Finding the Region of Origin for Users of Proverbs

Armando B. Mendes, CEEAplA and University of Azores, Department of Mathematics (amendes@uac.pt)

Áurea Sousa, CEAUL and University of Azores, Department of Mathematics (aurea@uac.pt)

Günther Matthias A. Funk, IELT and University of Azores, Department of Mathematics (mfunk@uac.pt)

Introduction:

Proverbs are a form of popular knowledge present in every culture and society. During a long project for proverbial sentences identification a data base was being constructed. This data base collects, today, information about 25.000 idiomatic sentences including more than one thousand valid answers for proverbial sentences recognition surveys. In this work a project is described aimed to extract knowledge from this data base in order to comprehend better the inquiries about their level of proverbial recognition and the influence of the locations they have been living.

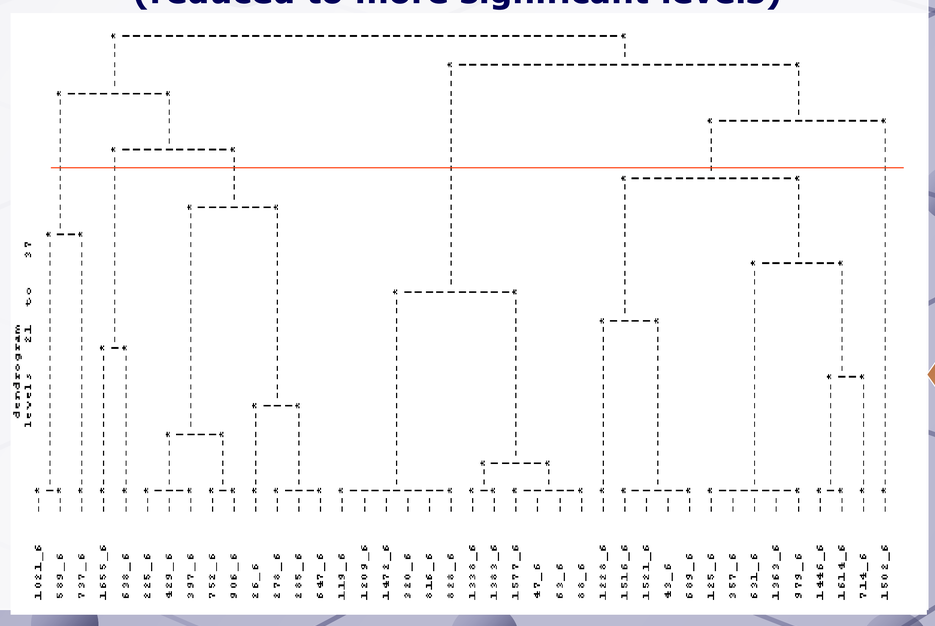
Data collection and preprocessing:

- **surveys**: the authors first analyzed the recognition of more than 22,000 expressions by the inhabitants of São Miguel, the biggest Azorean Island, which is a melting pot for the Azorean Archipelago. Circa 5,000 units were known by more than 10% of the these inquiries and serves as a *corpus* for a study of the proverbial knowledge over all islands and also the Azorean immigrants in USA. The respondents age was also controlled in the study.
- **data base restructuring**: the data was collected by 1,181 proverb recognition inquiries, each with a sample of over 1,500 proverbs, known as packages. The results were registered in data tables and organized in a relational data base. The data base restructuring was a long and hard phase. These included implementing relational normalization, referential integrity, procedures for assuring data quality, and pre-processing activities. Of major importance where the data reduction implemented. These excluded inquiries, proverbs or variables considered of low quality or less relevant for a particular analyze. These, also incorporated feature smoothing or concept climbing in several categorical variables as the places where the inquiries lived for more than 5 years.

Ascendant Hierarchical Cluster Analysis (AHCA) of Symbolic Data:

- We consider only inquiries who lived more than 5 years in a single island, excluding the ones who lived in two or more different places.
- 40 proverbs were selected (symbolic data units or symbolic objects), from the ones which were most well known, of the package 6 which are described by 6 symbolic modal variables (**Level of education**, **Age group**, **Level of recognition** of each proverb, **place of residence**, **Sex**, **Region/Island**), with, respectively, 6, 7, 7, 3, 2 and 9 modalities.
- Each entry of the symbolic data matrix contains a frequency distribution. The AHCA of the 40 symbolic data units (Proverbs) is based on the **weighted generalised affinity coefficient** (Nicolau and Bacelar-Nicolau, 1999), and the A.H.C.A of the 9 Islands is based on the **basic affinity coefficient** (Bacelar-Nicolau, 1980).
- These measures of comparison between elements have been combined with classical and probabilistic aggregation criteria (Nicolau, 1983).

Figure 1. Dendrogram obtained by the AVBmethod (reduced to more significant levels)

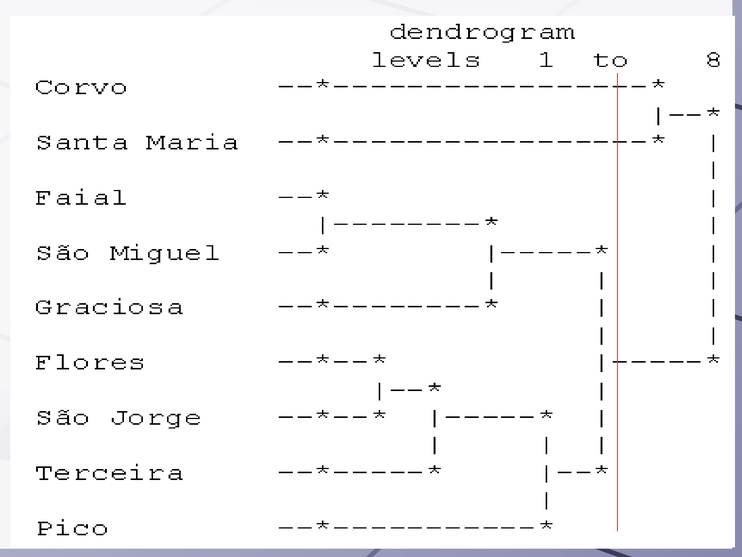


Six clusters (level 34)

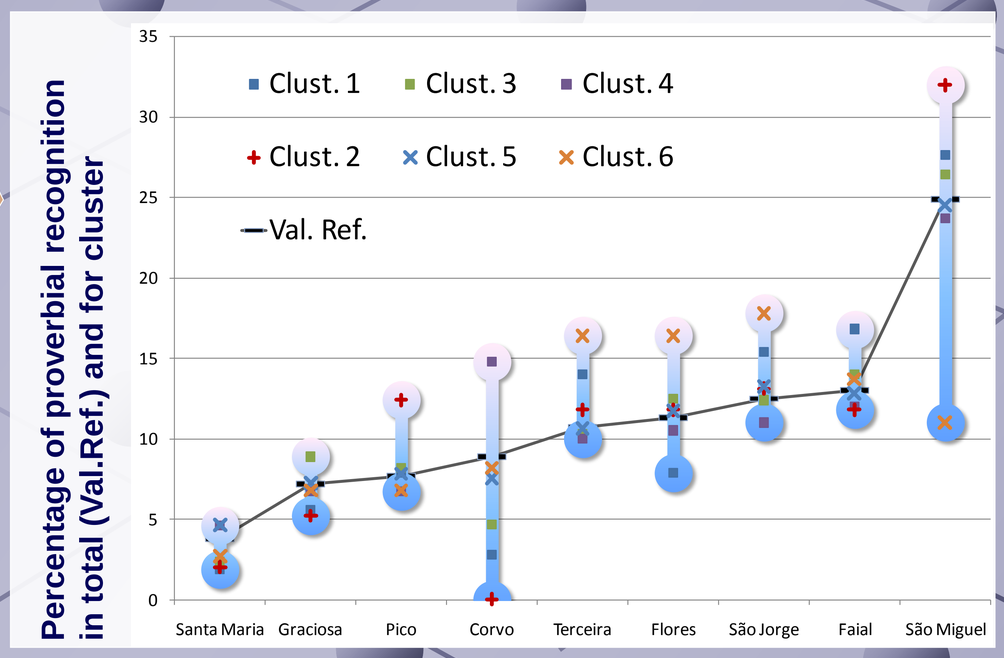
Table 1. A piece of the Symbolic Data Matrix

Proverbs	...	*Male*	*Region/Island*
816_6	...	Yes (0.38), No (0.62)	SM (0.24), F1 (0.11), P (0.06), G (0.08), T (0.08), SJ (0.11), C (0.11), SM (0.04), Fa (0.13)
....	...	...	...
737_6	...	Yes (0.39), No (0.61)	SM (0.31), F1 (0.04), P (0.09), G (0.03), T (0.15), SJ (0.15), C (0.03), SM (0.01), Fa (0.18)

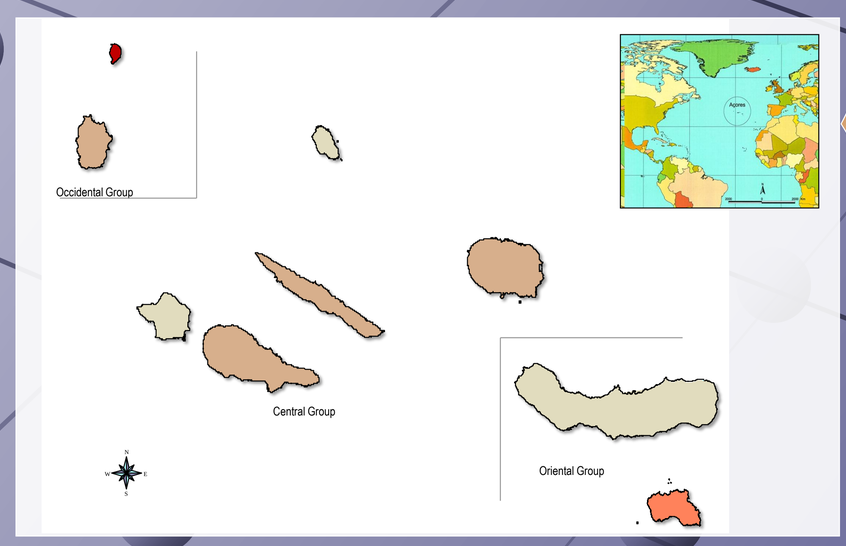
Figure 2. Dendrogram obtained by the AV1 method



three clusters (level 6)



Percentage of proverbial recognition in total (Val.Ref.) and for cluster



Conclusions:

- The results are in general coherent with domain knowledge and geographic position.
- The large similitude observed between **Faial** and **São Miguel** can be explained by the particular influxes between this islands, in spite of the geographic separation.
- The observed disconnection of **Corvo** and **Santa Maria** from the other islands is well known and due to small dimension and geographic isolation.
- Several proverbs clusters can be used as **characteristic of regions**. This is the case of cluster 2, very well known in São Miguel and Pico, or cluster 6, also well known in Terceira, Flores and São Jorge and poorly known in São Miguel.

Future Work:

several techniques are presently being applied to this data. A part from symbolic HCA, link Analysis, LAD for data reduction, lift curves and analogies to genetic transmission and social networks, are creating knowledge over the link between knowing a proverb and living in a place.

References:

Bacelar-Nicolau, H. (1980). *Contribuições ao Estudo dos Coeficientes de Comparação em Análise Classificatória*. Tese de Doutoramento, FCL, Universidade de Lisboa.

Nicolau, F.; Bacelar-Nicolau, H. (1999). *Clustering Symbolic Objects Associated to Frequency or Probability Laws by the Weighted Affinity Coefficient*. In: Applied Stochastic Models and Data Analysis. Quantitative Methods in Business and Industry Society, H. Bacelar-Nicolau, F. Nicolau and Jacques Janssen (Eds.), INE, Lisboa, Portugal, 155-158.

Nicolau, F. (1983). Cluster Analysis and Distribution Function. *Methods of Operations Research*, 45, 431-433.