A comparison of two tools for analyzing linguistic data: logistic regression and decision trees

David Eddington

The present paper compares logistic regression (referred to herein as its implementation in Varbrul) with another method for analyzing linguistic data-decision trees. Comparison of the two methods demonstrates that decision trees are able to find the same sorts of generalizations as Varbrul. However, decision trees provide more coarsely-grained output compared with Varbrul's more informative factor weights. In addition, decision trees often mistakenly overgeneralize. Nevertheless, decision trees can be used in tandem with Varbrul. Because decision trees automatically calculate interactions, they suggest interaction terms that may be considered in subsequent Varbrul analyses. Decision trees also allow continuous variables in contrast to Varbrul's instantiation of logistic regression which is limited to categorical variables. Therefore, decision tree analysis may help establish cutoff points when continuous data are converted into categories for Varbrul. Data sets containing knockouts and multinomial dependent variables, as well as those containing cells with zeros, are a challenge for Varbrul analysis. These are usually dealt with by recoding and reconfiguring the data. However, in some cases no amount of principled recoding is able to yield a parsimonious Varbrul analysis. Therefore, decision trees are suggested as an alternative method of analysis since they are not adversely affected by these factors. In order to contrast and compare the two methods, Varbrul and decision tree analyses of a number of linguistic data sets are presented. *

1. Introduction

A great deal of linguistic research involves analyzing competing structures, morphemes, or phones, and attempting to find the context in which one occurs rather than another. There are cases in which there is only one governing factor, as in English where it is only the initial phone of a word that determines whether the article *an* or *a* precedes the word. However, it is more often the case that several different factors play a part. For example, the choice of plural suffix in German is influenced by the word's gender, phonemic make-up, status

* Thanks to Robert Bayley, Mark Waltermire, and Catherine Travis for their input and critique of this paper.

 (Received June 2009)

as a proper name, etc. In these instances the goal is to find significant patterns in the data; we want to know which factors influence the choice of suffix, which factor is associated with each suffix, and the strength of each factor on the choice of suffix. In other words, we want to predict the value of a dependent variable on the values of a number of independent variables.

One tool for answering questions of this nature is logistic regression. Varbrul (Rousseau & Sankoff 1978), which in its more recent versions is known as GoldVarb (Rand & Sankoff 1990; Robinson, Lawrence & Tagliamonte 2001) and GoldVarb X (Sankoff, Tagliamonte & Smith 2005), is a program that is tailor-made for applying logistic regression to the sorts of data that linguists, especially sociolinguists, are confronted with. It evaluates the likelihood that each independent variable co-occurs with the dependent variable, and calculates the strength of each relationship. It models these relationships and allows different models to be evaluated in terms of how well they fit the data. During stepwise testing it determines whether a particular independent variable adds any additional predictive power to the model. Those that do not are eliminated.

There is no doubt that the fields of sociolinguistics and variationist linguistics would not have progressed to their present state were it not for the widespread application of the logistic regression analysis these programs have made freely available to the linguistic community.

In this paper, I wish to present a tool that is much less well-known than Varbrul in linguistic circles, but has potential in the quantitative analysis of linguistic data–decision trees. In particular, I contrast and compare Varbrul and decision tree analyses and show how the two programs may be used together. I also discuss how decision trees may help analyze data that do not lend themselves to Varbrul analysis.

2. Decision trees

Decision trees are often used in the field of machine learning. They are designed to mine a data set and find all possible generalizations or relationships between the independent variables (i.e. factor groups) and a categorical dependent variable. This is similar to the way that Varbrul calculates what factors favor or disfavor a particular dependent variable. Decision tree programs have many uses. For example, large medical databases contain information about patients'

symptoms, blood pathology, etc. along with the diagnosis of each patient. These data are computationally sifted through in order to discover what combination of symptoms is most likely to indicate which malady. In political polling, the political leanings of individuals may be predicted based on combinations of spending habits, educational level, television programs viewed, brand of wine recently purchased, etc. Decision tree algorithms provide the information on which such predictions are based.

A number of different decision tree programs have been developed (CART: Breiman, Freidman, Olshen & Stone 1984; C4.5: Quinlan 1993; R: Maindonald & Braun 2003; Venebles & Ripley 2002). The analyses reported on in the present paper were carried out using C4.5. However, details of how to perform decision tree analyses as well as the particular differences between the different decision tree algorithms are beyond the scope of the present paper. However, the essence of decision tree analysis is that they operate by partitioning the data into sets with the same values of a variable. They first find the independent variable that accounts for the largest majority of the variation in the dependent variable and partition the data into sets, called branches, based on that variable. Each of the resulting branches is further subdivided based on the values of other independent variables until all of the data is accounted for. The result is a tree structure that indicates what independent variable or combination of variables is associated with particular values of the dependent variable.

Decision trees often overfit the data; that is, they become overly complex and contain many branches that do not make good generalizations about the data. In order to overcome this difficulty a number of algorithms have been designed to determine which branches do not yield statistically significant predictions (e.g. Baayen 2008, Quinlan 1993). Branches that do not add any predictive value are pruned from the tree. This results in a tree that better models the data, and is easier to interpret.

2.1. Linguistic applications of decision trees

Decision trees have been used extensively in natural language processing and corpus linguistic tasks but their use as a tool outside of these computationally intense fields of linguistics is much more limited. However, Daelemans, Berck & Gillis (1997) used decision trees to study Dutch diminutive allomorphy, while Eddington and Lachler (forthcoming) mined Navajo verb stems for generalizations.

Ernestus & Baayen (2003, 2004) investigated neutralized segments and past tense morphology in Dutch with decision trees, while the English past tense was studied by Ling & Marinov (1993) using a decision tree algorithm. Decision trees have been applied to detection of stress in English (Xie, Andrae, Zhang & Warren 2004), and the discovery of complementary distribution of English stop phonemes (Randolph 1990). To my knowledge, only a handful of researchers (Akama, 2003; Mendoza-Denton, Hay & Jannedy 2003; Breiman, Freidman, Olshen & Stone 1984) have applied decision trees algorithms to sociolinguistic data. The study by Mendoza *et al.* is notable in that it compares the results of a decision tree analysis with that of Varbrul.

2.1.1. Labov's Department Store Study

In order to demonstrate the sort of outcome calculated by a decision tree, I compare it with a Varbrul analysis of Labov's classic department store study (1972; see also Paolillo 2002 who has done extensive analysis of these data using Varbrul). [1] In Labov's study, department store clerks were asked which floor a particular item of merchandise could be found on. The researcher always asked about merchandise located on the fourth floor in order to observe whether or not the clerk pronounced the /r/ in *fourth* and *floor*. In each instance, the researcher asked the clerk to repeat the floor number in order to elicit a second, more emphatic response. The independent variables used in this comparison are the particular store the clerk was employed in (Saks, Macy's, Klein's), the word (*fourth*, *floor*), and the speech type (first response-normal, second response-emphatic). The dependent variable was the presence or absence of /r/.

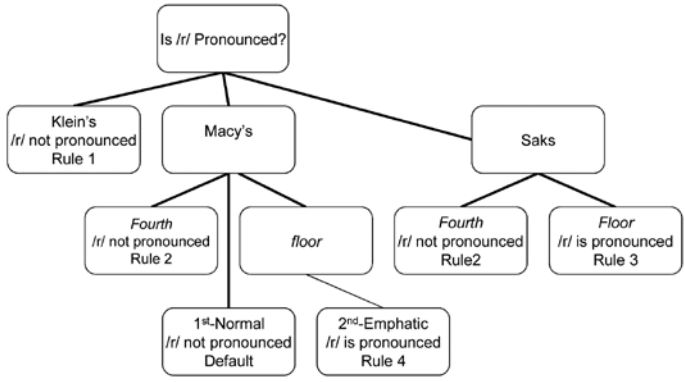
An initial Varbrul analysis was done in which each variable was added one at a time in order to determine whether they significantly add to the predictability of the model. This showed that speech type did not significantly affect the pronunciation of /r/, so that variable was eliminated and the analysis rerun with only the two remaining variables. In Table 1, higher factor weights indicate that a variable favors the pronunciation of /r/, while lower weights indicate that the pronunciation of /r/ is disfavored. Therefore, the Varbrul analysis indicates that clerks in Saks and Macy's favored the pronunciation of /r/, while those in Klein's favored its deletion. The word *fourth* favored deletion of /r/, while its retention was favored in *floor*.

Table 1. Varbrul analysis of factors contributing to pronunciation of /r/ in department stores.

		Factor weight
Store	Saks	.706
	Macy's	.602
	Klein's	.204
Word	*floor*	.626
	fourth	.385
Speech Type		n.s.

Total chi-square = 2.2123; Chi-square per cell = 0.3687; Input = 0.275; Log likelihood = -396.501

The decision tree of the same data appears in Figure 1. From this tree four rules are calculated, each with a success rate that indicates how often the rule renders the correct outcome. In any other case not covered by these rules, the default pronunciation is calculated to be r-less. The overall success rate of these rules and default is 73.0%. In the decision tree, a variable that appears higher up in the tree influences the dependent variable more than one below it. Therefore, the store is the most influential followed by the word. The speech type exerts the least amount of influence. The decision tree shows that clerks in Klein's disfavor /r/ in comparison to those in Macy's and Saks. The pronunciation of *floor* is also shown to be generally r-full in comparison with the r-lessness of *fourth*.



Rules:

1. Clerks in Klein's do not pronounce /r/ (195/216, 90.3% correct).
2. *Fourth* is pronounced without an /r/ (192/270, 71.1% correct).
3. Clerks in Saks pronounce the /r/ in *floor* (52/82, 63.4% correct).
4. Clerks in Macy's pronounce /r/ in *floor* as an emphatic second response (31/51, 60.8% correct).

Default: r-less (62/110, 56.4% correct)

Overall correct: 73%

Figure 1. Decision tree for Labov's department store data (1972).

One could protest that the decision tree analysis produces results that are obvious by merely 'eyeballing' the data. That is, the rules produced by the program are already apparent in the graphs and charts that Labov provides in his department store study (1972: 51-53), so what advantage does the method have? The answer is that charts and graphs are tools that researchers have at their disposal to make sense of raw data. I suggest that decision trees provide additional useful tools for linguistic research. The department store data are also relatively simple; there are 729 tokens and the analysis is limited to three variables. I chose this simple set in order to contrast and compare the two types of analysis. However, decision tree programs are extremely powerful and are able to efficiently sort through thousands and even tens of thousands of tokens, and hundreds of variables searching for generalizations. With more complex data such programs are adept at finding generalizations that are not readily apparent from a cursory inspection.

A number of differences between decision trees and Varbrul are apparent. First, Varbrul is able to distinguish between variables that do and do not add to the model to a statistically significant degree. The decision tree, in contrast, makes no claims about the significance of its rules beyond how often they correctly apply to the information in the database. However, pruning techniques (e.g. Baayen 2008; Quinlan 1993) may be applied to decision trees that remove branches that are less likely to generalize beyond the data at hand. While these methods increase a tree's generalizability, they do not provide levels of statistical significance.

A second difference is that the factor weights calculated by Varbrul allow graded comparisons to be made between the members of a variable. For example, employees in Macy's and Saks favored retention of /r/ to similar, but not equal degrees (.706 and .602 respectively), while clerks in Klein's highly disfavored retention (.204). The decision tree algorithm used in the analysis does not provide such fine-grained distinctions. Perhaps the major difference between the analyses is that the decision tree automatically generates interactions between the variables.

It is fairly common practice for variationist studies to consider only the main effects of the variables if the analysis yields a parsimonious model of the data. However, Sigley (2003) reanalyzed a number of studies and found that many relevant interactions between variables were present that had not been explored previously. A number of different methods for dealing with interactions are available for Varbrul analyses such as cross-product recoding,

dummy-interaction recoding, and data set partitioning (Paolillo 2002, Tagliamonte 2006). For example, Lucas, Bayley, Rose & Wulf (2002) utilized cross-product recoding in a Varbrul analysis of hand position variation in American Sign Language. At first, they obtained a great deal of interaction in their data. They eliminated the interaction by combining the separate variables for race (African American or Caucasian) and social class (middle class or working class). The new variable had four values: Caucasian middle class, Caucasian working class, African American middle class, and African American working class. [2] This recoding removed the interactions from the statistical analysis which allowed a valid model to be constructed in Varbrul. At the same time, it allowed the interactions themselves to be identified.

Although methods for finding interactions in Varbrul exist, they require the interactions to be found by hand. I would like to suggest that decision trees provide a tool for automatically determining interactions, which can then be reevaluated using Varbrul. For example, in spite of the fact that the main effects produce a model that fits Labov's department store data quite well, the decision tree in Figure 1 suggests two-way and three-way interactions. For this reason, I followed Paolillo's (2002) analysis of Labov's data and created new interaction variables for Varbrul: store by word, store by speech type, speech type by word, and store by word by speech type. These variables were added to the original variables and the analysis rerun.

Table 2. Comparison of Varbrul and decision tree analyses of the interactions between factor that contribute to the pronunciation of /r/.

Interaction variables	*Factor weight*	*Decision tree rules*	
Klein's normal fourth	0.113	Rule 1, 2	r-less
Klein's emphatic fourth	0.285	Rule 1, 2	r-less
Klein's normal floor	0.184	Rule 1	r-less
Klein's emphatic floor	0.361	Rule 1	r-less
Macy's normal fourth	0.520	Rule 2	r-less
Macy's emphatic fourth	0.419	Rule 2	r-less
Macy's normal floor	0.673	Default	r-less
Macy's emphatic floor	0.805	Rule 4	r-full
Saks normal fourth	0.522	Rule 2	r-less
Saks emphatic fourth	0.639	Rule 2	r-less
Saks normal floor	0.825	Rule 3	r-full
Saks emphatic floor	0.823	Rule 3	r-full

Total Chi-square = 0.0000; Chi-square per cell = 0.0000; Input = 0.273; Log likelihood = -390.197

Under these circumstances only the three-way interaction was found to be significant and the resulting model fit the data much better judging by the log likelihood which moved from – 396.501 in the main effects analysis to -390.197 in the interaction analysis (χ^2 (6) = 12.608, $p < 0.05$). The comparison of the Varbrul and decision tree analyses in Table 2 is revealing. While decision trees automatically calculate interactions, they produce a much coarser-grained binary output: /r/ is predicted to either be present or absent. Varbrul's factor weights, on the other hand, provide a gradient scale that indicates the degree to which a certain combination of variables favors or disfavors the pronunciation of /r/. Its numerically interpretable outcome makes it easier to make more detailed distinctions about the influence of each variable. In addition, the decision tree has overpredicted the degree of r-lessness in some cases. The tendency for decision trees to overfit the data has been noted previously by Maindonald & Braun (2003).

In the present study, overfitting manifests itself in the prediction that Macy's clerks' non-emphatic pronunciation of *floor*, as well as the emphatic pronunciation of *fourth* by Saks employees are r-less. This contradicts the corresponding feature weights (0.673, 0.679) that indicate that an r-full pronunciation is actually favored under those circumstances. In sum, the decision tree is helpful in pointing out interactions that need to be explored, but its output is not as detailed as the factor weights generated by Varbrul. The decision tree is adept at finding the broader generalizations in the data, but in some cases its output is inaccurate because it has overfitted the data.

2.1.2. Assibilation of /ʁ/ in Piripiri Portuguese

In the state of Piaui in northeast Brazil, the town of Piripiri is noted for its particular assibilated variety of /ʁ/. In most Brazilian dialects, /ʁ/ has a number of possible realizations (i.e. [r, h, x, χ, ʁ]), however, in Piripiri it is often pronounced as a voiceless apical alveolar fricative [ś] or voiceless alveopalatal fricative [ʃ] as in *quarta* [kwaśtə]~[kwaʃtə] 'fourth' when it is followed by /t/. A Varbrul analysis of the data (Taylor & Eddington 2006) found a number of social variables influencing the assibilation of /ʁ/ before /t/ (see Table 3). The upper class strongly favors a non-assibilated pronunciation as does the lowest age group. The middle class favors the non-assibilated variety to a lesser degree than the upper class. Participants 41 or older strongly disfavored the non-assibilated pronunciation.

Table 3. Varbrul analysis of factors contributing to non-assibilation of /ʁ/ in Piripiri.

		Factor Weight
Social Class	Upper	0.739
	Middle	0.501
	Lower	0.298
Age	18-40	0.744
	41-60	0.383
	61+	0.352
Gender		n.s.
Word		n.s.

Total Chi-square = 59.7453; Chi-square per cell = 0.8298; Input = 0.923 Log likelihood = -219.251

The first three rules derived from a decision tree analysis of the same data reflect the findings of the Varbrul. They are quite general in that they apply to a large number of items in the data set (239, 150, and 131 respectively). However, Rules 4-6 apply to a much smaller number of items (38, 29, and 5) and as a result may be considered spurious generalizations. Nevertheless, Rules 5 and 6 indicate interactions in the data, which suggest that interactions should be considered in a Varbrul analysis as well. To this end, the data were recoded to create new interaction variables which were evaluated alongside the main effects groups (Table 4).

Table 4. Decision tree rules for Piripiri /ʁ/.

1. 18-40 year olds do not assibilate /ʁ/ (231/239, 96.7% correct).
2. The middle class does not assibilate /ʁ/ (130/150, 86.7% correct).
3. The upper class does not assibilate /ʁ/ (126/131, 96.2% correct).
4. The /ʁ/ in the word *Fortaleza* is not assibilated (34/38, 89.5% correct).
5. 41-60 year old males do not assibilate /ʁ/ (25/29, 86.2% correct).
6. Lower class, 61+ year old males DO assibilate /ʁ/ in the word *forte* (3/5, 60% correct).

Default: no assibilation (83/118, 70.3% correct)
Overall correct: 89%

Table 5. Comparison of Varbrul and decision tree analyses of the interactions between variables related to non-assibilation.

Interaction variables	*Factor weight*	*Decision tree rules*	
Lower-class, 41-60 years	0.151		
Lower-class, 61 or older	0.218		
Middle-class, 61 or older	0.274	Rule 2	no assibilation
Middle-class, 41-60 years	0.476	Rule 2	no assibilation
Upper-class, 60 or older	0.563	Rule 3	no assibilation
Lower-class, 18-40 years	0.650	Rule 1	no assibilation
Upper-class, 18-40 years	0.723	Rule 1, 3	no assibilation
Middle-class, 18-40 years	0.730	Rule 1, 2	no assibilation
Upper-class, 41-60 years	0.837	Rule 3, 5	no assibilation

Total Chi-square = 105.9396; Chi-square per cell = 0.7357; Input = 0.924; Log likelihood = -214.467

The only variable that was found to be significant was the class by age interaction group (Table 5). The interactions found by the decision tree were unsurprisingly not chosen as significant given their limited applicability. However, the interaction analysis paints a clear picture in which speakers from the upper class, regardless of their age, favor a non-assibilated /ʁ/. In like manner, the youngest speakers, regardless of their class, also favor a non-assibilated /ʁ/. The decision tree also captures this state of affairs, but incorrectly extends the lack of assibilation to the older middle class speakers as well (Rule 2). As far as these data are concerned, the decision tree analysis does not point to significant interactions (although the analysis of the interactions in Varbrul yields important findings), and it gives a much less accurate portrayal of the social variables that influence the variation in question.

2.2. Continuous variables

One limitation of Varbrul is that it is designed to work only with categorical data (e.g. sex, race, vowel/consonant/pause, etc.) but not with continuous data such as age, yearly income, or formant frequency (Bayley 2002). This is not a drawback of logistic regression itself, but only of Varbrul since other statistical packages that perform logistic regression, such as SPSS, allow continuous variables. Of course, continuous data may always be converted into categorical data by dividing them into groups (e.g. Age: 30-39/40-49/50-59; Income: under $39,000, $40,000-69,000, $70,000 or higher). However, such predetermined divisions may obscure natural divisions that exist in the

data unless different groupings are experimented with, which again requires a good deal of data reconfiguration.

Determining natural divisions between continuous data is automatically carried out by decision trees. As an example, I used a subset of the data from the BYU Syllabification Survey (Eddington, Treiman & Elzinga *forthcoming*). In the survey, the participants were presented words such as *lemon* and asked to determine whether they would syllabify the word *le – mon* or *lem – on*. A third option was to mark *"I'm not sure"*. What is of interest for the present paper are what variables influenced the *"I'm not sure"* responses. Variables included the tenseness or laxness of the nucleus of the first syllable, whether the medial consonant was a sonorant or obstruent, whether the second syllable was stressed or not, and the log frequency of the word. The log frequency ranged between zero and 13.55. Since log frequency is continuous, it was divided into three groups for the Varbrul analysis: 0-4.3, 4.4-8.7, and 8.8 and higher. As Table 6 indicates, the frequency was the only significant variable chosen during the stepping up and down analysis with higher frequency words disfavoring *"I'm not sure"* responses and lower frequency words favoring it.

Table 6. Variables favoring the "I'm not sure" responses in the syllabification survey.

		Factor Weight
Log Frequency	0-4.3	0.727
	4.4-8.7	0.459
	8.8 and higher	0.394
Vowel Quality		n.s.
Consonant Quality		n.s.
Stress		n.s.

Total Chi-square = 18.8929, Chi-square per cell = 0.7872; Input 0.889; Log likelihood = -875.075

However, the middle frequency group falls in the middle which makes its influence hard to determine. In these circumstances, a decision tree analysis that includes the actual frequency rather than a categorized frequency was helpful. It yields a rule to the effect that words with a log frequency over 5.16 are not given *"I'm not sure"* responses. This rule is quite robust in that it applies to 1,781 of the 2,482 items and applies correctly in 92.5% of the cases. This cutoff point falls squarely in the middle of the middle frequency group used in the Varbrul analysis, which may account for its middle-of-

the-road factor weight (.459). Therefore, readjusting the frequency categories based on the figure calculated during the decision tree analysis should provide a Varbrul model with a better fit and more clearly interpretable results. As Table 7 indicates, by taking the value provided by the decision tree algorithm as a cutoff point the factor weights move away from the middle and yield a better fitting model. This is another example of how decision trees and Varbrul can work in tandem.

Table 7. Variables favoring the "I'm not sure" responses in the syllabification survey using cutoff points determined by decision tree analysis.

		Factor Weight
Log Frequency	0-2.58	0.769
	2.59-5.16	0.704
	5.17-7.74	0.427
	7.75 and higher	0.379
Vowel Quality		n.s.
Consonant Quality		n.s.
Stress		n.s.

Total Chi-square = 25.8611; Chi-square per cell = 0.8082; Input 0.895; Log likelihood = -855.608

2.3. Knockouts

Most researchers who have used Varbrul have encountered knockouts in their data. A knockout occurs when no tokens appear in a particular variable. Variables typically include factors such as age, sex, social class, etc. Consider a study in which the deletion or retention of word final consonants in some language is measured, and the variable for age has three categories: 20-29 year-olds, 30-39 year-olds, and 40-49 year-olds. If no one between the ages of 30 and 39 deleted any of the word final consonants in question, that would leave a zero or knockout in that cell. A knockout would also exist if all of the members of that group deleted all of the consonants. Knockouts indicate that a certain variable categorically influences the dependent variable, which in this case is consonant deletion or retention. The existence of a categorical influence itself is an important finding because it gives insight into variables that influence the phenomenon studied. Therefore, knockouts need to be discussed when writing up the results of a study. However, Varbrul cannot be run with knockouts in the data, for this reason, they must be eliminated.

One way of eliminating knockouts is by gathering more data in the hopes that some token of the variable will be found. However, the most common way of dealing with knockouts is by recoding variables (Bayley & Young *forthcoming*; Paolillo 2002; Sankoff 1988; Tagliamonte 2006; Young & Bayley 1996). In the above example, the 30-39 year-old group could be merged with either the 20-29 year-old group or the 40-49 year-old group. Assuming that there is some variation in the other group, the knockout would thus be eliminated. Of course, collapsing variables in this manner must not be done at random; there must be a principled reason for doing so. As far as age is concerned, the division between the groups into spans of ten years is initially done for convenience only, so combining different age groups is not an unwarranted step. However, the combination must make linguistic sense as well. Assume that there was a great deal more deletion for the 20-29 year-olds and very little for the 40-49 year-olds. The logical step would be to combine the two groups that tend not to delete (30-39 and 40-49) rather than to combine one group that deletes the consonants a great deal with another group that never does. In fact, one could argue that at times such conflation of groups actually gives more insight into the social divisions that affect the phenomenon. In this case, retention is more pronounced for speakers 30 and older.

When all other methods of eliminating knockouts have failed, a last resort is to add a fictitious token (Paolillo 2002: 165; Guy & Bayley 1995). Variation in language is such that the knockout variable must surely exist somewhere, so adding one assumed token is not totally absurd. Of course, the fictitious token does render the resulting Varbrul analysis somewhat fictitious as well.

As far as knockouts are concerned, the more difficult cases to resolve appear to involve linguistic rather than social variables. Consider Navajo verb stems which typically have a CVC structure. The vowel of the stem varies between nasal and oral, high tone or low tone, and long or short in duration. The final consonant also varies. These changes signal differences in modes and aspect, yet finding paradigmatic relationships between these variables and particular modes or aspects has been extremely difficult. [3] I attempted a Varbrul analysis in the hopes of shedding some light on the system. The goal of the analysis was to determine what vowel and consonant qualities are favored by the different modes (imperfective, perfective, iterative, future, optative) when the aspect is held constant. Since the momentaneous aspect is the most frequent, it was chosen for the analysis.

The variables for the analysis were vowel orality, vowel tone, vowel length, and final consonant. Numerous knockouts were obtained

when each of the five modes was chosen as the application value. Some, but not all of the knockouts could be eliminated by combining the final consonants into phonetic categories (fricative, stop, nasal, etc.). However, collapsing the consonants into these categories meant that it was not possible to determine whether stem-final /t/ favored one mode in contrast to /ʔ/, for example, which was the purpose of the analysis to begin with. No other kind of recoding of the data seemed possible on principled grounds, therefore, it appears that these data are simply not amenable to Varbrul analysis, and some other method is called for.

Since decision trees are not hampered by knockouts in the data they may be used as an alternative in such cases. A decision tree analysis of the Navajo verb stem data found the generalizations in Table 8 (Eddington & Lachler *forthcoming*).

Table 8. Decision tree rules for the Navajo momentaneous verb stems.

1. Stems ending in /l, z, ʒ, Ø/ have perfective mode (165/193, 85.5% correct).
2. Stems ending in /ʔ/ whose vowel is long have perfective mode (72/100, 72% correct).
3. Stems ending in /l, d, z, ʒ, Ø, ʔ, n/ with a low tone vowel have perfective mode (137/277, 49.5% correct).
4. Stems ending in /h, s, ʃ/ with long, oral vowels have imperfective mode (192/421, 45.6% correct).
5. Stems ending in /s, ʃ/ with long vowels have imperfective mode (16/38, 42.1% correct).
6. Stems ending in /h, s, ʃ/ with short vowels have iterative mode (174/397, 43.8% correct).
7. Stems ending in /h/ with long, nasal vowels have iterative mode (87/196, 44.4% correct).
8. Stems ending in /ɬ/ have future mode (228/384, 59.4% correct).

The low success rates, coupled with the fact that rules for only four of the 13 aspects are found may suggest that the generalizations calculated are of little value. However, Navajo verb stem morphology is notoriously difficult to analyze synchronically (Leer 1979). Most pedagogical grammars (e.g. Faltz 1998; Goossen 1995) simply do not discuss the verb stem alternations because it appears that no general paradigms exist and that each verb stem needs to be memorized individually. This fact explains why the decision tree analysis results in so few rules with low rates of applicability. On the other hand, the fact that some generalizations were found speaks to the usefulness of decision tree analysis in such difficult cases.

2.4. Limitation on values of the dependent variable

Since the late 1980s, different versions of Varbrul have been produced for different operating systems. The early DOS version (Pintzuk 1988) allowed for analysis of a dependent variable with more than two values. However, the subsequent versions that have gained much wider acceptance only allow binomial variables. For many analyses this is not an obstacle, but when more than two values are present in the data it presents a challenge. Consider an analysis of Spanish /s/ weakening (e.g. Cedergren 1978) in which syllable final /s/ has three possible realizations, [s, h, Ø], and whose goal is to determine what social and linguistic variables influence each of the three realizations. The traditional method for handling multinomial variables is to convert them in to several separate binomial analyses (Paolillo 2002). One analysis calculates what variables favor the retention of [s] versus [h] and [Ø] combined. Another measures the variables that influence the pronunciation of [h] as opposed to [s] and [Ø] combined, and a third pits [Ø] against [s] and [h] combined. Contrasting [s] against the reduced pronunciations [h] and [Ø] does not seem problematic, but if [Ø] is compared with the values of the non-reduced [s] and the reduced [h] collapsed together, it seems that a linguistically unmotivated move has been made in order to force ternary dependent variables into binary ones. Such conflations also have the disadvantage of making the outcomes of the different runs difficult to compare and may conceal some potentially interesting results.

Decision trees are not limited to binomial dependent variables, and as a result, they provide an alternative analysis for data that may not lend themselves to Varbrul. For example, Eddington (2002) identified 13 different relationships a Spanish base may have with its diminutive form (Table 9). These involve whether the diminutive suffix appears after the stem or the word, and which suffix is chosen. The question that arises is what characteristics of the base word influence which diminutive form is used. The data were comprised of 2,422 bases. In many instances, one base is related to more than one type of diminutive (e.g. *mano* 'hand' > *manecita, manita, manito*; *juego* 'game' > *jueguecito, jueguito*). Variables in the analysis were the gender of the base word, whether stress falls on the final syllable, the phoneme in the nucleus of the penultimate syllable, and the word final phoneme.

Table 9. Examples of Spanish base/diminutive relationships

1. *calvo* 'bald' > *calvito*
2. *galleta* 'cookie' > *galletita*
3. *quieto* 'peaceful' > *quietecito*
4. *piedra* 'stone' > *piedrecita*
5. *pastor* 'shepherd' > *pastorcito*
6. *llave* 'key' > *llavecita*
7. *normal* 'normal' > *normalito*
8. *Isabel* 'Isabella' > *Isabelita*
9. *rey* 'king' > *reyecito*
10. *luz* 'light' > *lucecita*
11. *lejos* 'far away' > *lejitos*
12. *garrapatas* 'tick' > *garrapatitas*
13. *patrona* 'patron saint' > *patroncita*

To begin with, Varbrul analysis was not possible given the large number of knockouts. Any recoding of variable values would result in losing the information most important to the study: the specific values in a variable that favor a particular type of relationship between a base and its diminutive. This situation is largely due to the fact that some types of diminutive formation apply to only five or six of the bases in the database, while others apply to several hundred. Nevertheless, even if the knockouts could be eliminated in some principled fashion, thirteen different Varbul runs would need done in which 12 of the dependent variable values would be combined and contrasted with the remaining one. Comparing the outcome of the different runs and arriving at some sort of overall conclusion would be extremely difficult if not impossible. In short, these data cannot be analyzed with Varbrul and some other method is called for.

Table 10. Decision tree rules for Spanish diminutives

Penult Nucleus	*Final Phoneme*	*Stressed Final Syllable?*	*Gender*	*Proportion Correct*	*Percent Correct*	*Diminutive Form (Relationships from Table 9)*
Anything but ew, je, Ø	o			874/892	98	add *-ito* to the stem; *calvito*; 1
a, e, i, o, u, ej, wi	i, e, o	No	Masc.	85/109	78	add *-ito* to the stem; *calvito*; 1
Anything but ew, we, Ø	a			965/985	98	add *-ita* to the stem; *cartita*; 2
e, aj	a, e, o			7/7	100	add *-ita* to the stem; *cartita*; 2
je, we	o			67/82	82	add *-ecito* to stem; *huevecito*; 3
we			Masc.	3/3	100	add *-ecito* to stem; *huevecito*; 3
je, we	a			50/64	78	add *-ecita* to stem; *viejecita*; 4
	n, r, u, d		Masc.	143/157	91	add *-cito* to the word; *pastorcito*; 5
o, aj, ja, wa	e, i	No	Masc.	39/50	78	add *-cito* to the word; *verdecito*; 5
a, e, i, o, u, ew	n, r, d		Fem.	22/25	88	add *-cita* to the word; *cancioncita*; 6
a, o, u	e		Fem.	21/27	78	add *-cita* to the word; *llavecita*; 6
	k, l, s, t, x, θ	Yes	Masc.	80/83	96	add *-ito* to word; *normalito*; 7
	θ, l		Masc.	7/7	100	add *-ito* to word; *normalito*; 7
a, e, i, o, u, ew	θ, l		Fem.	9/9	100	add *-ita* to word; *Isabelita*; 8
Ø			Fem.	6/7	86	add *-ecita* to word; *crucecita*; 10
	s	No	Masc.	5/6	83	add *-itos* to word minus *-Vs; Carlitos*; 11
i, e, a, o, u, ew	s		Fem.	4/5	80	add *-itas* to word minus *-Vs;apenasita*; 12

A decision tree analysis is able to accommodate the multinomial nature of the dependent variable, the existence of knockouts, and the interaction between variables. It yields easily interpretable rules that state the generalizations that exist (Table 10) for 11 of the 13 diminutive patterns from Table 9. For example, the first rule in Table 10 states that a base word with anything in the penultimate nucleus except /ew, je, Ø/ (where /Ø/ indicates the word is monosyllabic), and that ends in /o/ has a diminutive in which *-ito* appears after the stem. Therefore, the diminutive of *calv+o* 'bald' is *calvito*. This corresponds to the first diminutive/base relationship stated in Table 9. Overall, the decision tree correctly accounts for 93.7% of the diminutive forms. [4] Generalizations about 11 of the 13 diminutive relationships are found that give insight into what variables favor one type of diminutive formation over another.

3. Conclusions

Quantitative data are crucial for testing hypotheses in linguistics. In many instances, hypotheses involve determining what factors influence the use of one structure, pronunciation, or lexical item over another. Often, many factors combine their influence and it is of interest to know the strength of each factor, as well as the direction of their influence. Logistic regression analysis via Varbrul has been the analytic tool of choice for many years to answer questions of this nature, especially in sociolinguistics. In contrast, decision tree analysis is not as well-known in linguistics even though it may be applied to the same kinds of data as Varbrul.

The purpose of this paper has been to contrast and compare the two methods by using them to analyze a number of different sets of linguistic data. In cases in which the results of each could be compared side-by-side both methods prove adept at capturing general trends in the data. Both are able to eliminate variables that do not help explain the dependent variable. For example, neither finds biological gender or particular word important in the assibilation of /ʁ/ in Piripiri Portuguese. In like manner, both methods give insight into which values of a variable are more influential. Varbrul expresses this as factor weight that ranges from zero to one. Decision trees, on the other hand, show how often the generalization the tree makes is correct. For example, in the department store study of /r/ deletion, the decision tree indicates that Klein's clerks did not pronounce the /r/ in *fourth floor*. This is correct in 90.3% of the cases. In like man-

ner, Varbrul gives a low factor weight of 0.204 to Kleins showing that those clerks disfavor pronunciation of /r/. The situation in Saks is quite different. The decision tree makes a rule to the effect that Saks clerks do pronounce /r/, which is correct in 63.4% of the cases. This corresponds to the high factor weight (0.706) calculated by Varbrul which demonstrates that retention of /r/ is highly favored.

There are a number of advantages and disadvantages to each method of analysis. Varbrul uses statistical significance in determining which variables aid in explaining the variation in the dependent variable. On the other hand, decision trees make no claims about the significance of the generalization they find beyond how often they correctly apply to the information in the database, although pruning algorithms help eliminate relationships that may not prove valid beyond the data set they are derived from.

At times, the existence of knockouts, cells containing zeros, multinomial dependent variables, and multiple interactions make Varbrul analysis difficult to carry out or interpret. In some instances, a Varbrul analysis is impossible in spite of the researcher's best efforts at recoding and other types of data reconfiguration. In these cases, decision trees may be used as an alternative to Varbrul since they are not hindered by such variables.

Given the strengths and weaknesses of these two analytical tools, perhaps the best conclusion is to follow the example of Mendoza-Denton *et al.* (2003) who use the two in tandem. There are two instances in which decision trees and Varbrul work especially well together. As the analysis of the department store data shows, decision trees automatically calculate interactions between variables. This means that they may guide the recoding of variables in Varbrul to reflect the interactions suggested by the decision tree. Second, decision trees, unlike Varbrul, are able to handle continuous data. The data from the syllabification study demonstrate how decision tree analysis is useful in pinpointing cutoff points. This allows one to make more precise divisions of continuous into categorical variables which may then be used in a Varbrul analysis of the data.

Address of the Author

4064 JFSB, Brigham Young University, Provo, UT 84602
<eddington@byu.edu>

Notes

[1] The data may be found at: ella.slis.indiana.edu/~paolillo/projects/varbrul/data/ds.tok. It contains 729 token which differs from the 730 upon which Paolillo's (2002) analyses are based.
[2] Since the Caucasian groups did not differ significantly from each other, they were later collapsed into a single variable as well (Bayley & Young, *forthcoming*).
[3] However, Lachler (2000) demonstrates that many subparadigms can be identified.
[4] The model calculates a default base/diminutive relationship 2 (from Table 9) applies when none of the rules in Table 10 apply.

Bibliografical References

Akama Hiroyuki 2003. Probabilistic language processing in the form of a decision tree: Usage of the French impersonal subject pronoun 'on' (or 'l'on'). Tokyo: Tokyo Institute of Technology. Ms. www.dp.hum.titech.ac.jp/~akama/on.pdf

Baayen R. Harald 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Bayley Robert 2002. The quantitative paradigm. In Chambers Jack, Peter Trudgill & Natalie Schilling-Estes (eds.). *The handbook of language variation and change*. Malden, MA: Blackwell. 117-141.

Bayley Robert & Richard Young *in press*. VARBRUL: A special case of logistic regression.

Breiman Leo, Jerome H. Friedman, Richard A. Olshen & Charles J. Stone 1984. *Classification and regression trees*. Belmont, CA: Wadsworth International Group.

Cedergren Henrietta J. 1978. En torno a la variación de la s final de sílaba en Panamá: Análisis cuantitativo. In López Morales Humberto (ed.). *Corrientes actuales en la dialectología del Caribe Hispánico*. Hato Rey, PR: Universidad de Puerto Rico. 35-50.

Daelemans Walter, Peter Berck & Steven Gillis 1997. Data mining as a method for linguistic analysis: Dutch diminutives. *Folia Linguistica* 31. 57-75.

Eddington David 2002. Spanish diminutive formation without rules or constraints. *Linguistics* 40. 395-419.

Eddington, David, Rebecca Treiman & Dirk Elzinga *forthcoming*. The syllabification of American English: Evidence from a large-scale experiment. Brigham Young University. Ms.

Eddington David & Jordan Lachler *forthcoming*. A computational analysis of Navajo verb stems. In Newman John & Sally Rice (eds.). *Experimental and empirical methods*. Stanford, CA: Center for the Study of Language and Information.

Ernestus Mirjam & R. Harald Baayen 2003. Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language* 79. 5-38.

Ernestus Mirjam & R. Harald Baayen 2004. Analogical effects in regular past tense production in Dutch. *Linguistics* 42. 873-903.

FALTZ Leonard M. 1998. *The Navajo verb: A grammar for students and scholars*. Albuquerque: University of New Mexico Press.

GOOSSEN Irvy W. 1995. *Diné bizaad: Speak, read, write Navajo*. Flagstaff, AZ: Salina Bookshelf.

GUY Gregory R. & Robert BAYLEY 1995. On the choice of relative pronouns in English. *American Speech* 70. 148-162.

LABOV William 1972. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

LABOV William 2001. The anatomy of style-shifting. In ECKERT Penelope & John R. RICKFORD (eds.). *Style and sociolinguistic variation*. Cambridge: Cambridge University Press. 85-108.

LACHLER Jordan 2000. Verb Stem Ablaut in Navajo: A Regular Irregularity. In HENDERSON M. T. (ed.). *Proceedings of the 1999 Mid-America Linguistics Conference*. Lawrence, KS: University of Kansas Linguistics Department. 241-251.

LEER Jeff 1979. *Proto-Athabaskan Verb Stem Variation. Part One: Phonology*. Fairbanks: Alaska Native Language Center.

LING Charles X. & Marin MARINOV 1993. Answering the connectionist challenge: A symbolic model of learning the past tense of English verbs. *Cognition* 49. 235-290.

LUCAS Ceil, Robert BAYLEY, Mary ROSE & Alyssa WULF 2002. Location variation in American Sign Language. *Sign Language Studies* 2. 407-440.

MAINDONALD John & John BRAUN 2003. *Data analysis and graphics using R*. Cambridge: Cambridge University Press.

MENDOZA-DENTON Norma, Jennifer HAY & Stephanie JANNEDY 2003. Probabilistic sociolinguistics: Beyond variable rules. In BOD Rens, Jennifer HAY & Stephanie JANNEDY (eds.). *Probabilistic linguistics*. Cambridge, MA: Massachusetts Institute of Technology Press. 97-138.

PAOLILLO John C. 2002. *Analyzing linguistic variation: Statistical models and methods*. Stanford, CA: Center for the Study of Language and Information.

PINTZUK, Susan. 1988. *Varbrul programs for MS DOS*. Philadelphia: University of Pennsylvania.

QUINLAN J. R. 1993. *C4.5: Programs for machine learning*. San Mateo, CA: Morgan Kaufmann.

RAND, David and David SANKOFF. 1990. *GoldVarb: A variable rule application for MacIntosh*. Montréal: Centre de recherches mathématiques, Université de Montréal. Program available at http://individual.utoronto.ca/tagliamonte/goldvarb.htm

RANDOLPH Mark A. 1990. A data-driven method for discovering and predicting allophonic variation. In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, S14.10. 1177-1180.

RIETVELD Toni & Roeland VAN HOUT 1993. *Statistical techniques for the study of language and language behavior*. Berlin: Mouton de Gruyter.

ROBINSON John, Helen LAWRENCE & Sali A. TAGLIAMONTE 2001. GoldVarb 2001: A multivariate analysis application for Windows. Program available at http://individual.utoronto.ca/tagliamonte/goldvarb.htm

ROUSSEAU Pascale & David SANKOFF 1978. Advances in variable rule methodology. In SANKOFF David (ed.). *Linguistic variation: Models and methods*. New York: Academic Press. 57-69.

SANKOFF David 1988. Variable rules. In AMMON Ulrich, Norbert DITTMAR, and Klaus J. MATTHEIER (eds.). *Sociolinguistics*. Berlin: Walter de Gruyter. 984-997.

SANKOFF David, Sali A. TAGLIAMONTE & Eric SMITH 2005. *Goldvarb X: A multivariate analysis application*. http://individual.utoronto.ca/tagliamonte/Goldvarb/GV_index.htm

SIGLEY Robert 2003. The importance of interaction effects. *Language Variation and Change* 15. 227-253.

TAYLOR Michael & David EDDINGTON 2006. Negative prestige and sound change: A sociolinguistic study of the assibilation of /ʁ/ in Piripiri Portuguese. In SAGARRA Nuria & Almeida Jacqueline TORIBIO. *Selected Proceedings of the 9th Hispanic Linguistics Symposium*. Somerville, MA: Cascadilla. 320-325.

VENABLES W. N. & B. D. RIPLEY 2002. *Modern applied statistics with S*. 4th ed. New York: Springer-Verlag.

YOUNG Richard & Robert BAYLEY 1996. VARBRUL analysis for second language acquisition. In BAYLEY Robert & Dennis R. PRESTON (eds.). *Second language acquisition and linguistics variation*. Amsterdam: John Benjamins. 253-306.

XIE Huayang, Peter ANDRAE, Mengjie ZHANG & Paul WARREN. In: HOGAN J., P. MONTAGUE, M. PURVIS & C. STEKETEE, *Proceedings of the second workshop on Australasian information security, Data Mining and Web Intelligence, and Software Internationalisation*, 145-150. Darlinghurst, Australia: Australian Computer Society.