# JUSTIFICATION OF AMALGAMATED PATTERN PRIMITIVE VARIABLE FOR LANGUAGE DESCRIPTION BY THE APPLICATION OF HYPERGEOMETRIC DISTRIBUTION

By

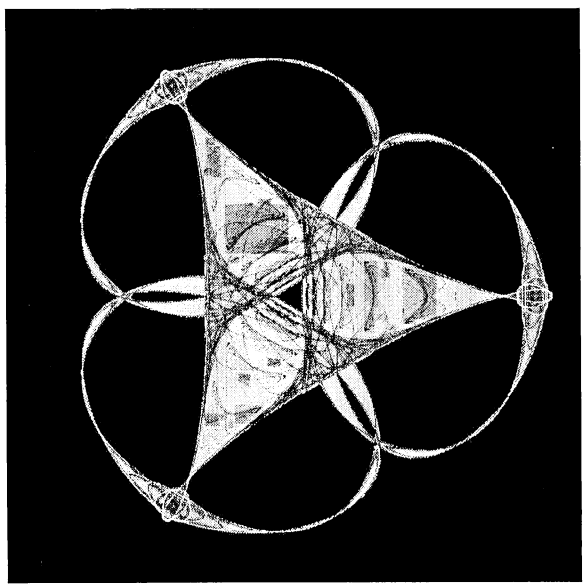
**H.S. Dhami**

and

**L.K. Verma**

**IMA Preprint Series # 1637**

August 1999



**INSTITUTE FOR MATHEMATICS AND ITS APPLICATIONS**

UNIVERSITY OF MINNESOTA

514 Vincent Hall

206 Church Street S.E.

Minneapolis, Minnesota 55455–0436

Phone: 612/624-6066 Fax: 612/626-7370

URL: http://www.ima.umn.edu

# JUSTIFICATION OF AMALGAMATED PATTERN PRIMITIVE VARIABLE FOR LANGUAGE DESCRIPTION BY THE APPLICATION OF HYPERGEOMETRIC DISTRIBUTION

H.S. Dhami & L.K. Verma
Dept. of Mathematics,
University of Kumaun,
Almora Campus,
ALMORA (U.P.) 263601 INDIA

In the present paper the amalgamated pattern primitive variable has been tested for actual sub-languages by the formation of linear expression for distribution function and the application of hypergeometric distribution.

Key words :- Standard language/linear expression/distribution function/amalgamated variable.

## INTRODUCTION

Primitives serve as basic pattern elements to provide a compact but adequate description of data in terms of the specified structural relations and this particular behaviour is reflected for language descriptions also. The five parameters identified in earlier studies [4], with reference to Kumauni language, namely, alphabetic counts, computer counts, phonemes, graphemes and allophones have been amalgamated by Dhami [2] for obtaining the standard form of the language by clustering process so as to obtain the value $X_1+X_2+3X_3+2X_4$, Where $X_1,X_2,X_3$ & $X_4$ correspond to first four primitives.

Here an attempt is being made to test this obtained value for the four geographical representative sub-languages of Kumauni (language under study) and for this purpose words demonstrating phonetic variations have been selected.

## CONCEPTUAL FOUNDATION FRAMEWORK

For the three matrices {of order 12x4} corresponding to pronounciation, palatization and verb forms, linear equations of distribution functions for actual values were obtained separately for alphabetic counts, computer counts, phonemes and graphemes, which when normalized gave rise to following three groups of matrix type representations

$$24.29E_1+25.27E_2+23.65E_3+26.76E_4$$
$$23.34E_1+27.28E_2+29.13E_3+20.59E_4$$
$$24.29E_1+24.59E_2+25.62E_3+25.49E_4$$
$$23.22E_1+23.09E_2+27.13E_3+23.18E_4;$$

$$23.43E_1+25.86E_2+28.35E_3+22.34E_4$$
$$24.65E_1+24.63E_2+29.96E_3+20.74E_4$$
$$24.72E_1+25.30E_2+25.71E_3+23.95E_4$$
$$24.18E_1+25.57E_2+27.14E_3+23.10E_4 \text{ and}$$

$$24.39E_1+24.56E_2+27.82E_3+23.23E_4$$
$$25.00E_1+25.26E_2+27.38E_3+22.35E_4$$
$$24.91E_1+24.47E_2+26.07E_3+24.55E_4$$
$$25.09E_1+24.41E_2+26.22E_3+24.26E_4,$$

Where $E_1$ represents total number of words exhibiting the first characteristic (derived on the basis of similarity in alphabetic counts in normality from the proposed standard form) and similarity for $E_2$,$E_3$ & $E_4$.

Weights for different character states have been obtained and for them the maximum possible weight to be applied to a system should have its value by formula $W_{EE}(max) = s/2$, where s stands for the total elements in the study (48 for our case). This theoretical perception when tested for the amalgamated value of the variable $[X_1+X_2+3X_3+2X_4]$ has yielded following expressions

$$24.46E_1+24.87E_2+26.07E_3+24.60E_4$$
$$24.58E_1+24.73E_2+25.88E_3+24.81E_4$$
$$\text{and} \quad 24.82E_1+24.83E_2+25.48E_3+24.87E_4,$$

Which exhibit the close agreement in respect of three representative sub-languages. The sub-language mentioned at serial number third of above matrix type representations is restricted to a limited region and so its unfamiliarity with others can be understood. Furthermore, the clustering has been around the first one and so the general opinion of philologists in favour of considering it as the representative language is also justified.

## APPLICATION OF HYPERGEOMETRIC DISTRIBUTION

For the four sub-languages, we have constructed three tables of type given below-

| Root Number | Western Kumauni | Eastern Kumauni | Northern Kumauni | Southern Kumauni |
|---|---|---|---|---|
| 1 | * | | | * |
| 2 | | * | * | |
| 3 | * | | | * |
| 4 | | | | |
| 5 | | * | | * |
| 6 | * | | | * |
| 7 | | | | |
| 8 | | | | |
| 9 | * | | * | |
| 10 | | | | |
| 11 | | | | |
| 12 | | | | |
| Total N=12 | $n_A$=4 | $n_B$=2 | $n_C$=2 | $n_D$=4 |

Where the numerals of root number column correspond to the rows of the resemblance matrix and * is the measure of closeness of any two sub-languages. The probability of obtaining at least R common * has been evaluated as upper Hypergeometric cumulative probability $P(R/n_B,n_A,N)$ between any two Sub-languages by the formulae given in the book of Johnson & Kotz[3], as

$$P(R) = P\{\text{at least R common *s}\}$$

$$= \bar{P}(R/n_B,n_A,N)$$

$$= \left\{ \frac{\binom{N-n_B}{n_A}}{\binom{N}{n_A}} F_R(-n_B,n_A; N-n_A-n_B+1; 1) \right\}$$

if $N \geq (n_A+n_B)$;

and in case if $N < (n_A+n_B)$ then

$$P(R) = \bar{P}(N+R-n_A-N_B/N-n_B,N-n_A,N)$$

$$= \left\{ \frac{\binom{N-(N-n_B)}{N-n_A}}{\binom{N}{N-n_A}} F_{N+R-n_A-n_B}(-N+n_B,-N+n_A; N-n_A-n_B+1;1) \right\}$$

Numerical values of the hypergeometric function have been obtained with the help of tables and formulae given in the books of Slater [5] and Abramowitz et al [1]. The consolidated form of numerical values for the upper hypergeometric cumulative probabilities for the table given in the beginning of this section is as follows-

| | I | II | III | IV |
|---|---|---|---|---|
| I | 1 | 0.090909 | 0.090909 | 0.002020 |
| II | 0.090909 | 1 | 0.424242 | 0.141414 |
| III | 0.090909 | 0.424242 | 1 | 0.141414 |
| IV | 0.002020 | 0.141414 | 0.141414 | 1 |

Similar techniques when applied to matrices corresponding to palatization and verb forms have resulted in the formation of two tables given below-

| | I | II | III | IV |
|---|---|---|---|---|
| I | 1 | 0 | 0 | 0 |
| II | 0 | 1 | 0.300000 | 0.151515 |
| III | 0 | 0.300000 | 1 | 0.151515 |
| IV | 0 | 0.151515 | 0.151515 | 1 |

and

| | I | II | III | IV |
|---|---|---|---|---|
| I | 1 | 0 | 0 | Non-existent |
| II | 0 | 1 | 0.333333 | 0 |
| III | 0 | 0.333333 | 1 | 0.045455 |
| IV | non-existent | 0 | 0.045455 | 1 |

This process of hypergeometric distribution application when applied to five primitives and then finally to amalgamated value of the variable shall display significant level of upper hypergeometric cumulative probability and thus the justification of results obtained in the study of Dhami [2].

## REFERENCES


[1] Milton Abramowitz & Irene A. Stegun (1972) Handbook of mathematical functions with formulas, Graphs and mathematical tables, Dover Publications, Inc., New York.

[2] H.S. Dhami (Feb 1999) Amalgamation of pattern primitives for the generation of standard form of language, I.M.A. preprint series # 1609, University of Minnesota, Minnesota 55 & 55 – 0436.

[3] Norman L. Johnson & Samuel Kotz (1976) Discrete distributions, John Wiley & Sons, New York.

[4] Jaya Kandpal (1992) Mathematical analysis of Kumauni language, unpublished Ph.D Thesis (Degree awarded under the supervision of Dr. H.S. Dhami).

[5] L.J. Slater (1966) Generalized hypergeometric functions, Cambridge University press, Cambridge.