# Inferring and Revising Theories with Confidence: Analyzing the 1901 Canadian Census

Chris Drummond Chris.Drummond@nrc-cnrc.gc.ca
Institute for Information Technology
National Research Council Canada
M-50, 1200 Montreal Road
Ottawa, Ontario, Canada K1A 0R6

Stan Matwin stan@site.uottawa.ca
School of Information Technology and Engineering
University of Ottawa
Ottawa, Ontario, Canada, K1N 6N5

Chad Gaffield gaffield@uottawa.ca
Institute of Canadian Studies
University of Ottawa
Ottawa, Ontario, Canada, K1N 6N5

**Abstract**

This paper shows how machine learning can help historians analyze and understand important social phenomena. Using data from the Canadian census of 1901, we discover the influences on bilingualism in Canada at beginning of the last century. The discovered theories partly agree with, and partly complement the existing views of historians on this question. Our approach, based around a decision tree, not only infers theories directly from data but also evaluates existing theories and revises them to improve their consistency with the data. One novel aspect of this work is the use of confidence intervals to determine which factors are both statistically and practically significant, and thus contribute appreciably to the overall accuracy of the theory. When inducing a decision tree directly from data, confidence intervals determine when new tests should be added. If an existing theory is being evaluated, confidence intervals also determine when old tests should be replaced or deleted to improve the theory. Our aim is to minimize the changes made to an existing theory to accommodate the new data. To this end, we propose a semantic measure of similarity between trees and demonstrate how this can be used to limit the changes made.

# 1 Introduction

The aim of this research is to develop a tool that will help historians explore the influences on the languages spoken in Canada at the beginning of the last century. At the time of Confederation in 1867, language was a secondary issue to other concerns, most notably, religion. By the turn of the century, however, language was becoming an increasingly significant concern in Canada as in other western countries, and during the following decades, it came to be seen as a principal indicator of an individual's identity. While much research has focused on the changing official views of language in Canada, little is known about the actual linguistic abilities of the Canadian population before the later twentieth century. Despite the central role that language has played in the origins of modern Canada, our current knowledge is limited to the political, religious and educational controversies that have erupted since the 1880s.

To address this problem, we apply a machine learning algorithm to the 1901 Canadian census. For the first time, the census asked all residents in Canada three language questions: mother-tongue, ability to speak English, and ability to speak French. Our research investigates a random five-percent sample of the 1901 enumeration that has been created by the Canadian Families Project. The sample is composed of all individuals living in households that were randomly selected from each microfilm reel of the census enumeration for that year. Households were selected to permit analysis of individuals with relevant social units. The resulting sample is a cluster sample but given the nature and large size of the sample, the design effect is not a concern is this study. For a detailed analysis of this question, see Ornstein (2000). The sample includes data on 231,909 individuals over the age of five, and it allows us to explore how factors such as ethnic origin, mother-tongue, place of birth and residence, age and sex influenced the frequency of bilingualism across Canada. We build upon research that focused on the interpretive implications of how the census questions were posed, and how the actual enumeration was undertaken (Gaffield, 2000). We now focus on the responses to these questions written down by the census officials at the doorsteps of individuals and families across the country.

We are certainly interested in inferring theories directly from the data. But we are also interested in testing existing theories, such as those representing the views of politicians of that era, to see if they are confirmed, or indeed contradicted, by the data. Confirmation, or contradiction, is likely to be a matter of degree and not all parts of the theory will be affected equally. It may be possible to reduce the contradiction and avoid abandoning the theory altogether. One advantage of the approach discussed below is that it minimizes the amount a theory is changed to bring it into accordance with the data. This should help historians not only evaluate an existing theory but also to identify any erroneous assumptions on which it was based.

The algorithm we use is a decision tree learner. Decision trees are important representations that have been used extensively in machine learning research. They are easy to understand, even by non-specialists, and have been used by domain experts in many diverse applications including agriculture and law (Murthy, 1998). They have not been used extensively in the historical research community. In research on social and cultural change in Canada, the statistical tool of choice has been logistic regression (Baskerville and Sager, 1998; Darroch and Soltow, 1994). But an important property of any learning algorithm "is that it not only produce accurate classifiers (within the limits of the data) but that it also *provide insight and understanding into the predictive structure of the data*" (Breiman et al., 1984, authors' italics). We would argue that logistic regression fails in this regard. Decision trees do provide such insight and when applied to the census data will help to identify interesting population subgroups whose linguistic abilities differed from the dominant group. Furthermore, recent empirical research (Perlich et al., 2001) indicates that when working with very large datasets (in millions of examples), decision trees perform better than logistic regression. Analyzing historical census data requires learning methods that achieve high performance on data sets of this order of magnitude.

In decision tree learning, an important issue is over-fitting avoidance. A complex tree that fits the training data well typically has unnecessary structure that does not contribute to the accuracy of the tree and may even degrade it. As the number of examples increases so does the problem (Oates and Jensen, 1997). Many different algorithms have been proposed for pruning away unnecessary structure (Murthy, 1998). In this application, we also regard structure that results in only a small increase in accuracy as unnecessary. The new or modified trees are intended to be used by historians, so comprehensibility is of paramount importance. Although accuracy on new data is the main way of determining the validity of a theory, a more complex theory with only a minor improvement in accuracy is unwarranted and the simpler theory would be preferred.

Some algorithms have a parameter that controls the amount of pruning. To make the trade-off between accuracy and tree size more principled, we use confidence intervals to prune the tree rather than one of these methods. Confidence intervals are closely related to statistical significance tests which have been used for pruning by a number of researchers (Quinlan, 1986; Frank, 2000; Jensen, 1991). In recent times significance tests have been subjected to increasing criticism (Harlow et al., 1997). It has long been known that statistical significance and practical significance are not the same thing. Statistical significance tests give no indication of the size of the effect. Even very small effects will be statistically significant if there is a sufficiently large amount of data. Using confidence intervals allows the determination of not only a statistically significant improvement in the accuracy of the tree, but also to quantify the size of the improvement. A test then will only be added to the tree if the expected accuracy gained is sufficiently large to justify it.

One way to evaluate an existing theory represented by a tree is to compare its accuracy to a tree grown directly from the data. If the difference is large the existing theory might be rejected. But even if a theory has relatively poor accuracy, a small change might improve its accuracy substantially. It would seem sensible to only reject the theory completely if the changes required were also large. To quantify the size of a change, a measure is needed of the difference between two theories. For decision trees, one measure might be the syntactic change in the form of tree, say the number of tests added or deleted. An alternative is a measure based on some notion of the semantics of a tree.

The semantic measure proposed here is based on viewing a decision tree as a particular labeling of the attribute space. Under this geometric view, the semantics of a tree is determined by how it would label all future instances. Trees with the same partitioning of the attribute space into classes would be equivalent semantically even though they might be syntactically different. The semantic difference between two trees can then be determined by the number of potential instances classified the same way. But even this does not seem to capture the full semantics of a decision tree designed by an expert. Not only are semantics dependent on which attributes are chosen but also on the order in which they are chosen. Attributes that are closer to the root are likely to be considered more important in classifying the instances. To include these two semantic influences in decision tree learning, we generate synthetic instances that are consistent with the expert's tree designed to represent an existing theory. When modifying an expert's tree, the performance of tests selected based on the synthetic data can be compared to those selected based on some mixture of the synthetic data and new instances gathered from the domain. Using a mixture biases the system towards using attributes from the existing theory. Confidence intervals then determine when old attributes should be replaced by new ones or deleted altogether to improve the theory.

This paper builds on previous work by the authors (Drummond et al., 2002). Section 2 explores in greater depth the use of confidence intervals to prune decision trees and thereby control the trade-off between comprehensibility and accuracy. In Section 3, this technique is now applied not only to decision trees but also to probability estimation trees. It is shown that both types of tree are useful in identifying interesting population subgroups from the 1901 Canadian census data. Section 4 discusses in detail our semantic measure. It shows how the measure is combined with confidence intervals and new data to control the revision of an existing tree. Section 5 shows how this approach is used to evaluate and revise an existing theory on the influences on bilingualism in Canada in 1901.

## 2 Inducing A Decision Tree

A binary tree is used to represent the theories induced from the data, the same representation used by the well known CART algorithm (Breiman et al., 1984). A binary tree may be deeper than a tree with a greater branching factor if the same attribute is tested multiple times for different values. Binary tests should, however, help historians determine not only what are the important attributes but also the critical values of those attributes. The tree is grown in the standard greedy manner, the best test, according a splitting criterion, is selected to be added to the tree. The main difference in our approach is that a test is actually added only when there is a high confidence that a worthwhile increase in accuracy will result.

$$f(a, v) = \max_{a,v} |P(L_{a,v}|+) - P(L_{a,v}|-)| \tag{1}$$

We use the splitting criterion proposed by Utgoff et al. (1997) and shown in equation 1. The continuous form of this criterion is a commonly used statistic, the Kolmogorov-Smirnov distance. The best split is the one with the greatest difference in the estimated probability of a positive instance going left $P(L_{a,v}|+)$ and a negative instance going left $P(L_{a,v}|-)$. The criterion is applied to each attribute and each value and the attribute-value with the greatest difference is selected. This value becomes the left branch of the split and the right branch represents the remaining values of the attribute. The difference in likelihood provides a measure of the probability that positive and negative examples come from different distributions. A large difference tends to produce branches with a large difference in class ratios. Further splits should ultimately lead to better accuracy. Likelihood difference is completely insensitive to the class distribution, but some research (Drummond and Holte, 2000) suggests this is an advantage rather than a disadvantage. Although it is prior insensitive, it has a close relationship to accuracy, which we exploit in generating a test statistic. It also does not have as strong a tendency as other criteria towards purity on one side of the split. This property should be useful in reducing bias when growing probability estimation trees (Zadrozny and Elkan, 2001).

## 2.1 Choosing a Statistic

Our aim is to only add tests that improve the accuracy of the tree by a useful amount. But when greedily growing a decision tree often adding a single test does not improve accuracy at all; tests on multiple attributes are needed. We need a measure which is correlated with accuracy but will not suffer from this problem. We could use chi-square or indeed the difference in likelihood that we use as a splitting criterion. Other measures used as splitting criteria such as information gain are also potential candidates. For this application, the fact that these measures nearly always improve with additional tests is a disadvantage. When testing an existing theory, we also want to determine if removing tests is likely to improve accuracy. As a compromise, we use a measure based on accuracy but with a modified class distribution. Accuracy often does not improve when a test is added due to the strong imbalance in classes away from the root node. Reducing this imbalance means the measure is more likely to show improvement when a single test is added but also produces negative values.

To meet the necessary independence assumptions, our statistic is applied to separate pruning data (Jensen, 1991). The test chosen by the splitting criterion partitions this data, producing a contingency table as shown in Figure 1. The rows represent the class of the instances. Looking at the numbers not in parentheses, there are 32 positive and 8 negative examples, for a total of 40 instances. The columns show the number of instances going to the left, 26, and right, 14. If each side is labeled according to the majority class, there is no increase in accuracy. Yet, more of the positives go to the left and more of the negatives to the right which seems desirable. If the number of positives and negatives was closer to equality, say 24 and 16 as indicated by the numbers in parentheses, the sides of the split would be labeled differently and there would be an increase in accuracy.

| | Left | Right | |
|---|---|---|---|
| Pos | 24 (18) | 8 (6) | 32 (24) |
| Neg | 2 (4) | 6 (12) | 8 (16) |
| | 26 (22) | 14 (18) | 40 |

Figure 1: A Contingency Table

When applying the test statistic, a confusion matrix is produced from the contingency table. Based on the training data, the side of split where the positive likelihood is greater than the negative likelihood is labeled positive and the other side negative. Equation 2 gives the accuracy of the split if the left and right hand sides are labeled positive and negative respectively. Here, the role of the probability of each class, $P(-)$ and $P(+)$, is evident. To make a statistic less sensitive to class distribution, the values are replaced by ones closer to 0.5. If the terms are rearranged and the class probabilities set to 0.5, equation 3 results. This is essentially the difference between the likelihood functions of the two classes. This is very similar to our splitting criterion except that it is not the absolute difference in likelihood, it depends on how the sides of the split are labeled

$$
\begin{aligned}
Acc &= P(L|+)P(+) + P(R|-)P(-) \qquad (2)\\
&= P(L|+)P(+) + (1 - P(L|-))P(-)\\
&= P(-) + (P(L|+)P(+) - P(L|-)P(-))\\
&= 0.5 * (1 + P(L|+) - P(L|-)) \qquad (3)
\end{aligned}
$$

A series of statistics can be produced by using equation 2 and applying the squashing function $P'(a) = (P(a)+\alpha)/(1+2\alpha)$ to the class probabilities. Sensitivity to the class distribution is control by $\alpha$. When growing decision trees, we use an $\alpha$ of one. This statistic can be viewed either as accuracy with a modified class distribution or as the linear combination of two statistics, accuracy and likelihood difference. When growing probability estimation trees, we set both class probabilities to 0.5. The statistic is then just a measure of the difference in likelihood. In both cases, the statistic is divided by the fraction of instances reaching the test, and thus estimates the overall improvement in performance.

## 2.2 Pruning with Confidence Intervals

In decision tree learning, the complexity of the tree is controlled by pruning. In post-pruning, the tree is first grown until it fits the training set well and then extraneous tests, not expected to improve accuracy, are pruned away. In pre-pruning, new tests are only added if they are likely to improve accuracy. Quinlan (1986) proposed using a chi-square significance test to pre-prune the tree. If the result of adding a new test is not statistically significant then noise might account for any apparent gain in using the test. Such a test is unlikely to generalize well. In C4.5, Quinlan (1993) adopted a post-pruning technique. He argued that pre-pruning suffers from the horizon effect, measuring the gain of a single test is insufficient as multiple new tests are needed to improve accuracy. Frank (2000) experimentally compared the two techniques based on significance tests and showed there was little performance difference. The horizon effect was not found to be a problem provided the significance value was chosen appropriately.

Frank (2000) also investigated pre-pruning using a nonparametric significance test, called a permutation test. The rejection region of the null hypothesis is estimated by generating new random samples based on the data. In this paper, pre-pruning is based on confidence intervals rather than significance tests. We use a technique called bootstrapping (Efron and Tibshirani, 1993) which has been used extensively to generate confidence intervals. We follow the basic procedure proposed by Margineantu and Dietterich (2000). They used bootstrapping to generate confidence intervals for the expected difference in cost between two classifiers. We apply the same technique, but for a different purpose, as each new test is added to the tree. Rather than discuss the method in detail we refer to their paper (Margineantu and Dietterich, 2000). If two tests are being compared, we use a three dimensional matrix as they proposed. If we are considering adding or deleting a test, we use the matrix and its row marginals. We apply our test statistic to 500 randomly generated matrices. After sorting the resulting values in ascending order, the fiftieth element will be the lower bound of a 90% one-sided confidence interval.

If this lower bound is greater than zero, we are confident that the gain is statistically significant. In Figure 2 a), H0 is the null hypothesis that the difference is less than or equal to zero, H1 is the alternative hypothesis that adding the test improves accuracy. Not only is the lower bound greater than H0, it is also greater than 0.5%. We can be confident that this test would improve the accuracy of the tree by 0.5%, so the test would be added. If the bound is smaller than the chosen percentage or smaller than H0, as shown in Figure 2 b), the test would not be added. When starting with an existing theory, we are also interested is deleting structure. Applying the same test allows us, as shown in Figure 2 c), to determine that we are confident that removing structure does not degrade performance.

The values used to decide when a test should be added where chosen by the authors to represent a reasonable confidence in a useful increase in

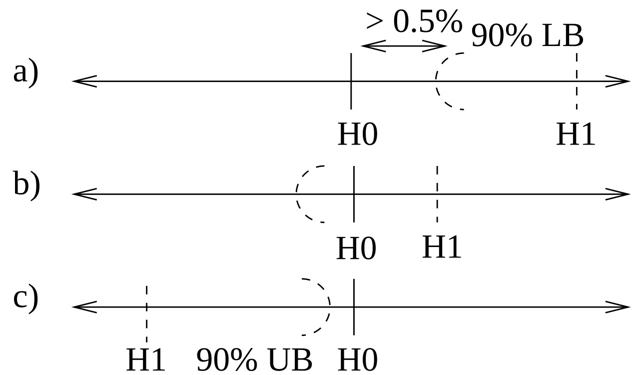


Figure 2: Using Confidence Intervals for Pruning

accuracy. Future work will investigate the effect of varying these values and changing the test statistic used to estimate the increase in accuracy.

## 3 Theories Induced from Data

In the next two sections, we explore theories generated directly from the data. We use eight attributes from the 1901 census data felt to be potentially relevant to the issue of bilingualism. Some of the nominal attributes have had their values combined into groups and the continuous attribute *age* has been divided into three intervals. To generate the class label *Bilingual*, we combined the attributes *Can speak English* and *Can speak French* but removed instances where one or both of the attributes were unknown. For the rest of this paper, unilingual will mean can speak French or English and bilingual will mean can speak both. To decide whether a new test should be added, an increase in performance of 0.5% is needed at a confidence level of

90%. Attributes will only be added if the number of instances on each side of the split is greater than 10. Numbers less than 10 might belong to a single family or a related group and be therefore of little interest. The instances are randomly split into a test and training set, 75% of the instances going to the training set. A pruning set is produced from a random 25% of the training set. The splits are all stratified to maintain the class ratios in each set.

### 3.1 Decision Trees

In this section, we look at theories in the form of decision trees. For our purposes, a good decision tree is one that accurately predicts if an individual is bilingual or unilingual from the combination of attribute-values. Figure 3 is the tree produced using the whole data set and represents the factors that affected bilingualism throughout Canada in 1901. At each leaf the classification is shown: bilingual is labeled “Y” and unilingual is labeled “N”. The most important attribute, at the root of the tree, is *mother-tongue*. The split is between those that have French as their mother-tongue “MTONGUE=FR”, and those that do not (divided into English, German, Gaelic and Others) “MTONGUE=oth”. Notably, for this latter category the tree terminates at a leaf immediately below the root. This classifies all people that do not have French as their mother-tongue as unilingual. The former category is further divided by *birth place*, those born in urban communities “BPLACE=UR” and can write are mostly bilingual. For rural communities "BPLACE=RU", this is only true for males aged 20 to 49. The accuracy gained by adding each attribute is shown to the left of the vertical line. To the right of the line, the total accuracy (80.17%) is labeled "A", the majority classifier accuracy (72.89%) is labeled "MC", the total gain in accuracy (7.28%) is labeled "G" and its 90% lower bound (7.00%) is labeled "LB". The lower bound is generated using bootstrapping on the overall confusion matrix.

```
MTONGUE=FR                            2.91|
|  BPLACE=RU                          1.66|
|  |  AGE=20-49                       1.79|
|  |  |  SEX=F: N                     0.54|A 80.17 MC 72.89
|  |  |  SEX=M: Y                         |G  7.28 LB  7.00
|  |  AGE=oth: N                          |
|  BPLACE=UR                              |
|     CANWRITE=N: N                   0.37|
|     CANWRITE=Y: Y                       |
MTONGUE=oth: N                            |
```

Figure 3: Decision Tree for Canada

We next explore how the factors that affected bilingualism varied across Canada. Figure 4 shows a map [1] of Canada in 1901 when the census was taken. The territories and districts were very sparsely populated at this time. So we combine the territories and districts into a single region, with a population size more in accordance with other regions. We also make a single region out of the eastern provinces; New Brunswick, Nova Scotia and PEI. We grow decision trees for each of the regions as shown in Figure 5.

For British Columbia, the tree consists of the single attribute *mother-*

[1] ©2001. Government of Canada with permission from Natural Resources Canada

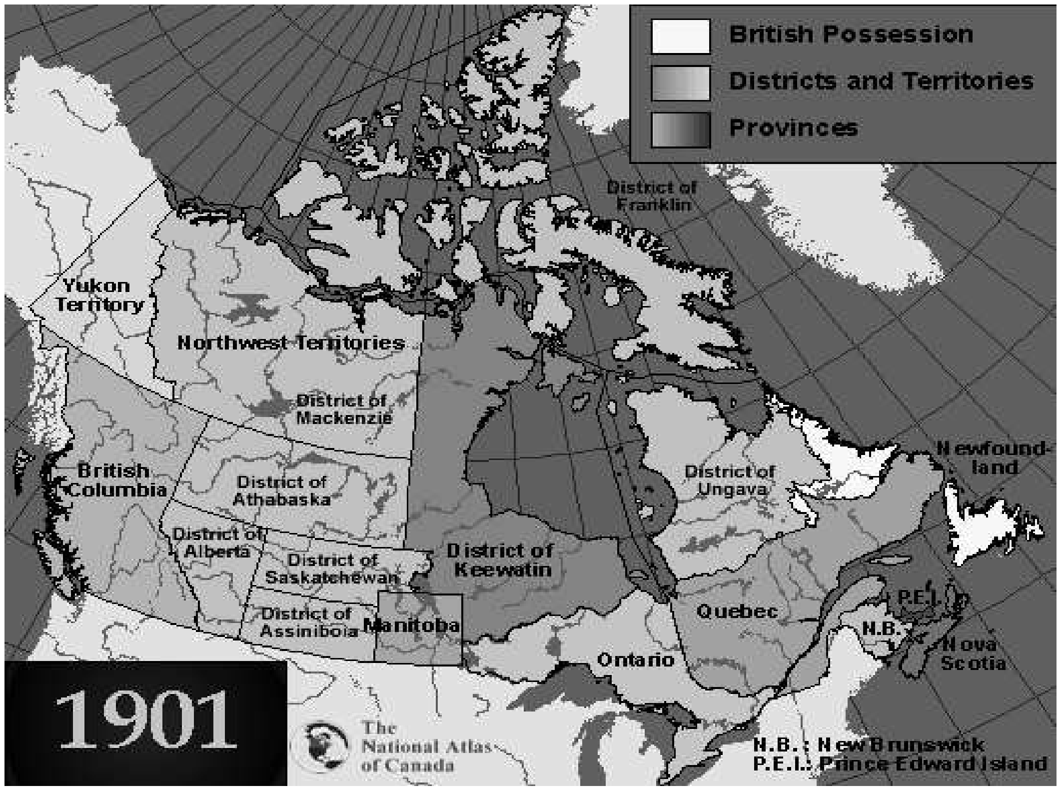


Figure 4: Map of Canada in 1901

*tongue* classifying all individuals with a mother-tongue of French as bilingual and all others as unilingual. The majority classifier is already quite accurate, see Figure 5, due the large preponderance of unilingual people in British Columbia. But using the attribute *mother-tongue* correctly predicts a bit over a third of the bilingual people without sacrificing much accuracy on the unilingual ones. Adding extra attributes produces no appreciable improvement. For the territories, the tree has the same root node, but an additional attribute *can read* improves accuracy when the mother-tongue is French.

For Manitoba, the tree also has the same root node, but the additional attribute is now *can write*. For Ontario, as for British Columbia, only the single attribute of *mother-tongue* is used. The Eastern provinces have a tree which is similar to Manitoba. *Mother-tongue* is again the most important attribute, adding the attribute *can write* is useful, although it does not improve accuracy on its own. However with an additional attribute excluding children "AGE=5-19" , accuracy is improved.

For Quebec, a quite different tree is produced. Although the attribute *mother-tongue* is used, it appears much further down the tree, close to the leaves. The most important attribute is *birth place*, whether the person was born in a rural or urban community. The attributes used on both sides of this split are very similar. However, of people born in rural communities children are immediately classified as unilingual. The overall tree is much less accurate than those of the other regions. But as there was a nearly equal number of bilingual and unilingual speakers in Quebec, it still a considerable improvement over the majority classifier.

From an algorithmic perspective, attributes seem generally to be added if, and only if, they result in an increase in accuracy at the leaves of a practically significant amount. For the larger trees this is not always the case. This might be due to using a 90% confidence limit, 10% of the time this limit will not be met. It might also be due to the test statistic not being a direct measure of accuracy. In the latter case, post-pruning using accuracy might address the problem, but this remains the subject of future

```
British Columbia
MTONGUE=FR: Y                     4.58|A 80.17 MC 72.89
MTONGUE=oth: N                        |G  4.58 LB  3.48

Territories
MTONGUE=FR                        4.52|
|  CANREAD=N: N                   0.64|A 93.54 MC 88.38
|  CANREAD=Y: Y                       |G  5.16 LB  4.24
MTONGUE=oth: N                        |

Manitoba
MTONGUE=FR                        7.66|
|  CANWRITE=N: N                  0.63|A 89.51 MC 81.22
|  CANWRITE=Y: Y                      |G  8.29 LB  6.72
MTONGUE=oth: N                        |

Ontario
MTONGUE=FR: Y                     7.06|A 94.28 MC 87.22
MTONGUE=oth: N                        |G  7.06 LB  6.69

Eastern Provinces
MTONGUE=FR                        8.64|
|  CANWRITE=N                     0.00|
|  |  AGE=5-19: N                 1.76|A 90.66 MC 80.26
|  |  AGE=oth: Y                      |G 10.40 LB  9.55
|  CANWRITE=Y: Y                      |
MTONGUE=oth: N                        |

Quebec
BPLACE=RU                         7.15|
|  AGE=5-19: N                    0.00|
|  AGE=oth                            |
|     SEX=F: N                    1.52|
|     SEX=M                           |
|        CANWRITE=N: N            0.82|
|        CANWRITE=Y                   |
|           MTONGUE=FR: Y         0.27|
|           MTONGUE=oth: N            |A 67.05 MC 53.96
BPLACE=UR                             |G 13.09 LB 12.34
   SEX=F                          0.15|
   |  CANWRITE=N: N               1.56|
   |  CANWRITE=Y                      |
   |     MTONGUE=FR: Y            0.78|
   |     MTONGUE=oth: N               |
   SEX=M                              |
      CANWRITE=N: N               0.84|
      CANWRITE=Y: Y                   |
```

Figure 5: Regional Decision Trees

work. For Quebec, it was possible to increase accuracy by about 0.7%, by reducing the confidence interval to 50% and removing the requirement for any gain. But to achieve this, the number tests went from 9 to 32 so is of debatable merit. With the test statistic we use, it is possible to produce a split where the majority class for each branch is the same. This makes no difference in accuracy and can be removed to make the tree smaller. In fact, for most of the trees this was unnecessary as no additional structure was added. The tree for Quebec had one extra test for mother-tongue being French and the Eastern provinces had one extra test for the individual's sex. Neither of these were shown in the trees given above.

From a historical perspective, the decision trees are in keeping with some, though not all, of the ways in which politicians, census officials, and other observers at the time discussed the question of bilingualism. The general assumption was that English was becoming an international language of commerce, and that if Canada were to continue developing, everyone in the country should be able to speak it. In contrast, no public figure stressed the importance of learning French. In this sense, the question of bilingualism was directed to two groups: French-language residents and immigrants (or "foreign elements" as they were called) who did not speak either French or English. The decision trees confirm that it was indeed mother-tongue francophones who accounted for much of the bilingualism in Canada. Similarly, individuals who were more likely to be involved in commerce were more bilingual as evident in the factors of literacy and birth place as well as age and sex. The importance of economic factors is also reflected in the greater tendency of middle-aged males in rural areas in Quebec (who would be more likely to be working in rural industries or in the forest economy) to be more bilingual. At the same time, this rural pattern shows one of the ways that the decision trees diverge from the theories that underlay the contemporary public debate. Specifically, the decision trees reveal an extent of diversity in language patterns that is not consistent with the ways in which observers characterized Canadian society. For the most part, for example, Quebec was assumed to be a quite homogeneous society especially in the countryside. The general picture was of a unilingual French-language rural world in Quebec that contrasted with the bilingual urban communities of Montreal and to a lesser extent Quebec City, and even more with the unilingual English-language cities outside Quebec such as Toronto. The decision trees reveal that Quebec was indeed a quite distinct part of Canada in terms of bilingualism but that within this distinction, there was considerable diversity depending upon sex, age, literacy and birth place.

### 3.2 Probability Estimation Trees

In this section, we also generate theories directly from the data, but using probability estimation trees rather than decision trees. Probability estimation trees identify sub-groups that differ in the proportion of bilingual individuals, but not in the majority class. The only difference in how these trees are grown is the test statistic used for pre-pruning, and discussed at the end of section 2.1. Figure 6 shows the probability estimation tree (PET) for Canada as a whole. The probability of being bilingual replaces the classification at each leaf, we also show these probabilities for each branch.

For measuring the overall performance of such trees, there is no clearly preferred metric such as accuracy. One measure popular in climatology for probabilistic forecasts is the Brier score. This is one of a number of measures used by Zadrozny and Elkan (2001) to assess the performance of probability estimators. The Brier score for two class problems can be written as in equation 4. It is the mean squared difference between the predicted probability of a positive instance, $P(x_i)$, and one or zero depending if the actual label $y_i$ of the instance is positive or negative respectively. As the Brier score is a measure of error, the best possible score is zero. The worst score is strictly one, but even a random guess, $P(x_i) = 0.5$, will give a score of 0.25.

$$BrierScore = \frac{1}{n} \sum_{i}^{n} (P(x_i) - y_i)^2 \tag{4}$$

We also use the Mann-Whitney-Wilcoxon statistic, shown in equation 5. This measure compares the rank of the positive and negative instances according to their probability estimates, $p_i$ and $q_j$ respectively. This is the probability than a randomly chosen positive example will rank higher than a randomly chosen negative one and is also equivalent to the area under the ROC curve (Hanley and McNeil, 1982). Here values range from zero to one, with one being the best value. In this case a random guess gives a value of

0.5.

$$
\begin{aligned}
Mann - Whitney - Wilcoxon &= \sum_{i}^{k}\sum_{j}^{l} I(p_i, q_j))^2 &\quad (5)\\
I(x,y) &= \begin{cases} 1, & \text{when } x > y \\ \frac{1}{2}, & \text{when } x = y \\ 0, & \text{when } x < y \end{cases} &\quad (6)
\end{aligned}
$$

We use the Brier score to assess the gain of adding each attribute, the numbers on the left side of the vertical line in figure 2.1. For the assessing the overall gain, shown to the right of the line, the total Brier score (0.1379) is labeled "B", the Brier score using the fundamental class frequency (0.1976) is labeled "CF", the gain in Brier Score (0.0597) is labeled "G" and its 90% lower bound (0.0583) is labeled "LB". The Mann-Whitney-Wilcoxon statistic is labeled "M"(0.8284) and its lower bound labeled "LB" (0.8253).

The base of the probability estimation tree for Canada is identical to the decision tree, but there are additional attributes, indicated by the plus signs. The original decision tree terminated in a leaf when the mother-tongue was not French. Some of these attributes are the same as those when the mother-tongue is French. Now, attributes such as *birth place*, *sex*, *age* are useful independent of the mother-tongue. The *can write* attribute seems more important when the mother-tongue is French whereas *ethnic origin*, particularly a French origin "ORIG=FR" (as opposed to English, Irish, German, Scottish and others) seems more important when it is not.

We also see that being of Irish origin "ORIG=IR" increases the probability of being bilingual.

For British Columbia, Figure 7, there is a single additional attribute *can write* for individuals whose mother-tongue is not French. It is nearly seven times more likely that someone is bilingual if they can write. This goes against the general trend in Canada where the *can write* attribute seems to be only important when the mother-tongue is French. For the territories, Figure 8, the probability estimation tree is identical to decision tree. The probability estimation tree for Manitoba, Figure 9, has no additional structure for individuals whose mother-tongue is French but adds a number of attributes for those whose mother-tongue is not. People born in rural communities are less likely to be bilingual, particularly children.

```
  MTONGUE=FR: 0.543            0.04613|
  |  BPLACE=RU: 0.468          0.00362|
  |  |  AGE=20-49: 0.576       0.00246|
  |  |  |  SEX=F: 0.452        0.00163|
  |  |  |  SEX=M: 0.675               |
  |  |  AGE=oth: 0.387                |
+ |  |     CANWRITE=N: 0.285   0.00081|
+ |  |     CANWRITE=Y: 0.435          |
  |  BPLACE=UR: 0.678                 |
  |     CANWRITE=N: 0.413      0.00185|
  |     CANWRITE=Y: 0.734             |
+ |        SEX=F: 0.648        0.00108|B 0.1379 CF 0.1976
+ |        SEX=M: 0.816               |G 0.0597 LB 0.0583
+ |           AGE=5-19: 0.703  0.00047|M 0.8284 LB 0.8253
+ |           AGE=oth: 0.873         |
  MTONGUE=oth: 0.098                  |
+    BPLACE=RU: 0.067          0.00077|
+    |  ORIG=FR: 0.338         0.00072|
+    |  ORIG=oth: 0.0587              |
+    |     AGE=5-19: 0.042     0.00011|
+    |     |  ORIG=IR: 0.064   0.00001|
+    |     |  ORIG=oth: 0.035         |
+    |     AGE=oth: 0.067             |
+    |        SEX=F: 0.051     0.00005|
+    |        SEX=M: 0.081             |
+    BPLACE=UR: 0.146                 |
```

Figure 6: PET: Canada

```
  MTONGUE=FR: 0.870            0.03587|B 0.0723 CF 0.1104
  MTONGUE=oth: 0.080                  |G 0.0382 LB 0.0297
+    CANWRITE=N: 0.0165        0.00228|M 0.7772 LB 0.7456
+    CANWRITE=Y: 0.110                |
```

Figure 7: PET: British Columbia

```
  MTONGUE=FR: 0.720            0.03970|B 0.0575 CF 0.1027
  |  CANREAD=N: 0.421          0.00551|G 0.0452 LB 0.0381
  |  CANREAD=Y: 0.872                 |M 0.7817 LB 0.7521
  MTONGUE=oth: 0.056                  |
```

Figure 8: PET: Territories

For Ontario, Figure 10, the *can write* attribute now divides individuals whose mother-tongue is French. In the decision tree, this test did not produce greater accuracy as both branches are bilingual. Again, when the mother-tongue is not French the *birth place* and *age* attributes are useful, as is *ethnic origin* being French. We also see that immigrants "IMMIG=Y" are slightly less likely to be bilingual. For the eastern provinces, Figure 11, most attributes are added to the branch for those whose mother-tongue is French. This is different from the western regions and Ontario, where most extra attributes are for the branch where French is not the mother-tongue. For individuals whose mother-tongue is French, adult males tend to be almost all bilingual particularly if they can write. Females who can write, particularly those born in urban communities, are also almost all bilingual. For those whose mother-tongue is not French, adults are twice as likely as children to be bilingual.

```
  MTONGUE=FR: 0.748             0.06040|
  |  CANWRITE=N: 0.404          0.00432|
  |  CANWRITE=Y: 0.848                 |B 0.0866 CF 0.15255
  MTONGUE=oth: 0.063                   |G 0.0660 LB 0.05475
+    BPLACE=RU: 0.048           0.00111|M 0.8662 LB 0.84462
+     |  AGE=5-19: 0.026        0.00011|
+     |  AGE=oth: 0.063                |
+    BPLACE=UR: 0.101                  |
```

Figure 9: PET: Manitoba

The probability estimation tree for Quebec, Figure 12, adds relatively little structure. But it does show one interesting feature: for those born in urban communities the division for females and males is based on the same

```
  MTONGUE=FR: 0.852             0.05860|
+ |  CANWRITE=N: 0.732          0.00068|
+ |  CANWRITE=Y: 0.902                 |
  MTONGUE=oth: 0.043                   |
+    AGE=5-19: 0.026            0.00016|
+    |  ORIG=FR: 0.203          0.00004|
+    |  ORIG=oth: 0.022                |B 5.1199 CF 0.11146
+    |     BPLACE=RU: 0.017     0.00001|G 0.0603 LB 0.05774
+    |     BPLACE=UR: 0.030            |M 0.8766 LB 0.86916
+    AGE=oth: 0.052                    |
+       BPLACE=RU: 0.040        0.00011|
+       |  ORIG=FR: 0.322       0.00038|
+       |  ORIG=oth: 0.036             |
+       BPLACE=UR: 0.067               |
+          ORIG=FR: 0.423       0.00027|
+          ORIG=oth: 0.061             |
+             IMMIG=N: 0.066    0.00001|
+             IMMIG=Y: 0.050           |
```

Figure 10: PET: Ontario

attributes in the same order, but the probabilities are lower for females. This suggests that the additional attributes are not highly context sensitive, i.e. largely independent of sex. For rural communities the order of attributes is different but again sex and the ability to write are strong predictors of bilingualism.

From an algorithmic perspective, the Brier score improves monotonically with extra structure for all the trees. But the differences between values for each added attribute are quite small, and much smaller than accuracy (allowing for the scaling factor of 100) when compared to decision trees. Thus how much gain is actually achieved is not clear, although visual inspection of the differences between the probabilities at each leaf suggests useful distinctions are being made. The Brier score can be decomposed a number of differ-

```
  MTONGUE=FR: 0.672           0.07625|
  |  CANWRITE=N: 0.529        0.00314|
  |  |  AGE=5-19: 0.298       0.00423|
  |  |  AGE=oth: 0.662               |
+ |  |     SEX=F: 0.507       0.00091|
+ |  |     SEX=M: 0.797              |
  |  CANWRITE=Y: 0.787               |B 0.0715 CF 0.1584
+ |     SEX=F: 0.730          0.00054|G 0.0870 LB 0.0822
+ |     |  BPLACE=RU: 0.686   0.00076|M 0.9095 LB 0.9017
+ |     |  BPLACE=UR: 0.949          |
+ |     SEX=M: 0.842                 |
+ |        AGE=5-19: 0.732    0.00109|
+ |        AGE=oth: 0.930            |
  MTONGUE=oth: 0.039                 |
+    AGE=5-19: 0.026          0.00004|
+    AGE=oth: 0.047                  |
```

Figure 11: PET: East

ent ways into various components. One such decomposition is discussed by Zadrozny and Elkan (2002). We are really only interested in the component that represents the difference between the estimated and true probability, typically called calibration. The small gain achieved may be partly due to the other components. They tend to be larger when the true probabilities, and therefore even well calibrated estimates, are far from zero or one. Our alternative, the Mann-Whitney-Wilcoxon statistic, clearly shows a useful gain for the all the complete trees. We informally experimented with using this to measure the gain in adding each attribute. But the gain, particularly towards the leaves, was still small. Another disadvantage of this metric is that it does not directly measure good calibration rather it measures the relative ranking of positive and negative examples. Clearly neither measure gives the intuitive insight supplied by accuracy. So further work is neeeded

```
  BPLACE=RU: 0.370             0.01120|
  |  AGE=5-19: 0.259           0.00410|
+ |  |  CANWRITE=N: 0.133      0.00140|
+ |  |  CANWRITE=Y: 0.297             |
+ |  |     SEX=F: 0.273       0.00005|
+ |  |     SEX=M: 0.321              |
  |  AGE=oth: 0.446            0.00400|
  |     SEX=F: 0.323                  |
  |     SEX=M: 0.548                  |
  |        CANWRITE=N: 0.451   0.00100|
  |        CANWRITE=Y: 0.587          |B 0.2119 CF 0.2484
  |           MTONGUE=FR:0.640 0.00056|G 0.0366 LB 0.0348
  |           MTONGUE=oth: 0.402      |M 0.7126 LB 0.7070
  BPLACE=UR: 0.584                    |
     SEX=F: 0.492              0.00333|
     |  CANWRITE=N: 0.198      0.00293|
     |  CANWRITE=Y: 0.539             |
     |     MTONGUE=FR: 0.587   0.00104|
+    |     |  AGE=20-49: 0.628 0.00004|
+    |     |  AGE=oth: 0.544          |
     |     MTONGUE=oth: 0.441         |
     SEX=M: 0.675                     |
        CANWRITE=N: 0.345      0.00345|
        CANWRITE=Y: 0.730             |
+           MTONGUE=FR: 0.800  0.00226|
+            |  AGE=5-19: 0.658 0.00124|
+            |  AGE=oth: 0.871          |
+           MTONGUE=oth: 0.573        |
```

Figure 12: PET: Quebec

to determine whether the small gain is due to our test statistic adding too much structure or is a consequence of the performance metric.

From a historical perspective, the probability estimation trees point to the importance of distinct language community sub-groups of the Canadian population some of which were the focus of public debate but most were not discussed. Overall, these trees make clear that both “English Canada” and “French Canada” were composed of quite different sub-groups both nationally and in different regions of Canada. One intriguing result that deserves further study is the identification of a sub-group of Irish-origin bilingual children in Figure 6. Previous research has shown that census officials generally attributed the father's ethnic origin to children. It appears that these children are the result of linguistically-mixed marriages. The role of such marriages in fostering bilingualism during Canada's formative decades was not emphasized at the time, and the extent of this pattern requires further investigation. It is also noteworthy that the probability estimation trees suggest that the relationship between various factors operates differently in different contexts. For example, the influence of literacy on bilingualism is inconsistent much as sex does not always appear to relate to language patterns in the same way as illustrated by urban similarity but rural dissimilarity in Quebec. The general picture that emerges from the trees is that different factors explain levels of bilingualism in different regions of Canada and among various sub-groups. In contrast, public debate at the time tended to divide Canadian society into fewer large groups, most notably, that of French Canada.

## 4 Revising an Existing Tree

In this section, we discuss how an existing tree can be modified such as to minimize the change to the underlying semantics of the theory its represents. This work has much in common with research concerned with theory revision

(Mooney, 1993; Towell and Shavlik, 1994). The main difference lies in how we quantify changes to the theory and how we use confidence intervals to decide when those changes are worth making.

## 4.1 Capturing A Tree's Semantics

In this section, we propose a way of capturing the semantics of a theory represented by a decision tree. This becomes important when trying to assess what impact new data should have on an existing theory. It is after all the effect on the semantics of a theory that we want to quantify. Using the syntactic change in a tree is an indirect way of measuring this. We felt a measure based on how a tree partitions the attribute space, that also takes into account the ordering of attributes, is a more direct semantic measure.

To capture the semantics based on how the tree partitions the attribute space, we generate instances consistent with the tree, reversing the normal process. The problem with directly converting a tree to data is the very size of attribute space. With no prior knowledge of the distribution of data, except for that directly represented in the existing tree, it would be necessary to generate instances covering the Cartesian product of the attribute values. To limit the number of instances, we generalize the notion of an instance so that the probability of an attribute having a particular value is specified. This is similar to the treatment of unknown values in C4.5 (Quinlan, 1993) except that with no knowledge of the distribution of values a uniform one is used. As in C4.5 when the attribute tested is not a single value, the instance is sent down multiple branches. By adding a weight to the instance we can simulate the effect of multiple examples without incurring the additional processing cost.

Our approach to producing an initial tree, representing a user's insights, has much in common with prior elicitation in Bayesian analysis (Madigan et al., 1995). In Bayesian analysis, an expert generates imaginary data which can be used to convert uninformed priors into more informed ones. In our approach, the user constructs a decision tree to classify a specified number of imaginary instances, say 1000. An example of what such a tree might look like is shown in Figure 13. It is based on the decision tree for Manitoba induced directly from the data and discussed in section 3. Each leaf is marked with the number of individuals from the original thousand that are bilingual and unilingual. For instance, when the mother-tongue is other than French, the number of unilingual individuals is 765 much larger than the number of bilingual individuals, just 52.

```
MTONGUE=FR
|  CANWRITE=N: N Y(17) N(25)
|  CANWRITE=Y: Y Y(120) N(21)
MTONGUE=oth: N Y(52) N(765)
```

Figure 13: A Simple Domain Theory

To generate instances consistent with the tree, each path through the tree is represented by as many instances as there are classes at the leaf. So six instances are needed to be able to regenerate this tree. Three are for the positive class, bilingual and three are for the negative class, unilingual. An instance following an upper branch has the probability of the attribute value associated with each specified test set to one. For the lower branch, the probability is a uniform distribution over the remaining values. Figure 14 shows the probability values for some of the attributes for the positive instances. The negative instances will be identical except for the weights shown at the bottom of Figure 14.

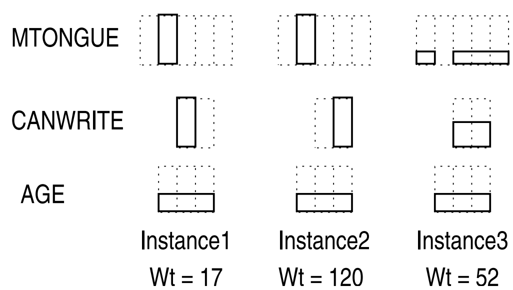


Figure 14: The Positive Instances

The attribute *mother-tongue* has five possible values, indicated by the dashed rectangles. The first two instances travel down the topmost branch of the decision tree. They have the probability of the mother-tongue being French set to one, indicated by the bold continuous rectangle. The third instance, which travels down the bottommost branch, has the probability of the mother-tongue being French set to zero and all other values of mother-tongue are set to a probability of one quarter. The first two instances travel different branches of the attribute *can write*. The first instance has a one for the "N" value, the second instance a one for the "Y" value. All unused attributes on a specific path, such as *age*, have a uniform distribution across all values.

Using these instances, it is now possible to change the order of the tests, or indeed to add a new test, and produce the same partition of the attribute space into classes. Figure 15 shows the effect of changing the root node from *mother-tongue* to *can write*. The same number of instances are classified as bilingual and unilingual. The distribution on the center branch is the same, but the top and bottom most branches have changed. As these two branches are a mixture of instances where the majority class was unilingual, they still classify instances as unilingual.

```
CANWRITE=N: N :- Y(43) N(407)
CANWRITE=Y
   MTONGUE=FR: Y :- Y(120) N(21)
   MTONGUE=oth: N :- Y(26) N(383)
```

Figure 15: Changing the Root Node

The topmost branch is made up of the first instance in Figure 14 plus half the third instance, as shown in Figure 16. The remaining half goes down the bottommost branch. The third instance had a uniform probability for *can write*. As this attribute is now the root node, this instance must be sent down both branches. This is achieved by making an additional copy of the instance. For the original instance, the probability of value “N” for *can write* is set to one, the same as the first instance. For the copy, the probability for value “Y” is set to one, the same as the second instance.

As there are only two values, the weight for both instances is set to half the original weight. If there were more, the weight is the original weight times the fraction of values represented by the branch. The number of positive instances at the leaf is now (17+52/2) or 43. There is no longer a uniform distribution for the attribute *mother-tongue*, which was different for the first and third instances. The splitting criterion would choose this attribute as a possible additional test. This would not, however, change the classification of instances. A linear scan across the instances indicates that the classification will not change if new tests are added, so no split is made.

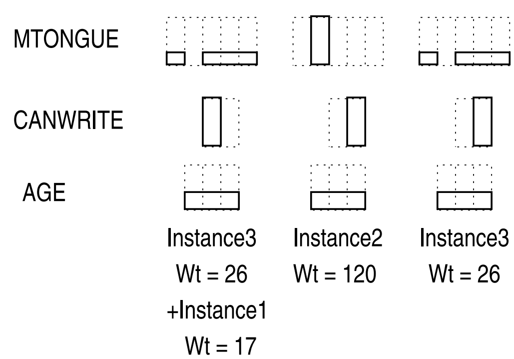


Figure 16: The Revised Instances

To construct a tree representing a user's intuitions, a single leaf node is first generated. The user then specifies the expected number of instances of each class. This leaf node can be converted to an internal node by selecting an attribute and one of its values. The system adds the selected test and generates two more leaf nodes. The user can again specify how the classes are distributed according to the test. To remove a test, the user selects an internal node converting it to a leaf. This is repeated until the user is satisfied with the structure of the tree. It is possible that the user's ordering of the attributes is not consistent with the distribution of instances the user has specified. The tree can be re-grown according to the instances or the distribution can be changed. One advantage of using the difference in likelihood as a splitting criterion is that it is relatively easy to determine how the distribution of instances must be changed to maintain the structure of the tree. At present, any adjustment must be carried out by hand by changing the distribution of classes. In future work, we intend at least to semi-automate the process by showing the alternatives to the user who can decide how the tree should be modified.

## 4.2 Updating the Tree

To update the tree at each existing test, the splitting criterion is applied to a combination of the old data generated to be consistent with the tree and the new data. By only considering tests suggested by this combination, we aim to minimise the changes made to the existing theory to accommodate the new data. If the original theory preferred certain attributes, any changes to the theory will tend to use those attributes, rather than introducing new attributes, say by promoting them higher up the tree. New tests will only be introduced if the new data has a strong preference for them. To achieve this, the splitting criterion is applied separately to the old and new data. The values returned are combined linearly to form a single value. The coefficients are determined by the number of instances, or weight, of the old data versus the number of new instances.

There are four possibilities that might occur. A new test might be added where the original tree had a leaf. The original test might be replaced by a different test. The original test might deleted altogether, or the old test maintained. To determine which takes place, confidence in the new best test is determined. If the original tree had a leaf at this node, a new test will be added to the tree if the lower bound of the confidence interval is greater than 0.5. This is the same as growing the tree directly from the data. If the new test is the same as the old test nothing will change. If the new test is different and its confidence interval exceeds the threshold it is compared to the new test. If the lower bound of the confidence interval for the difference exceeds the threshold, the test will be changed. If the new test does not exceed the threshold and the upper bound of the confidence interval on the difference does not include zero, the test is deleted.

The old and new data might also differ in how an instance should be classified at a leaf. A confidence interval can be used to decide which classification should be used. Again a bootstrapping technique is used, this time based on just the binomial ratios. At the leaf we can use lower bound of accuracy directly rather than our test statistic.

## 5 Evaluating an Existing Theory

In this section, we present an experiment showing how the method discussed in section 4 uses data to evaluate and revise an existing theory. The theory has been developed from analyses of debate in the House of Commons and newspaper coverage of political discussion about the language questions posed in the 1901 census. For a comprehensive analysis of the political debate about language, see Gaffield (2000). The decision tree representing the theory, see Figure 17, was designed to classify an imaginary 1000 people. The design exercise began by ranking attributes according to their importance in the debate that occurred in parliament and in the press at the turn of the century. Then each branch of the tree was assigned some proportion of the 1000 people, indicated by the numbers in parentheses. Next, each attribute was considered for its influence on bilingualism, and the number of unilingual and bilingual individuals was assigned. Politicians certainly did not all agree on the importance of various factors and their perceived influence on reported bilingualism, and therefore the experimental parameters represent a distillation of somewhat divergent views.

*Ethnic origin* was assessed to be the most important attribute, only those of French origin were expected to be bilingual, most other individuals were expected to be unilingual. The next most important attribute was assessed to be *birthplace*, being urban born was more strongly associated with bilingualism than being rural born. Attributes *sex*, *age* and *can write* were then added in that order. Once the tree was constructed, synthetic instances were generated to be consistent with the tree. The proportions of the classes at the leaves, indicated by the "Y()" and "N()" in the figure, were then adjusted so that the ranking of attributes was maintained, as discussed at the end of section 4.1. The tree is reasonably accurate (78.960%), only 1.2% less accurate than the tree grown directly from the data (80.169%).

```
ORIG=FR :- (400)
|  BPLACE=RU :- (212)
|  |  SEX=F : N :- 0.235 Y(20) N(65)
|  |  SEX=M :- (127)
|  |     AGE=20-49 :- (72)
|  |     |  CANWRITE=N : N :- 0.444 Y(12) N(15)
|  |     |  CANWRITE=oth : Y :- 0.556 Y(25) N(20)
|  |     AGE=oth : N :- 0.364 Y(20) N(35)
|  BPLACE=oth :- (188)
|     SEX=F :- (78)
|     |  AGE=20-49 :- (48)
|     |  |  CANWRITE=N : N :- 0.444 Y(8) N(10)
|     |  |  CANWRITE=Y : Y :- 0.667 Y(20) N(10)
|     |  AGE=oth :- : N :- 0.333 Y(10) N(20)
|     SEX=M :- (110)
|        AGE=>=50 :- (40)
|        |  CANWRITE=N : N :- 0.444 Y(8) N(10)
|        |  CANWRITE=Y : Y :- 0.545 Y(12) N(10)
|        AGE=oth : Y :- 0.714 Y(50) N(20)
ORIG=oth :- : N :- Y(50) N(550)
```

Figure 17: The Politicians' Theory

Figure 18 shows the politicians' theory after being revised using the data for the whole of Canada. This revised theory is more accurate than the politicians' theory. It is, indeed, slightly more accurate (80.204% lower bound 79.926%) than the decision tree grown directly from the data (80.169%) and shown in Figure 3. The base of the tree is identical to the tree grown from the data. Much of the structure from the theory has been deleted, but quite a lot remains, indicated by the "o" and "+" in figure 18. The most significant change to the theory is the first test, *mother-tongue* replaces *ethnic origin*. We experimented by growing some of the decision trees from section 3.1 but forcing the first test to be *ethnic origin*. The remaining attributes were identical but the resulting trees were between 1% and 2% less accurate. This suggests that replacing *ethnic origin* with *mother-tongue* accounts for most of the improvement in the revised theory. The additional structure, indicated by the "+'s" in figure 18, is the part of the politicians' theory which was not deleted when the tree was revised. It identifies two bilingual groups for people whose mother-tongue is not French. Urban males (labeled "+1") of French origin are predominantly bilingual, as are urban females (labeled "+2") of French origin, aged 20 to 49 who can write. This branch accounts for the slight increase in the accuracy of the tree. These groups were identified in the original theory. As the data supports this division, albeit very weakly, they have not been deleted.

The additional structure, indicated by the "o's", is not supported by the data even weakly. It was not deleted, however, as the tests did not indicate a statistically significant increase in accuracy. This structure does not change the classification of the tree and so could easily be deleted. The attribute sex labeled "o1" in the figure was in the probability estimation tree for Canada and did make a useful distinction between class ratios but is not useful for classification. The attribute sex labeled "o2" has a probability of being

```
   MTONGUE=FR :- 0.541
   |  BPLACE=RU :- 0.467
   |  |  AGE=20-49 :- 0.575
   |  |  |  SEX=F : N :- 0.451
   |  |  |  SEX=M : Y :- 0.674
   |  |  AGE=oth  : N :- 0.386
   |  BPLACE=UR :- 0.674
  |     CANWRITE=N :- 0.409
o  |      |  ORIG=FR :- 0.408
o2 |      |  |  SEX=F : N :- 0.303
o2 |      |  |  SEX=M : N :- 0.499
o  |      |  ORIG=oth :- : N :- 0.437
  |     CANWRITE=Y :- 0.732
o1 |        SEX=F : Y :- 0.647
o2 |        SEX=M : Y :- 0.813
   MTONGUE=oth :- 0.101
+     ORIG=FR :- 0.404
+     |  BPLACE=RU : N :- 0.343
+     |  BPLACE=UR :- 0.506
+     |     SEX=F :- 0.448
+     |     |  AGE=20-49 :- 0.581
+     |     |  |  CANWRITE=N : N :- 0.344
+2    |     |  |  CANWRITE=Y : Y :- 0.621
+     |     |  AGE=oth : N :- 0.295
+1    |     SEX=M :- : Y :- 0.570
+     ORIG=oth :- : N :- 0.089
```

Figure 18: The Revised Theory

bilingual for males of 0.499. As this is just less than 50%, all instances reaching this leaf are classified unilingual. If the classification is changed the accuracy improves very slightly to 80.024%, this is why our test did not delete it.

From an algorithmic perspective, it seems that attributes were modified and deleted when there was a clear advantage in doing so. But when the data did not support such deletion, the semantics of the original theory was maintained. From a historical perspective, the Canadian politicians of

1901 used mother-tongue to help clarify ambiguities among the labels used for ethnic groups; they did not see language as being a good identifier in and of itself. These theory revision experiments suggest that mother-tongue was more important that politicians believed at the time. But they were aware that times were changing, but probably not to the extent to which the data seems to suggest, and this led to addition of language questions to the census.

## 6 Limitations and Future Work

Limitations of this work come in two kinds: those related to the historical analysis of the census data and those related to the design of the algorithm. From a historical perspective, the census was designed to provide evidence of a single trend, the learning of English by French-language individuals and those who came to Canada speaking neither of the officially-recognized tongues. The trees point to the importance of this trend but they also show that a constellation of factors underlay the language patterns including age, sex, and rural-urban differences and this was not uniform across the country. It is for this reason more research is needed on specific geographic areas such as the so-called Bilingual Belt as well as on other data from the census including economic variables. The result of such work should be a greater appreciation for the complex ways in which language became a key feature of the making of modern Canada.

From an algorithmic perspective, the test statistic and other design choices have proven effective in practice on this data set but need to be experimentally validated on other data sets. It is worth exploring if there are alternative statistics or if the present one can be more strongly justified. Confidence in an existing theory might not constant for all parts of the theory. The existing theory determined the old tests and influenced the choice of new tests but did not affect the confidence value. An alternative would be to take a more Bayesian approach, perhaps using credible intervals rather confidence intervals, allowing locally defined confidence values. It is also worth exploring the trade-off in sizes of the pruning and training sets. A larger pruning set would give narrower confidence intervals but less data would be available to grow the tree.

# 7 Conclusions

From a historical perspective, the most compelling conclusions concern the extent to which the Quebec patterns appear to differ from those of the other regions of Canada, and the complexity in the patterns of bilingualism at the turn of the century. From an algorithmic perspective, this paper has demonstrated how confidence intervals can be used to identify factors that are both statistically and practically significant. It has also shown how combining a semantic measure of similarity between trees with confidence intervals can be used to evaluate and modify an existing theory.

# 8 Acknowledgements

We would like to thank the Natural Sciences and Engineering, and the Social Sciences and Humanities Research Councils of Canada for financial support.

# References

P. Baskerville and E. W. Sager. *Unwilling Idlers: The Urban Unemployed and Their Families in Late Victorian Canada*. University of Toronto Press, Toronto, 1998.

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, Belmont, CA, 1984.

G. Darroch and L. Soltow. *Property and Inequality in Victorian Ontario: Structural Patterns and Cultural Communities in the 1871 Census*. University of Toronto Press, Toronto, 1994.

C. Drummond and R. C. Holte. Exploiting the cost (in)sensitivity of decision tree splitting criteria. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 239–246, 2000.

C. Drummond, S. Matwin, and C. Gaffield. Inferring and revising theories with confidence: data mining the 1901 Canadian census. In *Proceedings of the ECML/PKDD'02 Workshop on "Mining Official Data" (MOD'02)*, pages 40–52, 2002.

B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, London, 1993.

E. Frank. *Pruning decision trees and lists*. PhD thesis, Department of Computer Science, University of Waikato, Hamilton, New Zealand, 2000.

C. Gaffield. Linearity, non-linearity, and the competing constructions of social hierarchy in early twentieth century Canada: The question of language in 1901. *Historical Methods*, 33(4):255–260, 2000.

J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating chracteristic (ROC) curve. *Radiology*, 143:29–36, 1982.

L. L. Harlow, S. A. Mulaik, and J. H. Steiger, editors. *What if there were no significance tests?* Lawrence Erlbaum Associates, 1997.

D. Jensen. Knowledge discovery through induction with randomization testing. In *Proceedings of the 1991 Knowledge Discovery in Databases Workshop*, pages 148–159, 1991.

D. Madigan, J. Gavrin, and A. Raftery. Enhancing the predictive performance of Bayesian graphical models. *Communications in Statistics – Theory and Methods*, 24:2271–2292, 1995.

D. D. Margineantu and T. G. Dietterich. Bootstrap methods for the cost-sensitive evaluation of classifiers. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 582–590, 2000.

R. J. Mooney. Induction over the unexplained: Using overly-general domain theories to aid concept learning. *Machine Learning*, 10(1):79–110, 1993.

S. K. Murthy. Automatic construction of decision trees from data: A multi-disciplinary survey. *Data Mining and Knowledge Discovery*, 2(4):345–389, 1998.

T. Oates and D. Jensen. The effects of training set size on decision tree complexity. In *Proceedings of the Fourteenth International Conference on Machine Learning*, San Francisco, 1997. Morgan Kaufmann.

M. D. Ornstein. Analysis of household samples: The 1901 census of Canada. *Historical Methods*, 33(4):195–198, 2000.

C. Perlich, F. Provost, and J. Simonoff. Tree induction vs. logistic regression: A learning-curve analysis. CeDER Working Paper IS-01-02, Stern School of Business, New York University, NY, NY 10012, 2001. To appear in the Journal of Machine Learning Research.

J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.

J. R. Quinlan. *C4.5 Programs for Machine Learning*. Morgan Kaufmann, San Mateo, California, 1993.

G. G. Towell and J. W. Shavlik. Knowledge-based artificial neural networks. *Artificial Intelligence*, 70:119–165, 1994.

P. E. Utgoff, N. C. Berkman, and J. A. Clouse. Decision tree induction based on efficient tree restructuring. *Machine Learning*, pages 5–44, 1997.

B. Zadrozny and C. Elkan. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 609–616, 2001.

B. Zadrozny and C. Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth International Conference on Knowledge Discovery and Data Mining*, pages 694–699, 2002.