this pilot study to explore the feasibility of automatically incorporating large lexicons into the ANGIE framework, and getting morphological decompositions. We only require as input the spellings of the words in the TIMIT corpus, along with their transcriptions in the TIMIT phoneme set. (Ten of the words have two pronunciations, such as "live", "project", and "read.") A description of the TIMIT phonemes is given in Appendix C.

TIMIT is a phonetically rich corpus. As described in [6], it is designed by researchers from MIT to have a good phonetic coverage of American English. The creators have included as many phonetic pairs as possible, especially those that are rare. The TIMIT corpus is a standard corpus used by many speech recognition researchers, which has been widely used to compare the performance of different systems.

3.4 Procedure

3.4.1 Overview

In order to obtain reliable parse trees, we take advantage of both the orthographic information from the spelling, as well as the phonemic information from the TIMIT phonemes. One important modification is that the TIMIT phonemes are placed in the terminal layer, which is usually occupied by phones. Thus they are treated as phones, or terminals in the ANGIE framework [1]. Thus the term "TIMIT phoneme" and "phone" will be exchanged freely in the rest of this chapter, and a "phonetic parse tree" refers to one with TIMIT phonemes as terminals.

There are many advantages to merging both the orthographic and phonemic sources of information. One is that the number of possible morphological decompositions can be constrained, reducing computational requirements. Accuracy should also improve, since information from two different sources is used. For example, the pronunciation based only on letters for a word like "diagonally" might be phonemically represented as **/d! iy ae+ g en el l! iy/**. Phonetic information would fail this theory immediately, due to the incorrect first **/iy/**. (See Appendix A for a description of these phonemes.)

Phonological information can be augmented by the orthography as well. [*ax l aw1 d*] is the TIMIT phoneme sequence for the word "allowed." The letters, specifically "ed", can be used to suggest that the word is in the past tense, and that the TIMIT phoneme [*d*] should be aligned with the ANGIE past tense phoneme **/d*ed/**[2].

Merging both the phonetic and orthographic information using a hierarchical framework such as ANGIE seems a daunting task, at first. This problem is simplified by using morphs to merge the phonological and orthographic information. ANGIE can not only constrain parse trees to map to some set of morphs in a lexicon, as in morph mode (subsection 2.3.3), but it can also force them to match a particular morph sequence.

This feature of constraining parse trees to match morph sequences can be used to combine phonetic and spelling information, by way of the following two steps.

1. Words are parsed by ANGIE based on the letters, and constrained to match morphs licensed in a morph-phoneme lexicon. Only the top four morph sequences which produce the spelling of the word are retained. The trees are parsed in Recognition mode (see subsection 2.3.2), because no phonemic transcriptions are available at this point.

2. The phones of the words are parsed, and constrained so that the morph sequence associated with the phoneme parse tree matches one of the four morphological decompositions extracted

[1] In this way ANGIE is utilized to produce mappings from ANGIE phonemes to TIMIT phonemes.
[2] The **/d*ed/** phoneme is used to capture the well known rule that the past tense affix only maps to either /d/, /t/, or /əd/

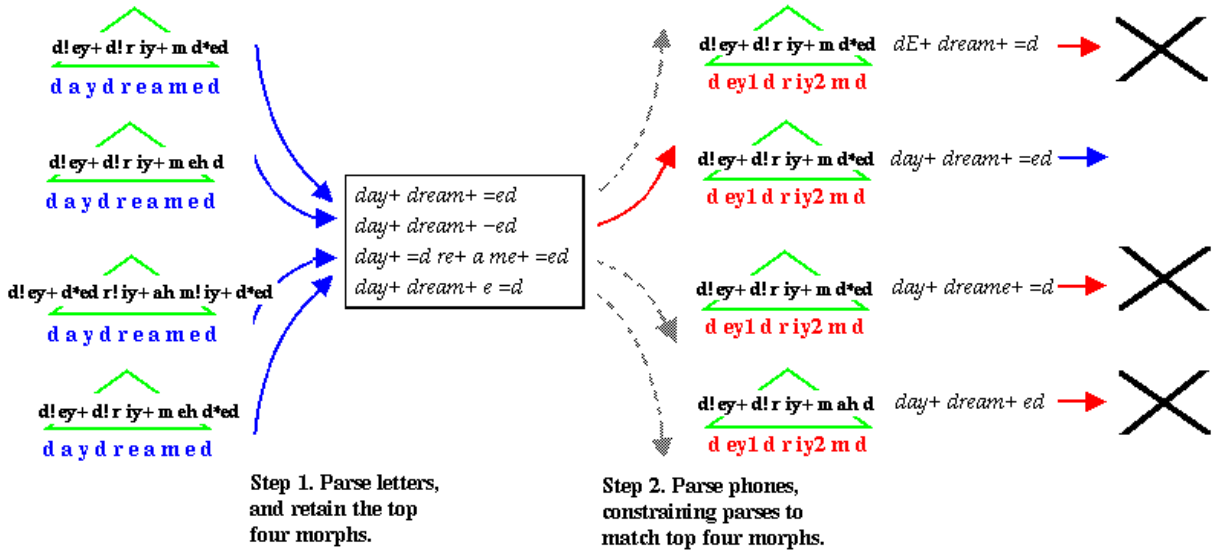

Figure 3-1: *In the first step, the word is letter parsed, and the top four morph decompositions are retained. In the second step, the word's* TIMIT *phonemes are parsed, while being constrained to match one of the top four morph sequences.*

from the letters. Train mode (see subsection 2.3.2) is used to ensure this by forcing the phonemes to match those of the given morphs. Some post-processing is required as well.

Any morph sequence that is generated from letters is guaranteed to correctly represent the spelling of the word. If the morphological sequence is also compatible with the TIMIT phoneme parse, it is likely to have the correct pronunciation as well. Morphs contain enough information to almost completely specify the upper layers of an ANGIE tree. By forcing the TIMIT phoneme and letter parse trees to have the same morphs, the upper layers of both parse trees should be identical.

An example of this procedure is shown in Figure 3-1. On the left side, the word "daydreamed" is parsed by letters, and the top four morphological decompositions (in the box, center) are retained. Then the TIMIT sequence [*d ey1 d r iy2 m d*] is parsed phonetically. These trees are forced to be consistent with the phonemics and morphologies of one of the top four morphs.

In this example, only the second TIMIT phoneme parse tree succeeds in matching one of the four letter morphs. The others would actually have failed at some point during parsing[3], but are included in their entirety for comparison. The first parse tree fails because the extracted morph sequence does not match one of the four given. The third and fourth fail for the same reason, even though the morphs do create the correct spelling in these cases.

The three phonetic parse failures, along with the rejection of the other three letter morphs, show how morph constraint merges information from both orthography and phonology, and rejects incorrect or sub-standard theories. We would like to think that the three failed phonemic trees are

[3] Those with phonemes that do not match one of the four top phoneme sequences would never have been generated.

rejected because they lack information that can be found from the spelling. For example, the morph sequence **dE+ dream+ =d** does not spell the word, even though phonologically it is correct. Even though the phonemes and morphs are consistent with a letter parse, **day+ dreame+ =d** fails, since the morphs' segmentation does not agree with the given four. The last TIMIT phoneme parse does not categorize the ending "ed" as a suffix (preferably an ISUF), but as a UROOT. All of the letter parses acknowledge the ending to be a suffix.

The phonological information also filters out incorrect letter parses. The non-optimal second and fourth letter parses are screened out, along with the bizarre third parse. Note that if the ANGIE phonemes were less stringent and were allowed to transition to more TIMIT phonemes, the second and third letter parse trees might have passed the second step. This shows that is is necessary to have a strict set of phoneme-to-phoneme rules, in order to prevent sub-standard theories from passing.

The order in which this procedure is performed (letters first, and then TIMIT phonemes) does not matter, theoretically. Empirically it is found that parsing with phonological information first and then orthography is less efficient. The desired morph sequences (those that match the phonology *and* spell the word) are often not in the top four morph sequences, and thus either more failures are possible, or more than four morph sequences must be retained.

One might assume that it would be simpler to parse the phonological information first and then just filter morphs to match the spelling, thus eliminating the extra computation from letter parsing. The weakness of this method is that sub-lexical information such as letter-specific endings and syllabification is not utilized. By this method, the suboptimal morph sequences **day+ dreame+ =d** and **day+ dream+ ed** would have passed, in addition to the preferred sequence **day+ dream+ =ed**[4].

3.4.2 Outline

The previous section explains how and why both orthographic and phonological information is used to extract sub-lexical information, using ANGIE. A basic algorithm in the context of ANGIE for merging this information is also provided. This next section outlines the course of extracting this information from TIMIT.

A block diagram of the procedure to extract sub-lexical information from TIMIT is shown in Figure 3-2. A tree showing how the data are divided according to this method is shown in Figure 3-3. In the first block, the 6,093 words in TIMIT are split depending on whether they already have entries in the ABH lexicon. If a word already has an entry in ABH, then it already has a morphological decomposition and we are done for that word. 3,593 words (at node A) overlap with the ABH in

[4] We consider **day+ dream+ ed** suboptimal because the ending "ed" is not recognized as an inflectional suffix (a past tense ending), but rather, a UROOT, which encodes less meaning. **day+ dreame+ =d** does recognize part of the ending as a suffix, but the segmentation is not preferred. Generally, we like to have SROOT morphs correspond to root words whenever possible. The SROOT **dreame+** is neither a common English word, nor a root form.

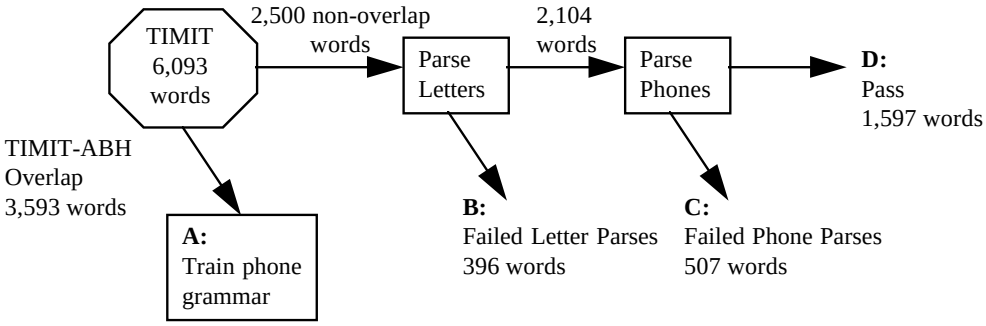

Figure 3-2: *A block diagram of the process of extracting sub-lexical information from* TIMIT *words.*

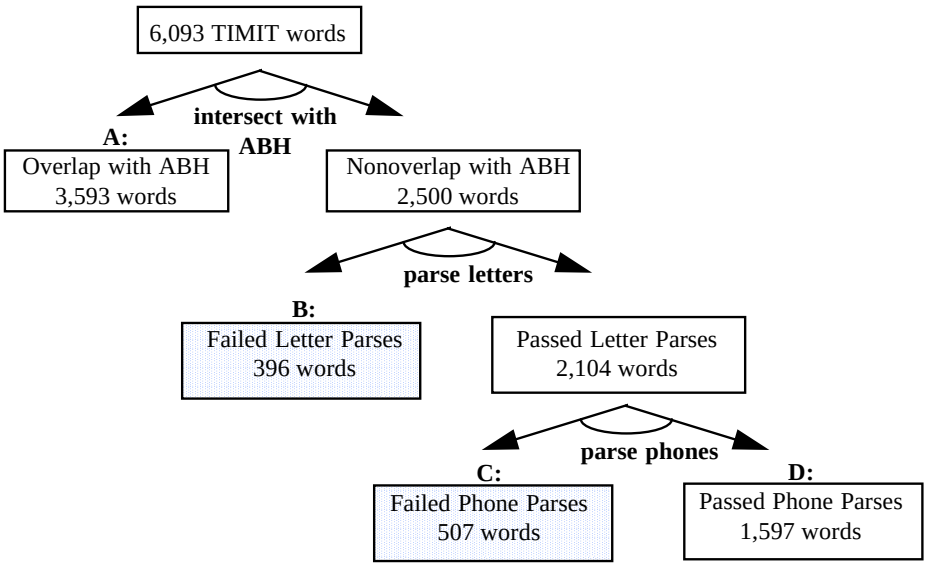

Figure 3-3: *This tree shows how the 6,093 words in* TIMIT *are divided between training data (3,593), failed letter parses (396), failed phonetic parses (507), and passed parses (1,597).*

this fashion, leaving 2,500 as fodder for our procedure. The words at node A are used to train an ANGIE-to-TIMIT phoneme grammar.

These 2,500 words are first parsed by letters, as described in subsection 3.4.1. We must account for some words which are rejected by the ANGIE framework, or do not get morphological decompositions. These failures are denoted by the node B. This set will be further evaluated.

The remaining words which pass the letter parse are then piped to the phonetic parsing unit, denoted by the third block. Again we allow for some words to fail, collecting them at node C. These words should also be useful for further study.

Finally, the resulting words which pass both letter and parsing steps land at node D. This set is evaluated based on the accuracy of the morphological decompositions. The quality of these words will ensure that our algorithm is sound, and that ANGIE is well suited to extract sub-lexical information.

The next section deals with each of these nodes in more detail, from A to D.

3.5 Experiments

3.5.1 TIMIT ABH Overlap

There are 3,593 TIMIT words (see node A in Figures 3-2 and 3-3) that already have been carefully transcribed in the ABH corpus. This overlap set serves two purposes. First it reduces the number of words that still have to be transcribed[5]. More importantly, these overlap words are used to develop the low level phone rules which map ANGIE's phonemes to TIMIT phonemes. Then they are used to train ANGIE's probabilities for ANGIE phoneme to TIMIT phoneme mappings, which are needed in the phone parsing step of our process. This subsection explains how these overlap words, along with those words in the ABH corpus, are used to generate knowledge bases required by ANGIE.

The 3,593 overlap words have both TIMIT transcriptions and ANGIE phoneme transcriptions. They are used to train ANGIE's models for transitions from ANGIE phonemes on the fifth layer to TIMIT phonemes on the sixth. Before a trained grammar can be generated, rules are required to guide the creation of these probabilistic models. The hand-written high level rules, developed from ABH, already exist. Low level ANGIE-to-TIMIT rules are hand written, so that all 3,593 overlap words can parse into the ANGIE framework. Then the counts are collected from these parse trees, normalized, and stored as probabilities, just as described in subsection 2.2.3.

Five different sources of knowledge are needed by ANGIE to parse the orthography of a word into an ANGIE parse tree, and obtain morphological decompositions. There are letter rules, a letter trained grammar, meta rules, a morph-phoneme lexicon, and a word-morph lexicon. These

[5] It is possible that some of these overlap words have an alternate pronunciation that is not transcribed in the ABH corpus, but has this alternate transcription in TIMIT. We do not consider such cases here.

Table 3.1: *Morphological distribution by category, of the 5,168 morphs used to cover the ABH corpus.*

Morph Type	Count	Percentage
dsuf	613	11.9%
fcn	82	1.6%
isuf	44	0.8%
pre	255	4.9%
spre	11	0.2%
sroot[2,3]	3,850	74.5%
uroot	313	6.1%
Total	5,168	100.0%

five knowledge bases are derived from all 9,083 words in the ABH corpus. For completeness, a description and origin of each source is listed next.

The high and low level letter rules are written by hand, so that all 9,083 words in ABH parse. Just as for the phone trained grammar, the 9,083 words are all parsed using these rules, and then the probabilities collected and stored. Meta rules are used to pre-process the spelling, in order to improve the accuracies. These meta rules, developed on the ABH corpus, are crafted by an expert.

The remaining two knowledge bases are needed to implement ANGIE's morph feature. The morph-phoneme lexicon is a list of morphs into which a word may be decomposed. These morphs, with their phoneme realizations, have been largely hand crafted for all 9,083 words in the ABH corpus. There are 5,168 different morphs, with the morph distribution by category tabulated in Table 3.1.

The word-morph lexicon is used to provide additional constraint to the letter parsing. Subsection 2.3.2 relates how, in Recognition mode, parses can be constrained. In this mode, the only phoneme-to-phoneme transitions allowed are those that exist in a word-phoneme lexicon. Since a morph-phoneme lexicon is already available, it is acceptable to provide a word-morph lexicon, and convert the morphs to phonemes by direct lookup.

3.5.2 TIMIT Letter Parses

2,500 words remain that are in TIMIT but not in ABH. These words are all parsed in letter mode, and constrained so that each morphological node (on the third layer) is consistent with a morph in the 5,168 morph-phoneme lexicon, as described in subsection 2.3.3. 396 of these words fail to parse, and 2,104 succeed with at least one morphological sequence. These failure modes show how general and encompassing the ANGIE framework is, and whether our morphs can cover a large set of words.

This section examines the 396 words, denoted by node B in Figure 3-2, in detail, so as to improve ANGIE's knowledge bases. There are many reasons these words could fail, and knowing the exact cause for each is complicated by the fact that many of these errors occur simultaneously.

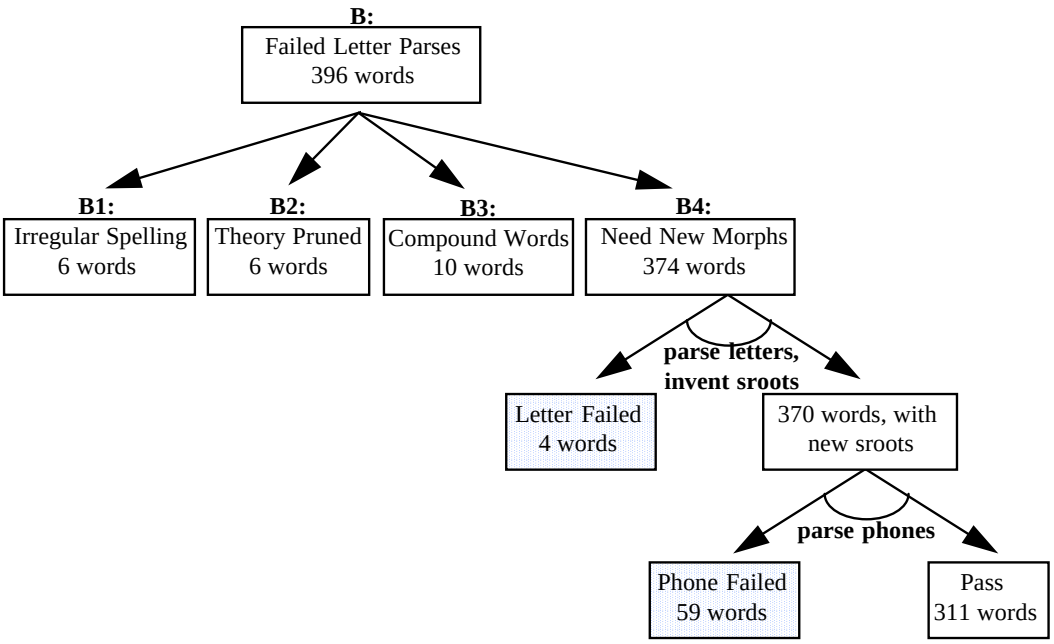

Figure 3-4: *This tree shows how the 396* TIMIT *words which fail the letter parsing step are subdivided into four failure modes.*

The failures can be grouped under four main categories, as listed below and depicted in Figure 3-4. Only the last category's failures are due to any morph constraint; the first three groups stem from the probabilistic framework.

1. The spelling of the word is irregular. (6).
2. The correct theory is pruned. (6).
3. The word is a compound word. (10).
4. To correctly transcribe the word, a new morph, not in the 5,168 morph lexicon, is required. (374)

The only way to tabulate these failure modes is to have the answers at hand for comparison. 390 words (not including the six that are rejected by the framework) have had their morphological decompositions handwritten by an expert. New morphs needed to transcribe these 390 are also added to the morph-phoneme lexicon. A tool used to expedite the procedure of hand-writing morphs is described in Chapter 6.

Irregular Spellings

All of these words (See node B1 in Figure 3-4) do not conform to standard English spelling, and should be rejected. ANGIE rejects these words because the probabilities do not allow a transition,

Table 3.2: *Six words with irregular spellings, rejected by the framework.*

Word
fjords
pneumonia
schnooks
somebody+ll
today+ll
tsunami

Table 3.3: *Six words whose correct theory is pruned.*

Word
interchangeably
oceanographic
photochemical
rearrange
transact
unoccupied

and not because of any morphological constraints. (If these words are parsed in letter mode, without morph constraint, they still fail.)

The six words are listed in Table 3.2. The underlined letter is the position at which the parse fails. The two contractions (a "+" in the word's orthography represents an apostrophe) fail simply because "+ll" contractions are not allowed after two-syllable words in our framework, since they are not real words[6]. The remaining four have odd letter sequences ("fj", "pn", "hn", and "ts") that have not been encountered previously by ANGIE.

Failures Due to Pruning

Six words (node B2 in Figure 3-4) fail because the correct theory is pruned. These words are listed in Table 3.3. When the number of maximum theories is increased to a very large number (1000), the correct theory passes. There are two reasons why the correct theory is pruned. For lengthy words such as the first three in Table 3.3, it is likely that a large number of theories are entertained. In the presence of the many other competing theories, the correct theory falls past the cutoff and is pruned. Another explanation for all six words is that the correct sub-lexical structure is not seen often enough in the training data, so that it is probabilistically less likely, and gets pruned.

Table 3.4: *Ten compound words that fail in letter mode due to sparse training data.*

Word	Correct Morph Sequence	Explanation
beefsteak	**beef+ steak+**	"ea" is not encountered under the SROOT2 category
greenness	**green+ =ness**	No examples of two "n"s across a syllable boundary.
meanness	**mean+ =ness**	No examples of two "n"s across a syllable boundary.
outgrow	**out+ grow+**	"w" not encountered at the end of a SROOT 2
overthrow	**o+ ver throw+**	"w" not encountered at the end of a SROOT 2
overweight	**ov+ er weight+**	No examples of "ei" as a SROOT2
paperweight	**pap+ er weight+**	No examples of "ei" as a SROOT2
rattlesnake	**ratt+ le snake+**	No "sn" at the beginning of an SROOT2
stopwatch	**stop+ watch+**	No "tch" at the beginning of an SROOT2
weatherproof	**weath+ er proof+**	No examples of "f" following "oo", as an SROOT2

Compound Words

These ten words, denoted by node B3 in Figure 3-4, are all compound words, which are usually transcribed with two stressed morphs. The second stressed root (SROOT2) is rarely encountered in the training data, so that it receives zero probabilities for many transitions. This causes the parse to fail.

These words are listed in Table 3.4, along with an explanation of the failure. The underlined sequence is approximately the position at which the failure occurs. Of these ten words, eight fail because of sparse training data involving the second stressed syllable (SROOT2). In anticipation of this problem, the ANGIE parsing algorithm is adapted so that if a parse with an SROOT2 fails, that parse can be attempted again with the SROOT2 category treated as a first stressed syllable (SROOT). The eight words in question then pass with this added back-off. By adding the SROOT2 back-off, we have smoothed ANGIE so that it handles compound words.

Failures due to New Morphs

Finally, the most interesting failures involve the 374 words at node B4 of Figure 3-4, which require new morphs. Since the set of stressed morphs might be limitless, it would be helpful if ANGIE can parse words, and not require SROOT morphs to be licensed in the lexicon. This option of allowing all stressed roots is available only in letter mode, where the invented morphs can be created from the parse tree. The letters under each morphological node in the third layer can be grouped to form a morph. Then the phoneme translation for each morph can be read off the tree. Finally, the morphological tag can be extracted from the category on the third layer.

The next experiments try to automatically invent new SROOT morphs so that these 374 words may parse. The 374 words are parsed in letter mode, and allowed to invent new SROOTs, and

[6] A special set of contractions such as "that+ll", "you+re", and "would+ve" are allowed and treated as function words.

Table 3.5: *Fifteen words that fail because the correct morph sequence is incompatible with the letter, phone, or high level rules. The missed alignments are underlined.*

Failure Due to High Level Rules		
Word	ANGIE Phonemes	Explanation
bulged	**/b! ah+ l jh d*ed/**	**/l jh/** is not allowed to end a syllable
thwarted	**/th! w aor+ t d*ed/**	**/th! w/** is not allowed to begin a syllable

Failure Due to Letter Rules		
Word	ANGIE Phonemes	TIMIT Phonemes
bivouac	**/b! ih+ v w! ae k/**	*[b ih1 v w ae2 k]*
couldn+t	**/k! uh+ d en t/**	*[k uh1 d en t]*
diarrhoea	**/d! ay+ er r! iy+ ah/**	*[d ay2 axr iy1 ax]*
divorcee	**/d! ih v! aor+ s ey/**	*[d ax v ao2 r s ey1]*
drought	**/d! r aw+ t/**	*[d r aw1 t]*
jeopardize	**/jh! eh+ p er d! ay+ z/**	*[jh eh1 p axr d ay z]*
leopards	**/l! eh+ p er d s*pl/**	*[l eh1 p axr d z]*

Failure Due to ANGIE-to-TIMIT Rules		
Word	ANGIE Phonemes	TIMIT Phonemes
acquiescence	**/ae+ k k! w iy eh+ s en s/**	*[ae2 k w iy eh1 s ix n t s]*
boomerang	**/b! uw+ m eh r! ae+ ng/**	*[b uw1 m axr ae2 ng]*
giraffes	**/jh! ih r! ae+ f s*pl/**	*[jh axr ae1 f s]*
kayak	**/k! ay+ y! ae+ k/**	*[k ay1 ae2 k]*
scowled	**/s! k aw+ l d*ed/**	*[s k aw1 el d]*
tyranny	**/t! ih+ r en n! iy/**	*[t ih1 r ae n iy]*

then they are further constrained by parsing again with the TIMIT phones. Four words fail to parse in letter mode, and an additional 59 fail when the phones are parsed. Two of the four words ("bootleggers" and "butterscotch") still fail because the training data for SROOT2 is sparse. The other two, "sheriff" and "sheriff+s", actually require a new DSUF morph, namely **-iff**.

We would like to provide some insight into why the words fail to parse in phone mode. We try parsing the expert transcriptions of these 59 words using the same rules, grammar, and lexicons, except this time we force ANGIE to match the expert morph sequence. Then we can discover why the words fail.

One word, "cloverleaf" is found to be incorrectly transcribed in TIMIT phonemes, into *[ao l ow1 v axr l iy2 f]*[7]. A total of fifteen words would have been rejected by the framework if they had the morph transcriptions given by the expert. These words, along with an explanation for failure, are given in Table 3.5.

Of the remaining 43 words from the 59, ten need new DSUF morphs, four new UROOTs, and another four new PRE morphs. The remaining 25 words require new SROOT morphs. These words

[7] The source of this error is related to the one-character keyboard mapping to the phoneme set used by the TIMIT transcribers. The TIMIT phoneme *[ao]* was represented as "c."

Table 3.6: *Tabulation of* ANGIE *derived morphological decompositions from 311* TIMIT *words, with invented* SROOT*s, compared to morphs transcribed by an expert.*

Words	Percentage	Category
166	53.4%	Have one morph transcription which is identical to hand transcribed
90	28.9%	The most likely of multiple morph transcriptions is identical to hand transcribed
21	6.8%	One of the multiple morph transcriptions is identical to hand transcribed
6	1.9%	The segmentations of the transcriptions are the same
28	9.0%	Do not match the hand transcriptions, or their segmentations
311	100.0%	Total

Table 3.7: *Tabulation of the phonemes from the top* ANGIE *theory from 311* TIMIT *words, with invented* SROOT*s, compared to phonemes transcribed by an expert.*

Words	Percentage	Category
269	86.5%	The phonemic transcription is identical to hand transcribed
19	6.1%	The phonemic transcriptions, without the onset and stress markers ("!" and "+") are identical
23	7.4%	Do not match the phonemic transcriptions, even without "!", and "+"
311	100.0%	Total

should have been able to invent their required SROOT morph and parse. They probably failed through a combination of errors, including having the correct theory be pruned.

It would be informative to know how accurate the morph sequences are for the remaining 311 words which are allowed to have invented SROOTs, and pass the phonetic parse. This is possible since the hand-transcribed morphs and phonemes for the words are available. It is important to check the phonemic transcriptions as well as the morph sequences, since morphs as well as their *phonemic* sequences are allowed to be invented. This means that the morphological sequences for the word may be identical, but if the morphs' phonemic transcriptions are different, the word's phonemes are as well. Or, it is possible that the phonemes are the same, while the morphs are slightly different.

Table 3.6 shows the comparisons by morphs. The 28 morphs that do not match the hand transcriptions, along with the six that only match by segmentation are remarkably close to the expert transcription. Most of the differences lie in morph markers, and the lack of "_e" to denote long vowels. Also, when a new morph is needed to transcribe a word, the transcriber and ANGIE might not use the same label to represent the same morph. For example, a morph may not be capitalized (our method for differentiating between morphs), but still be the correct morph, such as "nat+" versus "nAt+."

Table 3.8: *Composition of 357 morphs that are needed to parse the 311 letter failed* TIMIT *words.*

Category	Morphs	Percentage
DSUF	19	5.3%
PRE	7	2.0%
SROOT	326	91.3%
UROOT	5	1.4%
Total	357	100.0%

It is important to realize that the hand-transcribed morphs contain higher level information that ANGIE can not be expected to decipher. This information includes the "_e", and function word markings. Furthermore, even among the experts there are some inconsistencies about transcription, exemplified by the examples **in+ o va+ tion** or **in- o va+ tion**.

The phoneme transcriptions of the 311 words are compared in Table 3.7. Only one ANGIE generated phoneme sequence is compared against the hand-transcriptions per word. This sequence is extracted from the most likely phonetic parse theory[8]. The 23 words whose phonemics do not match the hand transcriptions, even when the onset and stress markings are removed, are still remarkably close to the expert phonemes. All of them are acceptable as phonemic representations.

It would also be interesting to know the types of new morphs that need to be added. The information about the 357 new morphs needed to cover the set of 374 words that require them is shown in Table 3.8. It was assumed earlier that most of the words are SROOTs, which is true.

Based on these results, we can be fairly confident that the resulting morphological and accompanying phonemic transcriptions for the 1,597 words are accurate. These results might provide a lower bound, since these morphs are invented, without the benefits of ANGIE's constraint.

3.5.3 TIMIT Phonetic Parses

After parsing with letters to get morphs, the phonemic transcriptions are parsed, and forced to match one of the top four morph sequences derived from the letters. 2,104 words pass the letter parsing step, with morphological decompositions. From this set, 1,597 words, or 1,598 pronunciations, pass the phonetic parsing, while 507 words fail. This 507 word failure set corresponds with node C in Figure 3-2, and is analyzed in detail in this subsection. The 1,597 words are analyzed in the next subsection.

Because of the uncertainty of the letter made morphs, it is difficult to know under what category these failures fall. A missing ANGIE phoneme-to-TIMIT phoneme rule can only be detected if the morphs are sure to be correct, which is not the case here. A preliminary perusal of these 507 parse

[8] We could obtain the phonemes by looking up the morphs phonemic realization. However, this is inefficient because one morph can have multiple phoneme realizations, and there is no easy way to know which is more likely.

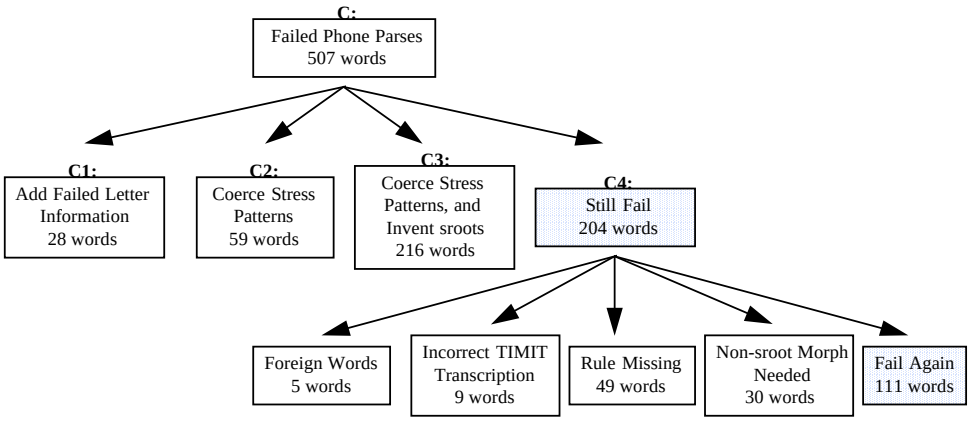

Figure 3-5: *This tree shows how the 507* TIMIT *words which fail the phone parsing are subdivided into three different failure modes.*

failures indicates that many of the words are missing new morphs. Another common error is that many of the transcriptions derived from letters are stress shifted.

We have added a feature similar to the invented SROOT property to make our procedure more robust. Many of the letter morphs of the 507 parse failures begin on the wrong stress, which shifts the stress pattern and throws off the entire pronunciation. Our solution to this is to force a letter parse to have the first syllable stressed, and retrieve the top four morphs. The top four morphs with the first syllable unstressed are also collected. Then the phonetics of the word are parsed, with the constraint of these eight top morphs. Because of the structure of ANGIE, this stress coercion feature is straightforward to implement.

We use a combination of this stress coercion feature, the knowledge we have gained from the 374 words that failed in the letter parse, and invented sroots to try to extract our morphs. We try this in three cumulative steps, where words that pass are set aside and failures are piped to the next step. In this way we try to loosen constraint gradually, allowing words that only need the extra information to pass. The figure in parentheses is the numbers which pass. A tree in Figure 3-5 illustrates the division of these words.

1. Parse again, with the knowledge derived from the 374 letter failed words. (28).

2. Force different syllable stress patterns in the letter parsing step to extend coverage. (59).

3. Force syllable stress, and also allow new sroots. (216)

The remaining 204 fail. The remaining sections deal with each of these four groups.

Table 3.9: *Tabulation of* ANGIE *derived morphological decompositions from 28* TIMIT *words, with information learned from letter failures, compared to morphs transcribed by an expert.*

Words	Percentage	Category
5	17.8%	Have one morph transcription which is identical to hand transcribed
6	21.4%	The most likely of multiple morph transcriptions is identical to hand transcribed
1	3.6%	One of the multiple morph transcriptions is identical to hand transcribed
0	0.0%	The segmentations of the transcriptions are the same
16	57.1%	Do not match the hand transcriptions, or their segmentations
28	100.0%	Total

Table 3.10: *Tabulation of the phonemes from the top* ANGIE *theory from 28* TIMIT *words, with information learned from letter failures, compared to phonemes transcribed by an expert.*

Words	Category	
12	42.8%	The phonemic transcription is identical to hand transcribed
5	17.8%	The phonemic transcriptions, without the onset and stress markers ("!" and "+") are identical
11	39.3%	Do not match the phonemic transcriptions, even without "!", and "+"
28	100.0%	Total

Parsing with information from the Failed Letter Parses

The 507 words are parsed again, this time with information derived from the 396 parse failures. This information includes 374 new morphs gleaned from the failed letters. In addition, the most likely morph decomposition of the 1,597 words that pass both steps is added to the word-morph lexicon, along with those of the 390 hand transcribed words from the 396 that failed[9]. Also, the letter, TIMIT phoneme, and high level rules are expanded to allow transitions that are necessary for the 390 words to parse correctly. With this extra information, 28 words (See node C1 in Figure 3-5) parse. Table 3.9 analyzes the morphs, compared to transcriptions written by an expert. The phoneme comparisons are shown in Table 3.10.

These results are not as good. One of the reasons is that nine of the non-matching sixteen words actually need a new morph. (Seven need SROOTs.) ANGIE gets around parsing these words without the recommended morph by segmenting the words a bit differently, which gives morphological decompositions that are not wrong, but not favored by expert transcribers. The phoneme comparisons are more heartening, for the eleven words which are different vary in small ways that are still acceptable,

[9] The words that are not included are "bleu", "cloverleaf", "fjords", "somebody+ll", "today+ll", and "tsunami." "cloverleaf" is discarded because its TIMIT pronunciation was incorrect. The other five are thrown out because they are not considered to be correctly formed, English words.

Table 3.11: *Tabulation of* ANGIE *derived morphological decompositions from 59* TIMIT *words, with information learned from letter failures, as well as robust stress coercion, compared to morphs transcribed by an expert.*

Words	Percentage	Category
19	32.2%	Have one morph transcription which is identical to hand transcribed
10	16.9%	The most likely of multiple morph transcriptions is identical to hand transcribed
2	3.4%	One of the multiple morph transcriptions is identical to hand transcribed
2	3.4%	The segmentations of the transcriptions are the same
26	44.1%	Do not match the hand transcriptions, or their segmentations
59	100.0%	Total

Table 3.12: *Tabulation of the phonemes from the top* ANGIE *theory from 59* TIMIT *words, with information learned from letter failures as well as robust stress coercion, compared to phonemes transcribed by an expert.*

Words	Percentage	Category
31	52.5%	The phonemic transcription is identical to hand transcribed
10	16.9%	The phonemic transcriptions, without the onset and stress markers ("!" and "+") are identical
18	30.5%	Do not match the phonemic transcriptions, even without "!", and "+"
59	100.0%	Total

such as /**ay d! iy+ ah s*pl**/ and /**ay d! iy+ ah s**/ for "ideas", or /**aw+ r s! el+ v s*pl**/ versus /**aw+ er s! el+ v s*pl**/ for "ourselves."

Parsing with Coerced Stress Patterns

As mentioned before, one stress sequence is often favored over the desired pattern, in the 507 letter failed words. Furthermore, the pattern depends entirely on the stress of the first syllable. Hence we cover all bases by forcing both patterns. We extract the top four morphs with the first syllable stressed, and another four with the syllable unstressed, for the 479 which fail in the previous experiment. Then these eight morph sequences are parsed again with phones. The knowledge used in the previous step is also used. 59 more transcriptions are retrieved with this process, depicted by node C2 in Figure 3-5. The usual morph and phoneme tables are included in 3.11 and 3.12.

The quality of the 28 words from the 59 which do not match the transcribed versions, or only have equal segmentations, vary. A common difference is that often a prefix is proposed, instead of an initial stressed syllable, or vice versa, as for **pI+ an+ O** versus **pI- an+ O**. The other common disagreement is on syllable boundaries at ambisyllabic consonants, such as **reS+ o lute+**

Table 3.13: *Tabulation of* ANGIE *derived morphological decompositions from 216* TIMIT *words, with information from letter failures, stress coercion, and invented* SROOT*s, compared to morphs transcribed by an expert.*

Words	Percentage	Category
66	30.6%	Have one morph transcription which is identical to hand transcribed
32	14.8%	The most likely of multiple morph transcriptions is identical to hand transcribed
15	6.9%	One of the multiple morph transcriptions is identical to hand transcribed
6	2.8%	The segmentations of the transcriptions are the same
97	44.9%	Do not match the hand transcriptions, or their segmentations
216	100.0%	Total

Table 3.14: *Tabulation of the phonemes from the top* ANGIE *theory from 216* TIMIT *words, with information from letter failures, stress coercion, and invented* SROOT*s, compared to phonemes transcribed by an expert.*

Words	Percentage	Category
117	54.2%	The phonemic transcription is identical to hand transcribed
31	14.4%	The phonemic transcriptions, without the onset and stress markers ("!" and "+") are identical
68	31.4%	Do not match the phonemic transcriptions, even without "!", and "+"
216	100.0%	Total

and **res+ ol -ute**. Most of these transcriptions are somewhat acceptable. Fourteen of these words actually require new morphs, according to the expert transcription.

Parsing with Coerced Stress Patterns, and Invented SROOTs

420 words still remain that do not get morphs. We add extra robustness by allowing them to parse with invented SROOTs, along with the coerced stress, and the letter knowledge that is used in node C1. 216 words (node C3) get morphs in this fashion, while 204 (node C4) fail. The morph and phoneme comparisons are listed in Tables 3.13 and 3.14.

The 97 morphs which do not match the expert transcriptions are also reasonable. The most common errors include a missing "_e" as in **hav_e+ =ing**, which requires some lexical knowledge which ANGIE does not have. Other differences include disagreements over syllable boundaries (the expert **in- gre+ dI -ent** versus ANGIE's **in- gred+ -ient**). With invented SROOTs, there is also the possibility of finding a new morph that has the same spelling as an existing morph, but a different pronunciation. Then, the morph is counted as incorrect, as in **E- rot+ -ic** and **e- rot+ -ic**. One final observation is that 24 of these words need a new DSUF, PRE, and UROOT in addition to a new

Table 3.15: *A list of six improper or non English* TIMIT *words that are "masquerading" behind known morphs in letter mode.*

Word	Morphs	Phonemes
bayou	**bay+ ou**	**/b! ey+ ow/**
	bay+ ou+	**/b! ey+ aw+/**
bourgeois	**bour+ ge+ O -is**	**/b! er+ jh! iy+ ow ih s/**
	bour+ ge+ O iS+	**/b! er+ jh! iy+ ow ih+ z/**
chablis	**cha+ bli =s**	**/k! ey+ b! l iy s*pl/**
connoisseur	**conn+ O is+ -se ur+**	**/k! aa+ n ow ih+ s s! iy er+ /**
	conn+ O is+ -se U+ =r	**/k! aa+ n ow ih+ s s! iy yu+ er/**
coyote	**co+ -y o+ tE**	**/k! ow+ iy ow+ t! iy/**
	co+ -y ot+ e	**/k! ow+ iy aa+ t eh/**
ya	**ya+**	**/y! ey+/**

SROOT morph.

Failures

Even after these three tactics are applied to extract morphological decompositions, 204 (node C4) still fail. When these words are examined further we find that some of them (93) would have had trouble finding the correct morph, for four main reasons. The reasons why the other 111 words fail are unclear.

1. They are foreign words that do not conform to standard English pronunciation rules. (5)

   These five words are not well-formed English words and should have been thrown out at the start. They pass the first test of getting letter morphs by "masquerading" behind known morphs. Parsing by phones helps strain out these decoys. Table 3.15 shows the words, along with their letter-derived morphs and phonemes.

   An English speaker unfamiliar with these words would pronounce them very similarly to ANGIE's proposed pronunciations.

2. The TIMIT transcription is incorrect. (9)

   There are nine words in the 204 failure set which are transcribed incorrectly, or at least strangely. A list, with their transcriptions, is included in Table 3.16. One side benefit of our procedure is that it helps to strain out disparities like these in the given corpus.

3. The ANGIE framework is missing a rule. (49)

   49 words cannot find their correct transcriptions because a rule specifying a transition is missing. Examples of some of these rules are included in Table 3.17. Some of the letter rules are for strangely spelled words, such as "silhouette" or "hemorrhage." Other letter rules are

Table 3.16: *A list of twelve words with incorrect* TIMIT *transcriptions.*

Word	TIMIT Transcription
castorbeans	[*k ae1 s axr b iy1 n z*]
countryside	[*ao ah1 n t r iy s ay2 d*]
ellipsoids	[*ax l ih1 p s oy d*]
emphysema	[*eh2 m f ax z iy1 m aa*]
infectious	[*ih n f eh1 k sh uw ax s*]
musical	[*m uw1 z ih k el*]
nancy+s	[*n ae1 n ao iy z*]
unwaveringly	[*ah n w ey1 v axr ix ng*]
vietnamese	[*v iy eh t n aa m iy1 z*]

Table 3.17: *Some missing rules needed to parse the 507 phone failed* TIMIT *words.*

Letter Rules	
Rule	Example
/**er**/ → o r r	hem**orr**hage
/**z**/ → s t h	a**sth**ma
/**sh**/ → c h e	musta**che**
/**aor**/ → u o r	autofl**uor**escence
/**aar**/ → a r r e	biz**arre**

TIMIT Phoneme Rules	
Rule	Example
/**ih**/ → [*el*], /**l**/ → [*-el*]	cart**il**age
/**d**/ → [*jh*]	a**dj**ourned
/**g**/ → [*jh*]	su**gg**estion
/**aar+**/ → [*aa2 r*]	**ar**ticulation
/**aa+**/ → [*ah2*]	everyb**o**dy
/**iy**/ → [*ih2*]	**d**esegregate

for rare sound-to-letter rules, such as "asthma." Many of the missing phone rules serve to merge ANGIE phonemes. The ANGIE phonemes for a word might split a syllabic l across a syllable boundary, as in /**eh l!**/, but the TIMIT transcription might have it as the unit [*el*]. Other phone rules support deletion, as for "adjourned" or "suggestion." As ANGIE phonemes only distinguish between two levels of stress, while TIMIT employs three, sometimes a possible stress alignment is left out. Other rules, as for "everybody", try to capture common variations in pronunciations.

4. A new morph is required, other than an SROOT. (30)

   There are 30 words that actually need a new morph other than an SROOT, so that they should not be expected to parse. Sixteen of these require a DSUF, eight a UROOT, and the rest, PRE morphs.

Table 3.18: *Composition of the new 321 morphs that are needed to parse the 507 phone failed* TIMIT *words.*

Category	Morphs	Percentage
DSUF	37	11.5%
PRE	18	5.6%
SROOT	247	76.9%
UROOT	19	5.9%
Total	321	100.0%

There are many reasons why the 111 words might be failing. The expert transcriptions for these words either do not require a new morph (40), or need an SROOT morph (71). The failures in these cases are probably due to pruning in two ways. Either the correct parse theory is pruned, or this theory is not one of the top four theories that contribute morphological decompositions. Preliminary analysis indicates that smoothing for compound words does not help.

New Morphs

In the previous section, 357 new morphs are needed to cover 374 letter failed words. From the 507 phone failed words, there are 488 words from the expert transcriptions that do not parse with this extra knowledge, and require an additional 321 morphs. The distribution of these morphs is related in Table 3.18.

In comparison with the morphs derived for letter failures (Table 3.8), there is a smaller percentage of new SROOTs, and many more affix-type morphs. This might suggest that parsing with letters is effective at discovering morphs which are known syllables but have different spellings. In contrast, words that need an affix may pass the letter parsing step by "borrowing" a morph that does not fit phonetically, in which case it is caught by the phonetic parsing.

3.5.4 TIMIT Resulting Parses

The quality of these 1,597 words is remarkably high. Only a subset of these are formally checked, since it is too time consuming to verify all of the morphological decompositions by hand. Fifty words are randomly chosen and transcribed by an expert, and then compared with the ANGIE generated versions. Tables 3.19 and 3.20 relate the necessary statistics (morph and phoneme accuracies) for these fifty words.

The nine morph sequences that do not match, or only have their segmentations match, are still reasonable. There are three phoneme sequences that do not match or only match when the onset and stress markers are removed. These phoneme sequences also are acceptable.

Table 3.19: *Tabulation of a random 50-word subset of* ANGIE *derived morphological decompositions from a set of 1,597* TIMIT *words that pass, compared to morphs provided by an expert.*

Words	Percentage	Category
18	36.0%	Have one morph transcription which is identical to hand transcribed
21	42.0%	The most likely of multiple morph transcriptions is identical to hand transcribed
2	4.0%	One of the multiple morph transcriptions is identical to hand transcribed
3	6.0%	The segmentations of the transcriptions are the same
6	12.0%	Do not match the hand transcriptions, or their segmentations
50	100.0%	Total

Table 3.20: *Tabulation of a random 50-word subset of* ANGIE *derived morphological decompositions' phonemes, from set of 1,597 words, compared to phonemes provided by an expert.*

Words	Percentage	Category
30	60.0%	Have one morph transcription which is identical to hand transcribed
10	20.0%	The most likely of multiple morph transcriptions is identical to hand transcribed
7	14.0%	One of the multiple morph transcriptions is identical to hand transcribed
2	4.0%	The phonemic transcriptions, without the markers "!", and "+", are identical
1	2.0%	Do not match the phonemic transcriptions, even without "!", and "+"
50	100.0%	Total

3.6 Chapter Summary

This chapter details the two-step letter and phone parsing algorithm developed on the TIMIT corpus. TIMIT is used because of its size, and phonetic variability. The purpose of this chapter is to attempt to extract morphological decompositions for all 2,500 words in this corpus. (The remaining 3,593 are already transcribed in the ABH corpus.) Along the way the data are analyzed both for parse failures, and to measure accuracy.

TIMIT is a corpus developed in part at MIT. It is used as a standard database by many speech recognition researchers. One quality of this corpus is that extra care has been taken to integrate a wide variety of phonetic combinations into this corpus.

The procedure to extract sub-lexical information takes two steps. In the first step, the orthography of a word is parsed into the ANGIE framework. From this framework, the morphological decompositions of the top four parses are retrieved for each word. In the second step, the phones or phonemes of the word are parsed, while being constrained to fit one of the top four morph sequences

derived from letters. In this way, both orthographic and phonological information is merged into the hierarchical parse tree. The sharing of this information helps eliminate sub-standard parses.

This process is applied to the TIMIT corpus. Of the 6,093 words, 3,593 overlap with the ABH corpus. These words are used to train ANGIE's probability model for ANGIE phoneme to TIMIT phoneme transitions. The sub-lexical extraction procedure is applied to the remaining 2,500 words.

When the 2,500 are first parsed with letters, 396 words fail to parse. The remaining 2,104 are then parsed by their TIMIT phonemes, and this time 507 fail. The four reasons why the 396 fail are that the probabilistic framework rejects the spelling of a word, the correct theory is pruned, some probabilities are missing due to sparse data, or new morphs are needed. Most of the letter failed words require new morphs, and these morphs are mainly stressed roots (SROOTs).

The 507 that obtain letter morphs but flunk the phonetic parsing are re-parsed using three different methods. 28 words pass when the knowledge derived from the 390 hand transcribed, letter failed words is added. Another 59 pass if two different stress patterns for morphs are coerced. When this stress coercion feature is combined with allowing new SROOTs, 216 more words pass. The remaining 204 words still fail, for various reasons.

The 1,597 words that pass both steps have remarkable transcriptions. If the most likely transcription is compared against the experts transcription, 78.0% are identical, according to a random fifty word subset. This metric for measuring accuracy undervalues the quality of the decompositions. Many times the morphological decompositions disagree in terms of segmentations, due to ambisyllabic consonants. Human experts usually segment the morphs based on etymology. Generally, ANGIE's transcriptions are more consistent than a human's because of this fact. As a result, when the system and expert disagree, we actually might prefer the system's choice.

Chapter 4

Experiments with the COMLEX Corpus

4.1 Motivation

In Chapter 3 we develop a method for extracting sub-lexical information from 2,500 words in the TIMIT corpus. We would like to use the same method to extract information from the much larger COMLEX corpus. Dealing with such a large lexicon will again test the limits of our procedure, but in a different way from TIMIT. TIMIT is purposely injected with many possible phonetic combinations, which must all be captured by the ANGIE framework. On the other hand, COMLEX encompasses the pronunciations and various structures of over 30,000 words, covering sub-lexical variability in another way. Evaluating the morphs' coverage of words in COMLEX will help ascertain whether morphs are a good, compact representation for words in English.

4.2 Goals

This chapter fulfills two main objectives. One is to test how well our morph extraction procedure, which has been developed on a 2,500 word subset of TIMIT, can apply to a lexicon of 30,000 words. There are four main criteria by which our procedure can be judged:

1. Accuracy of the morphological decompositions.
2. Coverage of the paradigm.
3. Consistency of morphological decompositions.
4. Information should be extracted with as little human effort (as automatically) as possible.

We plan to test our algorithm on the 34,484 words from the COMLEX corpus. These 34,484 words have 36,673 pronunciations, since alternate pronunciations are allowed[1].

The other purpose is to measure how far a little extra knowledge can take us. We have learned about 37 new rules (added to a total of approximately 1,400 letter and high level rules total), and 740 new morphs, from the 2,500 words in TIMIT.[2] We have trained an ANGIE letter grammar on the total 11,571 words from ABH and TIMIT. It will be interesting to see how many more words in COMLEX find morphological decompositions with this information. It would be rewarding to find an asymptotic accumulation of morphs, which suggests that a finite set of our morphs can compactly represent a much larger set of words.

4.3 Corpus Description

What we refer to as "COMLEX" is actually the pronouncing dictionary for the words in the COMLEX lexicon, known as PRONLEX. COMLEX is a lexicon intended for natural language processing. It is produced and distributed by the Proteus Project at New York University, under the auspices of the Linguistic Data Consortium[3]. The word list for this corpus is based on words from the WSJ30K, WSJ64K, and Switchboard corpora. WSJ30K and WSJ64K are lexicons derived from several years of the *Wall Street Journal*. These two lexicons are used in ARPA Continuous Speech Recognition corpora. The Switchboard corpus is a collection of telephone conversations, totaling three million words.

The motivation behind PRONLEX is to provide a consistent transcription, from which dialectical and other variations can be generated. The corpus is hand-transcribed. Transcribers follow a set of rules in order to maintain consistency among the transcriptions.

The entry for each word in our 66,135 word PRONLEX/COMLEX dictionary consists of the word, phonemic transcription, and class (as in NAME, ABBREV, etc.) Multiple pronunciations are included when they vary by part of speech, such as for "abstract." There are three levels of stress used in the phonemic transcriptions, "main stress", "non-main-stress", and "lack-of-stress." A listing of the phonemes used in COMLEX may be found in Appendix D.

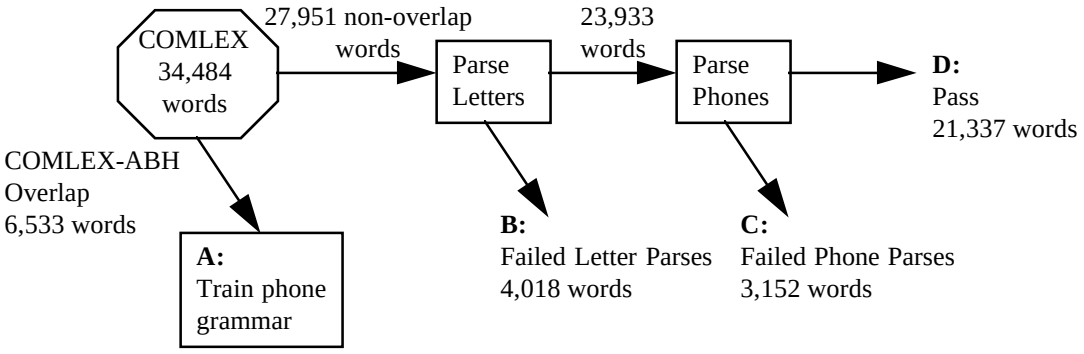

Figure 4-1: *A block diagram of the process of extracting sub-lexical information from* COMLEX *words, without* TIMIT *information.*

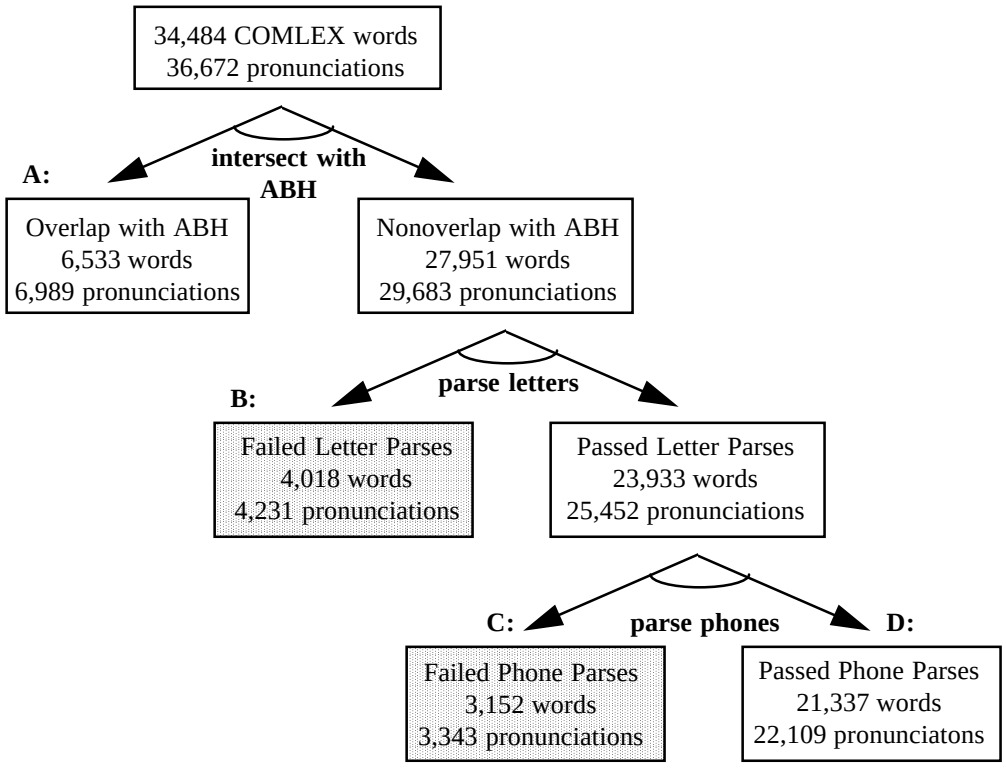

Figure 4-2: *This tree shows how the 34,484 words in* COMLEX *are divided between training data (6,533), failed letter parses (4,018), failed phonetic parses (3,152), and passed parses (21,337).*

4.4 Procedure

The procedure that is defined in section 3.4 is re-applied to the words in COMLEX. A block diagram is included in Figure 4-1 to illustrate the high-level process. A tree showing how the words in COMLEX are partitioned according to our algorithm is shown in Figure 4-2. We plan to first extract the morphs for the 34,484 words in COMLEX *without* using the knowledge gained from TIMIT. After we gather all the results, we try the same set of experiments, this time with this knowledge. Then the results can be compared.

The experiments without TIMIT knowledge are described in the next section. Section 4.6 relates the results of the same experiments augmented with TIMIT-derived morphs and rules.

Figures 4-1 and 4-2 illustrate the process and division of data for COMLEX. There are again four different groups of words. We begin with 34,484 words from COMLEX, of which 6,533 (node A in Figure 4-1) overlap with our ABH corpus. This set is used to discover and then train the allowed transitions between ANGIE's phonemes and COMLEX's phonemes. Again, even though ANGIE's framework traditionally uses phones or letters as parse tree terminals, we can also use the paradigm to determine ANGIE phoneme-to-COMLEX phoneme mappings.

The non-overlapping 27,951 words are parsed into letters, and transcribed into morphs. Only the trained probabilities and morphs derived from ABH (and not TIMIT) are used in this suite of experiments. There are 4,018 words that fail to letter parse and get morphs, denoted by node B in Figure 4-1.

When the morphs corresponding to these 23,933 words are parsed with the COMLEX phonemes, 21,337 words (node D) pass. This leaves 3,152 words (node C) that obtain letter morphs but do not parse by phones. Each of the four sets, denoted by node A through D in Figure 4-2, are examined in the next section.

4.5 Experiments, without TIMIT Knowledge

4.5.1 COMLEX ABH Overlap

6,533 words overlap with the 9,083 words in our ABH set (node A). These words are first used to determine ANGIE phoneme-to-COMLEX phoneme mappings, and then again to train ANGIE's probabilities. Subsection 3.5.1 relates how the ABH-TIMIT overlap set are used to determine rules

[1] We do not use all 66,135 words available. Names, foreign words, abbreviations, and other deviants are kept out of the set. Then the remaining 43,330 words are divided into train, development, and test sets, leaving us with only 34,484 words.

[2] If the number of morphs is tabulated from the two failed steps, it totals 678. The extra 62 morphs are added by the transcriber in anticipation of new words, while transcribing the 390 + 507 words that failed in TIMIT. For example, "grownup" was transcribed as **grow+ nup+**. Even though these new morphs are incorrect for this word, they are added because they can be used in other words such as "prenuptial" and "grow."

[3] `http://www.ldc.upenn.edu`

and then a phone trained grammar. The rules and grammar are similarly created for COMLEX, except that this time the rules are not created by hand but automatically generated. The next two subsections describe the creation of rules and grammar. The five sources of knowledge needed to obtain a letter parse are the same ABH ones that were used for TIMIT.

Automatic Rule Induction

First ANGIE-to-COMLEX phoneme rules have to be derived, and then ANGIE can be trained to produce a trained grammar. These rules are employed on the 6,533 overlap set of words, which have both ANGIE phoneme sequences, and the COMLEX transcription. Hand-writing these rules is a mechanical process. First, obvious mappings are determined beforehand, and then words are parsed into ANGIE to discover missing rules. Because this procedure is time-consuming, a method for automatically deriving these rules has been implemented and applied by Meng [8].

The algorithm begins with a set of obvious mappings, which it uses to anchor ANGIE and COMLEX phonemes. The technique is to relabel the COMLEX phonemes to their ANGIE phoneme equivalent to provide these mappings. The stress and secondary stressed vowels ("non-main-stress") are both transcribed as stressed ANGIE phonemes. Then, the ALIGN program [2] is used to align the ANGIE and COMLEX phonemic transcriptions for each word. If there are alternate transcriptions, every combination of ANGIE and COMLEX transcriptions is aligned. There are 7,580 different ANGIE-COMLEX transcription pairs.

The ALIGN program then tabulates the mis-alignments. The ANGIE phoneme-to-COMLEX phoneme rules are created, based on these mis-alignments, which include errors such as substitutions, deletions, insertions, merges and splits. When these newly generated rules are applied to the 7,580 different pairs, all but 65 (which implies 99.1% coverage) parse into the ANGIE framework. Ten more hand-written rules enable these words to parse.

One consequence of automatically generating rules by aligning *every* combination of alternate spellings is that the rules can become overly general. In order to combat this, an expert usually looks over the rules for incorrect transitions. Some human-engineered constraints are added to this process to further reduce the number of incorrect rules.

ANGIE to COMLEX Trained Grammar

After these rules are created, the COMLEX phonemes of all the overlap words are parsed through the ANGIE framework. The counts are collected, normalized as probabilities, and then stored in a trained grammar, just as described in subsection 2.2.3.

4.5.2 COMLEX Letter Parses

We parse the spellings of the 27,951 words which do not overlap with ABH. Of these 27,951 words, 23,933 (85.6%) parse into the ANGIE framework and obtain morphological decompositions. 4,018 words (node B) do not parse. These words fail either because the rules do not allow a certain transition, or a morph is missing. We parse these 4,018 words without morph constraint, to see how many require new rules. Only 245 fail, and most of these (208) use an apostrophe (symbolized as "+") in constructions such as "musicians+" or "must+ve". These words are not accommodated by the ANGIE framework. Most of the other words (37) which are rejected by the letter rules are names like "abramowitz", or words like "razzmatazz" and "svelte." These words are either not real English words, or borrowed ones.

Table 4.1: *A random sample from the 3,319* COMLEX *words, their morphs, and phonemes, with invented* SROOTs*, that failed letter parsing.*

Word	Morphs	Phonemes
aphrodisiac	**aph+ ro diS+ -i ac+,**	**/ae+ f r! ah d! ih+ z iy ae+ k/**
biochemical	**bi+ -o chem+ -ic =al**	**/b! ay+ ow k! eh+ m ih k el /**
blitzed	**blitz+ =ed**	**/b! l ih+ t s d*ed /**
crooner	**croon+ =er**	**/k! r uw+ n er /**
dramatizing	**dram+ a tiz_e+ =ing**	**/d! r aa+ m ah t! ay+ z ing /**
jockeying	**jocke+ -y =ing**	**/jh! aa+ k iy ing /**
jostling	**jost+ -ling**	**/jh! aa+ s t l! ing /**
nooks	**nook+ =s**	**/n! uh+ k s*pl /**
overzealousness	**ov+ er zea+ -lous =ness**	**/aa+ v er z! iy+ l! ah s n! eh s /**
stout	**stout+**	**/s! t aw+ t /**
styrene	**sty+ rene+**	**/s! t ay+ r! iy+ n /**
sycophantic	**syc+ o phant+ -ic**	**/s! ih+ k ah f! ae+ n t ih k /**
tawdry	**tawd+ r -y**	**/t! ao+ d er iy /**
xenophobic	**xen+ o phob+ -ic**	**/z! eh+ n ah f! aa+ b ih k /**

Parsing with Invented SROOTs

When the 4,018 words are all allowed to invent new SROOTs, 3,319 pass both letter and phonetic parses, as shown in Figure 4-3. 709 words (758 pronunciations) fail. This suggests that many more new SROOT morphs are needed to cover COMLEX.

The quality of these morph transcriptions is surprisingly high. "_e" morphs are used, as in **dron_e+ =ing**. Correct endings are detected, as for **gust+ =s**. Cross-examination of the results reveals that there are a few segmentations that are not preferred; one example (which involves an ambisyllabic consonant) is **stab+ -lest**. A random sample of these words with their morphs is provided in Table 4.1. Some impressive morph decompositions include **bi+ -o chem+ -ic =al**, **syc+ o phant+ -ic**, and **xen+ o phob+ -ic**. The phonemes of the transcriptions for the 3,291

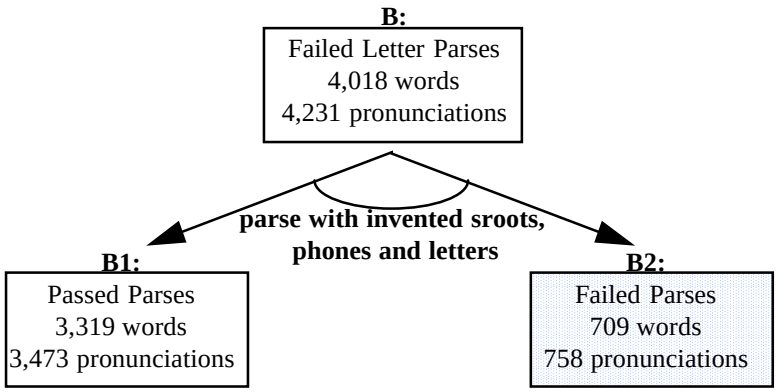

Figure 4-3: *This tree shows how the 4,018* COMLEX *words which fail the letter parsing step are further processed.*

words are also fairly good. Some mistakes include that for "overzealousness," and "xenophobic." The others are well done.

4.5.3 COMLEX Phonetic Parses

We now move on to the 23,933 words which do not fail the letter parse. They have among them 25,452 pronunciations, since some words have alternate pronunciations. Of these words, 21,337 words pass the phonetic parse, while 3,152 fail. This failed set corresponds to node C in Figure 4-1. The division of the failures is shown in Figure 4-4.

Parsing with Coerced Stress Patterns

If we were emulating our steps in TIMIT exactly, we would add the information we gathered from the letter failed set and see how many more pass. This information, in terms of new invented morphs from the COMLEX failed letter parses, cannot be used because it has not been evaluated. We do not want to contaminate our knowledge base with false morphs.

The next step is to parse the failed words, forcing both stress patterns when deriving morphs from the letters. When this experiment is carried out on the 3,343 pronunciations that fail, an additional 398 pronunciations pass.

Many of these 379 words have acceptable transcriptions. As shown in Table 4.2, many other are stress shifted, such as "disarmingly." Others have the wrong phoneme transcriptions, as in "inspirational." It is likely that these incorrect transcriptions pass through because the automatic, ANGIE-to-COMLEX rules are too forgiving.

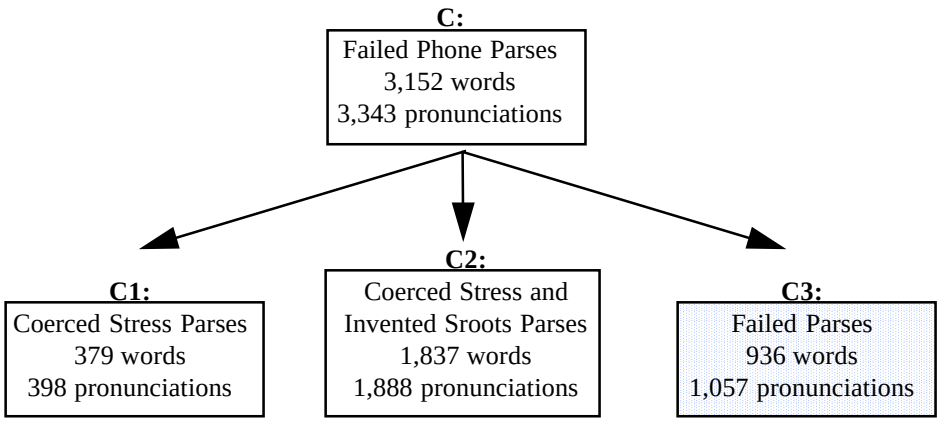

Figure 4-4: *This tree shows how the 3,152* COMLEX *words which fail the phone parsing are further processed.*

Parsing with Coerced Stress Patterns, and Invented SROOTs

2,945 pronunciations (2,778 words) do not parse, even with coerced stress patterns. We assume that they might need new SROOTs, and so we allow SROOT invention as well as force stress patterns.

With this, 1,888 pronunciations, belonging to 1,837 unique words pass. 1,057 pronunciations fail, encompassing 986 words. A random sample of some of the words, along with their top phoneme sequence, is included in Table 4.3. Even with the loosened constraints of invented sroots, and the relatively lax ANGIE-to-COMLEX rules, these transcriptions are quite remarkable.

4.5.4 COMLEX Resulting Parses

22,109 pronunciations, representing 20,934 words, survive with at least one morphological decomposition after passing both the letter and phonological parsing (node D in Figure 4-1). A random sample is included in Table 4.4. Overall, it is acceptable to add these words into our framework. Only two words from this sample, "sniped" and "acidity," have unacceptable transcriptions. These casualties result from over-generalizations in the rules. The ANGIE phonemes /**ih**+/ and /**ay**+/ are allowed to transition to either the COMLEX [*ay+*] or [*ih+*]. Even if the rules allow questionable transitions, the probabilities should have filtered it out. Table 4.10 summarizes the results of this section.

Table 4.2: *A random sample from the 379* COMLEX *words which pass with stress coercion, with their morphs and top phoneme sequence.*

Word	Morphs	Phonemes
anyplace	**An+ -y place+**	**/ey+ n iy p! l ey+ s /**
armada	**ar- ma+ da**	**/er m! ae+ d! ah /**
delete	**de+ -lete**	**/d! iy+ l! iy t /**
disarmingly	**dis- Ar- ming+ =ly**	**/ d! ih+ s aar m! ih+ ng l! iy /**
inspirational	**in- spi- rati+ on =al**	**/ih+ n s! p ih r! ae+ sh en el /**
isometrics	**is+ O met+ -ric =s**	**/ih+ s ah m! eh+ t r! ih k s*pl /**
oilfield	**oi+ îl fi+ -el =d**	**/oy+ el f! iy+ el d*ed /**
reflexes	**ref+ lex+̂ -es**	**/r! eh+ f l! eh+ k s s*pl /**
tornado	**tor- na+ do+**	**/t! er n! ey+ d! ow /**
transatlantic	**tran- sat+ lant+ -ic**	**/t! r ae n s! ae+ t l! ae+ n t ih k /**
unprepared	**un- pre- par+ =ed**	**/ah+ n p! r iy p! ehr+ d*ed /**
unreality	**un- re- al+ i -ty**	**/ah+ n r! iy ae+ l ih t! iy /**
unrealized	**un- re+ al -ize =d**	**/ah n r! iy+ el ay z d*ed/**

4.6 Experiments, with TIMIT Knowledge

4.6.1 COMLEX ABH Overlap

In this section we explore what happens to our results when the information gained from TIMIT is added to the ANGIE framework. Hopefully this knowledge can improve the results substantially. The total new knowledge is encoded in terms of 51 new high level, low level letter, and low level phone rules, as well as 754 new morphs. It also includes a new letter grammar, and a new ANGIE phoneme-to-COMLEX phoneme grammar. We hope to show that our system is reaching an asymptotic state in the amount of knowledge it must acquire to parse English words.

There are two sources of these TIMIT-derived knowledge bases. As described before, we want to measure the change in performance when information from another corpus (TIMIT) is added to ANGIE's knowledge base. Thus we add the new letter and high level rules (37 total) and new morphs (740) that are needed in order to properly "absorb" TIMIT into ANGIE's framework.

Then we must consider the overlap set between the base lexicon and the target lexicon. In the previous section, the base lexicon is ABH, and the target is COMLEX, so that the overlap set consists of some 6,533 words, which are used to train the ANGIE-COMLEX grammar. In this set of experiments, our base lexicon is now ABH and TIMIT (ABHT), and the target remains COMLEX. Our overlap set now consists of 8,265 words, which includes 1,732 TIMIT-COMLEX overlap words added to the original 6,533. To follow the normal convention, we use this larger set to train an ANGIE phoneme-to-COMLEX phoneme grammar. In the course of training the new 1,732 words, 14 new phone rules and 14 new morphs are added to the knowledge bases[4].

[4] It seems erroneous to add new morphs at this stage, as the words in TIMIT are supposed to have found correct

Table 4.3: *A random sample of 1,888* COMLEX*words which pass with stress coercion, and invented* SROOT*s, with their morphs and top phoneme sequence.*

Word	Morphs	Phonemes
anesthesia	**an+ -es thes+ -ia**	**/ae+ n eh s th! eh+ z iy ah /**
bicentennial	**bi+ cent+ en ni+ al**	**/b! ay+ s! eh+ n t en n! iy+ el /**
cavalcade	**cav+ al cade+**	**/k! ae+ v el k! ey+ d /**
cruddy	**crudd+ -y**	**/k! r ah+ d iy /**
expanse	**ex^ panse+**	**/eh k s p! ae+ n s /**
extraordinarally	**ex^+ tra ord+ in ar+ al =ly**	**/eh+ k s t! r ah aor+ d en aor+ el l! iy/**
fey	**fey+**	**/f! ey+ /**
gene+s	**gene+ =+s**	**/jh! iy+ n s*pl /**
gumshoe	**gum+ shoe+**	**/g! ah+ m sh! uw+ /**
hamstrung	**ham+ strung+**	**/h! ae+ m s! t r ah+ ng /**
harping	**harp+ =ing**	**/h! aar+ p ing /**
leaking	**leak+ =ing**	**/l! iy+ k ing /**
pheasant	**pheas+ -ant**	**/f! eh+ z en t /**
phosphates	**phos+ phate+ =s**	**/f! aa+ s f! ey+ t s*pl /**
salvage	**salv+ -age**	**/s! ae+ l v eh jh /**
transpac	**trans+ pac+**	**/t! r ae+ n s p! ae+ k /**
usurped	**u- surp+ =ed**	**/yu s! er+ p d*ed /**
weaving	**weav+ =ing**	**/w! iy+ v ing /**

The information from all of TIMIT and the new set of TIMIT-COMLEX overlaps is combined to form a set of 754 new morphs, which are added to the 5,168 morph-phoneme lexicon, and 51 new rules. As usual, we create our letter grammar from the 11,571 word base (ABHT) lexicon, and the phone grammar from the 8,265 ABHT-COMLEX overlap words.

The same process denoted in Figure 4-1 is used with TIMIT knowledge, and is shown in Figure 4-5. A tree illustrating the division of data is shown in Figure 4-6.

4.6.2 COMLEX Letter Parses

We begin with the non-overlap set of 26,219 words (27,847 pronunciations). When we parse with the knowledge gained from TIMIT, 2,834 words fail on the letter parse, and 23,385 pass. This set is depicted by node B in Figure 4-5, or 4-6.

There are two reasons why a parse can fail. Either the word cannot be fit into the framework, which is usually due to an irregular spelling, or the word has trouble finding matching morphs. To rule out morph failure we can try parsing the spellings of the 2,834 words without morph constraint.

morphological decompositions at this point. Although these 1,732 words have found the correct morphological decompositions for their TIMIT pronunciations, some of them have alternate COMLEX pronunciations, which are not represented by their morphs, or by the ANGIE phoneme-to-COMLEX rules. Hence we must add the new morphs and rules to our sources for them to parse. As there are only about twenty words in this set, it is more convenient to hand-write the morphs and rules, instead of generating them automatically.

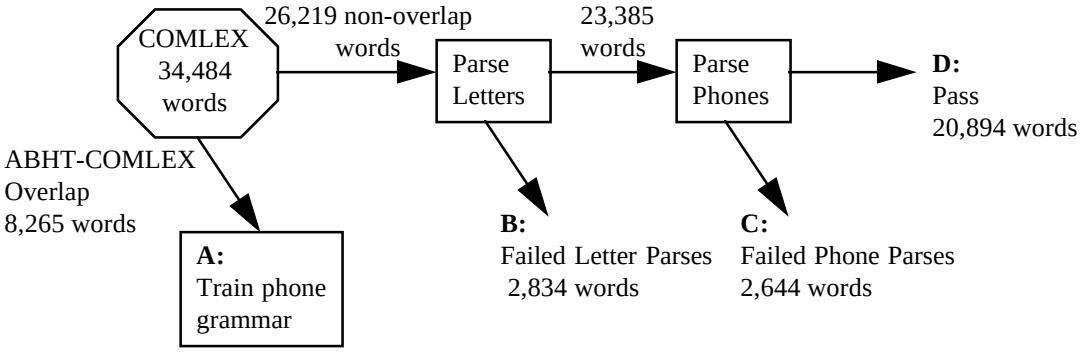

Figure 4-5: *A block diagram of the process of extracting sub-lexical information from words, for* COMLEX, *with* TIMIT *knowledge.*

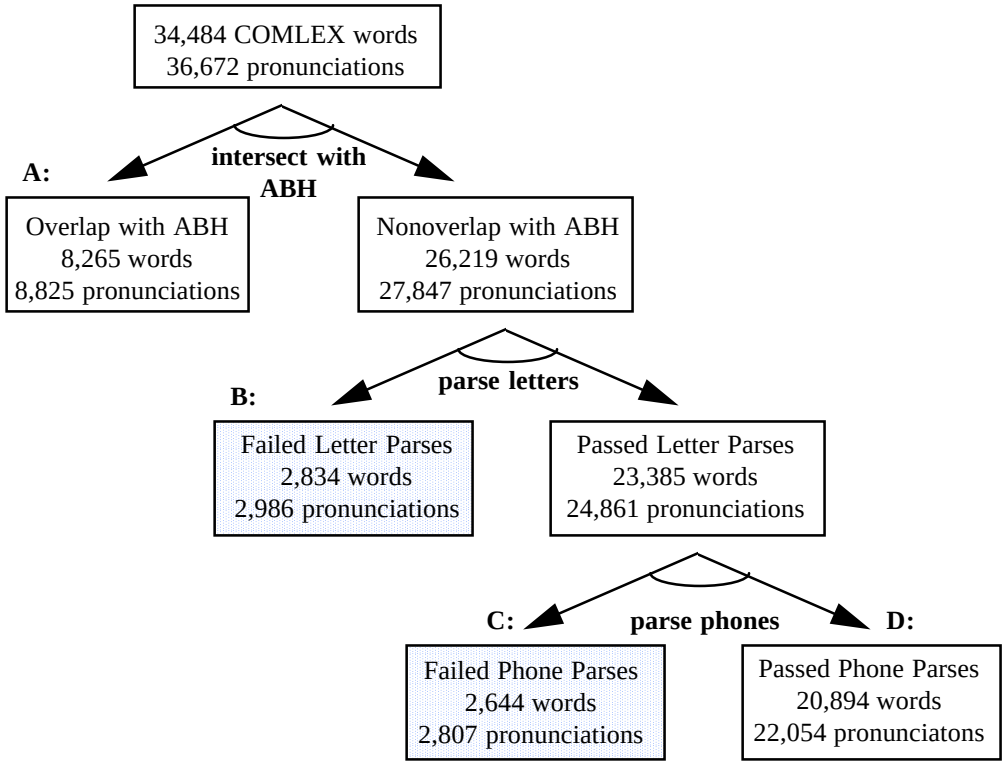

Figure 4-6: *This tree shows how the 34,484 words in* COMLEX *are divided between training data (8,265), failed letter parses (2,834), failed phonetic parses (2,644), and passed parses (20,894).*

Table 4.4: *A random sample of* COMLEX *words their morphs, and phonemes, from the set of 22,109 pronunciations which pass both letter and phone parsing steps.*

Word	Morphs	Phonemes
acidity	**a- ci+ di -ty**	**/ah s! ay+ d! ih t! iy /**
agile	**ag+ ile**	**/ae+ g el /**
corollary	**cor+ ol -lary**	**/k! aor+ el l! ehr iy /**
feted	**fe+ tE =d**	**/f! iy+ t! iy d*ed /**
fullback	**full+ back+**	**/f! uh+ l b! ae+ k /**
gopher	**go+ -pher**	**/g! ow+ f! er /**
gratifying	**grat+ i fy+ =ing**	**/g! r ae+ t ih f! ay+ ing /**
holistic	**ho+ list+ -ic**	**/h! ow+ l! ih+ s t ih k /**
interwoven	**int+ er wov+ en**	**/ih+ n t! er w! ow+ v en /**
megalomaniac	**meg+ al -o man+ -i ac+**	**/m! eh+ g el ow m! ae+ n iy ae+ k/**
motionless	**mo+ tion =less**	**/m! ow+ sh! en l! eh s /**
optometric	**op+ to met+ r -ic**	**/aa+ p t! ow m! eh+ t er ih k /**
palpable	**pal+ pa -ble**	**/p! ae+ l p! ah b! el /**
plentitude	**plen+ ti tude+**	**/p! l eh+ n t! ih t! uw+ d /**
sniped	**snip+ =ed**	**/s! n ih+ p d*ed /**
spits	**spit+ =s**	**/s! p ih+ t s*pl /**
supports	**sup- port+ =s**	**/s! ah p p! aor+ t s*pl /**
synopsis	**syn- op+ -sis**	**/s! en aa+ p s! ih s /**
toasted	**toast+ =ed**	**/t! ow+ s t d*ed /**
underprice	**un+ der price+**	**/ah+ n d! er p! r ay+ s /**
unincorporated	**un- in- cor+ por ate+ =d**	**/ah+ n ih n k! aor+ p! er ey+ t d*ed/**
uttered	**utt+ er =ed**	**/ah+ t er d*ed /**

The words that fail in this step definitely need new rules in order to parse.

When parsing through the letter mode without morph restrictions, only 234 words, instead of the 245 from before, have trouble. Thus eleven new words now pass letter parsing. Some interesting sets include "pneumo", and "pneumocystis", which are rescued by the "pn" rule learned from "pneumonia." "psalm" and "psalms" pass because a rule allowing the "l" to be silent (as in "almonds") is added. Finally, "schnauzers", and "schnoodle" are aided by the word "schnooks." A curious failure is the word "attermann," which passes with the old rules but not with the extended set. One possibility is that the TIMIT knowledge helps to choose a better parse, which is not entirely supported by the morphs. This parse directly competes against the sub-standard passable theory, which ends the same way, and so is pruned. Three of the eleven words that pass are the contractions "could+ve," "would+ve," and "should+ve."

Parsing with Invented SROOTs

These 2,834 letter failed words are now allowed to parse with invented SROOTs. 318 words fail to parse with invented SROOTS in letter mode. 245 pronunciations fail when parsing the COMLEX

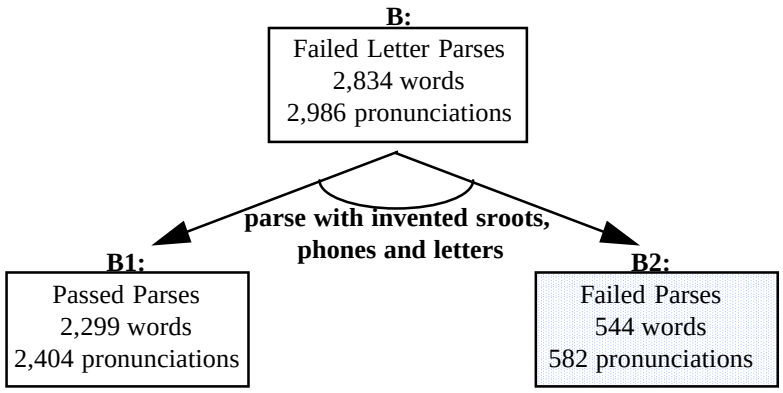

Figure 4-7: *This tree shows how the 2,852* COMLEX *words which fail the letter parsing step are further processed.*

Table 4.5: *A random sample from the 2,299* COMLEX *words, with their morphs and phonemes, which need invented* SROOT*s to parse.*

Word	Morphs	Phonemes
craps	**crap+ =s**	**/k! r ae+ p s*pl /**
deftly	**deft+ =ly**	**/d! eh+ f t l! iy /**
dispersant	**dis- pers+ -ant**	**/d! ih s p! er+ z en t /**
doss	**doss+**	**/d! ao+ s /**
glock	**glock+**	**/g! l aa+ k /**
huffing	**huff+ =ing**	**/h! ah+ f ing /**
juggler	**jugg+ -ler**	**/jh! ah+ g l! er /**
kish	**kish+**	**/k! ih+ sh /**
puddle	**pudd+ le**	**/p! ah+ d el /**
reentry	**reen+ -try**	**/r! iy+ n t! r iy /**
whine	**whine+**	**/w! ay+ n /**
zing	**zing+**	**/z! ih+ ng /**

phonemes, so that 2,404 pronunciations (2,299 words) obtain parses. A random sample of the words, with their morphs and top phonemes, is included in Table 4.5. Parsing with invented SROOTs again yields excellent results.

4.6.3 COMLEX Phonetic Parses

The remaining 23,385 words, with 24,861 pronunciations, are parsed with their COMLEX phonemes, and constrained to match one of the morphs obtained from the previous step. 2,807 pronunciations, or 2,664 words, fail to parse, as shown by node C, Figure 4-5. 22,054 pronunciations, corresponding to 20,894 words, pass (node D). These words are inspected in subsection 4.6.4. This subsection emulates the same steps taken in subsection 4.5.3, when COMLEX words are parsed without TIMIT

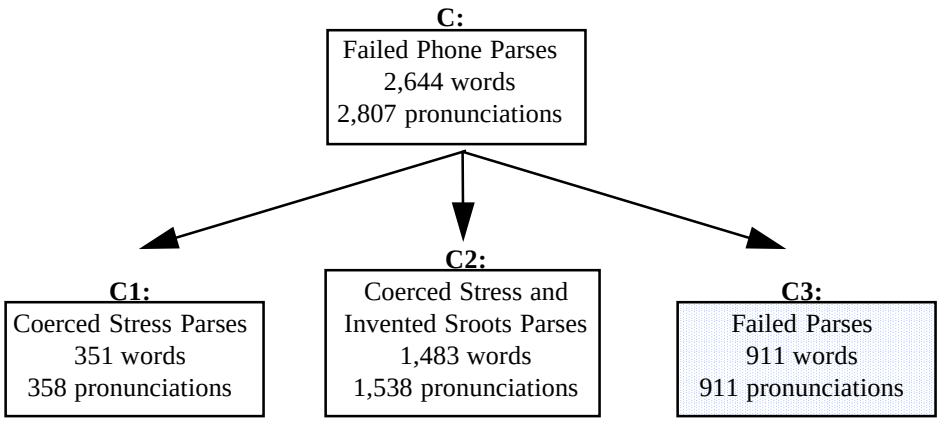

Figure 4-8: *This tree shows how the 2,771* COMLEX *words which fail the phone parsing are further processed.*

knowledge.

Parsing with Coerced Stress Patterns

Table 4.6: *A random sample from the 341* COMLEX *words parsed with stress coercion, with their morphs and phonemes.*

Word	Morphs	Phonemes
appetizing	**app+ -et iz_e+ =ing**	**/ae+ p eh t ay+ z ing /**
aristocracy	**ar+ is+ to cra+ -cy**	**/aar+ ih+ s t! ow k! r ey+ s! iy /**
biomaterials	**bi- o+ ma teR+ -ial =s**	**/b! ay ow+ m! ah t! ihr+ iy el s*pl /**
concertos	**con+ cer+ to =s**	**/k! aa+ n s! ehr+ t! ow s*pl /**
congresspersons	**cong+ -ress per+ son =s**	**/k! aa+ ng g r! eh s p! er+ s! en s*pl /**
erudition	**er+ u diti+ on**	**/ehr+ yu d! ih+ sh en /**
fidelity+s	**fi- del+ -ity =+s**	**/f! ay d! el+ ih t! iy s*pl /**
hominem	**ho- min+ em+**	**/h! ow m! ih+ n eh+ m /**
improprieties	**im+ pro pri+ e tI+ =es**	**/ih+ m p! r ah p! r ay+ eh t! iy+ s*pl /**
parachuted	**pAr+ -ach ute+ =d**	**/p! ehr+ ah ch yu+ t d*ed /**
renegades	**ren+ e gad+ =es**	**/r! eh+ n eh g! ae+ d s*pl /**
semiconductor	**sem+ I+ con duc+ tor**	**/s! eh+ m iy+ k! en d! ah+ k t! er /**
unopposed	**un- op- pose+ =d**	**/ah+ n ah p p! ow+ z d*ed /**

The 2,644 words (with 2,807 pronunciations) that fail in the above step are parsed again with coerced stress patterns. With this procedure, 358 more pronunciations (341 words) parse. A random sample of the morphs and phonemes is shown in Table 4.6. Some of the transcriptions are acceptable. Some errors shown in the table include a confusion between the short and long vowels (**/ae/** and **/ey/**) for "a", as for "renegades" and "aristocracy." Many words unfortunately choose the incorrect

stress pattern, such as "hominem."

The remaining 2,449 pronunciations (2,311 words) are piped to the next step.

Parsing with Coerced Stress Patterns, and Invented SROOTs

Table 4.7: *A random sample from the 1,483* TIMIT *words parsed with stress coercion and invented* SROOT*s, with their morphs and phonemes.*

Word	Morphs	Phonemes
affidavit	**aff+ i dav+ -it**	**/ae+ f ih d! ey+ v ih t /**
applause	**ap- plause+**	**/ah p p! l ao+ z /**
barge	**barge+**	**/b! aar+ jh /**
browse	**browse+**	**/b! r ow+ z /**
converged	**con- verge+ =d**	**/k! en v! er+ jh d*ed /**
crimping	**crimp+ =ing**	**/k! r ih+ m p ing /**
fathomable	**fath+ om =able**	**/f! ae+ th em ah b! el /**
forbade	**for+ bade+**	**/f! aor+ b! ae+ d /**
ginseng	**Gin+ seng+**	**/jh! ih+ n s! eh+ ng g /**
hamstrung	**ham+ strung+**	**/h! ae+ m s! t r ah+ ng g /**
launderer	**laun+ der Er+**	**/l! ao+ n d! er er+ /**
liquidities	**lI+ quîd+ -iti =es**	**/l! iy+ k! w ih+ d ih t! iy s*pl/**
neurofibromatosis	**neur+ O fib+ rO ma+ to -sis**	**/n! yu+ r ow f! ay+ b r! ow m! ay+t! ow s! ih s /**
pastiche	**pas+ tiche+**	**/p! ae+ s t! ih+ ch /**
poked	**poke+ =d**	**/p! ow+ k d*ed /**
poops	**poop+ =s**	**/p! uw+ p s*pl /**

The 2,449 pronunciations, or 2,311 words are parsed again, this time with SROOT invention. Curiously, nine words fail when the letters are parsed for morphs. Another 911 words fail the phone parse. The remaining 1,483 words, with 1,538 pronunciations, pass with a morph analysis. A random sample of these are included in Table 4.7. The transcriptions of all of these examples are well done. The only exceptions are the words "forbade" and "liquidities," where it seems as if a letter's long and short vowels are confused. Especially impressive is the analysis of "neurofibromatosis."

A curious failure is that a set of nine words fail when trying to get letter morphs with the invented morphs capability. However, they pass if the SROOT morph invention is turned off. A list of these words is shown in Table 4.8. The reason that they fail with invention but not without must involve pruning. What appears to be happening is that the possible morph sequence that is realized without invented morphs is ranked lower than those theories which allow invented morphs, and eventually gets pruned. The other theories with invented morphs are rejected later on.

Table 4.8: *Nine words which pass with coerced stress, but fail when they are allowed to invent new* SROOT*s.*

Word
archaeological
archaeologist
archaeologists
archaeology
creativeness
earmuffs
metabolisms
statesmanlike
tumult

4.6.4 COMLEX Resulting Parses

In the end, 22,054 pronunciations, or 20,894 words, from the original set of 26,219 words pass the procedure and get morphological decompositions (node D). Some examples are shown in Table 4.9. Most of these are accurate transcriptions. Vowel confusions appear in words like "admirals", "applicable" and "promenade." "ounces" is parsed very strangely. Someone not familiar with the word "unconscionable" might pronounce it similarly to the transcription given in Table 4.9. Overall the quality of these transcriptions is impressive. The results of these experiments are summarized in Table 4.10.

4.7 Chapter Summary

In this chapter the procedure used to derive morphs from TIMIT is applied to a much larger corpus known as COMLEX. If the morphs from a set of over 10,000 words can adequately cover a set of 30,000 words, we can be assured that morphs are a valid sub-word model. We would like to also measure how much the extra knowledge derived from TIMIT can improve parse coverage. This extra knowledge includes 51 new high and low level letter rules, 754 new morphs, and new letter and phone trained grammars.

COMLEX is a corpus intended for natural language processing. PRONLEX, the pronouncing dictionary for COMLEX, is what is actually used, but we call it COMLEX for simplicity. COMLEX is a 66,135 word corpus with a phonemic baseform for each word. Before utilizing this set, we eliminate about half of the words in the lexicon, including foreign words and names, which may not obey the spelling and phonological rules of English.

We apply the letter and phone parsing procedure developed in section 3.4 to 34,484 words in COMLEX, both without and with information learned from TIMIT. Table 4.10 summarizes the results of this chapter. The nodes given in the table are consistent with those in Figures 4-3, 4-4, 4-7, and 4-

Table 4.9: *A random sample from the 20,894* COMLEX *words which pass both steps.*

Word	Morphs	Phonemes
admirals	**ad- mir+ al =s**	**/ae d m! ihr+ el s*pl /**
appendages	**ap- pend+ -age =s**	**/ah p p! eh+ n d eh jh s*pl /**
applicable	**ap- pli+ ca+ -ble**	**/ah p p! l ay+ k! ey+ b! el /**
basics	**ba+ -sic =s**	**/b! ey+ s! ih k s*pl /**
blackmailed	**black+ ma+ il =ed**	**/b! l ae+ k m! ey+ el d*ed /**
defrost	**de- frost+**	**/d! iy f! r ao+ s t /**
fellas	**fell+ -as**	**/f! el+ ah s /**
fertilizes	**fer+ til ize+ =s**	**/f! er+ t! ih l! ay+ z s*pl /**
immovable	**im- mov+ =able**	**/ih m m! uw+ v ah b! el /**
moderns	**mod+ -ern =s**	**/m! aa+ d er n s*pl /**
ounces	**o+ un ces+**	**/ow+ en s! eh+ s /**
promenade	**pro- men+ ade+**	**/p! r ow m! eh+ n ey+ d /**
publics	**pub+ -lic =s**	**/p! ah+ b l! ih k s*pl /**
pulverize	**pUl+ ver ize+**	**/p! ah+ l v! er ay+ z /**
shameful	**shame+ =ful**	**/sh! ey+ m f! el /**
thirty_five	**thirt+ y five+**	**/th! er+ t ih f! ay+ v /**
timpani	**tim+ pa -ni**	**/t! ih+ m p! ah n! iy /**
unbranded	**un- brand+ =ed**	**/ah n b! r ae+ n d d*ed /**
unconscionable	**un- cons+ Ci on+ a -ble**	**/ah n k! aa+ n s sh! iy aa+ n ah b! el /**
unspectacular	**un- spec+ tac+ u -lar**	**/ah n s! p eh+ k t! ae+ k yu l! er /**
wallop	**wall+ -op**	**/w! aol+ ah p /**
widower	**wid+ ow =er**	**/w! ih+ d ow er /**

8.

The quality of the morphological decompositions is on a whole, acceptable. The most accurate sets include not only those that pass completely through both steps, but those letter failed words that are allowed to invent their own stressed morphs. It seems that the failed phonetic, coerced stress set are a little worse in quality than those that are allowed to invent SROOTs, but this belief has not been rigorously tested.

Table 4.10: *Tabulation of results for* COMLEX, *both with and without* TIMIT*-derived knowledge, in terms of pronunciations. The numbers are somewhat incomparable since the overlap group changes.*

Category	COMLEX	COMLEX + TIMIT
ABH(T) Overlap	6,989	8,825
Failed Letters, Invented SROOTs (node B1)	3,473	2,404
Failed Letters, Completely (node B2)	758	582
Failed Phones, Coerced Stress (node C1)	398	358
Failed Phones, Coerced Stress and Invented SROOTs (node C2)	1,888	1,538
Failed Phones, Completely (node C3)	1,057	911
Passed, Completely (node D)	22,109	22,054
Total	36,672	36,672
Percentage not recovered, of all words (nodes B2 + C3)/(node D)	4.9%	4.1%

Chapter 5

Analysis and Comparisons

This chapter analyzes the results obtained in previous sections. The first section provides information about the coverage of our procedure on the three data sets (TIMIT, COMLEX, and COMLEX with TIMIT knowledge). The improvements in coverage due to knowledge gained from TIMIT is also analyzed. The quality of the morphological transcriptions is explored in the next section. Then, an evaluation of the constraint provided by the letter and phone parsing step is provided. Another section compares hand-written rules to automatically generated ones. The consistencies of the morphological decompositions are briefly discussed in the final section.

5.1 Coverage

In this section we summarize the results from the three lexicons, and explore whether adding information from TIMIT improves parsing coverage in COMLEX. In order to make a valid, direct comparison, we must first normalize the data to exclude the TIMIT-COMLEX overlap set, and then normalize by the number of words parsed by the algorithm, not the total number of words in the corpus. Table 4.10 summarizes the division of the data when COMLEX is parsed with and without knowledge obtained from TIMIT[1]. "Percentage not recovered" is the sum of words that either "Failed Letters, Completely," or "Failed Phones, Completely." For convenience, the COMLEX failure modes have been matched with their associated nodes, depicted in Figures 4-3 and 4-4 for the COMLEX set parsed without TIMIT information, and Figures 4-7 and 4-8 for the COMLEX set parsed with TIMIT information.

We want to remove the 1,732 TIMIT-COMLEX overlap words from both COMLEX data sets so that a direct comparison can be made. The COMLEX data that are already parsed with TIMIT knowledge

[1]We tally our distributions in terms of pronunciations instead of words because one word may have multiple pronunciations, which makes counting words more difficult and less meaningful.

Table 5.1: *Tabulation of results for* COMLEX*, without* TIMIT*-derived knowledge, in terms of pronunciations. The distribution of the 1,732* TIMIT *words (1,836 pronunciations) which overlap with* COMLEX *in the first experiment are included as a separate column.*

Category	COMLEX	TIMIT Overlap	COMLEX - TIMIT
Failed Letters, Invented SROOTs (node B1)	3,473	228	3,245
Failed Letters, Completely (node B2)	758	35	723
Failed Phones, Coerced Stress (node C1)	398	25	373
Failed Phones, Coerced Stress and Invented SROOTs (node C2)	1,888	98	1,790
Failed Phones, Completely (node C3)	1,057	50	1,007
Passed, Completely (node D)	22,109	1,400	20,709
Total	29,683	1,836	27,847
Percentage not recovered (nodes B2 + C3)/Total	6.1%	4.6%	6.2%

already exclude this set, which has been assigned to the overlap (node A) group to train a phone grammar. The version of COMLEX without TIMIT knowledge still contains this set.

Table 5.1 separates these overlap words from our COMLEX distribution. The first column shows the distribution of the COMLEX words parsed without TIMIT knowledge. In the second is the distribution of the 1,836 pronunciations that are both in TIMIT and COMLEX. Once the sets are grouped into the proper categories, the overlap set can be subtracted from the COMLEX set, leaving the same set of 26,219 words (27,847 pronunciations) that are used as the non-overlap group when parsing with ABH and TIMIT.

These words which are in COMLEX and *not* in ABHT consist of 27,847 pronunciations. The distribution of these 27,487 pronunciations, or 26,219 words, can be directly compared, as shown in Table 5.2. The percentages are normalized not by the total number of pronunciations in the lexicon, but by the total number of pronunciations that are actually parsed (as opposed to being used as training data for rules and the phone trained grammar).

The results in Table 5.2 indicate that the extra knowledge did improve performance to some extent. 4.8% (79.2% - 74.4%) more COMLEX pronunciations (1,345) pass with the added information, and 0.8% fewer words (237 pronunciations) are unrecoverable. Another observation is that the number of words that require invented SROOTs drops by 3.0% (841 pronunciations), possibly because the words in question have found their morphs in new TIMIT morphs. Table 5.3 shows the results for TIMIT, with a similar distribution among the different groups.

Because more words are passing as more knowledge is incorporated into the ANGIE framework, as shown in Table 5.2, we can hope that parse coverage can slowly approach 100%. Unfortunately,

Table 5.2: *Tabulation of results for* COMLEX*, both with and without* TIMIT*-derived knowledge, in terms of pronunciations. The percentages are normalized only for the words which do not overlap with ABH, or with* TIMIT*. (The 1,723 word subset has been removed.)*

Category	COMLEX	COMLEX + TIMIT
Failed Letters, Invented SROOTs (node B1)	11.6%	8.6%
Failed Letters, Completely (node B2)	2.6%	2.1%
Failed Phones, Coerced Stress (node C1)	1.3%	1.3%
Failed Phones, Coerced Stress and Invented SROOTs (node C2)	6.4%	5.5%
Failed Phones, Completely (node C3)	3.6%	3.3%
Passed, Completely (node D)	74.4%	79.2%
Total (pronunciations)	27,847	27,847
Percentage not recovered (nodes B2 + C3)/(node D)	6.2%	5.4%

Table 5.3: *Tabulation of results for* TIMIT*, in terms of pronunciations.*

Category	TIMIT
Failed Letters, Invented SROOTs	12.4%
Failed Letters, Completely	3.4%
Failed Phones, Letter Information	1.0%
Failed Phones, Coerced Stress	2.4%
Failed Phones, Coerced Stress and Invented SROOTs	8.5%
Failed Phones, Completely	8.0%
Passed, Completely	63.9%
Total (pronunciations)	2500
Percentage not recovered	11.4%

there are not enough data points to make such an extrapolation. If we can add the new information from COMLEX into ANGIE, then parse coverage should increase. ANGIE has proposed at least 2,401 new SROOT morphs for the words in COMLEX, even after the morphs from TIMIT are added to its knowledge base. This is almost 50% of the size of the current morph lexicon! On the other hand, the number of new SROOT morphs is more than an order of magnitude smaller than the total size of the COMLEX lexicon.

5.2 Evaluation of Accuracy, in TIMIT

While coverage is an important feature of our morphs, accuracy of the decompositions is also very much desired. It is difficult to measure the accuracy of the COMLEX transcriptions, as any adequate sample size would require intensive effort to analyze. As a *very* rough and informal estimate, we can

Table 5.4: *A rough measure of accuracy derived from the* TIMIT *corpus. We measure accuracy by considering the mostly likely* ANGIE*-generated morph, and comparing it against a hand-written transcription. Those sequences that are identical are counted as correct. Phoneme accuracy is computed in a similar fashion.*

Category	Sample Size	Morph Accuracy	Phoneme Accuracy
Failed Letters, Invented SROOTs	311	82.3%	92.6%
Failed Phones, Letter Information	28	39.3%	60.7%
Failed Phones, Coerced Stress	59	49.2%	69.5%
Failed Phones, Coerced Stress and Invented SROOTs	216	45.4%	68.5%
Passed, Completely	50	78.0%	80.0%

measure TIMIT's parse accuracy, collecting the statistics expressed in Chapter 3. These statistics are those tables, such as Tables 3.6 and 3.7, which tabulate the similarity of ANGIE's generated morph decompositions between those from an expert. These results are summarized in Table 5.4. The accuracy is the percentage of words whose top (or only) morphological/phoneme decomposition matches that of the expert.

This information is based on very informal analysis, but it does match with the perceptions of the author. All of these observations were informally observed in COMLEX as well. The decompositions that were recovered from failed phone parses were generally not as good as those from the failed letters or passed words. Strangely enough, the failed letter words that were allowed to invent their own SROOTs seemed to have fewer errors than those that passed without any back-offs!

It could be the case that the failed letter set is a self-selecting group – they definitely need a new morph, and thus are able to create one that is a best fit. The words that pass through the procedure may have some words that find their optimal morphs, but there might also be some that really need a new morph, but masquerade behind a less-than-perfect morph, which is good enough to keep the word from failing a step, but is not the best morph available. For this reason one future suggestion might be to screen words based on the probabilities of the parse tree, instead of parse failures. Low scoring parse trees probably are not being aligned optimally. These words could be fed into a mechanism to back-off and try variations such as the compound word smoothing, SROOT invention, or stress coercion.

5.3 Interpretation of Constraints

The combination of the letter and phone parsing steps may be described in terms of the "generate and test" paradigm. The first letter parsing step generates possible morphological transcriptions. Then these morphological transcriptions are tested against the phonological information, and are

either accepted or discarded.

One feature not noted until now is the number of different morphological decompositions, or alternate morph sequences per word. (This is different from the number of morphs per morphological decomposition.) The distribution of these sequences can range from zero (a failure) to four. (Four is the maximum number of alternative morphological decompositions allowed. It can be set to an arbitrary number.) The number of alternates should be related to how constraining a process is – one would expect the average number of alternate morph sequences per word to be larger after the letter parsing step than after the phone step, since many of the decompositions are pruned.

Figure 5-1 graphically illustrates this measure of constraint. The morph distributions for all 2,500 words in TIMIT are plotted after letter parsing and then after phone parsing. A word that fails to parse has zero alternate morphs. The average number of alternate morphs sequences drops from 2.8 to 1.1.

The distribution is also available for the 29,683 pronunciations in COMLEX in Figure 5-2. The trend is similar, but not as pronounced as in TIMIT. We suspect that this is entirely due to the strictness of the phone rules. Since the automatically generated rules are more lax than hand written ones (about 850 in number for automatic compared to 370 for hand-written), more morph sequences are allowed to pass the phonetic parsing step. Here the average number of morphological decompositions almost halves, from 3.0 to 1.7.

One feature of both these histograms is that the distribution of alternate morph sequences shifts, demonstrating the constraining nature of each parsing step. In the letter parsing step, the distribution is heavily skewed toward many alternates. The constraining feature of phone parsing is illustrated by the shift of the distribution towards only one morphological decomposition.

5.4 Hand-Written versus Automatic Rules

In order to parse COMLEX into our framework, we need rules to specify the allowed transitions from ANGIE's phoneme set to COMLEX's base units. We have developed a procedure to automatically derive rules with the help of the ALIGN program [2]. A description of this procedure is given in subsection 4.5.1.

Hand-written rules may be more restrictive and accurate, since they are written by a human, who has some knowledge of possible generalities to include as well as some restrictions to apply. However, automatically generated rules, which allow almost all possible alignments, should cover many more words almost instantly. Creating a set of automatically generated rules takes under five minutes. We estimate that an expert writing the rules by hand may require several hours or even days, for a corpus as large as COMLEX.

The trade-off is between coverage and accuracy. Our procedure can be fit into the "Generate and

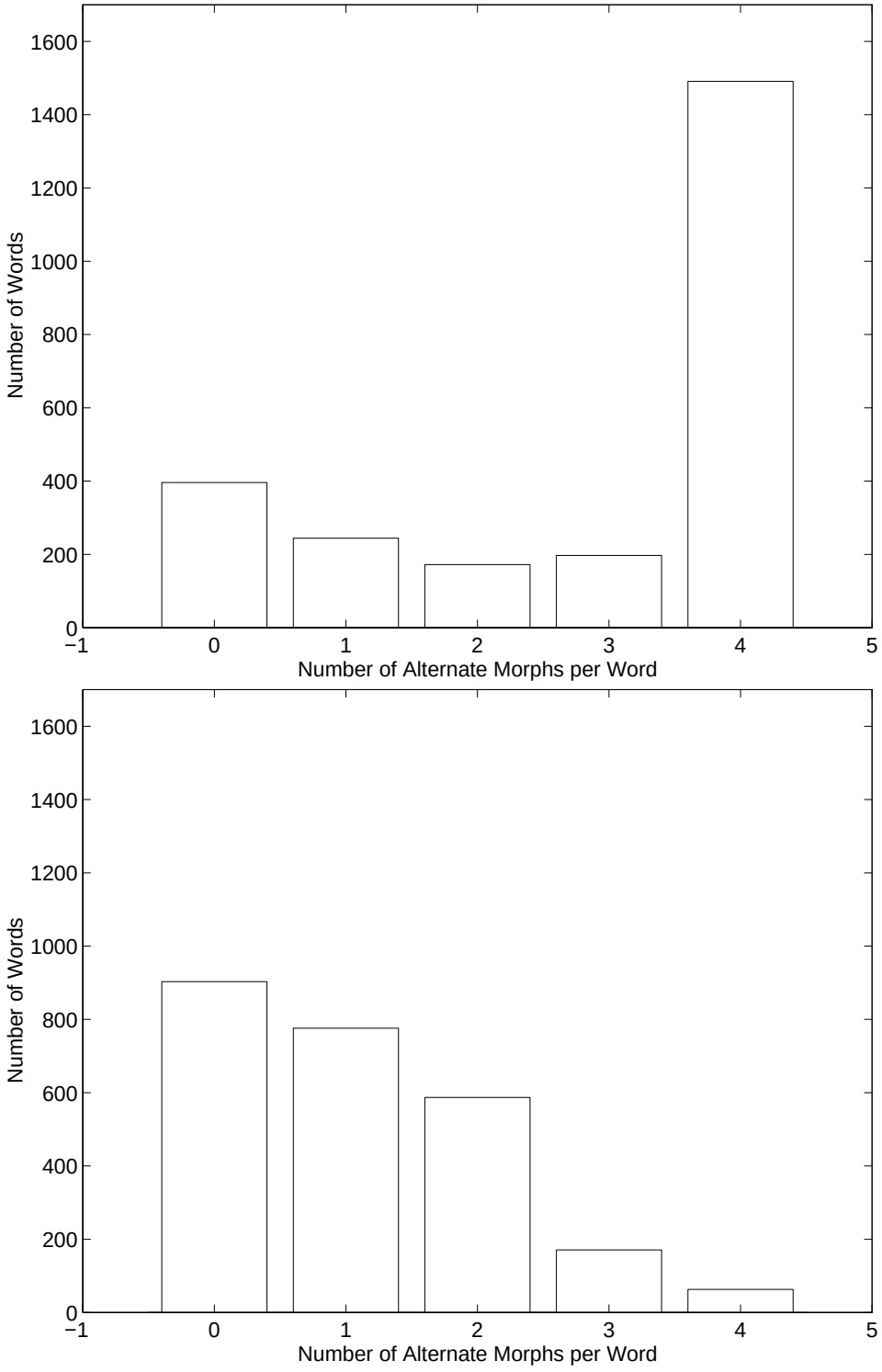

Figure 5-1: *Two histograms are plotted for the number of alternate morphs per word, after letter parsing and then after phone parsing, for all 2,500* TIMIT *words.*

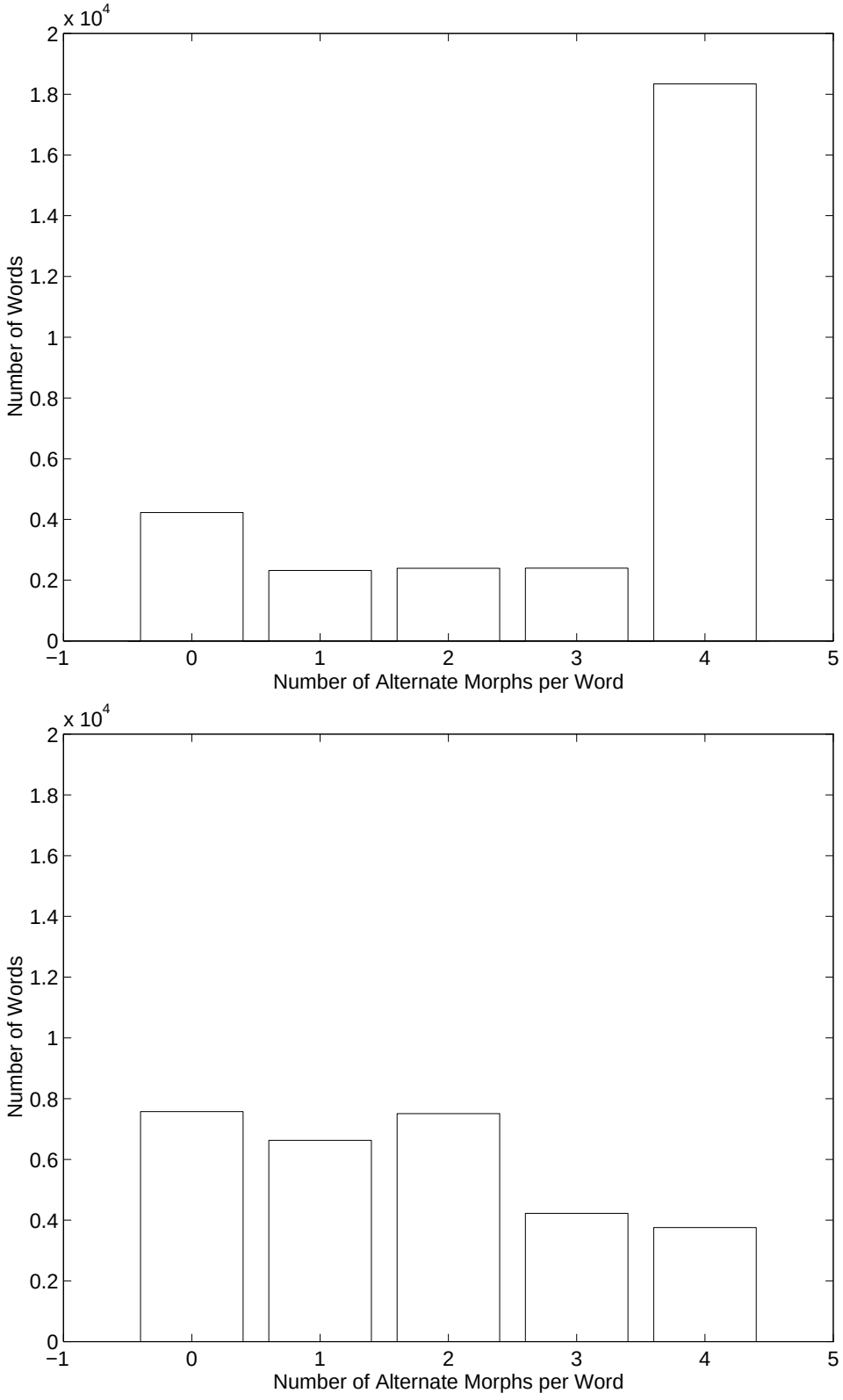

Figure 5-2: *Two histograms are plotted for the number of alternate morphs per word, after letter parsing and then after phone parsing, for the 29,683* COMLEX *pronunciations.*

Table 5.5: *Comparison of morph decompositions generated from automatically generated rules, compared to those generated from hand-written* TIMIT *to* ANGIE *phoneme rules, per pronunciation.*

Number	Percentage	Type
1,351	79.6%	Have one morph transcription which is identical to that from hand transcribed rules
86	5.1%	The most likely of multiple morph transcriptions is identical to that from hand-written rules
144	8.5%	One of the morph transcriptions matches one of those from hand-written rules
2	0.5%	The segmentations of the transcriptions are the same.
105	6.2%	Only parsed with automatic rules (failed with hand-written rules)
1698	100.0%	Total

Test" paradigm. The letter parsing step of our algorithm generates multiple morph sequences, but then the more restrictive phonological parsing step screens out most of the incorrect morphs. The graphs in Figures 5-1 and 5-2 demonstrate this filtering in terms of morphological decompositions, as the distributions shift to the left as phonetic parsing is applied.

Relaxing this second step of parsing ANGIE phonemes with the target phone set can allow substandard morphological decompositions to pass through. We would like to evaluate the differences in performance when these two different types of rules are used. A metric to keep in mind is that the estimated number of rules in the automatic set is 852, while the hand-written version only has about 373.

To compare the two sets, we focus on the test of phone parsing, where the letter parsed 2,104 TIMIT words are parsed with TIMIT phoneme constraint, as shown before node C in Figures 3-2 and 3-3. We have already obtained morphs for 1,597 of these words (1,598 pronunciations) using hand-written rules. This procedure is redone, this time with automatically generated ANGIE phoneme to TIMIT phoneme rules.

The results are summarized in Table 5.5. When we parse with automatically generated rules, 1,698 pronunciations obtain phonetic transcriptions. These words can be divided into three classes. The first group of words have one of their morphological decompositions the same as one from the hand-generated case (1,351+ 86 + 144 = 1,591, or 93.7%). The second group contains the 105 words which do not get morphological transcriptions with the hand-written rules, but do with the automatic ones. Finally the third group consists of two words, "fingerprints", and "flowerpot", which only match in segmentations.

The 105 words which are rejected by the hand-written rules but not by the automatic make an interesting set to study. They are either words that should have been rejected (false acceptance), or words that needed extra help from the automatic rules to parse through the framework (false rejection). This set is divided about evenly between these two cases. Many of them are in-

correct, through a stress shift, as in **jo- cu+ lar**. Others have the incorrect pronunciation, as in **/s! t iy+ f/** for "steph." Another common error is that the morphs are segmented in strange ways, such as **ven+ dI =ng**. Some words do appear to be correct, however.

The two words which only match in segmentations are "fingerprints", and "flowerpot." Their morph sequences are **fing+ er print+ =s** and **fing+ =er print+ =s**, and **flow+ er pot+** and **flow+ =er pot+**. Any of these morphs appear to be correct. The reason why the **er** is sometimes exchanged for the **=er** is that somehow the two parses are ranked in different orders, so that when similar columns are pruned (subsection 2.4.2), only one of them survives.

The astute reader will note that five words that parse with hand-written rules do not pass when automatically generated rules are used. These words are "aristocratic," "chestnuts," "elongation," "marshmallows," and "moistened." It appears that all of these words have either a rule that splits one ANGIE phoneme to two TIMIT phonemes, or merges two ANGIE phonemes to one TIMIT. These splits and merges are handled by the automatic generation procedure, but are probably so rare that they were not found in the overlap set.

Overall, the automatic rules appear to be comparable to hand-written rules. 84.6% of the words filtered through these rules have their most likely or only transcription match that from hand-written rules. Very few (less than 6.2%) pass through which have incorrect morphological decompositions.

It takes much more effort to hand-write ANGIE phoneme to, in this case, TIMIT phoneme rules. Depending on the task, allowing a few sub-standard morphological transcriptions to pass through the framework might be acceptable, if one desires high coverage quickly. One example is the case of COMLEX, where hand-parsing all 6,533 overlap words to create a ANGIE phoneme to COMLEX phoneme rule set would be a very time-intensive task.

However, we could balance the coverage and ease of automatically generated rules with the accuracy of hand-written rules. Automatically generated rules could be hand edited to yield an intermediate set that would be more effective then either fully automatic or fully manual rules.

5.5 Consistency of Morphological Decompositions

Besides accuracy, the morphological decompositions for words are entirely self-consistent. Table 5.6 shows two instances of this. The ABH transcriptions that all contain "motion" have the same root decompositions, including stress. A similar conclusion can be drawn for the words containing "support", which were found in the COMLEX morph lexicon.

This is one advantage of using a computer to derive sub-lexical information from words. A deterministic algorithm should always return self-consistent transcriptions such as those in the Table 5.6, regardless of whether it has been days or years since it has last transcribed lexicons. The same cannot be said of a human transcriber.

Table 5.6: *Examples of consistent morphological decompositions for words containing the fragment "motion", and those for words containing "support."*

Word	Morphs
commotion	**com- mo+ tion**
demotion	**de- mo+ tion**
emotionalism	**e- mo+ tion =alism**
motionless	**mo+ tion =less**
promotional	**pro- mo+ tion =al**
promotions	**pro- mo+ tion =s**
supportable	**sup- port+ =able**
supportive	**sup- port+ -ive**
supports	**sup- port+ =s**
unsupportable	**un- sup- port+ =able**
unsupported	**un- sup- port+ =ed**

5.6 Chapter Summary

In this chapter we analyze various dimensions of our parsing paradigm. The first matter we study is whether our morphs are a good sub-word representation, and if they can be extracted accurately and consistently from large corpora. Based on the decreasing percentage of "unrecoverable words," or words that fail to parse even in the face of back-offs, we can hope that as ANGIE's knowledge base expands, morph coverage will increase, perhaps even to 100%.

Accuracy is the other vital issue that must be addressed. Hand analyzing the transcriptions, whether they are from TIMIT or COMLEX, is an arduous task. Instead we loosely estimate the accuracies using the 390 letter failed and 507 phone failed words we hand-wrote for TIMIT as a guide. Based on these informal estimates and the author's own observations about the data, it appears that words that fail when phone parsing generally have worse back-off generated transcriptions than those which pass through the framework, or letter failed words. Another intriguing observation is that words allowed to invent their own SROOT morphs seem to have fewer errors than the words which pass through the framework without failing.

The amount of constraint the letter and phone parsing steps apply to the data is readily seen in a histogram of the number of alternate morphological decompositions. In the letter parsing step, many alternate morphological decompositions (up to four) are created, with only orthographic and ANGIE constraint. When the morphs are constrained to parse with phonemes, the average number of alternate decompositions drops appreciably. This is true both for TIMIT and COMLEX. The shift in decompositions is not as abrupt for COMLEX, probably because the phone rules are relatively lax.

The difference in performance resulting from using hand-written versus automatic rules is explored next. When automatically generated ANGIE phoneme to COMLEX phoneme rules are substituted in place of hand-written rules, an additional 100 words filter through the phonetic parsing

step, totaling 1,698 passed words instead of 1,598. 84.7% of the automatically generated rule words match those that came from hand-written rules. Thus automatically generated rules appear to degrade performance somewhat, since this set is more than double the size of the hand-written set. We can hope that the COMLEX transcriptions, which currently employ automatically generated rules, would have comparable performance to that of TIMIT if the time were taken to create and then use hand-made rules.

The last issue touched upon in this chapter is consistency of transcriptions. Two examples are offered to show how sub-word patterns are stored and applied by our procedure. Words containing the sub-word sequence "motion" all have similar decompositions, as do those matches for the word "support".

Chapter 6

A Tool for Morphological Transcription

6.1 Motivation

In the analysis of TIMIT in Chapter 3, it was necessary to hand transcribe the morphological decompositions of over 800 words. A morphological decomposition contains rich linguistic information, including stress, morphology (in terms of affixes and roots), and phonemes. Morph transcription is a much more complex process than transcription of other units, since multi-dimensional constraints apply. For example, the morph transcription must match the spelling of the word, and the associated phonemes must produce an acceptable phoneme sequence for the word. Furthermore, the morphs must be segmented to represent the syllabification and stress patterns of the word. This chapter describes a tool which helps a person transcribe morphological decompositions efficiently and effectively.

6.2 Description

To aid in this procedure, a tool has been implemented that integrates these different linguistic levels into one system. There are four main features which aid transcription. The first feature is that either words, morphs, or phonemes can be indexed using either a regexp, alphabetical, or spelling-equivalent search. Secondly, easy access to other word examples is provided to help the transcriber keep transcriptions consistent. Third, transcribers can listen to the phonemic transcriptions using the Dectalk synthesizer, to acoustically verify phonemic transcriptions. Finally, the association between morphs and phonemes is automatically generated so that the relationship between word, morphs, and phonemes is readily apparent.

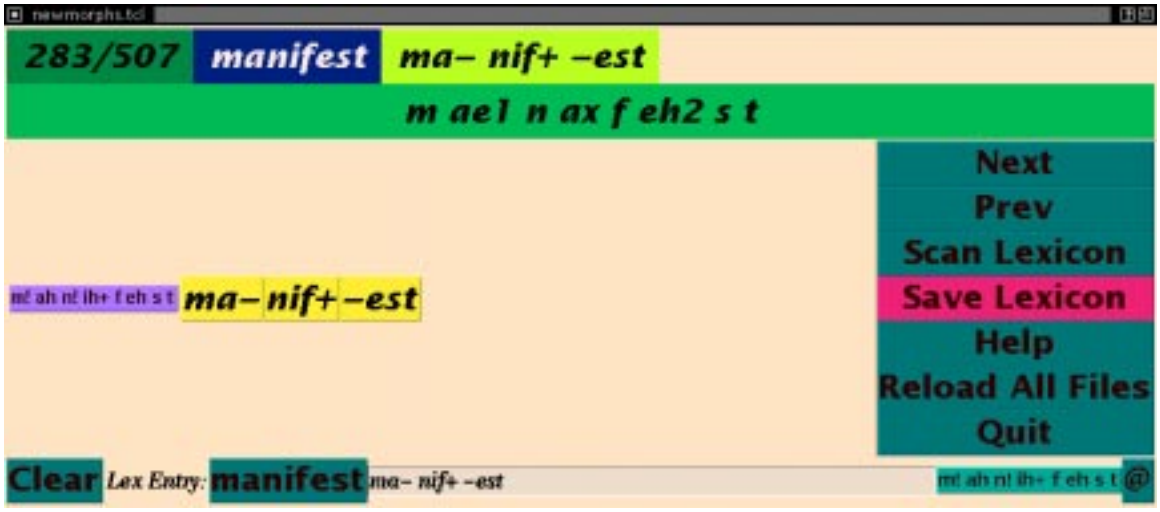

Figure 6-1: *"newmorphs.tcl" is the main window through which morph transcriptions are entered.*

6.3 Operation

The tool takes as input a word-morph lexicon and morph-phoneme lexicon. The word-morph lexicon usually contains proposed morphs that need to be edited. A list of words is a valid substitute if no proposed transcriptions are available. Optional arguments include a file that contains the phonetics for each word.

There are a total of six different modules. The main module, labeled *newmorphs.tcl*, is the main entry area for new morph pronunciations. The other modules, labeled *fwm*, *fmp*, *morphfinder*, *scanlex*, and *sim_spell* are all used to help the transcriber search for morphs or words. The tool is mouse-driven to reduce keyboard activity.

The windows all communicate with one other. There are two types of communication. A morph selected in one module can be appended to the morph transcription in the "Lex Entry:" Box of *newmorphs.tcl*. The other type of communication also involves morphs. When morphs are updated in the morph-phoneme lexicon, all of the other modules are updated with this knowledge.

The next sections describe each of the modules.

6.3.1 Newmorphs.tcl

newmorphs.tcl, the main window, is depicted in Figure 6-1. Information shown in this window includes the word being transcribed, the proposed morphological decomposition given in the word-morph file, and a phonetic transcription, if available. Also in the top part of the window is a count of the total number of words in the lexicon, along with the current index.

The row of buttons and the entry box at the bottom window allow one to enter morphological decompositions manually. "Clear" clears the entry box. The button to the right of "Lex Entry:" is

the word button (in this case, "manifest"). Clicking this returns the Dectalk pronunciation of the word. Next to that is the lexical entry box, into which morphological decompositions can be typed.

If a morphological sequence is typed into the box, the button to the right of the entry box is updated to produce the associated phoneme sequence of the morphs. This feature can handle multiple morphological decompositions (separated by "@" signs) as well as multiple pronunciations for a single morph. Clicking this phoneme button sends the phoneme string, translated into Dectalk phones, to the synthesizer. The synthesizer then returns the acoustic signal of the phonemes to the user.

The buttons in the center represent the morphs in each of the proposed morphological decompositions. These proposed decompositions are found in the word-morph lexicon that is loaded at run-time. To the left of each morphological decomposition is the phoneme equivalent. These phoneme buttons can also be clicked for the acoustic pronunciation. The morph buttons, if clicked with the middle mouse button, cause the associated morph to be appended to the morph sequence already in the lexical entry box. Using the right button instead causes the particular morph, with its phoneme sequence, to be highlighted in the *fmp* module. This feature greatly reduces the amount of typing that must be done.

The buttons on the rightmost side serve other high level functions. "Next" and "Prev" go to the next or previous word. "Scan Lexicon" pops up the *scanlex* window which is described in detail in subsection 6.3.5. "Save Lexicon" saves the current word-morph lexicon, with any changes. When changes have been made to this lexicon, this button changes color to alert the transcriber. When this button is invoked, a window appears with the morph lexicon to be saved, so that the transcriber can browse through before saving. After the save is complete, the "Save Lexicon" button disappears. "Help" and "Quit" should be self-explanatory. "Reload All Files" reloads all of the original input files, in case changes have been made to them outside of the tool since the tool was initiated.

The morphological decompositions are also listed in the center of the window, as buttons. Clicking on these buttons adds the morph to the lexical entry box, which is at the bottom of the window. Buttons to the left show the phoneme sequence of the morph transcription, which is automatically generated from the morphs and the morph-to-phoneme dictionary. Clicking on this button sends the Dectalk translation of the phonemes to the Dectalk server, and a synthesized pronunciations is returned.

6.3.2 Fwm

This window, shown in Figure 6-2, keeps track of what word is being transcribed, along with past transcriptions and words to be transcribed. They can be searched alphabetically. If a word-morph entry is double-clicked, the *newmorphs.tcl* window displays that word so that it can be transcribed. This indexing method is much more efficient than pushing "Next" or "Prev" several times.

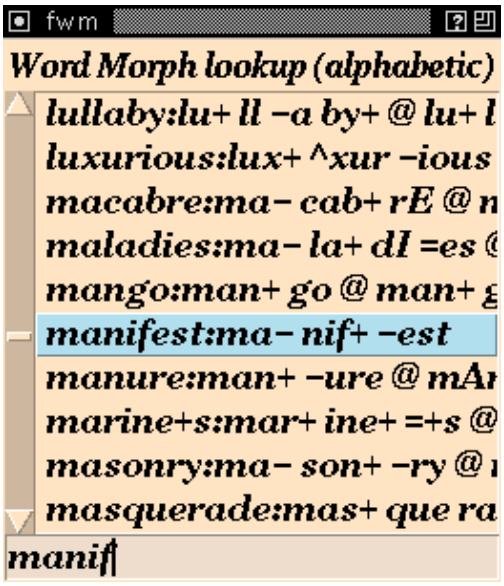

Figure 6-2: *"fwm" keeps track of all the words and their morph transcriptions.*

6.3.3 Fmp

fmp contains a list of all the morphs with their phoneme transcriptions, as illustrated in Figure 6-3. They can be searched alphabetically. If a morph entry is double-clicked, the morph is appended to the morphological sequence in the lexical entry box in *newmorphs.tcl*.

There is also a place to add new morph pronunciations. A morph and its phonemes can be entered in the boxes under "Insert New Entry." Since this tool does not provide a facility for morph deletion, the module must robustly manage errors from the user. There is one constraint that the new morph must definitely obey. If a morph is labeled as a stressed root morph, the phoneme sequence must contain a stressed vowel phoneme. This simple feature in practice has screened out many incorrect morph transcriptions. It is also possible to use the feature to only allow morph sequences which match the spelling of the word. This capability was discarded because it was not as useful in practice.

Once the new morph and phonemes are added, all other modules that deal with morphs are updated to include this morph. Before the morph is added, it along with the phonemes can be sent to the Dectalk Server for an acoustic confirmation by pressing "Listen."

The button "New Guys" is used to list all the morphs that have been added to the morph-phoneme lexicon since it was last saved. The phoneme transcriptions are included.

6.3.4 Morphfinder

While *fmp* (Figure 6-4) is used to search morphs alphabetically, *morphfinder* can search for them using a regexp search. This can be useful for looking up all morphs that, in this case, contain the

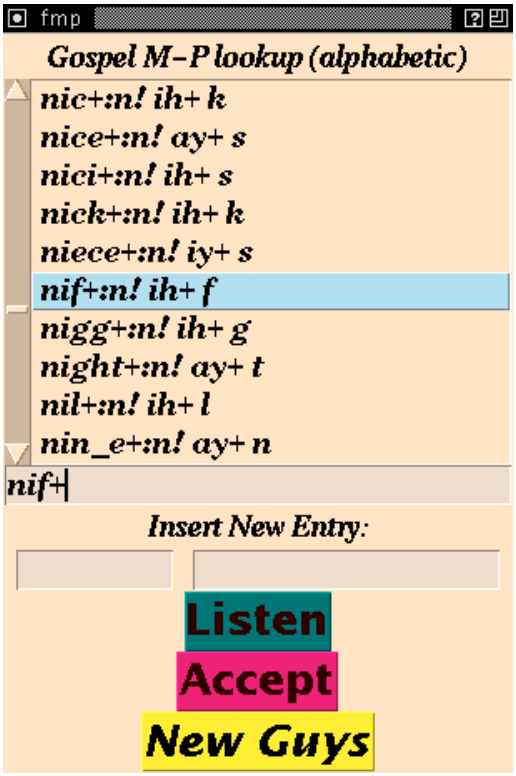

Figure 6-3: *One can search "fmp" for existing morphs, or add new ones.*

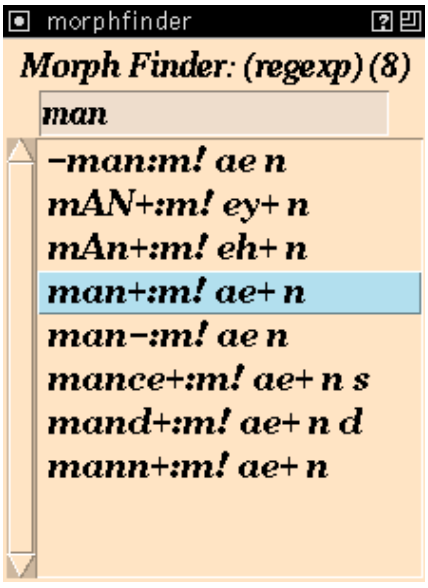

Figure 6-4: *"morphfinder" is a utility to search morphs using a regexp search.*

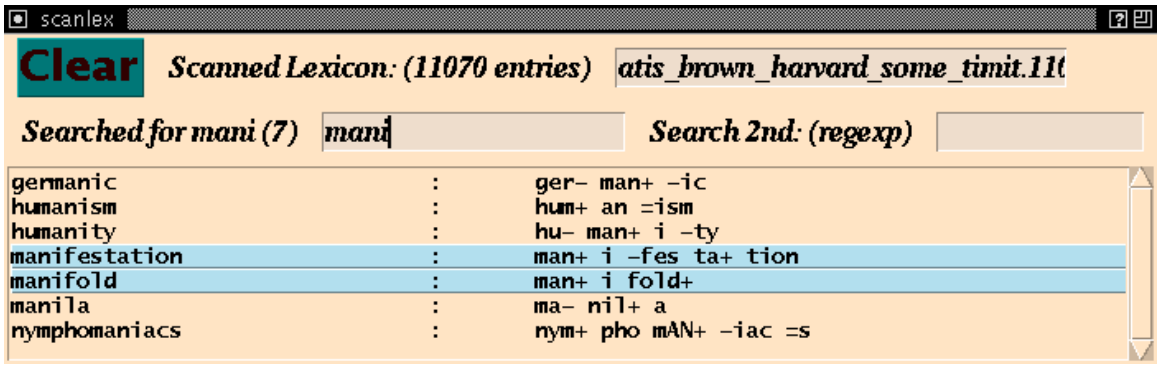

Figure 6-5: *A word-morph or morph-phoneme lexicon can be regexp searched on either field with the "scanlex" module.*

fragment "man". Similar to *fmp*, double-clicking any of the entries adds the morph to the entry in the entry box in *newmorphs.tcl*.

6.3.5 Scanlex

A *scanlex* window is shown in Figure 6-5. *scanlex* is used to load any word-morph or morph-phoneme lexicon. The lexicon can be searched using a regexp search on either of its two fields. It is very useful when trying to create a consistent transcription for a word, for it can be used to search for words that have similar spellings. The second field, which searches morphs for a word-morph lexicon, can be used to find word instances of a particular morph.

If a morph-phoneme lexicon is loaded, *scanlex*'s features are very similar to *morphfinder*, except that either the morphs *or* phonemes can be searched using a regexp search.

6.3.6 Sim_spell

One of the constraints a morph transcription has to satisfy is that the concatenation of the morphs must produce the word. Thus, it is very helpful to find all instances of morphs which have a particular spelling. *sim_spell*, shown in Figure 6-6, provides this type of search. The *morphfinder* module can also search for morphs, but it is not as efficient. *sim_spell* returns only those morphs which, if tags are removed, return the spelling of the desired morph. Again, double-clicking on entries appends the morph to the morph sequence in *newmorphs.tcl*.

6.3.7 Additional Features

There is a feature similar to the "Save Lexicon" button. If the morph-phoneme lexicon is changed by adding a new morph, a new button in *newmorphs.tcl* appears. Named "Save Gospel" for historical reasons, this button lets the transcriber know that the morph-lexicon has been altered and must be

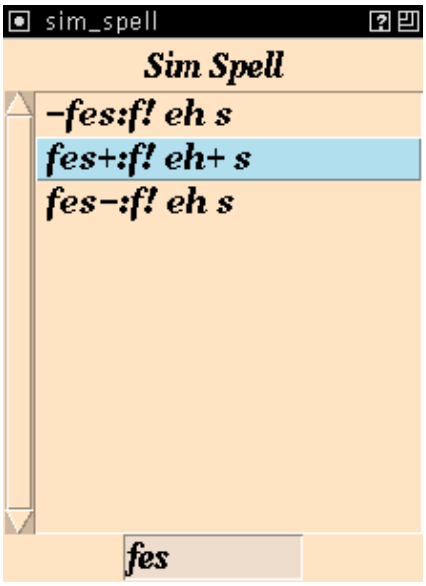

Figure 6-6: *All morphs that can generate a specific spelling are returned using the "simspell" search.*

saved away. Like the "Save Lexicon" button, a new window appears with the lexicon to be saved, so that the transcriber can check the information before saving.

6.4 Implementation

This tool was developed while hand-writing the 390 letter failed and the 507 phone failed TIMIT words, so that only necessary, practical functions were added. Because word-morph and morph-phoneme lexicons can get very large, implementational details become very important, especially with those parts that deal with search.

A traditional search through a sorted list can take O(n) time to complete. When dealing with a lexicon on the order of a thousand entries, this factor transforms to a very noticeable delay. Hence, all alphabetic searches are implemented as a binary tree search, which takes O($log_2 n$).

Another implementational feature involves the *sim_spell* module. This module caches results for a given query, and returns them only if the morph-phoneme lexicon has not been changed.

6.5 Evaluation

As mentioned before, this tool has been used to transcribe 897 words, adding new morphs as necessary. Using this tool is a large improvement to the way morph transcriptions had formerly been generated. Transcription used to take place via an emacs buffer, with windows for the word-morph lexicon to be transcribed, a reference morph-phoneme lexicon, and a reference word-morph lexicon. Any searches were accomplished by using the emacs search command, or `grep`-ing lines from the lex-

icons at the shell command prompt. By incorporating all the different source files and adding various search mechanisms into one tool, the task of transcription has been made much easier. Many of the tasks are mouse-driven (clicking on a button adds a morph and updates the phonemic transcription automatically) which reduces typing strain.

6.6 Future Work

One feature that would be very helpful is being able to load in the ANGIE rules, to ensure that morphological decompositions will parse into the ANGIE framework. One problem that surfaced when this tool was first used was that UROOT morphs were invented with a coda phoneme. According to the high level rules, a UROOT can not end in a consonant. All of the words transcribed with these incorrect UROOTs had to be rewritten, and the offending morphs had to be removed from the lexicon.

6.7 Chapter Summary

This section describes a tool that has been developed to aid hand-transcription of words into our morphs. This tool is useful because morph transcription is a much more complicated process than usual phone or phoneme transcription. Not only must the morphs match the phonemes of the word, but also accurately represent the syllabification, stress, and spelling.

Four capabilities of this tool include being able to search words, morphs, or phonemes using various methods. Words with similar sequences can be found so that transcriptions can be kept consistent. A word or phoneme string can be synthesized for acoustic confirmation of a transcription. Finally, the relationship between a word, its transcribed morph sequence and phonemes is automatically available.

Six modules, each with its own specialization, are used to implement the above features. Morph sequences are entered and analyzed in the *newmorphs.tcl* module. *fwm* provides an index to all the words that are transcribed. Morphs can be searched in three different ways in the *fmp*, *morphfinder*, and *sim_spell* modules. *fmp* also allows entry of new morphs. Finally, *scanlex* allows either field of a lexicon (either word-morph or morph-phoneme) to be searched with regular expressions. A helpful feature is a flag to indicate when either the word-morph or morph-phoneme lexicon have been altered, and need to be saved.

Because this tool searches large lexicons, the searches need to be optimized. One solution is to search using a binary-tree algorithm. Another optimization involves using a cache to store query results.

This tool has been used to transcribe over 800 words into their morphological decompositions. It is an improvement over the way words used to be transcribed. All of the information is integrated

into one framework, and most tasks are mouse-driven.

One feature that would be helpful is closer integration with the ANGIE framework, in terms of the context-free rules. Sometimes a transcription is generated which cannot parse into the ANGIE framework. One improvement is to add the rule information to screen out morphs which do not fit in the ANGIE framework.

Chapter 7

Conclusions and Future Work

7.1 Thesis Summary

This thesis introduces a new semi-automatic procedure for representing words in terms of a sub-word representation, which we have named "morphs." Sub-word modeling, which includes morphology, syllabification, stress, and phonemes, has been shown to improve performance in speech applications. This has motivated us to use these sub-lexical units in terms of our "morphs" to represent words in the English language.

We would like to know if our representation is extensible, and if it is possible to automatically or semi-automatically extract these sub-lexical units from large corpora. Thus we have proposed a procedure that can extract these morphs accurately and efficiently. Then we evaluate both the procedure, as well as how well our morphs can serve as a basic unit for capturing sub-lexical constraints.

ANGIE is a system that can parse either spellings or phonetics into a probabilistic hierarchical framework. We used this formalism to generate and test our morphs. Since this system is a central feature of this thesis, an entire chapter is dedicated to its description and operation.

We decided to develop our procedure based on a medium-sized corpus known as TIMIT. We began with a grammar that had been developed and trained on a corpus we call "ABH" (a combination of corpora including ATIS, Brown, Harvard List), consisting of 9,083 words. We then applied the knowledge we had gained from ABH, both with and without that from the TIMIT experiment, to the much larger COMLEX lexicon[1]. In this way we tested how well a set of morphs derived from 10,000 words can be applied to a much larger set of 30,000. If morphs are a good representation, good coverage should be attainable.

We have some encouraging signs that our set of morphs is large enough to encompass most or

[1] We omitted proper nouns and abbreviations.

all English words. Coverage of TIMIT is 88.6%, and for COMLEX it ranges from 93.8% to 94.6%. The parse coverage of our procedure is quite good, considering the large size of the COMLEX corpus. The accuracy of the morphological decompositions is reasonable as well. According to an informal evaluation, morphological decompositions of words that pass through both letter and phone parsing steps have a 78.0% probability of exactly matching the expert transcription. Of course this metric does not take into account alternate decompositions which may also be correct, or more consistent than human generated ones.

We performed an analysis and comparison of the experiments performed on TIMIT and COMLEX. The topics covered include constraint, hand-written versus automatic rules, and consistency of morphological decompositions. Constraint can be measured by the average number of alternate morphological decompositions per word. The average number of morphs generated from the letter parsing is about three, both for TIMIT and COMLEX. After parsing with phones, this figure drops from 1.1 for TIMIT to 1.7 for COMLEX. Automatically derived rules provide a quick alternative to hand-written rules, with more coverage, but at a price of some performance loss. Morphological decompositions produced by our procedure also appear to be self-consistent.

A separate chapter gives a brief description of a new tool that was developed, in parallel with this thesis, to simplify the task of morph transcriptions. Morph transcription is a more difficult task than phone or phonemic transcriptions, simply because constraints have to be satisfied on more than one level. The morph spelling must form the spelling of the word, and syllabification must be correct, as well as the combined phonemics. Also, morphs with similar spellings but different pronunciations must be distinguished through selected capital letters, such as the examples **nat+** and **nAt+**. This tool aids the transcriber by providing easy access to many different sources of knowledge.

7.2 Improvements to the Algorithm

This thesis describes many inventive ways to improve or at least try to improve the efficiency of the morph extraction process. We have used such maneuvers as allowing SROOT invention, smoothing for compound words, and forcing stress patterns. Potentially we could also adjust the maximum number of theories allowed in parsing (forty), or the maximum number of alternative morphological decompositions produced by letter parsing (four), to improve coverage and accuracy. We would ultimately like to build a system that can combine all of these mechanisms into one integrated framework, allowing the user to set some of the free parameters. Then, the system could automatically apply one on these procedures, depending on criteria such as failure, or parse probability. We could imagine a scenario where the system fails to parse the letters of a word. The system could back-off, and parse letters again, but without morph constraint. If the word passes, then it could be assumed

that an SROOT morph is missing, and to parse again with the SROOT invention. On the other hand, if the word fails to letter parse, even without morphs, it assumes that a letter rule is missing. Then the system could try to discover this missing rule, perhaps by applying a much less constraining set of rules, such as the automatically generated rules we have previously described. Once the missing rule is found it could be added to the rule set, with or without human supervision.

Even if this grander goal cannot be immediately realized, there are other more simple improvements that can be made. In these experiments, inaccurate morphological decompositions are marked by whether they fail to parse. There is no way to catch those invalid morphological decompositions that manage to pass through the framework. A potentially better metric for evaluating the accuracy of a morph sequence is by the parse tree probability, rather than whether it fails to parse. This metric serves the purpose of a rejection threshold. We could refine this metric, depending on which set is being evaluated. This could be accomplished by gathering the average probability on each set of failed groups (those that only pass with invented SROOTs, or stress coercion, etc.).

Another avenue to explore involves the automatic generation of ANGIE phoneme-to-COMLEX or TIMIT phoneme mappings. In the first step, an obvious set of mappings between the phone sets could be provided manually. In parallel, the automatic generation procedure creates a general set of rules. The words that fail with hand-written rules but parse due to the addition of automatically generated rules are then examined. The "mis-firing" rules that allow incorrect alignments to parse through ANGIE are removed. This can ensure a more restrictive, and perhaps more accurate, set of semi-automatically generated rules.

For example, in section 5.4, automatically generated ANGIE phoneme-to-TIMIT phoneme rules are applied to the 2,104 letter passed TIMIT words. 1,698 (80.7%) pass, compared to 1,597 (75.9%) when hand-written rules are employed. Table 5.5 shows that a set of 105 words pass through which would have been rejected by the hand-written rules. The words in this set that should not have passed the phonetic parsing step may be analyzed, in order to determine which rule allowed them to pass. Then that rule can be deleted from the set.

The procedure could be reversed to gradually expand a set of hand-written rules. A very basic set of mappings can be included in the hand-written rules. Then the automatically generated rules can be used to parse the words which do not parse with the more restrictive hand-written set. ANGIE provides a graphical interface to view parse trees. Viewing these parse trees can help humans discover new rules, by means of this visual model. We could implement a graphical user interface to select rules from the tree, and add them to the rule set automatically.

It is likely that the morph (SROOT) inventory of English forms a large set with a distribution that includes a long tail of rare or uniquely occurring morphs[11]. The morphs could potentially be divided into two sets, one of which is a relatively small set used frequently on many words, and the other composed of uncommon morphs which usually only apply to one word (and possibly its derivatives).

In this case it would be nearly impossible to enumerate the entire set of morphs. However, we could build a back-off into our system which allows SROOTs to be invented to accommodate the rare set, in the case of parse failures, or low probability according to the grammar. Given our rough measure of accuracy (82.3%) on invented SROOT morphs, performance should not degrade, and might even improve.

7.3 Improving our Knowledge Base

Part of our work has led to the transcription of over 26,000 different words from the COMLEX corpus. Adding these transcriptions to our base ABHT lexicon, and then retraining ANGIE's models could lead to even better performance on letter-to-sound applications for general English. However, even though we are assured of reasonably high transcription accuracies, it would be good to confirm that these data are "clean," and screen out sub-standard morphological decompositions that should not have passed.

One of the advantages of our procedure is that the morphological decompositions are generally consistent among similar words. Examples are shown for the word families containing the word "support" and those with the word "motion," as shown in Table 5.6. We can extract correct morphological decompositions by taking advantage of this consistency. We can find a word in our clean ABHT lexicon, such as "support", and then look for words in the 26,000 set with this same sub-sequence. Those that have similar morphological decompositions can be automatically accepted into a new base ABHTC, or "BATCH" lexicon.

7.4 Phone-to-Phone Translation of Corpora

Available corpora are all transcribed in different phone sets, so that it is problematic to merge them to obtain one very large lexicon. Research groups must translate the phone set to some other phone set with which they are more comfortable. The method of phone-to-phone translation is usually accomplished by simple methods such as rewrite rules, which might include some context. However this method would not harness the power of higher level transcription conventions.

We can use the procedure defined in this thesis to parse these words into the ANGIE framework and obtain morphological decompositions. A dictionary of morph-phone transcriptions can be created semi-automatically beforehand in the target phone set, using some of the techniques developed in this thesis. Then the transcriptions of the words in the target phone set are acquired through direct dictionary lookup of the morphs. Morphs would then preserve higher level transcription conventions across lexicons.

7.5 Adding Morph Features

It might be possible to add features to morphs to provide higher level linguistic information about a word [11]. In a speech recognition system, such information could be passed to a language model, to facilitate or prune the search.

Some examples of higher level information include part of speech, negation, or tense. However, since this information is derived just from the spelling and phonetics of a word, we cannot expect the proposals to be correct. The certainty of the features could be tied into ANGIE's probabilistic framework to provide a confidence-measure metric. For example, if a word ends with the morph suffix -ANCE we could assume that the word is a noun. Negation is easily detected by the "un" prefix, as in "undo". A simple application of this rule, without higher level references, could breed errors, as for the word "uncle." ANGIE, with its storehouse of morphological information, would not make this error. In the segmentation **un+ -cle**, the **un+** could not be treated as a negated prefix, since then there is no root morph for it to modify. ANGIE can already detect inflectional suffixes such as "ed." It is also possible that stress patterns are related to part of speech. For example, there are many noun/verb pairs with the same spelling, but different stress patterns. Examples are "abstract" and "produce".

7.6 Letter-to-Sound/Sound-to-Letter Generation

This thesis shows a way to "absorb" a new lexicon, in a different phone(me) set, into the ANGIE framework. In this process, we enforce our own labeling conventions onto the words of a lexicon, semi-automatically. At the same time, we are extending our own conventions, as we learn new morphs and rules. This procedure (which we have shown to have about 90% coverage) can be applied to all existing lexicons, in order to quickly compile a very large amount of annotated data. This information can then be reused by ANGIE to improve training weights. This, in turn, could improve ANGIE's performance on its original task, letter-to-sound/sound-to-letter generation.

The new morphological feature of ANGIE enables quick and reliable sound/letter generation. Previously, this task was accomplished by parsing either phones or letters into ANGIE, extracting the phoneme layer, and then performing an A* search using the phonemes to generate the other terminal set. Instead, a morph-phone dictionary can be constructed for the target phone set, and then words can be converted to phones by fast lookup of their associated morphs. If the morph transcription of a word does not exist, it can be generated using the procedure developed in this thesis.

7.7 More Exploratory Data Analysis

We can use our procedure to semi-automatically generate morph transcriptions for words, with high accuracy. These morph transcriptions are very rich in linguistic knowledge. With morphs, we automatically obtain a proposed syllabification of a word, in terms of phonemes, but we receive additional features such as stress and orthographic segmentation. There are markers to indicate a syllable might be a prefix, suffix, root, or function word. All of this information could be used in all sorts of analyses, either as an isolated set, or in conjunction with acoustic waveforms.

For example, we mentioned earlier that stress pattern may indicate part of speech for homographs, as for the words "abstract" or "produce". From these two examples we might generalize that when the first of two syllables is stressed, the word in question is a noun. Stressing the second syllable might indicate that it is a verb. With morphs, and a syntactic dictionary such as COMLEX, we can easily determine this statistic for homographs. We could also discover whether this knowledge would improve the likelihood of getting stress correct for other nouns/verbs besides the original homographs.

This linguistically rich information could be used in conjunction with acoustic waveforms to tease out regularities. Chung [1] has found a correlation between durations of phones and pre-pausal words with the aid of our morphs. Also, we already know the very frequent function words have much more reduced pronunciations than other words. The morphological information we provide can be used to find other consistencies like these, which can then be utilized in speech applications.

7.8 Improvements to our Speech Recognizer

This thesis demonstrates a means to semi-automatically annotate large corpora with sub-lexical information. This body of linguistic information, compressed into a simple string realization (the morphs), can help our local SUMMIT speech recognizer improve its performance, by incorporating levels of constraints between the phone and word level.

Using morphs along with ANGIE, transcriptions could be generated "in real time." This can improve performance of a recognizer on unknown words. If an unknown word is detected by a recognizer (say by a low score), the phonetic transcription of the word could be passed to ANGIE to generate a spelling hypothesis. Then the vocabulary could be dynamically updated with this word, providing a seamless method for handling unknown words in speech recognition.

Another scenario is when a user uses our GALAXY system [4] to find bookstores in Cambridge. A list of bookstores, all with names unfamiliar to the recognizer, is returned. Since the recognizer does not know the words, it will be impossible for the user to refer to them by name. ANGIE, with a morphological knowledge base, can remedy this problem by automatically deriving a phonetic transcription for the names, only from the spelling, and then passing that to the recognizer.

7.9 A Pronunciation Server

The previous section describes how ANGIE can be used in conjunction with GALAXY to seamlessly derive pronunciations for words unknown to our recognition system. This idea could be described in terms of a pronunciation server, where queries with the spelling of a word are sent to the server, which then returns a phonemic or phonetic transcription. Since ANGIE can employ any terminal set, the pronunciation server could return its queries in any phone or phoneme set that is desired. This pronunciation server could not only be used in the context of our speech recognizer, but could be coupled with a concatenative speech synthesizer, in order to produce better synthesized speech.

The pronunciation server could be turned inside out, in order to perform sound-to-letter generation, rather than letter-to-sound. One application of a sound-to-letter generation system is a spelling aid. Someone who wants to know a word's spelling could input the pronunciation into our recognizer, or by hand-transcription. The phonological information can be sent to ANGIE, which can hypothesize associated spellings along with a measure of certainty. In order to guarantee accuracy, ANGIE could be linked to a dictionary. One of the strengths of being able to generate spellings, rather than looking them up in a dictionary, is that there is an unlimited coverage. All combinations of roots, with prefixes and suffixes, can be created in the context of ANGIE's morphs. Dictionary lookup restricts the number of words that can be spell checked to only those it contains.

7.10 A New Generation of Recognizers

The traditional method of deriving a word sequence from a phone lattice could be changed in deference to a morph based model, bringing in a new generation of speech recognizers. Morphs can be used to model intra-syllabic constraint. This unit can provide a window of context that is not available on most phone to word recognizers. Also, unknown words can be handled gracefully within the same framework, since they can be created from the set of morphs.

Morph units could be better basic units for spontaneous speech, where over 50% of the speech consists of word fragments. These fragments could be modeled as syllables, which can also model segments of words.

In a preliminary experiment, a morph based recognizer was built using an unsmoothed bigram language model to capture both intra- and inter-word morph constraints. The domain of this recognizer was relatively small, consisting of about 1,300 words, which mapped to 1,700 morphs.

The SUMMIT recognizer using the word models and an N-best approach achieves a total error rate of 5.7%, and a sentence error of 24.8%, with about 8.8 words per sentence. The morph-based model attains a 12.6% error rate, and a 46.7% sentence error rate, with 15.4 morphs per sentence. It is difficult to compare these models since they are based on different units. The example recognized sentence in Figure 7-1 should give a compelling reason why one should prefer morph-based models

```
REF: ...  chance+ of* snow+ mon+ -day in* UNKNOWN ** swit+ zer -land pause2
HYP: ...  chance+ of* snow+ mon+ -day in* BAS+ EL    swit+ zer -land pause2
```

Figure 7-1: *An example where a morph recognizer recognizes the unknown word "Basel."*

to a word-based one. The word based model does not have "Basel" in its vocabulary, and so it misses the word. However, in this morph-based case, the word is hypothesized from a sequence of morphs.

The potential behind a morph-based recognizer is enormous. As recognition units they are not much worse than words in terms of performance, but they are so much more versatile. Not only would they make the idea of large-vocabulary speech recognition possible, they might also serve as better models for spontaneous speech. Although they are very similar to syllables, morphs carry more information, including stress and morphology, as well as a possible spelling representation. All of this information greatly improves the language model perplexity, as well as providing potentially useful information to the acoustic models. It is our hope that the extra layers of linguistic information embedded in the ANGIE framework, coupled with morphs, can one day bring significant improvements to speech applications.

Appendix A

ANGIE Categories

Table A.1: *Sentence layer categories used by* ANGIE.

Layer 1	
Node	Description
SENTENCE	Root node, with an unlimited number of WORD nodes.

Table A.2: *Word layer categories used in* ANGIE.

Layer 2	
Node	Description
WORD	Basic unit for a word.

Table A.3: *Morphological layer categories used in* ANGIE.

Layer 3		
Node	Name	Description
DSUF	Derivational Suffix	Changes the part of speech of a word.
FCN	Function Word	High frequency words are pronounced differently.
FP	Filled Pause	Accounts for p auses between words.
ISUF	Inflectional Suffix	Endings such as plural, past tense, etc.
PRE	Prefix	An unstressed prefix.
SPRE	Stressed Prefix	A stressed prefix.
SROOT	Stressed Root	The first stressed syllable in a word.
SROOT2	Stressed Root	The second stressed syllable in a word.
SROOT3	Stressed Root	The third stressed syllable.
UROOT	Unstressed Root	An unstressed syllable.

Table A.4: *Subsyllable layer categories used in* ANGIE.

Layer 4		
Node	Description	Associated Morphs
ÂBLE	Suffix "able"	ISUF
ÂD	Stressed prefix "ad/ab"	SPRE
ÂL	Stressed prefix "al"	SPRE
ÂLL	Stressed prefix "all"	SPRE
ĈOM	Stressed prefix "com"	SPRE
D̂IS	Stressed prefix "dis"	SPRE
ÊR	Inflectional Suffix "er"	ISUF
ÊST	Inflectional Suffix "est"	ISUF
F̂UL	Inflectional Suffix "ful"	ISUF
ÎN	Stressed prefix "in"	SPRE
ÎNG	Inflectional Suffix "ing"	ISUF
ÎR	Stressed prefix "ir"	SPRE
ÎSM	Inflectional Suffix "ism"	ISUF
L̂ESS	Inflectional Suffix "less"	ISUF
L̂Y	Inflectional Suffix "ly"	ISUF
M̂ENT	Inflectional Suffix "ment"	ISUF
N̂ESS	Inflectional Suffix "ness"	ISUF
N̂ON	Stressed prefix "non"	SPRE
P̂AST	Inflectional Suffix for past tense	ISUF
P̂L	Inflectional Suffix for plural	ISUF
R̂E	Stressed prefix "re"	SPRE
ŜUB	Stressed prefix "sub"	SPRE
T̂H	Inflectional Suffix "th"	ISUF
ÛN	Stressed prefix "un"	SPRE
Ŷ	Inflectional Suffix "y"	ISUF
CODA	Final Syllable Consonant	SPRE, SROOT[2,3]
DNUC	Unstressed Nucleus (Vowel) for Derivational Suffixes	DSUF
FCODA	Final Syllable Consonant for Function Words	FCN
FNUC	Unstressed Nucleus (Vowel) for Function Words	FCN
FONSET	Initial Syllable Consonant for Function Words	FCN
FSUF	Suffixes for Function Words	FCN
GLOTTAL	Glottal Stop in Pause	FP
LCODA	Final Syllable Consonant Following an LNUC+	SROOT[2,3]
LNUC+	Stressed Long Vowel	SROOT[2,3]
NUC	Unstressed Nucleus (Vowel)	DSUF, PRE, UROOT
NUC+	Stressed Nucleus (Vowel)	SROOT[2,3]
NUC_LAX+	Stressed Short Vowel	SROOT[2,3]
ONSET	Initial Syllable Consonant	SROOT[2,3]
PAU	Pause, between Word Boundaries	FP
UCODA	Final Syllable Consonant Following Unstressed Nuclei	DSUF, PRE
UMEDIAL	Consonant between two Nuclei (DSUF only)	DSUF
UONSET	Initial Syllable Consonant	DSUF, PRE, UROOT

Table A.5: *Phoneme layer categories used by* ANGIE. *Vowel phonemes marked with a "+" are stressed, while those without it are unstressed. The "!" marker for consonants forces the phoneme to be in onset position (the beginning of a syllable). Phonemes lacking this onset marker are constrained to be in coda position, at the end of a syllable. Some phonemes are only used for one word, such as* **/ah_does, ix_in, ux_you/** *and* **/ay_i/**.

Layer 5		
Node	Symbol	Example
/aa/	/ɑ/	"prosperity"
/aa+/	stressed /ɑ/	"topic"
/aar/	/ɑr/	"arcade"
/aar+/	stressed /ɑr/	"smart"
/ae/	/æ/	"fantastic"
/ae+/	stressed /æ/	"grand"
/ah/	/ʌ/	"amaze"
/ah+/	stressed /ʌ/	"wonder"
/ah_does/	/ʌ/	"does"
/ao/	/ɔ/	"augment"
/ao+/	stressed /ɔ/	"strong"
/aol+/	stressed /ɔl/	"all"
/aor/	/ɔr/	"original"
/aor+/	stressed /ɔr/	"morph"
/aw+/	stressed /ɑʷ/	"our"
/ay/	/ɑʸ/	"idea"
/ay+/	stressed /ɑʸ/	"right"
/ay_i/	/ɑʸ/	"I"
/b/	/b/	"superb"
/b!/	/b/	"book"
/ch/	/č/	"rich"
/ch!/	/č/	"chop"
/d/	/d/	"aid"
/d!/	/d/	"dream"
/d*ed/	[/ʌ/] /d/	"dreamed"
/dh/	/ð/	"lathe"
/dh!/	/ð/	"this"
/eh/	/ɛ/	"rocket"
/eh+/	stressed /ɛ/	"tennis"
/ehr/	/ɛr/	"binary"
/ehr+/	stressed /ɛr/	"query"
/el/	/l̩/	"angel"
/el+/	stressed /ɛl/	"celebrate"
/em/	/m̩/	"poem"
/en/	/n̩/	"beacon"
/en_and/	/n̩/	"and"
/er/	/ɝ/	"water"
/er+/	stressed /ɝ/	"circle"
/ey/	/e/	"friday"

Table A.6: *Phoneme layer categories used by* ANGIE. *Vowel phonemes marked with a "+" are stressed, while those without it are unstressed. The "!" marker for consonants forces the phoneme to be in onset position (the beginning of a syllable). Phonemes lacking this onset marker are constrained to be in coda position, at the end of a syllable. Some phonemes are only used for one word, such as* **/ah_does, ix_in, ux_you/** *and* **/ay_i/**.

Layer 5		
Node	Symbol	Example
/ey+/	stressed /e/	"eight"
/ey_a/	/e/	"a"
/f/	/f/	"laugh"
/f!/	/f/	"friend"
/g/	/g/	"dog"
/g!/	/g/	"garden"
/h!/	/h/	"house"
/ih/	/ɪ/	"ethic"
/ih+/	stressed /ɪ/	"ribbon"
/ihr+/	stressed /ɪr/	"year"
/ing/	/ɪŋ/	"running"
/ix_in/	/ɪ/	"in"
/iy/	/i/	"party"
/iy+/	stressed /i/	"swedish"
/iy_the/	/i/	"the"
/jh/	/ǰ/	"knowledge"
/jh!/	/ǰ/	"judge"
/k/	/k/	"speak"
/k!/	/k/	"king"
/l/	/l/	"helen"
/l!/	/l/	"letter"
/m/	/m/	"frame"
/m!/	/m/	"mike"
/n/	/n/	"phone"
/n!/	/n/	"new"
/ng/	/ŋ/	"pink"
/ow/	/o/	"auto"
/ow+/	stressed /o/	"coder"
/oy+/	stressed /ɔʸ/	"point"
/p/	/p/	"group"
/p!/	/p/	"parse"
/q/	/ʔ/	glottal stop
/r/	/r/	"far"
/r!/	/r/	"ray"
/ra_from/	/rʌ/	"from"
/s/	/s/	"boss"
/s!/	/s/	"stephanie"
/s*pl/	/s/	"systems"
/sh/	/š/	"crash"
/sh!/	/š/	"michelle"

Table A.7: *Phoneme layer categories used by* ANGIE. *Vowel phonemes marked with a "+" are stressed, while those without it are unstressed. The "!" marker for consonants forces the phoneme to be in onset position (the beginning of a syllable). Phonemes lacking this onset marker are constrained to be in coda position, at the end of a syllable. Some phonemes are only used for one word, such as* **/ah_does, ix_in, ux_you/** *and* **/ay_i/**.

Layer 5		
Node	Symbol	Example
/t/	/t/	"ha<u>t</u>"
/t!/	/t/	"<u>t</u>ree"
/th/	/θ/	"ba<u>th</u>"
/th!/	/θ/	"<u>th</u>ank"
/uh/	/ʊ/	"c<u>ou</u>ld"
/uh+/	stressed /ʊ/	"w<u>oo</u>d"
/uw/	/u/	"t<u>o</u>day"
/uw+/	stressed /u/	"s<u>u</u>per"
/uw_to/	/u/	"t<u>o</u>"
/ux_you/	/ü/	"y<u>ou</u>"
/v/	/v/	"sa<u>v</u>e"
/v!/	/v/	"<u>v</u>ictor"
/w/	/w/	"s<u>w</u>an"
/w!/	/w/	"<u>w</u>ork"
/wb/		word boundary
/y/	/y/	"merc<u>u</u>ry"
/y!/	/y/	"<u>y</u>acht"
/yu/	/yu/	"tiss<u>ue</u>"
/yu+/	stressed /yu/	"<u>u</u>nit"
/z/	/z/	"wi<u>s</u>e"
/z!/	/z/	"<u>z</u>oo"
/zh/	/ž/	"ra<u>j</u>"
/zh!/	/ž/	"a<u>s</u>ia"

Table A.8: *Graphemes used in* ANGIE. *In the* ANGIE *framework, the terminal layer can be composed of either letters, phones, or even other phonemes. We only list the set of grapheme used by* ANGIE *in letter parsing. The context-dependent graphemes ($-x) are not included. New graphemes (doubletons) can be used as well.*

Layer 6							
+	bu	el	hs	m2	ot	s_e	ut
+a	c	en	hu	m_e	ou	sc	uy
+d	c2	eo	i	mb	ow	se	v
+m	c_e	er	i2	me	oy	sh	v2
+s	cc2	es	ia	mi	p	si	v_e
_	ce	et	ie	mm2	p2	ss2	ve
a	ch	eu	ii2	mn	p_e	st	w
aa2	ci	ew	ir	n	pe	su	we
ab	ck	ey	is	n+	ph	sw	wh
ae	cq	f	iy	n2	pp2	sy	wr
ah	cs	f2	j	n_e	ps	sz	x
ai	cu	fe	ju	nd	pt	t	y
al	cz	ff2	k	ne	q	t2	yl
an	d	g	k2	ng	qu	t_e	yu
ao	d2	g2	k_e	ni	r	te	z
ar	d_e	ge	ke	nn2	r2	th	z2
as	dd2	gg2	kn	o	r_e	ti	z_e
au	de	gh	l	o2	re	tt2	ze
aw	dh	gi	l+	oa	rh	tu	zi
ay	di	gn	l2	oe	ri	tw	zz2
b	e	gu	l2e	oh	ro	u	
b2	ea	gy	l_e	oi	rr2	u2	
b_e	ec	h	le	ol	rt	uc	
bb2	ed	h2	lh	on	s	ue	
be	ee2	hi	ll2	oo	s+	ui	
bt	ei	ho	m	oo2	s2	ul	

Appendix B

ANGIE Morphological Categories

This Appendix describes the categories found on the third (morphological) layer of an ANGIE parse tree, and how they relate to the tags found on morphs.

No morphs are associated with the node FP. Morphs are associated with the remaining nine morphological categories via on the following tags.

For SPRE, there is no marking to distinguish it from a PRE. The only difference is that the vowel of the SPRE is stressed.

Layer 3		
Node	Tag	Example Morphs
DSUF	"*-morph*"	**-gence, -ine, -ly**
FCN	"*morph**"	**and*, have*, to***
ISUF	"*=morph*"	**=ed, =est, =s**
PRE	"*morph-*"	**com-, ex-, un-**
SPRE	"*morph-*"	**Ad-, non-, re-**
SROOT	"*morph+*"	**apt+, clar+, sciss+**
SROOT2	Same as SROOT	
SROOT3	Same as SROOT	
UROOT	"*morph*" (no marking)	**for, lyn, vie**

Appendix C

TIMIT Phonemes

Table C.1: *These are the consonant phonemes in* TIMIT.

Consonants					
TIMIT phoneme	Symbol	Example	TIMIT phoneme	Symbol	Example
b	/b/	"bat"	*ng*	/ŋ/	"sing"
ch	/č/	"child"	*p*	/p/	"pod"
d	/d/	"dog"	*r*	/r/	"rain"
dh	/ð/	"lathe"	*s*	/s/	"saw"
f	/f/	"find"	*sh*	/š/	"shed"
g	/g/	"go"	*t*	/t/	"top"
hh	/h/	"happy"	*th*	/θ/	"thank"
jh	/ǰ/	"job"	*v*	/v/	"vase"
k	/k/	"kite"	*w*	/w/	"woman"
l	/l/	"lion"	*z*	/z/	"zoo"
m	/m/	"man"	*zh*	/ž/	"regime"
n	/n/	"neon"			

Table C.2: *Phonemes marked with a "1" have primary stress, while those with a "2" have secondary stress. Phonemes without a number are not stressed.*

Vowels					
TIMIT phoneme	Symbol	Example	TIMIT phoneme	Symbol	Example
aa, aa1, aa2	/ɑ/	"crops"	*ey, ey1, ey2*	/e/	"vain "
ae, ae1, ae2	/æ/	"back"	*ih, ih1, ih2*	/ɪ/	"lid"
ah, ah1, ah2	/ʌ/	"done"	*iy, iy1, iy2*	/i/	"free"
ao, ao1, ao2	/ɔ/	"tall"	*ow, ow1, ow2*	/o/	"show"
aw, aw1, aw2	/ɑʷ/	"town"	*oy, oy1, oy2*	/ɔʸ/	"joy "
ay, ay1, ay2	/ɑʸ/	"cry"	*uh, uh1, uh2*	/ʊ/	"full "
eh, eh1, eh2	/ɛ/	"men"	*uw, uw1, uw2*	/u/	"glue"
er, er1, er2	/ɝ/	"church"			

Table C.3: *These are the other phonemes in* TIMIT.

Other Vowels					
TIMIT phoneme	Symbol	Example	TIMIT phoneme	Symbol	Example
ax	/ə/	"op̲en"	*en*	/n̩/	"gard̲en"
axr	/ər/	"teach̲er"	*ix*	/ɪ/	"wait̲ed"
el	/l̩/	"appl̲e"	*y*	/y/	"law̲yer"
em	/m̩/	"at̲om̲"			

Appendix D

COMLEX Phonemes

Table D.1: *A listing of the phonemes used in* COMLEX.

COMLEX Phonemes					
Long	Short	Examples	Long	Short	Examples
iy	i	heed, heat, he	n	n	no
ux	u	(used by TI for /u/)	en	N	button(2)
ih	I	hid, hit	nx	G	hang
ey	e	aid, hate, hay	p	p	pot
eh	E	head, bet	b	b	bed
ae	@	had, hat	t	t	tone
aa	a	hod, hot	d	d	done
aax	a	(Brit: father, alms)	dx	?	Peter(2)
ao	c	law, awe	k	k	kid
ow	o	hoed, oats, owe	g	g	gaff
uh	U	could, hood	q	q	(Glottal stop)
uw	u	who'd, hoot, who	ch	C	check
ay	Y	hide, height, high	jh	J	judge
oy	O	Boyd, boy	f	f	fix
aw	W	how'd, out, how	v	v	vex
er	R	father(2); herd, hurt, her	th	T	thin
ax	x	data (2)	dh	D	this
ah	A	cud, bud	s	s	six
ix	X	credit(2)	z	z	zoo
wh	H	which	sh	S	shin
w	w	witch	zh	Z	pleasure(2)
y	y	yes	hh	h	help
r	r	Ralph	'1	'	main stress
l	l	lawn	'2	+	non main stress
m	m	me	'3	+	non main stress
em	M	(syllabic m)	'0	.	no stress

Bibliography

[1] Chung, G., "Hierarchical Duration Modelling for Speech Recognition Using the ANGIE Framework," *Proceedings of Eurospeech-97*, Rhodes, Greece, September 1997, (forthcoming).

[2] Garris, M. and S. Janet, "NIST String Alignment and Scoring Program," Release 3.0, National Institute of Standards and Technology, 1987.

[3] Geutner, P., "Using Morphology Towards Better Large-Vocabulary Speech Recognition Systems," *Proceedings of ICASSP-95*, Detroit, Michigan, May 1995, pp. 445-448.

[4] Goddeau, D., E. Brill, J. Glass, C. Pao, M. Phillips, J. Polifroni, S. Seneff, and V. Zue, "GALAXY: A Human Language Interface to On-Line Travel Information," *Proceedings of ICSLP-94*, Yokohama, Japan, September 1994, pp. 707-710.

[5] Kemp, T., and A. Jusek, "Modelling Unknown Words in Spontaneous Speech," *Proceedings of ICASSP-96*, Atlanta, Georgia, May 1996, pp. 530-533.

[6] Lamel, L., R. Kassel, and S. Seneff, "Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus," *Proceedings of the Speech Recognition Workshop*, Palo Alto, California, February 1986, pp 100-109.

[7] Lau, R., and S. Seneff, "Providing Sublexical Constraints for Word Spotting within the ANGIE Framework," *Proceedings of Eurospeech-97*, Rhodes, Greece, September 1997, (forthcoming).

[8] Meng, H., personal communication.

[9] Meng, H., "Phonological Parsing for Bi-directional Letter-to-Sound/Sound-to-Letter Generation." MIT Ph.D. Thesis, 1995.

[10] Meng, H., S. Hunnicutt, S. Seneff, and V. Zue, "Reversible Letter-to-Sound/Sound-to-Letter Generation Based on Parsing Word Morphology," *Speech Communication*, Vol. 18, pp. 47-63, 1996.

[11] Seneff, S., personal communication.

[12] Seneff, S., R. Lau, and H. Meng, "ANGIE: A New Framework for Speech Analysis Based on Morpho-Phonological Modelling," *Proceedings of ICSLP-96*, Philadelphia, PA, October 1996, pp. 110-113.

[13] van Leeuwen, H. C, "Speech Maker Formalism: A Rule Formalism Operating on a Mutli-level, Synchronized Data Structure," *Computer Speech and Language,* Volume 7, Number 4, October 1993.

[14] Wu, S., M. Shire, S. Greenberg, and N. Morgan, "Integrating Syllable Boundary Information into Speech Recognition," *Proceedings of ICASSP-97*, Munich, Germany, April 1997, pp. 987-990.