Chapter 3

A Feature Selection Approach in the Study of Azorean Proverbs

Luís Cavique
Universidade Aberta, Portugal

Armando B. Mendes
Universidade dos Açores, Portugal

Matthias Funk
Universidade dos Açores, Portugal

Jorge M. A. Santos
Universidade Évora, Portugal

ABSTRACT

A paremiologic (study of proverbs) case is presented as part of a wider project based on data collected among the Azorean population. Given the considerable distance between the Azores islands, the authors present the hypothesis that there are significant differences in the proverbs from each island, thus permitting the identification of the native island of the interviewee based on his or her knowledge of proverbs. In this chapter, a feature selection algorithm that combines Rough Sets and the Logical Analysis of Data (LAD) is presented. The algorithm named LAID (Logical Analysis of Inconsistent Data) deals with noisy data, and the authors believe that an important link was established between the two different schools with similar approaches. The algorithm was applied to a real world dataset based on data collected using thousands of interviews of Azoreans, involving an initial set of twenty-two thousand Portuguese proverbs.

1. INTRODUCTION

Proverb, “proverbium” in Latin, can be defined as a condensed saying with popular roots, recorded by an anonymous author and expressed by a minimal text, which is generally known and is based on oral tradition of a particular region, such as: “an apple a day keeps the doctor away”.

This study is based on several paremiologic works. Paremiology is the science that deals with the description, classification, etymology and pragmatics of proverbs. One of these works is the relevant collection of three books about the “Pearls
of the Portuguese Popular Wisdom” (Funk, Funk 2001a), (Funk, Funk 2001b), (Funk, Funk 2003).

In a series of interviews, several million records were collected from thousands of people denoting whether or not they recognized Portuguese proverbs, based on an initial set of twenty-two thousand proverbs. This constitutes a unique source for a socio-cultural analysis of the transmission mechanisms involved in oral culture in geographically separated places.

Two forms of knowledge validation were used: passive and active. In passive recognition, the interviewer read the proverb and the interviewee stated whether he recognized it. In active recognition, the interviewer read only the initial part of the proverb and the interviewee completed it. For example, the interviewer began by reading “An apple a day…” and the interviewee completed it with “…keeps the doctor away”.

This case study is based on data collected in eleven geographically separated areas inside the Azorean community cultural space. This community lives in the Portuguese Azorean archipelago located in the mid-Atlantic rift. In this particular case, it is interesting to analyze the relationship between local and overall knowledge within a common linguistic and cultural space. On the one hand, there is the geographical distance and isolation brought about by the natural sea barrier of this archipelago composed of nine inhabited islands. On the other hand, this archipelago not only extends over 2,330 km2, but it also spreads over a 630 km rectangle in a west-east direction and a 130 km in a north-south direction. The original groups can be seen in three main geographical clusters composed of the occidental, central and oriental groups as can be seen in Figure 1.

We can find some geographical continuity in the Central group, which is composed of 5 islands that are relatively close to each other: Faial (15,063 inhabitants), Pico (14,806), São Jorge (9,674), Terceira (55,823) and Graciosa (4,780). The same

Figure 1. The Azores Archipelago showing the geographical distribution of the nine islands