Predicting Intonational Boundaries Automatically from Text: The ATIS Domain

Michelle Q. Wang
Churchill College
Cambridge University
Cambridge UK

Julia Hirschberg
AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill NJ 07974

February 11, 1991

Abstract

Relating the intonational characteristics of an utterance to other features inferable from its text is important both for speech recognition and for speech synthesis. This work investigates techniques for predicting the location of intonational phrase boundaries in natural speech, through analyzing a utterances from the DARPA Air Travel Information Service database. For statistical modeling, we employ Classification and Regression Tree (CART) techniques. We achieve success rates of just over 90%.

1 Introduction

Intuitively, intonational phrasing divides an utterance into meaningful 'chunks' of information [3]. Variation in phrasing can change the meaning hearers assign to tokens of a given sentence. For example, '*Bill doesn't drink because he's unhappy*' is likely to be interpreted one way when uttered as a single phrase (i.e., Bill drinks, but not because he's unhappy) and another when uttered with a boundary between *drink* and *because* (the cause of Bill's failure to drink is his unhappiness).

While phrase boundaries are perceptual categories, they are associated with certain acoustic features. Generally, phrases may be identified by one of more of the following features: pauses (which may be filled or not), changes in amplitude and in the pitch contour, and lengthening of the final syllable in the phrase (sometimes accompanied by glottalization of that syllable and perhaps preceding syllables). Major phrase boundaries tend to be associated with longer pauses, more pronounced contour excursions, and greater amounts of final lengthening than minor boundaries.

2 Inferring Phrasing from Text

How the intonational phrasing of an utterance is related to aspects of the text uttered is potentially an important source of information for speech recognition, to constrain the set of allowable hypotheses by identifying boundary locations in both the recognized text and the acoustic signal or to moderate durational information at likely boundary locations. However, to date, syntactically-based prediction of intonational boundaries has met with limited success. While considerable work has been done on the relationship between some particular syntactic configurations and intonational boundaries [12, 2, 6, 9], the prediction of boundaries in unrestricted and spontaneous speech rarely been attempted [1].[1] Predicting boundaries solely from information available automatically from text analysis presents a further challenge, which must also be addressed if predictions are to be useful in real spoken language systems.

To address these issues, we experimented with the prediction of intonational boundaries from text analysis, using 298 utterances from 26 speakers in the Air Travel Information Service (ATIS) database for training and testing.[2] To prepare data for analysis, we labeled the speech prosodically by hand, noting location and type of intonational boundaries and presence or absence of pitch accents, using both the waveform and pitchtracks of each utterance. Although major and minor boundaries were distinguished in the labeling process, in the analysis presented below these are collapsed. Each data point in our analysis consists of a potential boundary location in an utterance, defined by a pair of adjacent words $< w_i, w_j >$. There are 3677 potential boundary locations $< w_i, w_j >$ in the ATIS sample analyzed here.

For each potential boundary site, we examine the predictive power of a number of textual features whose values can be determined from orthographic transcriptions of the ATIS sentences, as well as a number of phonological categories features available from our hand-labeling, to see, first, how well boundary locations can be predicted automatically from text,

[1] Bachenko and Fitzpatrick classify 83.5-86.2% of boundaries correctly for a test set of 35 sentences; Ostendorf et al report 80-83% correct prediction of boundaries only on a different 35 sentence test set. Altenberg models only major boundaries for a portion of his training data, 48 minutes of partly-read, partly spontaneous speech from a single speaker.

[2] These sentences were selected from the 772-odd utterances in the original TI collection.

and, second, whether prediction using fuller information, currently available only via hand-labeling, can improve performance significantly.

Temporal variables used in the analysis include utterance and phrase duration, and distance of the potential boundary from various strategic points in the utterance. Although it is tempting to assume that phrase boundaries represent a purely intonational phenomenon, it is possible that processing constraints help govern their occurrence. So, for example, longer utterances may tend to include more boundaries. 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 also appears likely to be correlated with boundary location. The tendency to end a phrase may also be affected by the position of the potential boundary site in the utterance. For example, positions very close to the beginning or end of an utterance may well be unlikely positions for intonational boundaries. We measure this variable too, both in seconds and in words. The importance of phrase length has also been proposed [6, 2] as a factor in boundary location. Simply put, it may be that consecutive phrases have roughly equal length. To test this, we calculate the elapsed distance from the last boundary to the potential boundary site, divided by the length of the last phrase encountered, both in time and words. To obtain this information from text analysis alone would require us to factor prior boundary predictions into subsequent predictions. While this would be feasible, it is not straightforward in our current analysis strategy. To see whether this information is useful, therefore, we currently use observed boundary location.

Syntactic constituency information is widely considered a major factor in phrasing [6, 14, 11, 15]. That is, some types of constituents may be more or less likely to be broken up into phrases, and some constituent boundaries may be more or less likely to coincide with intonational boundaries. To test the former, we examine the class of the lowest node in the parse tree to dominate both w_i and w_j, as determined by Hindle's parser, Fidditch [7]. To test the latter we determine the class of the highest node in the parse tree to dominate w_i, but not w_j, and similarly for w_j but not w_i. Word class is often used to predict boundary location, particularly in text-to-speech, where simple parsing into function/content word groupings generally controls the generation of phrase boundaries. To test the importance of word class, we examine part-of-speech in a window of four words surrounding each potential phrase break, using Church's part-of-speech tagger [5].

Informal observation suggests that phrase boundaries are more likely to occur in some PITCH ACCENT contexts than in others. For example, phrase boundaries between words that are DEACCENTED seem to occur much less frequently than boundaries between two accented words. To test this, we look at the pitch accent values of w_i and w_j for each $< w_i, w_j >$, comparing observed values with predicted pitch accent information obtained from [8].

Finally, in a multi-speaker database, an obvious variable to test 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. Since we have found no significant increase in predictive power when this variable is used, results presented below are speaker-independent.

3 Analysis and Results

For statistical modeling, we employ Classification and Regression Tree (CART) analysis [4] to generate decision trees from sets of continuous and discrete variables. At each stage in growing the tree, CART determines which factor should govern the forking of two paths from that node. Furthermore, CART must decide which values of the factor to associate with each path. Ideally, splitting rules should choose the factor and value split which minimizes the prediction error rate. The rules in the implementation employed for this study [13] approximate optimality by choosing at each node the split which minimizes the prediction error rate on the training data. In this implementation, all these decisions are binary, based upon consideration of each possible binary partition of values of categorical variables and consideration of different cut-points for values of continuous variables.

Stopping rules terminate the splitting process at each internal node. To determine the best tree, this implementation uses two sets of stopping rules. The first set is extremely conservative, resulting in an overly large tree, which usually lacks the generality necessary to account for data outside of the training set. To compensate, the second rule set forms a sequence of subtrees. Each tree is grown on a sizable fraction (80%) of the training data and tested on the remaining portion. This step is repeated until the tree has been grown and tested on all of the data. The stopping rules thus have access to cross-validated error rates for each subtree. The subtree with the lowest rates then defines the stopping points for each path in the full tree. Results presented below all represent cross-validated data.

Prediction rules label label the terminal nodes. For continuous variables, the rules calculate the mean of the data points classified together at that node. For categorical variables, the rules choose the class that occurs most frequently among the data points. The success of these rules can be measured through estimates of deviation. In this implementation, the deviation for continuous variables is the sum of the squared error for the observations. The deviation for categorical variables is simply the number of misclassified observations.

In analyzing our data, we employ four different sets of variables. The first includes observed phonological information about pitch accent and prior boundary location, as well as automatically obtainable information. The success rate of boundary prediction from this set is quite high, with correct cross-validated classification of 3330 out of 3677 potential boundary sites — an overall success rate of 90% (Figure 1). Furthermore, there are only five decision points in the tree. Thus, the tree represents a clean, simple model of phrase boundary prediction, assuming accurate phonological information.

Turning to the tree itself, we that the ratio of current phrase length to prior phrase length is very important in boundary location. This variable alone (assuming that the boundary site occurs before the end of the utterance) permits correct classification of 2403 out of 2556 potential boundary sites. Occurrence of a phrase boundary thus appears extremely unlikely in cases where its presence would result in a phrase less than half the length of the preceding phrase. The first and last decision points in the tree are the most trivial. The first split indicates that utterances virtually always end with a boundary — rather unsurprising news. The last split shows the importance of distance from the beginning of the utterance in boundary location; boundaries are more likely to occur when more than 2 $\frac{1}{2}$ seconds have elapsed from the start of the utterance.[3] The third node in the tree indicates that noun phrases form a tightly bound intonational unit. The fourth split in 1 shows the role of accent context in determining phrase boundary location. If w_i is not accented, then it is unlikely that a phrase boundary will occur after it.

The importance of accent information in Figure 1 raises the question of whether or not automatically inferred accent information (via [8]) can substitute effectively for observed data. In fact, when predicted accent information is substituted, the success rate of the classification remains approximately the same, at 90%. However, the number of splits in the resultant tree increases — and fails to include the accenting of w_i as a factor in the classification! A look at the errors in accent prediction in this domain reveals that the majority occur when function words preceding a boundary are incorrectly predicted to be deaccented. This appears to be an idiosyncracy of the corpus; such words generally occurred before relatively long pauses. Nevertheless, classification succeeds well in the absence of accent information, perhaps reflecting a high correlation between predictors of accent and predictors of phrase boundaries. For example, both pitch accent and boundary location are sensitive to location of prior intonational boundaries and part-of-speech context.

In a third analysis, we eliminate the dynamic boundary percentage measure. The result remains nearly as good as before, with a success rate of 89%. This analysis reconfirms the usefulness of observed accent status of w_i in boundary prediction. By itself (again assuming that the potential boundary site occurs before the end of the utterance), this factor accounts for 1590 out of 1638 potential boundary site classifications. This analysis also 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 our analysis by producing a classification tree that uses text-based information alone. For this analysis we use predicted accent values and omit information about prior boundary location. Figure 2 shows results of this analysis, with a successful classification of 90% of the data. In Figure 2, more variables are used to obtain a classification percentage similar to the previous classifications. Here, accent predictions are used trivially, to indicate sentence-final boundaries (*ra*='NA'), a function performed in Figure 1 by distance of potential boundary site from end of utterance (*et*). The second split in 2 does rely upon temporal distance — this time, distance of boundary site from the beginning of the utterance. Together these measurements correctly predict 38.2% of the data. The classifier next uses a variable which has not appeared in earlier classifications — the part-of-speech of w_j. In 2, in the majority of cases (88%) where w_j is a function word other than 'to,' 'in,' or a conjunction (true for about half of potential boundary sites), a boundary does not occur. Part-of-speech of w_i and type of constituent dominating w_i but not w_j are further used to classify these items. This portion of the classification is reminiscent of the notion of 'function word group' used commonly in assigning prosody in text-to-speech, in which phrases are defined, roughly, from one function word to the next. Overall rate of the utterance and type of utterance appear in the tree, in addition to part-of-speech and constituency information, and distance of potential boundary site from beginning and end of 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 similar rates of success.

4 Discussion

The experiments described above indicate that it is indeed possible to relate intonational boundaries to the text of an utterance with fair success,[4] using information available automatically using current NLP technology. This application of CART techniques to the problem of predicting phrase boundaries increases our understanding of the importance of several among the numerous variables which might plausibly be related to boundary location. Future word will extend the set of variables for analysis to include distance metrics defined in terms of stressed syllables, automatic NP-detection [5], MUTUAL INFORMATION, GENERALIZED MUTUAL INFORMATION scores can serve as indicators of intonational phrase boundaries [10]. We will also examine possible interactions among the statistically important variables which have emerged from our initial study. CART's step-wise treatment of variables, optimization heuristics, and dependence 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 need to understand just how these variables interact.

While we have not yet attempted the parallel classification of boundary sites from acoustic information for the ATIS sample, previous research [12] and our own preliminary analysis of a a smaller set of training data collected for the VEST (Voice English-Spanish Translation) project, suggest that in-

[3] This fact may be idiosyncratic to our data, given the fact that we observed a trend towards initial hesitations.

[4] For purposes of comparison with classification efforts that measure only success of boundary prediction (not success of non-boundary prediction as well), the best cross-validated prediction from the analyses done for this study has a 79.5% success rate and the best prediction from a full tree classifies 89.7% correctly.

tonational boundaries can be identified with some success from simple measures of final lengthening (inferred from relative word or syllable duration) and of pausal duration. For the VEST data, for example, boundary location can be inferred correctly from such metrics in 92% of cases. In future work, these features, as well as amplitude and other potential boundary indicators will be examined in the ATIS database.

References

[1] B. Altenberg. *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, 1987.

[2] J. Bachenko and E. Fitzpatrick. A computational grammar of discourse-neutral prosodic phrasing in English. *Computational Linguistics*, 1990. To appear.

[3] D. Bolinger. *Intonation and Its Uses: Melody in Grammar and Discourse.* Edward Arnold, London, 1989.

[4] L. Brieman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Wadsworth & Brooks, Monterrey CA, 1984.

[5] K. W. Church. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second Conference on Applied Natural Language Processing*, pages 136–143, Austin, 1988. Association for Computational Linguistics.

[6] J. P. Gee and F. Grosjean. Performance structures: A psycholinguistic and linguistic appraisal. *Cognitive Psychology*, 15:411–458, 1983.

[7] D. M. Hindle. Acquiring disambiguation rules from text. In *Proceedings of the 27th Annual Meeting*, pages 118–125, Vancouver, 1989. Association for Computational Linguistics.

[8] J. Hirschberg. Assigning pitch accent in synthetic speech: The given/new distinction and deaccentability. In *Proceedings of the Seventh National Conference*, pages 952–957, Boston, 1990. American Association for Artificial Intelligence.

[9] I. Lehiste, J. Olive, and L. Streeter. Role of duration in disambiguating syntactically ambiguous sentences. *Journal of the Acoustical Society of America*, 60:1199–1202, 1976.

[10] D. M. Magerman and M. P. Marcus. Parsing a natural language using mutual information statistics. In *Proceedings of AAAI-90*, pages 984-989. American Association for Artifical Intelligence, 1990.

[11] M. P. Marcus and D. Hindle. A computational account of extra categorial elements in japanese. In *Papers presented at the First SDF Workshop in Japanese Syntax.* System Development Foundation, 1985.

[12] M. Ostendorf, P. Price, J. Bear, and C. W. Wightman. The use of relative duration in syntactic disambiguation. In *Proceedings of the DARPA Speech and Natural Language Workshop.* Morgan Kaufmann, June 1990.

[13] M. D. Riley. Some applications of tree-based modelling to speech and language. In *Proceedings.* DARPA Speech and Natural Language Workshop, October 1989.

[14] E. Selkirk. *Phonology and Syntax.* MIT Press, Cambridge MA, 1984.

[15] M. Steedman. Structure and intonation in spoken language understanding. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, 1990.

tt	utterance length in seconds
tw	utterance length in words
st	seconds from start to w_j
et	seconds from w_j to end
sw	words from start to w_j
ew	words from w_j to end
la	is w_i accented or not/ cliticized, deaccented, accented, NA
ra	is w_j accented or not/ cliticized, deaccented, accented, NA
per	[# words from last boundary] / [# words in of last phrase]
tper	[seconds from last boundary] / [seconds from last phrase]
j1-4	part-of-speech for $w_{i-1}, w_i, w_j, w_{j+1}$: v:verb b:copula m:modifier f:function word n:noun p:preposition w:*wh*-word
fs,l,r	category of: s:smallest constituent dominating w_i and w_j l:largest constituent dominating w_i not w_j r:largest constituent dominating w_j not w_i m:modifier d:determiner v:verb p:preposition w:*wh*-word n:noun s:sentence f:function word

Table 1: Key to Node Labels in Figures

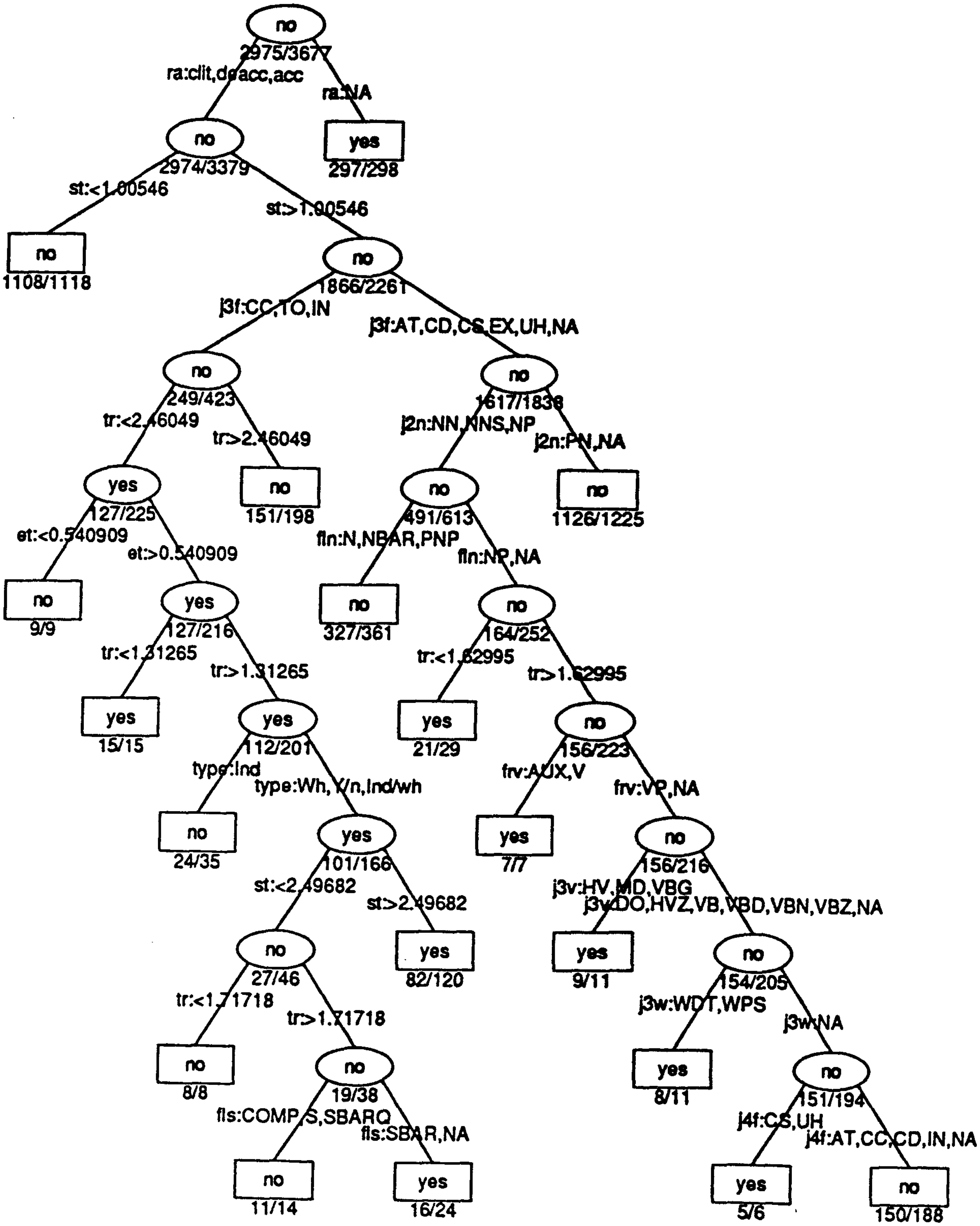

Figure 2: Phrase Boundary Predictions from Text Analysis Alone, 90%

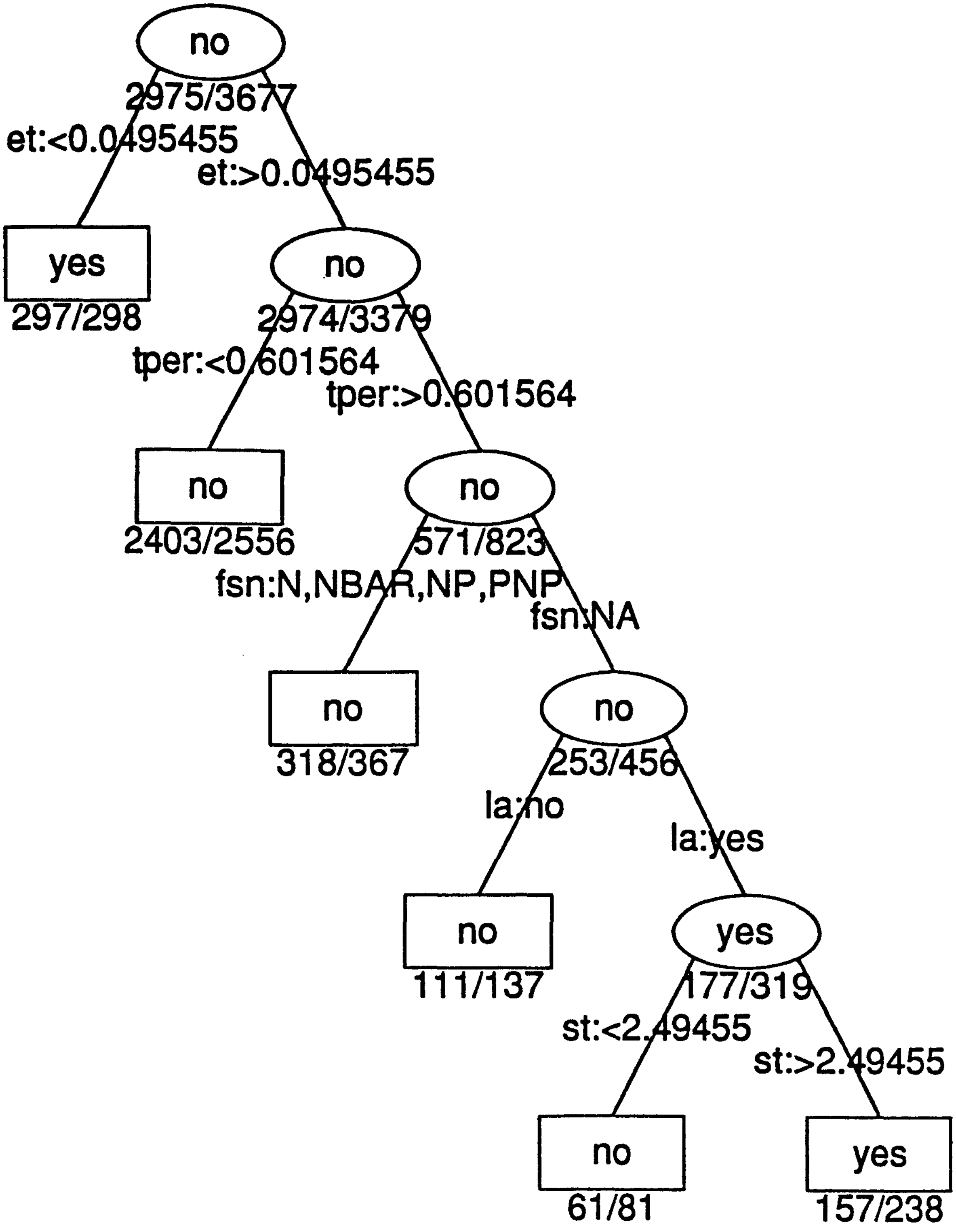

Figure 1: Phrase Boundary Predictions from Text and Observed Accents, 90%