Computer Speech and Language (1992) **6**, 175–196

Automatic classification of intonational phrase boundaries

Michelle Q. Wang* and Julia Hirschberg†

**Department of Computer Science, Stanford University, Stanford, California 94305, U.S.A. and †AT&T Bell Laboratories, 2D-450, 600 Mountain Avenue, Murray Hill, New Jersey 07974, U.S.A.*

Abstract

The relationship between the intonational characteristics of an utterance and other features inferable from its text represents an important source of information both for speech recognition, to constrain the set of allowable hypotheses, and for speech synthesis, to assign intonational features appropriately from text. This work investigates the usefulness of a number of textual features and additional intonational features in predicting the location of one particular intonational feature—intonational phrase boundaries—in natural speech. The corpus for this investigation is 298 utterances from the 774 in the DARPA-collected Air Travel Information Service (ATIS) database. For statistical modeling, we employ classification and regression tree (CART) techniques. We achieve success rates of just over 90%, representing a major improvement over previous attempts at boundary prediction for spontaneous speech.

1. Introduction

Understanding the relationship between an utterance's intonational features and features which can be inferred from the text uttered provides important information both for speech recognition and for speech synthesis. In recognition, the association of intonational features with syntactic and acoustic information can be used to reduce the number of candidates under consideration. In synthesis, better understanding of the relationship between intonational features and syntactic and discourse-level characteristics of the text in natural speech provides the basis for algorithms to assign intonational features more naturally.

In this study we investigate the usefulness of a number of textual features and additional intonational features in predicting the location of one particular intonational feature—intonational phrase boundaries—in natural speech. We select potential boundary predictors based upon hypotheses derived from linguistic theory, as well as examination of previous studies of phrase boundaries. Our corpus for this investigation is 298 sentences from the 774 sentences of the DARPA-collected Air Travel Information Service (ATIS) database (DARPA, 1990). For statistical modeling, we employ classification and regression tree techniques (CART).

0885–2308/92/010175+22 $03.00/0

2. Intonational theory

Intuitively, the intonational phrasing of an utterance divides it into meaningful "chunks" of information (Bolinger, 1989). For example, variation in a sentence's intonational phrasing can change the meaning hearers are likely to assign to an individual utterance of the sentence. Consider the utterances illustrated in (1), where the presence of a comma denotes a phrase boundary:

(1a) Bill doesn't drink because he's unhappy.
(1b) Bill doesn't drink because he's unhappy. (He drinks because he's an alcoholic).
(1c) Bill doesn't drink, because he's unhappy. (He believes that alcohol will amplify his depression).

If a speaker utters the sentence (1a) as a single phrase, as in (1b), hearers are likely to infer that Bill does indeed drink—but not from unhappiness. If the speaker instead utters the sentence as two phrases, as in (1c), hearers are likely to understand that Bill does *not* drink and that the reason for his abstaining is his unhappiness.

In describing intonation phrasing in particular and intonational features in general, we adopt Pierrehumbert's theory of intonational description for English (Pierrehumbert, 1980). According to this view, two levels of phrasing are significant in English intonational structure. Both types of phrases are composed of sequences of high and low tones in the *fundamental frequency* (F_0) contour. An *intermediate* (or minor) *phrase* consists of one or more *pitch accents* (local F_0 minima or maxima) plus a *phrase accent* (a simple high or low tone which controls the pitch from the last pitch accent of one intermediate phrase to the beginning of the next intermediate phrase or the end of the utterance). *Intonational* (or major) *phrases* consist of one or more intermediate phrases plus a final *boundary tone*, which may also be high or low, and which occurs at the end of the phrase. Thus, an intonational phrase boundary necessarily coincides with an intermediate phrase boundary, but not vice versa. Differences in phrasing are illustrated in the F_0 contours depicted in Fig. 1, illustrating the utterance of (1b), and Fig. 2, illustrating (1c).[1]

In empirical studies of phrase boundary identification, perceived boundaries correlate with certain physical characteristics of the speech signal. In addition to the tonal features described above, phrases may be identified by one of more of the following features: pauses (which may be filled or not); changes in amplitude; and lengthening of the final syllable in the phrase (sometimes accompanied by glottalization of that syllable and perhaps preceding syllables). In general, major phrase boundaries tend to be associated with longer pauses, greater tonal changes and more final lengthening than minor boundaries.

3. Previous research

Previous research on the location of intonational boundaries has focussed primarily on the relationship between prosodic boundaries and syntactic constituency boundaries (Downing, 1970; Bresnan, 1971; Selkirk, 1978; Cooper & Paccia-Cooper, 1980). While most researchers now acknowledge the role that semantic and discourse-level information play in boundary assignment (Bolinger, 1989), it is still generally assumed that

[1] These contours are produced by the Bell Labs Text-to-Speech System (Olive & Liberman, 1985).

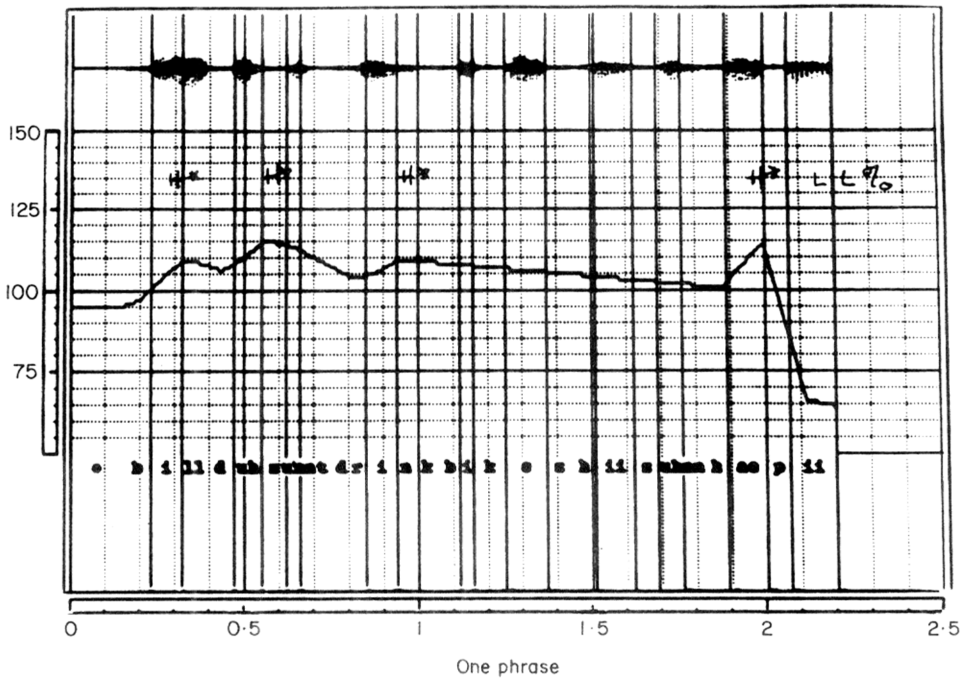

Figure 1. Bill doesn't drink because he's unhappy

syntactic configuration provides the basis for prosodic "defaults" that may be overridden by semantic or discourse considerations. The considerable experimental literature has concentrated mainly on the relationship between intonational phrasing (and its acoustic correlates) and syntactic structure (Lea, 1972; Lehiste, 1973; O'Malley, Kloker & Dara-Abrams, 1973; Klatt, 1975; Cooper & Sorenson, 1977; Streeter, 1978; Wales and Toner, 1979; Grosjean, Grosjean & Lane, 1979; Umeda, 1982; Gee & Grosjean, 1983). While current applications for such reward are dominated by attempts to develop phrasing methods for text-to-speech or message-to-speech systems (Altenberg, 1987; Bachenko & Fitzpatrick, 1990; Schnabel & Roth, 1990; Bruce, Granstrom & House, 1990), there is also a resurgence of interest in the location of intonational boundaries in speech recognition (Ostendorf *et al.*, 1990; Steedman, 1990), both from acoustic indicators and from syntactic analysis of text.

The success of corpus-based methods developed to associate intonational boundaries with syntactic or other features of the transcribed speech has varied based upon the genre of the corpus modeled. Most successful have been studies of read speech by professional speakers. For example, using only features that can be obtained automatically from an analysis of transcribed text Bachenko and Fitzpatrick's (1990) algorithm correctly classifies 83·5–86·2% of potential boundary sites for a test set of 35 citation-form sentences (E. Fitzpatrick, 10 February 1991, pers. comm.). Preliminary results on the prediction of intonational boundaries in radio speech by researchers at Boston University (M. Ostendorf and N. Veilleux, 29 May 1991, pers. comm.) indicate that over 87% correct boundary location can be achieved (averaging per cent correct for boundaries and for null boundaries, see below, Section 4.5), when scoring for each sentence is based upon the closest match for any one of four speakers' renditions of that

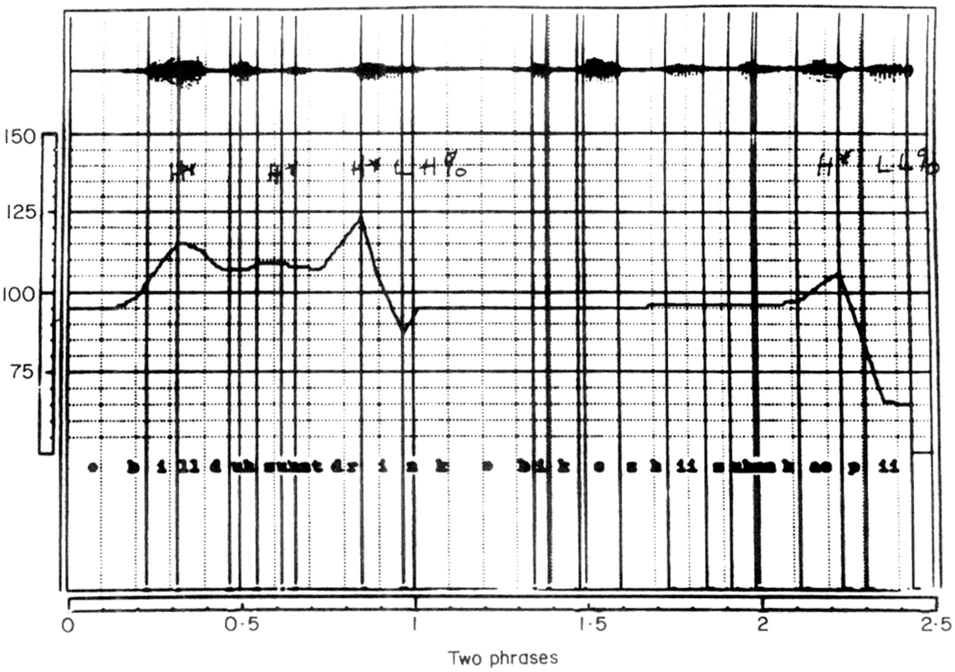

Figure 2. Bill doesn't drink because he's unhappy

sentence to the predicted phrasing. However, the robustness of such results remains to be seen. Limitations of training corpus size and coverage (in terms of syntactic and other potential predictors of phrasing), of genre and of scope of evaluation all suggest further investigation.

The modeling of phrasing in natural speech has been more difficult. For example, Altenberg's (1987) hand-crafted phrasing rules correctly classify an average of 72% of boundaries (tone units) in 48 minutes of his training set, a sample from the London–Lund Corpus of partly-read, partly spontaneous speech from a single non-professional speaker.[2] And this performance assumes the availability of syntactic, semantic and discourse-level information well beyond the capabilities of automatic text analysis to provide; the training data were tagged and parsed by hand, and subtle semantic and discourse-level distinctions, such as distinguishing restrictive from non-restrictive relative clauses, and determining "focus of attention", were available.

To address some of the issues of corpus size and type, of evaluation procedures, and of the nature and computational availability of potential predictors of phrasing, we examine a corpus of spontaneous (elicited) multi-speaker speech. Our immediate goals are, first, to model phrasing in a relatively large amount of spontaneous, non-professional speech. Second, we want to discover how well simple computational text analysis techniques can perform at predicting observed intonational boundaries. Third, we need to compare these predictions with predictions based upon the availability of information available currently only through hand labeling, to see how much predictive power will be lost by our computational approach. Fourth, we hope to acquire the relationships between phrasing and its predictors automatically from our corpus, rather

[2] These figures are derived from Altenberg's report of 95% coverage, 93% success at the sentence level; 94% coverage, 70% sucess at the basic clause level; and 78% coverage, 80% success in expanded phrases.

than by building a set of rules by hand, so that we will be able to model potentially different phrasing styles in future databases rapidly.

Our long-term goal in this investigation is to develop a model of intonational phrasing which can be applied to speech synthesis or to speech recognition tasks. For synthesis the application is relatively straightforward. By relating phrasing decisions to textual features in natural speech, we can hope to model human performance in synthetic speech. The simpler the analysis techniques we can employ, while maintaining a reasonable level of prediction, the easier it will be to incorporate such modeling in real-time speech synthesis. Furthermore, if such correlations can be acquired automatically from labeled speech, then we will be able to model variations in phrasing style as well. For recognition, the application is more speculative. We hope at most to be able to use our text-based phrasing model in conjunction with an acoustic model of phrasing to constrain the set of possible recognition hypotheses; at least, we hope that a good model of phrasing, based on a recognizer's training set, can be used to improve the recognizer's model of durational likelihoods, by factoring in phrasal position.

4. The study

4.1 Data

The corpus used in this analysis consists of 298 utterances (24 minutes of speech from 26 different speakers) collected by Texas Instruments for use in the DARPA ATIS (Air Travel Information System) spoken language system evaluation task. These 298 tokens represent all of the (syntactic) questions currently in the ATIS database.[3] In a Wizard-of-Oz simulation, subjects were given a travel scenario and asked to make travel plans accordingly, providing spoken input and receiving output at a terminal. Speakers were told that their utterances were being recognized by machine; vocabulary constraints were enforced by error messages. The full ATIS TI corpus currently consists of 774 sentences, culled from these experiments. The quality of the ATIS corpus is extremely diverse. Speakers range in fluency from close to isolated-word speech to exceptional fluency. Many utterances contain hesitations and other disfluencies, as well as long pauses (greater than 3 s in some cases). The number of such pauses may reflect the complexity of the task confronted by speakers. A scenario typical of those presented to the subjects appears below:

> A small gymnastic team from overseas will be touring several cities in this country. Pick the cities they will visit and plan their itinerary. Please be sure: (1) the team will not miss meals; (2) there will be ground transportation to and from downtown; and (3) they do not travel on weekends; (4) they fly on large (presumably safe) aircraft.

In addition, the novelty of communicating with a computer appears to have influenced the verbal styles chosen by some speakers.

To prepare this data for analysis, we labeled the speech prosodically by hand, noting location and type of intonational boundaries and presence or absence of pitch accents. Labeling was done from both the waveform and pitchtracks of each utterance. Each

[3] Although the corpus is currently limited to questions, we have little reason to believe that non-questions will present significant idiosyncracies.

label file was checked by 2–3 labelers. We labeled two levels of boundary, intonational phrase boundaries and intermediate phrase boundaries; in the analysis presented below, however, these are collapsed to a single category. More data will be needed for analysis before this distinction can be productively analysed.

4.2. Selection of variables for analysis

For the current analysis of boundary location, we define our data points to consist of all potential boundary locations in an utterance, defined as each pair of adjacent words in the utterance $<w_i, w_j>$, where w_i represents the word to the left of the potential boundary site and w_j represents the word to the right. We selected these potential variables for analysis based upon their theoretical potential as boundary predictors and the feasibility of acquiring values for these variables automatically.

Given the variability in performance we observed among speakers, an obvious variable to include in our analysis is speaker identity. While for applications to speaker-independent recognition this variable would be uninstantiable, we nonetheless need to determine how important speaker idiosyncracy may be in boundary location. We have found no significant increase in predictive power when this variable is used; so, results presented below are speaker-independent. However, when we have more data from each individual speaker this feature may become more useful.

One class of variable which is readily obtainable involves temporal information. Temporal variables include utterance and phrase duration, and distance of potential boundary from various strategic points in the utterance. Although it is tempting to assume that phrase boundaries represent a purely intonational phenomenon, we must allow for the possibility that processing constraints help govern their occurrence. That is, longer utterances may tend to be produced with more boundaries for mechanical production reasons. Accordingly, we measure the length of each utterance both in seconds and in words. The distance of the boundary site from the beginning and end of the utterance is another variable which appears likely to be correlated with boundary location. We speculate that the tendency to end a phrase might be affected by the position of the potential boundary site in the utterance. For example, it seems likely that positions very close to the beginning or end of an utterance might be less likely positions for intonational boundaries—again, perhaps, on grounds of production constraints. We measure this variable too, both in seconds and in words.

Gee and Grosjean (1983) and Bachenko and Fitzpatrick (1990) *inter alia* have noted the importance of phrase length in determining boundary location. Simply put, it seems possible that consecutive phrases often have roughly equal lengths. To capture this notion, we calculate the following fraction: the elapsed distance from the last boundary to the potential boundary site, divided by the length of the last phrase encountered. Unfortunately, to obtain this information from text analysis alone would require us to factor prior boundary predictions into subsequent predictions. As a first step, therefore, to test the utility of this information we have used observed boundary locations in our current analysis. However, using prior boundary prediction would be feasible if we discover that this information has predictive power. We employ both temporal and word count prediction for this variable.

Previous researchers have long considered syntactic information to be a good predictor of phrasing information (Gee & Grosjean, 1983; Selkirk, 1984; Marcus & Hindle, 1985; Steedman, 1990). To investigate such possibilities we consider first the sentence type to which the potential boundary site belongs. It is possible that certain classes of sentences, e.g. indirect questions, will contain more phrase boundaries than others, possibly due to some disambiguating function that phrasing might serve. We accommodate this possibility by dividing questions into three main categories—*wh*-questions, direct yes/no questions and indirect yes/no questions. Note however that, while (syntactic) yes/no questions may be automatically distinguishable from *wh*-questions, indirect questions cannot be reliably distinguished from direct ones using current computational techniques. Certainly, if utterance "type" turns out to be an important predictor for presence or absence of intonational boundary, results of our future examination of declaratives should prove even more interesting.

Part-of-speech information is another factor widely believed to predict boundary location, particularly in text-to-speech. For example, the belief that phrase boundaries rarely occur after function words forms the basis for most algorithms used to assign intonational phrasing for text-to-speech. Furthermore, we might expect that heads of phrases (e.g. prepositions and determiners) do not constitute the typical end to an intonational phrase. We explore these ideas by examining a window of four words surrounding each potential phrase break. We obtain part-of-speech information via Church's part-of-speech tagger (Church, 1988), whose output has been modified slightly to predict preposition/particle distinctions.

In addition to part-of-speech information, we also investigate the importance of syntactic constituency to the prediction of boundary location. Intuitively, we want to test the notion that some constituents may be more or less likely than others to be internally separated by intonational boundaries. To test which constituents are *less* likely to be intonationally separated, we examine the class of the lowest node in the parse tree to dominate both w_i and w_j. To test which constituents are *more* likely to occur at boundary edges, we determine the class of the highest node in the parse tree to dominate w_i, but not w_j, and the class of the highest node in the tree to dominate w_j but not w_i. In this way we can determine whether constituency relations are important to intonational phrasing. We use Hindle's parser, Fidditch (Hindle, 1989), to provide information on syntactic constituency for this analysis.

Recall that each intermediate phrase is composed of one or more pitch accents plus a phrase accent, and each intonational phrase is composed of one or more intermediate phrases plus a boundary tone. Clearly, then, pitch accents form the main building blocks for phrases. In addition, informal observation suggests that phrase boundaries are more likely to occur in some accent contexts than in others. For example, phrase boundaries between words that are de-accented seem to occur much less frequently than boundaries between two accented words. To test this observation, we look at the pitch accent values of w_i and w_i for each $< w_i, w_j >$. Since our prosodic labeling of the speech under analysis includes pitch accent location, we can use observed data values for the accent variables. However, since we would prefer to predict boundary location simply from information collected from text, we also substitute predicted pitch accent information for the observed data. We obtain accent predictions from text analysis procedures described in Hirschberg (1990). We use predicted accent as a binary feature (accented or not) or as a four-valued feature (cliticized,[4] de-accented,[5] accented or "NA", for end of sentence), to see whether or not finer-grained information will prove useful in predicted boundary location. On one hand, we want to discover from actual accent information how useful

[4] Deaccented, with a reduced vowel in the stressable syllable and lacking at least one word boundary.

[5] Lacking a pitch accent.

accent location can be in predicting boundary location. On the other hand, we want to determine whether information available from text analysis can help to automate our analysis.

In the analyses described below, we employ varying combinations of these variables to predict intonational boundaries. We use classification and regression tree techniques to generate decision trees automatically by selecting among variables provided to minimize error in a step-wise procedure.

4.3. Classification and regression tree analysis

Classification and regression trees (CART) techniques (Brieman *et al.*, 1984) permit the automatic generation of decision trees from sets of continuous or discrete variables. At each node in the generated tree, CART selects the statistically most significant variable to minimize prediction error at that point. In the implementation of CART used in this study (Riley, 1989), all of these decisions are binary, based upon consideration of each possible binary split of values of categorical variables and consideration of different cut-points for values of continuous variables.

Generation of CART decision trees depends on a set of splitting rules, stopping rules and prediction rules. These rules affect the internal nodes, subtree height and terminal nodes, respectively. At each internal node the program must determine which factor should govern the forking of two paths from that node. Furthermore, the program must decide which values of the factor to associate with each path. Ideally, the splitting rules ought to choose the factor and value split which minimize the prediction error rate. The splitting rules described in Riley (1989) use a heuristic which approximates optimality by choosing at each node the split which minimizes the prediction error rate on the training set of data.

Stopping rules are necessary for terminating the splitting process at each internal node. Thus, these rules govern the height of each subtree. To determine the best tree, this implementation uses two sets of stopping rules. The first set attempts to classify all the data, resulting in a tree which commonly lacks the generality necessary to account for data outside of the training set. To estimate a more reliable tree, the implementation uses a second set of rules. First, a sequence of subtrees is formed from the initial tree. Then, each tree is grown on a (different, randomly selected) portion (90%) of the training data and tested on the remaining portion. This step is iterated, so that the stopping rules have access to cross-validated error rates for each subtree. The subtree with the lowest rate then defines the stopping point for each path in the full tree. Trees presented below are labeled with their cross-validated estimates. These estimates have been tested by manually separating the data into training and test sets and obtaining error rates for 10 of the 55 feature sets used in this study. In none of these cases was the difference between CART's cross-validated prediction and the actual performance of the tree statistically significant; i.e. in every case both prediction and performance on a hand-separated test set were within a 95% confidence interval.

The prediction rules work in a straightforward manner to add the necessary labels to the terminal nodes. In the case of a categorical dependent variable, such as is the case here, the rules simply choose the class that occurs most frequently among the data points. The success of these rules can be measured through estimates of deviation. The deviation for categorical variables is simply the number of misclassified observations.

4.4. Procedure and results

In analysing boundary locations in our data we have two goals in mind. First, we want to discover the extent to which boundaries can be predicted, given information which can be generated automatically from the text of an utterance. Second, we want to learn how much predictive power can be gained by including additional sources of information which, at least currently, cannot be generated automatically from text. That is, how well can automatically inferable information be used to predict intonational boundaries—and how much better can we do using additional information available currently only through human observation, such as information about other intonational features? So, for example, while we can predict pitch accent information with reasonable success from text (Hirschberg, 1990), what is the relationship between the use of such approximated information and the actual observations in permitting inference of boundary location? In effect, how good a substitute is approximate information for these purposes? In discussion of our results below, we will compare predictions based upon automatically inferable information with those based upon observed data.

One decision which must be made in labeling intonational boundaries is how to treat speaker hesitations, false starts and other disfluencies when determining boundary location. While such phenomena may be accompanied by many of the acoustic correlates of true prosodic boundaries, it is not clear theoretically that they should be included as true intonational boundaries for purposes of description or prediction. Although disfluencies themselves may be to some extent predictable in terms of syntactic configuration or surface position (Hindle, 1983), there is no evidence that regularities they may exhibit will be similar to the regularities observable in the location of intonational boundaries. Since this issue is difficult to resolve, we perform our analyses on two data sets: one in which disfluencies are classified (for scoring purposes) as intonational boundaries and one in which they are classified as null boundaries.

We employ four different sets of variables during the first phase of analysis. The first data set (corresponding to the results in Fig. 3; key to node labels in Table IV) includes observed phonological information on pitch accent and on location of prior intonational boundaries, as well as automatically inferred information. Including observed values for accent and prior boundary location allows us to obtain a clear look at the contributions of the automatically generated analysis information by avoiding errors introduced by estimation of the phonological parameters. The success rate of boundary prediction for this data set is extremely high. The tree illustrated in Fig. 3 correctly classifies 3330 out of 3677 potential boundary sites—an overall success rate of 90%. Furthermore, the number of decisions in this tree is only five. Thus, the tree represents a clean and simple model of phrase boundary prediction rules, assuming accurate phonologial information about pitch accent and prior boundary location.

Turning to the decision nodes represented in the tree itself, we immediately note the importance of length of current phrase compared with that of prior phrase in boundary location. This variable alone (assuming that the boundary site occurs before the end of the utterance) allows us correctly to classify 2403 out of 2556 potential boundary sites. We conclude that occurrence of a phrase boundary is extremely unlikely if its presence would result in a phrase with less than half the length of the preceding phrase.

The first and last decision points in the three are easily the most trivial. The information garnered from the first split is that utterances virtually always end with a boundary—rather unsurprising news. Similarly, the last split in the tree shows the

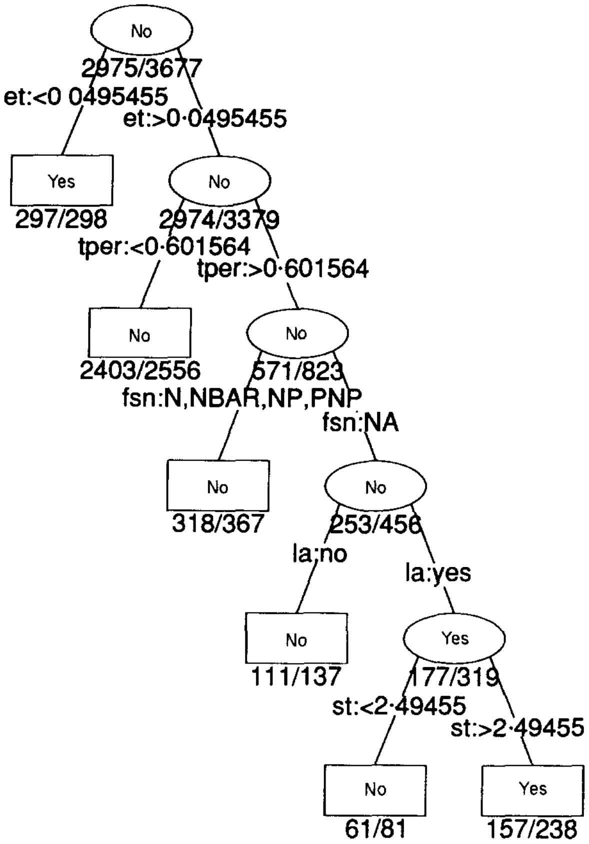

Figure 3. Boundary prediction from hand- and automatically-labeled features.

importance of distance from the beginning of the utterance in boundary location. According to this node, boundaries are more likely to occur (subject to the constraints found higher in the tree) when more than 2·5 seconds have elapsed from the start of the utterance.

Both internal tree splits bear much more theoretical importance than do the first and last splits. The third node in the tree indicates that when w_i and w_j are co-constituents in a nominal category (including proper and common nouns at different bar levels), they are usually both members of a single intonational phrase as well. In other words, noun phrases form a tightly bound unit intonationally.

The fourth split in the tree shows the role of accent context in determining phrase boundary location. The accenting of w_i is clearly an important factor. If w_i is not accented, then it is unlikely that a phrase boundary will occur after it.

The significance of accenting in the phrase boundary classification tree leads to the question of whether or not predicted accents will have a similar impact on the paths of the tree. In the second analysis, we substituted predicted accent values for the actual values (see Fig. 4). Interestingly, the success rate of the tree remains approximately the same, at 90%. However, the number of splits in the tree increases to nine and fails to include the accenting of w_i as a node in the decision tree. A closer look at the accent

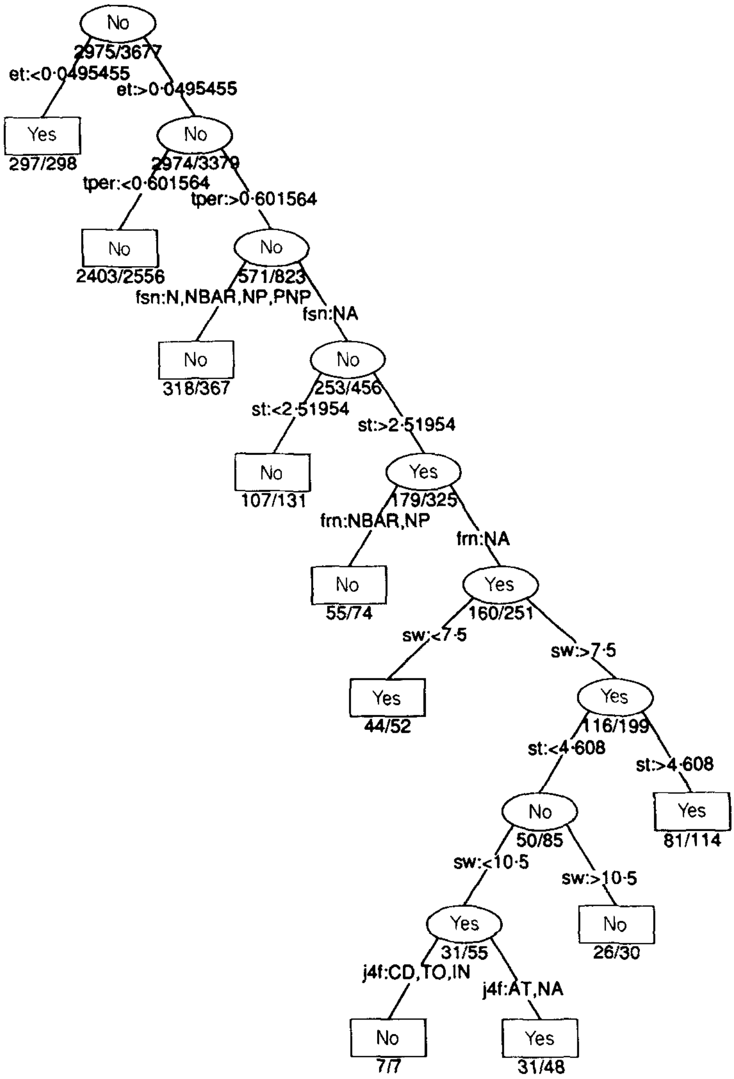

Figure 4. Boundary prediction using predicted instead of observed accent.

predictions themselves reveals that the majority of misclassifications come from function words that precede a boundary. Although the accent prediction algorithm we used to generate accent predictions expects these words to be cliticized, they are accented in this context. This phenomenon appears to be an artifact of the corpus; such accented function words generally occur before relatively long pauses in an utterance. Nevertheless, it is interesting to note that the classification tree is able to compensate quite well for the loss of the accent information. An implication of this result is that accent values may be predicted by some of the same factors that predict phrase boundaries. Preliminary investigation indicates that both pitch accent and boundary location are sensitive to location of prior intonational boundaries as well as part-of-speech tags and syntactic constituency of context.

In the third analysis, we eliminate the dynamic boundary percentage measure. The result remains nearly as good as before, with a success rate of 89%. The proposed decision tree (see Fig. 5) confirms the usefulness of information about the accent status

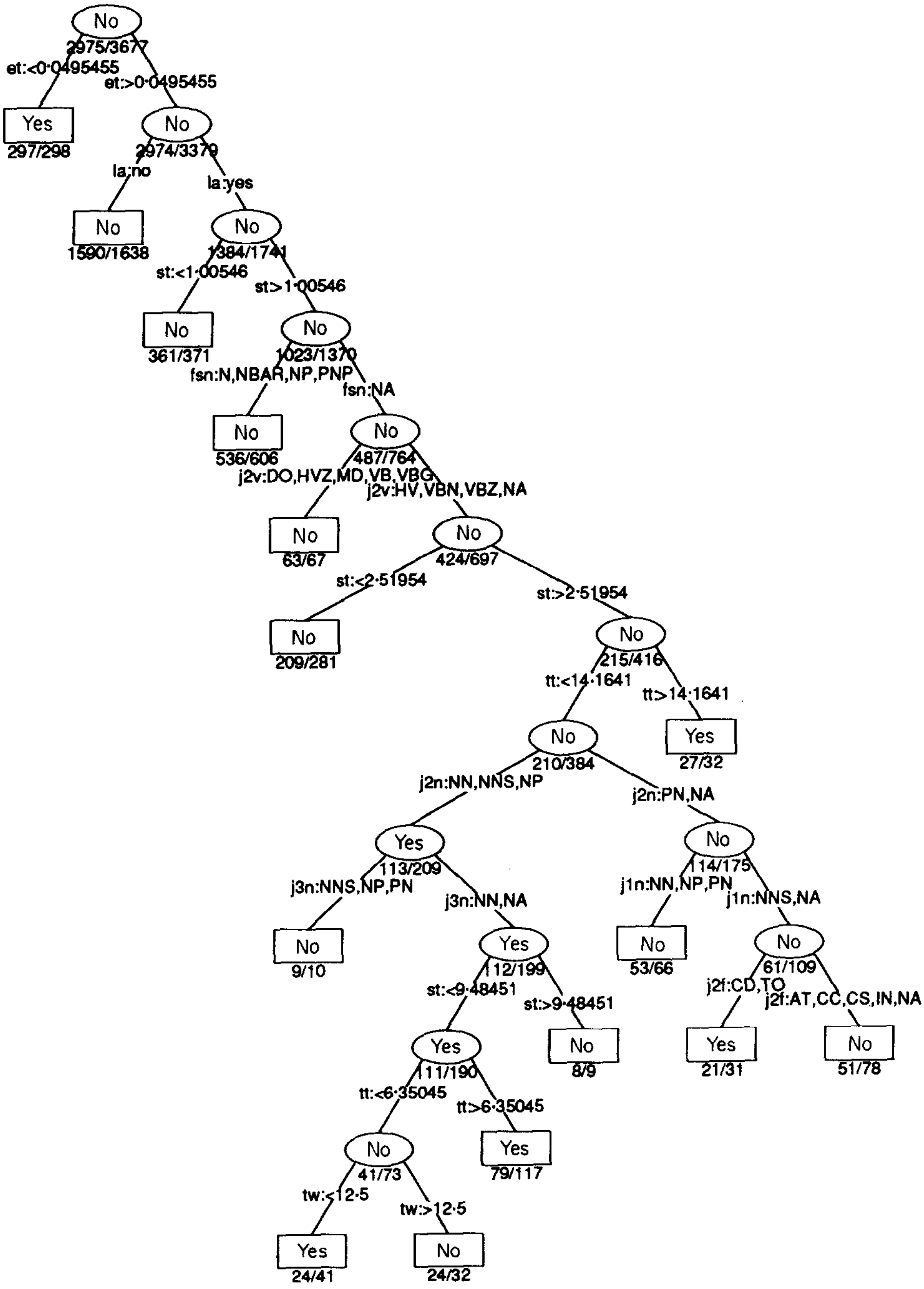

Figure 5. Boundary prediction without dynamic computations.

of w_i in boundary prediction. By itself (again assuming that the potential boundary site occurs before the end of the utterance), this factor can account for 1590 out of 1638 potential boundary site classifications. Similarly, this tree confirms the strength of the intonational ties among the components of noun phrases. In this tree, 536 out of 606 potential boundary sites receive final classification from this feature.

We conclude this phase of our analysis by producing a classification tree that uses text-based information alone. For this analysis we use predicted accent values instead of observed values and omit all continuous temporal variables, including overall rate and boundary distance percentage measures. Using our four-valued distinction (cliticized, de-accented, accented, "NA") for predicted accent for left and right context we achieve a (cross-validated) prediction score of 89%. The tree in Fig. 6 presents these results. This tree contains considerably more nodes than those appearing in previous figures to achieve a similar cross-validated clasification rate. Features that appear most frequently in the tree are part-of-speech variables and word-based distance measures, with predicted accent appearing sparsely. Part-of-speech information for all four positions appears in this tree, as does distance from beginning and end of utterance and total words in utterance. In general, results of this first stage of analysis suggest—encouragingly—that there is considerable redundancy in the features predicting boundary location; when some features are unavailable, others can be used with equal predictive accuracy.

Above we have seen that locating phrase boundaries from which disfluencies are excluded can be accomplished fairly successfully by calculating distance measures for each potential boundary site $<w_i, w_j>$, determining the utterance's overall rate and making reference to part-of-speech and constituence information for w_i and w_j. We now must determine whether or not these variables remain important when we include "boundaries" which arise from disfluency.

In the second phase of this study, we repeat the previous analyses, but extend our definition of a phrase boundary to include hesitations and disfluencies (see Figs 7–10). The result is a drop in success rate. The first analysis is successful 85% of the time, as compared with 90%. Similarly, the rates for the second analysis drop from 90% to 83%; for the third analysis from 89% to 85%; and for the fourth analysis from 89% to 83%. However, the same factors which are important in the first phase of the study continue to be significant in the second phase. Major splitting nodes in the four analyses include proximity to the end of the utterance, accent value of w_i, occurrence of a function word as w_j and membership of w_i and w_j in a noun phrase. This comparison confirms our belief that such disfluencies should not in fact be classified as true phonological boundaries, but should undergo separate analysis.

4.5. Boundaries vs. null boundaries

Since approximately 80% of the data points represent actual "null boundaries", it is important to look at whether these data points are being predicted more successfully than data points which represent actual boundaries, or, possibly, vice versa. That is, if 80% of data points are correctly classified as "null boundary", then one can achieve 80% success simply by classifying every data point as such.[6] The decision tree most successful

[6] The confusion matrices presented here and below over-estimate success rates slightly. As noted in Section 4.3, CART cross-validated error is averaged over multiple trees. These figures are calculated from a tree whose cross-validated length is chosen on this basis. The tree itself varies a few percentage points from the average. So, percentages should be considered a best approximation.

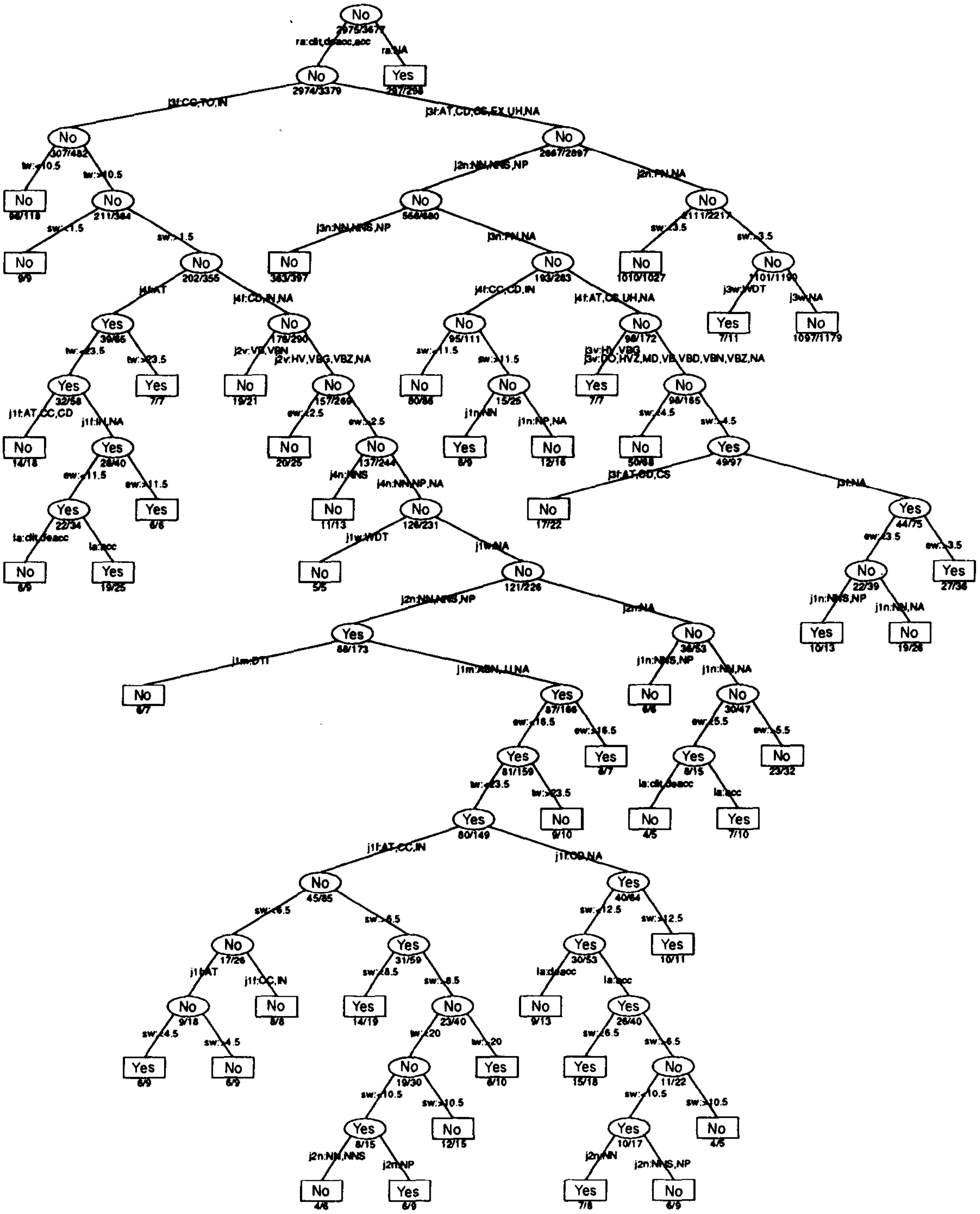

Figure 6. Boundary predictioin from automatic text analysis alone.

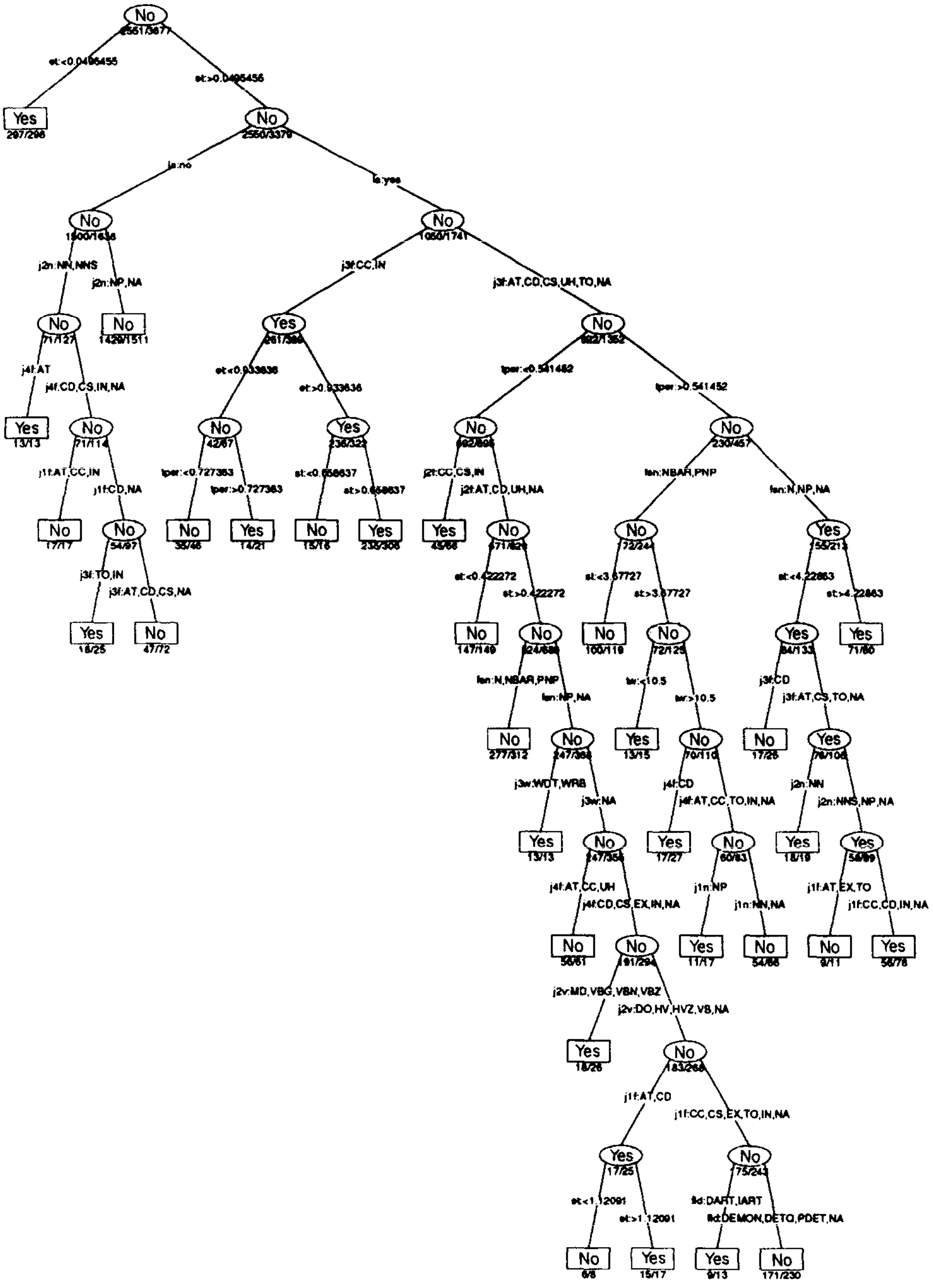

Figure 7. Prediction from hand- and automatically-labeled features, disfluencies classed as boundaries.

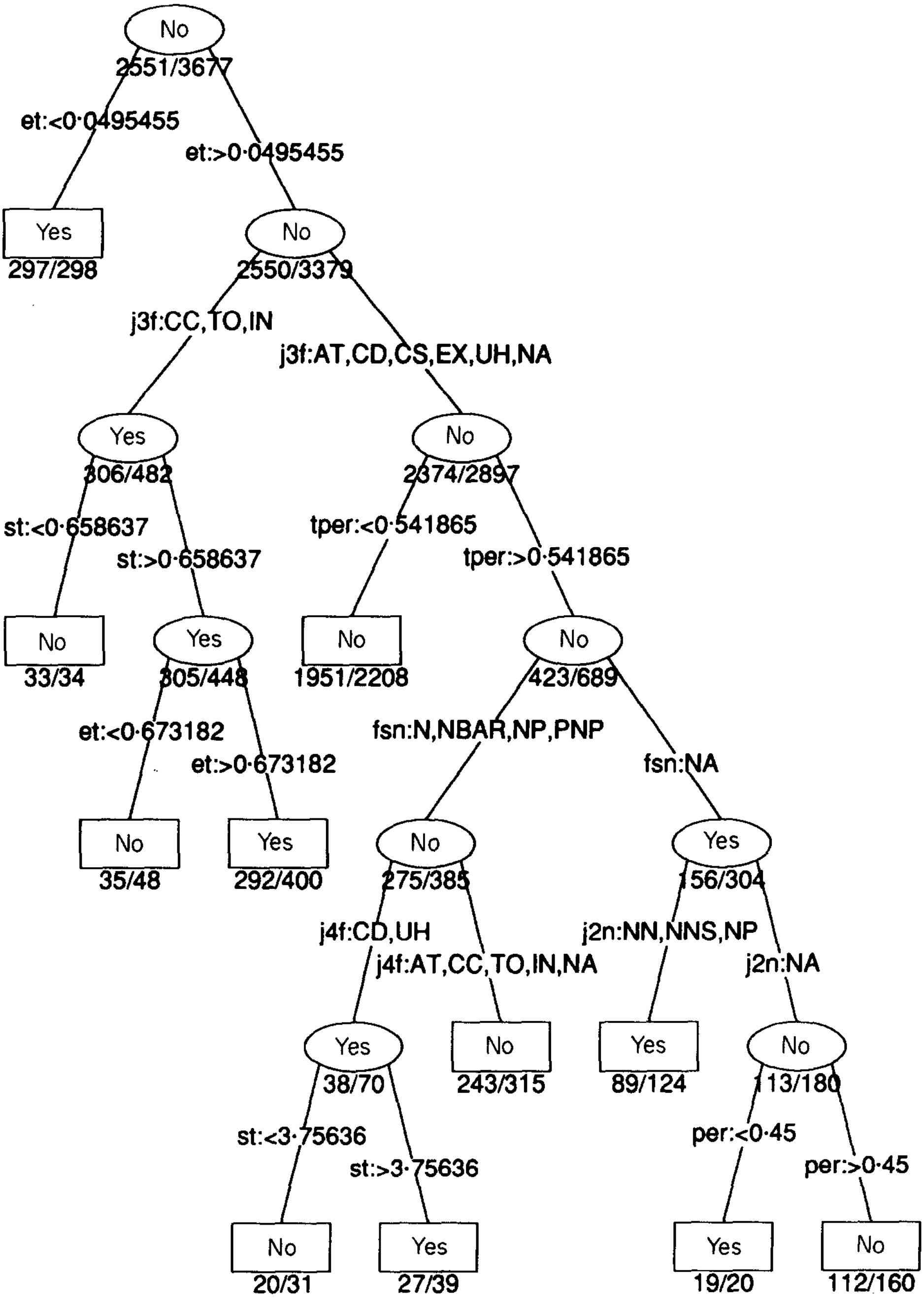

Figure 8. Prediction using predicted accent, disfluencies classed as boundaries.

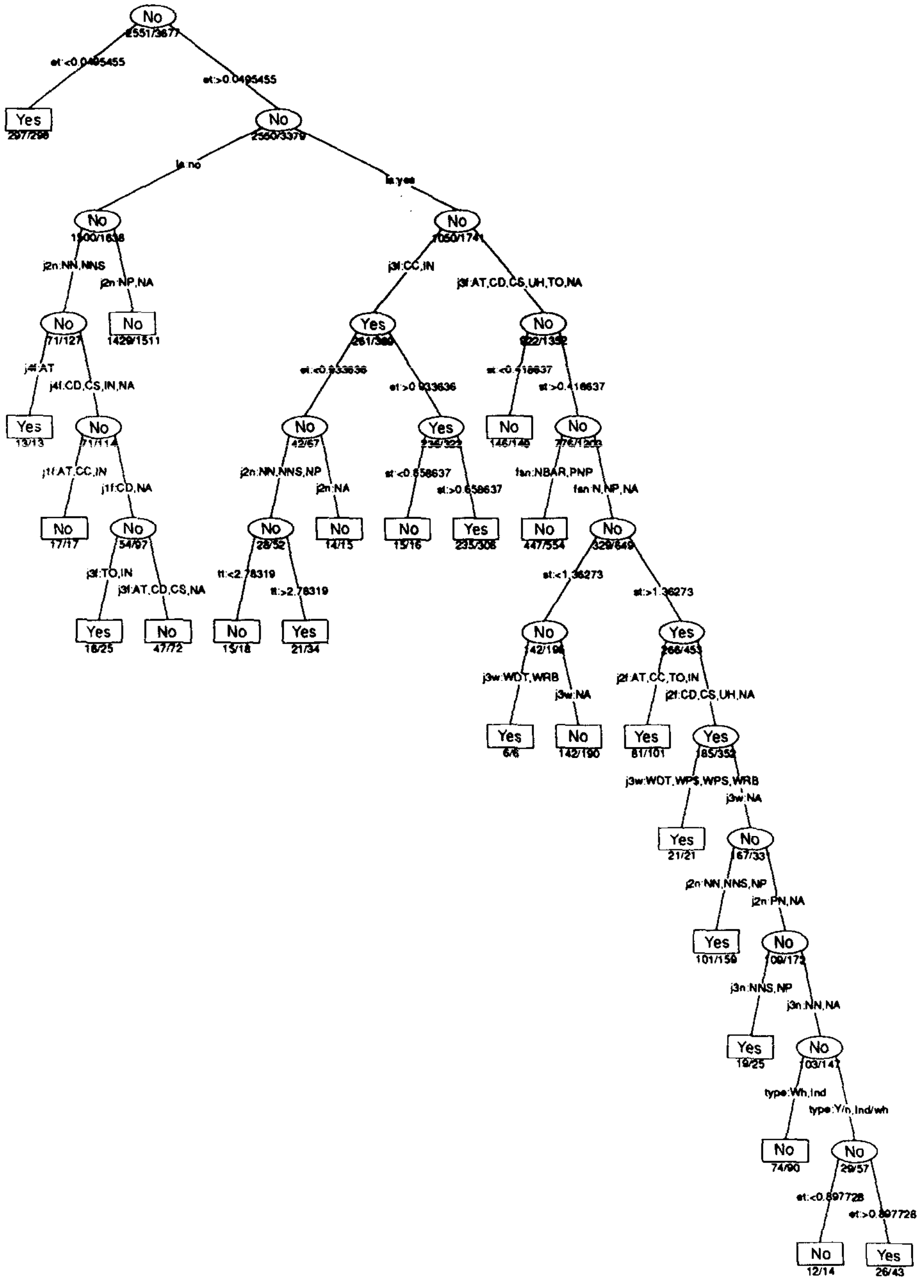

Figure 9. Boundary prediction without dynamic computations, disfluences classed as boundaries.

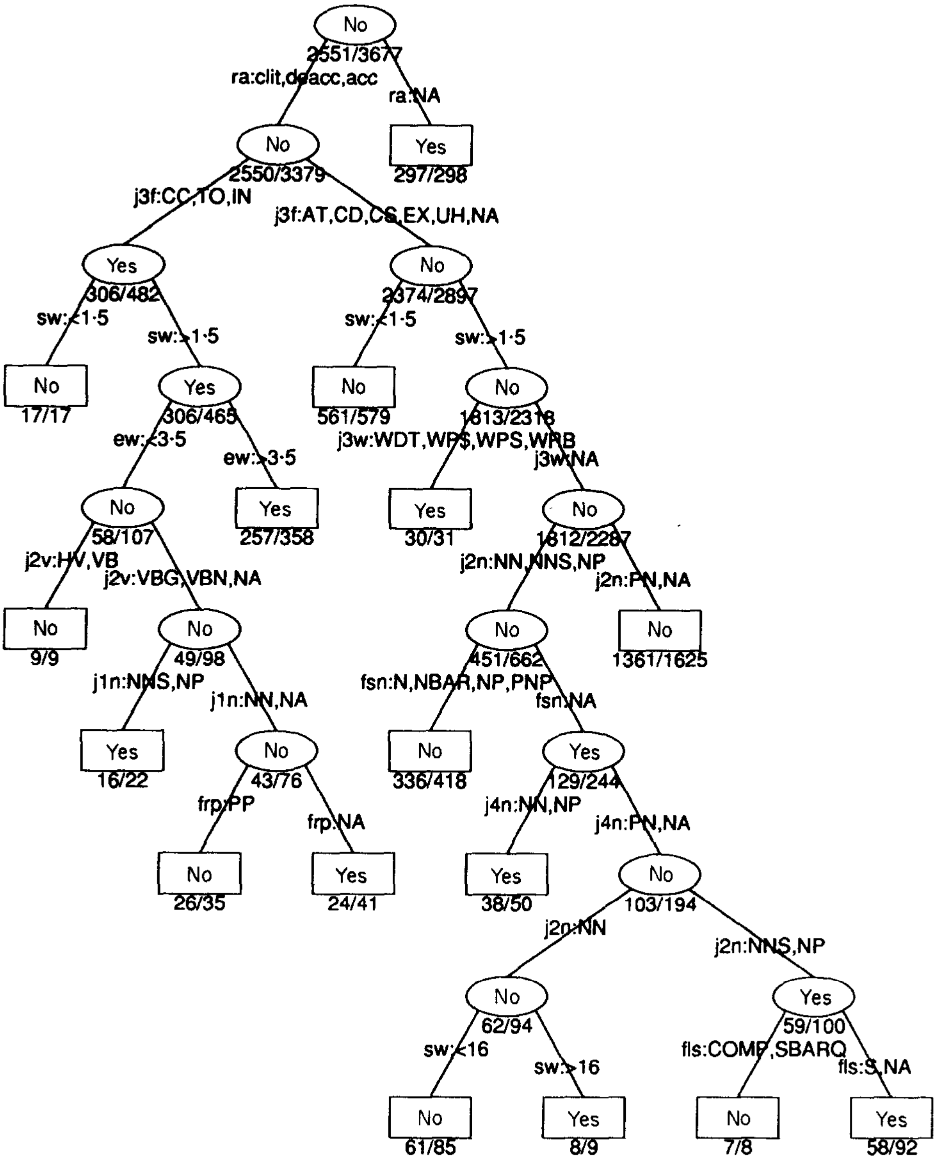

Figure 10. Boundary prediction from automatic text analysis alone, disfluences classed as boundaries.

in classifying observed intonational boundaries does so correctly in about 80% of cases (see Table I); this tree classifies "null boundary" cases correctly in around 93% of cases. However, the tree which performs best at classifying null boundaries, with 99·2% successfully classified, classifies observed boundaries correctly only 62% of the time (see Table II).

So, to correct for the imbalance in the data—and in prediction performance on boundaries vs. null boundaries—the average of the two success rates is taken to represent the overall success of a prediction tree. Thus, for the tree whose classification

TABLE I. Confusion matrix for best boundary classification

	Boundary	No boundary	Per cent correct
Boundary	895	231	79·5
No boundary	187	2364	92·7

TABLE II. Confusion matrix for best null boundary classification

	Boundary	No boundary	Per cent correct
Boundary	435	267	62·0
No boundary	25	2950	99·2

performance is given in Table I, this score is 86·1%.[7] In fact, this represents the best performance of the trees considered in the study under this new metric. So, the tree which classifies boundary data points most accurately also performs best overall. This tree uses some observed acoustic features (in particular, observed pitch accents values rather than predicted), as well as automatically inferrable feature values, and also classifies disfluencies as boundary data points. Thus, this tree was trained on more data points classified as "boundary" than other trees. The best averaged overall performance from automatically inferrable information alone is 81·7%, obtained when syntactic constituency is considered along with other variables; however, similar performance (over 80% average correct) can be obtained when constituency information is omitted.

In sum, while cross-validated results of around 90% were obtained from the original analysis, which does indeed represent a significant improvement over the 80% correct a blind assignment of "null boundary" to every data point would score, normalized scores in the eighties are probably more representative of the best performance of a given predictor tree on boundaries as well as null boundary prediction. Note that scores of 86·1% (for the best tree) and 81·7% (using only automatically inferrable information) represent major improvements over the 40% success rate a blind assignment of "null boundary" would score under the new metric (which would average 80% correct for null boundaries and 0% correct for boundaries).

As an alternative correction for the skewedness of the original training data, the null boundary data points were downsampled (by selecting random null boundary data points) to roughly the same size as the boundary data, bringing the total sample size to approximately 40% of the original. New trees were then trained on this balanced sample, using nine of the feature sets previously tested, including sets with only automatically inferrable features and other sets with acoustic features as well. In every case, prediction of observed boundaries improved while prediction of observed null boundaries declined, when compared to predictions made with the same set of features values on the full training corpus. The best mean score obtained was 91·1% correct (Table III). This score was obtained using all automatically inferred features values except for the inclusion of continuous timing variables (distance from beginning and end of utterance in seconds)—

[7] Again, this is calculated on an actual subtree whose cross-validated length is minimal in terms of classification error. The cross-validated average may vary by a few percentage points.

although not distance from prior boundaries. However, it should be noted that the CART-generated cross-validated score for this tree is only 82·2%.

5. Discussion

The application of CART techniques to the problem of predicting and detecting phrasing boundaries not only provides an automatic classification procedure for predicting intonational boundaries from text, but it increases our understanding of the importance of several among the numerous variables which might plausibly be related to boundary location. We have seen that, in fact, similar predictive accuracy can be obtained from features obtaining from automatic text analysis alone as from text

TABLE III. Confusion matrix for best classification, balanced sample

	Boundary	No boundary	Per cent correct
Boundary	621	81	88·5
No boundary	46	697	93·8

TABLE IV. Key to phrase boundary tree labels[8]

For each potential boundary, $<w_i, w_j>$	
type	Utterance type
tt	Total seconds in utterance
tw	Total words in utterance
tr	Speaking rate
st	Distance (sec.) from start to w_j
et	Distance (sec.) from w_j to end
sw	Distance (words) from start to w_j
ew	Distance (words) from w_j to end
la	Is w_i accented or not/or, cliticized, de-accented, accented
ra	Is w_j accented or not/or, cliticized, de-accented, accented
per	[Distance (words) from last boundary]/[length (words) of last phrase]
tper	[Distance (sec.) from last boundary]/[length (sec.) of last phrase]

j{1–4}	Part-of-speech of $w_{i-1,i,j,j+1}$
	v = verb b = be-verb
	m = modifier f = fn word
	n = noun p = preposition
	w = WH

f{slr}	Category of
	s = smallest constit dominating w_i, w_j
	l = largest constit dominating w_i, not w_j
	r = largest constit dominating w_j, not w_i
	m = modifier d = determiner
	v = verb p = preposition
	w = WH n = noun
	s = sentence f = fn word

analysis combined with hand-labeled features, such as observed pitch accent and timing information. However, we have also seen that performance rates are in general lower for prediction of boundary data points than for prediction of null boundaries, perhaps because our training data is skewed toward the latter. We would hope that a larger training set will povide the opportunity to sample boundary data points more accurately; our experiment in sampling the current training set indicates that trees trained on balanced data will indeed perform better on boundary data points.

In order to confirm the generality of our analysis, we are extending our corpus to include all of the sentences in the ATIS database. In addition to the current set of variables we are using for boundary prediction, we will test additional factors in our extended study. Currently, we are investigating whether or not *mutual information* scores can help to predict intonational boundary location. We will also study whether simple NP-detection provides a viable substitute for the full syntactic constituency information we have employed to date, or whether other minimal parsing strategies will provide useful constituency information for boundary prediction.

We will also examine possible interactions among the statistically important variables which have emerged from our initial study. Although CART techniques have worked extremely well at classifying phrase boundaries, our results remain something of a "black box". For an initial study like ours, CART offers a simple and immediate overview of the phenomenon. However, CART's stepwise treatment of variables, optimization heuristics and dependency on binary splits obscure the possible relationships that exist among the various factors. Now that we have discovered a set of variables which do well at predicting intonational boundary location, we must refine our understanding of just how these variables interact.

We thank Michael Riley for helpful discussions. Code implementing the CART techniques employed here was written by Michael Riley and Daryl Pregibon. Part-of-speech tagging employed Ken Church's tagger, and we used Don Hindle's parser, Fidditch, for constituency analysis.

References

Altenberg, B. (1987). *Prosodic Patterns in Spoken English: Studies in the Correlation between Prosody and Grammar for Text-to-Speech Conversion, Volume 76 of Lund Studies in English.* Lund University Press, Lund.

Bachenko, J. & Fitzpatrick, E. (1990). A computational grammar of discourse-neutral prosodic phrasing in English. *Computational Linguistics*, **16**, 155–170.

Bolinger, D. (1989). *Intonation and Its Uses: Melody in Grammar and Discourse.* Edward Arnold, London.

Bresnan, J. (1971). Sentence stress and syntactic transformations. *Language*, **47**, 257–281.

Brieman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984). *Classification and Regression Trees.* Wadsworth & Brooks, Monterrey, California.

Bruce, G., Granstrom, B. & House, D. (1990). Prosodic phrasing in Swedish speech synthesis. In *Proceedings of the Tutorial and Research Workshop on Speech Synthesis.* Autrans, France, pp. 125–128. September.

Church, K. W. (1988). A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second Conference on Applied Natural Language Processing.* Austin, Texas, pp. 136–143.

Cooper, W. & Paccia-Cooper, J. (1980). *Syntax and Speech.* Harvard University Press, Cambridge, Massachusetts.

Cooper, W. E. & Sorenson, J. M. (1977). Fundamental frequency contours at syntactic boundaries. *Journal of the Acoustical Society of America*, **62**, 683–692.

DARPA. (1990). *Proceedings of the DARPA Speech and Natural Language Workshop.* Hidden Valley, Pennsylvania. June.

Downing, B. (1970). Syntactic structure and phonological phrasing in English. PhD thesis, University of Texas, Austin.

Gee, J. P. & Grosjean, F. (1983). Performance structures: A psycholinguistic and linguistic appraisal. *Cognitive Psychology*, **15**, 411–458.

Grosjean, F., Grosjean, L. & Lane, H. (1979). The patterns of silence: Performance structures in sentence production. *Cognitive Psychology*, **11**, 58–81.

Hindle, D. (1983). Deterministic parsing of syntactic non-fluencies. In *Proceedings of the 21st Annual Meeting*. Cambridge, Massachusetts, pp. 123–128.

Hindle, D. M. (1989). Acquiring disambiguation rules from text. In *Proceedings of the 27th Annual Meeting*. Vancouver, Canada, pp. 118–125.

Hirschberg, J. (1990). Assigning pitch accent in synthetic speech: The given/new distinction and deaccentability. In *Proceedings of the Seventh National Conference*. Boston, Massachusetts, pp. 952–957.

Klatt, D. (1975). Vowel lengthening is syntactically determined in connected discourse. *Journal of Phonetics*, **3**, 129–140.

Lea, W. A. (1972). Intonational cues to the constituent structure and phonemics of spoken English. PhD thesis, Indiana University.

Lehiste, I. (1973). Phonetic disambiguation of syntactic ambiguity. *Glossa*, **7**, 197–222.

Marcus, M. P. & Hindle, D. (1985). A computational account of extra categorial elements in Japanese. In *Papers presented at the First SDF Workshop in Japanese Syntax*. La Jolla, California.

Olive, J. P. & Liberman, M. Y. (1985). Text to speech—an overview. *Journal of the Acoustical Society of America*, **Suppl. 1, 78**, S6.

O'Malley, M. M., Kloker, D. & Dara-Abrams, B. (1973). Recovering parentheses from spoken algebraic expressions. *IEEE Transactions on Audio Electroacoustics*, **AU-21**, 217–220.

Ostendorf, M., Price, P., Bear, J. & Wightman, C. W. (1990). The use of relative duration in syntactic disambiguation. In *Proceedings of the DARPA Speech and Natural Language Workshop*. Hidden Valley, Pennsylvania. June.

Pierrehumbert, J. B. (1980). The Phonology and phonetics of English intonation. PhD thesis, Massachusetts Institute of Technology. Distributed by the Indiana University Linguistics Club.

Riley, M. D. (1989). Some applications of tree-based modelling to speech and language. In *Proceedings*. DARPA Speech and Natural Language Workshop. Cape Cod, Massachusetts.

Schnabel, B. & Roth, H. (1990). Automatic linguistic processing in a German text-to-speech synthesis system. In *Proceedings of the Tutorial and Research Workshop on Speech Synthesis*. Autrans, France, pp. 121–124. September.

Selkirk, E. O. (1978). On prosodic structure and its relation to syntactic structure In *Nordic Prosody II* (Fretheim, T., ed.) TAPIR, Trondheim.

Selkirk, E. (1984). *Phonology and Syntax*. MIT Press, Cambridge, Massachusetts.

Steedman, M. (1990). Structure and intonation in spoken language understanding. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*. Pittsburgh, Pennsylvania.

Streeter, L. (1978). Acoustic determinants of phrase boundary perception. *Journal of the Acoustical Society of America*, **63**, 1582–1592.

Umeda, N. (1982). Boundary: Perceptual and acoustic properties and syntactic and statistical determinants. *Speech and Language*, **7**, 333–371.

Wales, R. & Toner, H. (1979). Intonation and ambiguity. In *Sentence Processing: Psycholinguistic Studies Presented to Merrill Garrett* (Cooper W. E. and Walker E. C., eds). Halsted Press, New York.