The Use of Quantitative Methods in a Few Case Studies in the Indian Context

Amitav Choudhry
English Language Department
Faculty of Languages and Linguistics
University of Malaya

Introduction

Linguistics and statistics: Early inroads

It was in the early nineteen forties that Professor Prashanta Chandra Mahalanobis who was basically a physicist and a renowned statistician and also interested in other disciplines planned to bridge linguistics with statistics. In this project he collaborated with Suniti Kumar Chatterji, a well-known comparative philologist. This project culminated into a gigantic project of counting word frequencies of Bangla (Indic) literature with a view to map the stylistic variations of different authors in order to supply vocabulary entries for a standard dictionary, and similar projects. The project also had ambitious plans to observe the changing mode and development of the Bangla language by analyzing selected and representative literature, which included Bankimchandra Chattopadhyay, an accomplished Bengali novelist and Rabindranath Tagore, the well-known poet, novelist and Nobel laureate. Another notable fact is that Iswarchandra Vidyasagar, a well known educationist, who collected Bangla words with their frequency of occurrences in as early as late 19 C. with a view to write a primer, was the first person in India to conceive the idea of Statistical Linguistics and language planning (Bandyopadhyay and Debnath,1996). It was in the late nineteen sixties that the name text linguistics was coined, though references to this branch are naturally rooted in the early history of linguistics. Along with the new referent came a new substantial idea, namely the idea of *text coherence* based on *text references*

The rise of text linguistics and its constitution as a scientific branch certainly does not mean the end of the classical branches dealing with texts, e.g. stylistics of literally criticism. The same is valid for the quantitative

approaches in linguistics. Essayistic treatment of language and its phenomena is not touched by stressing the methodological sciences.

According to Altmann (1991.33) "texts or their parts have physical, sociological, psychological, linguistic, semiotic and information-theoretical aspects, in addition, during the course of text generation emergent textological patterns arise, e.g., rhyme, chaotic sequences, frequency distributions, aggregates, style, etc. a proper examination of the various facets makes up a vast scientific area in which researchers from several disciplines are engaged." Altmann *ibid* adds that there is a wealth of mathematical methods with the aid of which, specific problems could be solved, or which could trigger the construction of partial theories or add new dimension to already existing ones.

Though earlier efforts to apply quantitative methods in the Indian context was restricted to text analysis, gradually other areas like language teaching and testing, sociolinguistics, especially language attitudinal studies, psycholinguistics, historical linguistics and stylistics, started to use statistical methods. Computational linguistics has emerged as a very strong discipline in recent years especially to bridge the gap between major Indian languages, with the development of morphological analysers, operating systems in Indian languages, spellcheckers, electronic dictionaries, language modeling systems and the like. Research was also initiated in the fields of speech processing, pattern recognition and image processing.

The paper looks at the use of quantitative methods by the author and other quantitative linguists in the Indian multilingual context in a few domains of linguistics like text analysis, parameters of vocabulary balance, measuring language distance using the *Mahalanobis* D^2, stylistics, quantification in language attitudinal studies and the measurement of bilingualism.

Quantitative analysis of text: An Indian experience: *Choudhry and Debnath (2001)*

In this study based on data from a complete word count of Rabindranath Tagore's "Galgaguccha" (short stories. Parts I to IV), the hypothesis of vocabulary balance was tested.

According to Zipf (1949:22) "we obviously do not know whether there is in fact such a thing as vocabulary balance between our hypothetical forces of *unification* and *diversification* since we do not know whether human beings invariably economise with the expenditure of their effort; for that, after all, is what we are trying to prove" The paper looks at the evidence of vocabulary balance in Tagore's "Galpaguccha"· 39,145 different

words in the 315,850 running words were counted. The words are ranked in the decreasing order of their frequency of occurrence. The 10th most frequent word ($r = 10$) occurs 1,854 times ($f = 1,854$) The 100th word ($r = 100$) occurs 355 times ($f = 355$) The relationship between r and f in the vocabulary in "Galpaguccha" is to a large extent uniform except for the 10th and 20th rank and tends to regularize after the 30th rank. The reason for this is likely to be that the percentage of common words between the 10th and the 30th rank is not very high and the different words are more evenly distributed over the other ranks as is evident from the following Table 1, which contains the rank distribution in percentage of the 100 most common words in "Galpaguccha" (Parts 1-4).

Table 1: The rank distribution in percentage of the 100 most common words in "Galpaguccha" (Parts 1-4)

World rank in Galpaguecha	Frequency of occurrence of word	Percentage of total words
1-10	30319	9.6
11-20	14276	4.5
21-30	10847	3.4
30-40	9320	3.0
41-50	8004	2.5
51-60	6561	2.1
61-70	5303	1.7
71-80	4492	1.4
81-90	4109	1.3
91 100	3644	1.2
Total	96875	30.7

In conclusion we may say that Tagore in "Galpaguccha" manifests a trend that whenever a person uses words to convey meanings he will automatically try to get his ideas most efficiently by seeking a balance between the economy of a small wieldy vocabulary of more general reference on the one hand and the economy of a larger one of more precise reference on the other, in accordance with Zipf's (1949) prediction of a vocabulary balance between our theoretical forces of *unification* and *diversification*

Mahalanobis as a language planner: *Bandyopadhyay & Debnath (1996)*

This paper argues that statistical counting of texts would help to develop primers for neo-literates, printing technology and other related areas apart from the scope of doing stylistic analysis. In this connection, this paper looks if Mahalanobis' economic planning (India's second five-year plan) is related to his language planning. According to the authors, this type of language planning is called quantitative planning.

Mahalanobis distances can be used in analyzing cases in discriminant analysis. For instance, one might wish to analyze a new, unknown set of cases in comparison to an existing set of known cases. Mahalanobis distance is the distance between a case and the centroid for each group (of the dependent) in attribute space (n-dimensional space defined by n variables). A case will have one Mahalanobis distance for each group, and it will be classified as belonging to the group for which its Mahalanobis distance is smallest. Thus, the smaller the Mahalanobis distance, the closer the case is to the group centroid and the more likely it is to be classed as belonging to that group. Since Mahalanobis distance is measured in terms of standard deviations from the centroid, therefore a case which is more than 1.96 Mahalanobis distance units from the centroid has less than 05 chance of belonging to the group represented by the centroid, 3 units would likewise correspond to less than 01 chance.

An Analogue of the WARING – HERDAN formulae for lexical distributions: *Sircar (1972)*

In this paper Sircar (1972) states that the constantly decreasing progression (in a lexical distribution) of the number of words used once, twice, thrice, in any sufficiently large sample via mathematical expression in giving the theoretical probability values accounts for the pattern of decrease of the size of classes found in lexical distributions. Sircar (*ibid*) is of the opinion that no theoretical formulae can hope to tally exactly with lexical distribution. The reason for this is that lexical distributions are considerably influenced by stylistic differences and stylistic traits. Nevertheless, the Waring-Herdan formulae and the analogue both seem to hold promise of future improvement and greater insight into the nature of language.

Quantification of stylistic traits: A statistical approach. *Bagavandas and Manimannan (2004)*

This paper is an attempt to identify distinct stylistic features of three Tamil scholars belonging to a contemporary period and also to try to quantify the writing styles of these authors using eighteen stylistic features. These stylistic features have been categorized as eleven morphological variables, four habitual words and three function words. In terms of methods, ANOVA technique, two sample *t-statistic* and *Factor analysis* are used to measure the given stylistic traits and also to identify traits that have a higher frequency of occurrence.

Language attitudes of a linguistic minority in a regional area: Choudhry (1982)

This paper examines the language attitudes of the Bangla community in Hyderabad, south India. The respondents were asked to fill in a questionnaire based on the *Likert method* The subjects' attitudes towards Telugu Hindi, Bangla and English were tested. Statements were mainly on [a] communicative choice, [b] vocational importance, [c] medium of instruction in higher education, and regarding Bangla [d] retention of the mother tongue and ethnic identity

With a view to examining the extent to which the responses given to various statements were interrelated, we examined the association between responses to a few pairs of statements. To this end we prepared joint frequency distribution of responses to each of several pairs of statements and applied the Chi-square test ($\chi 2$ to check whether the responses to the two statement items were completely independent or not. The following statements were considered for analysis.

Statement No:	**Statement**
11.1	In our country English and not Hindi is the best language as the medium of instruction in school and for higher education.
11.2	Knowledge of English is essential in offering better future prospects in the vocational field.
11.3	Establishing universities with the regional language as the medium of instruction should be encouraged.

11.4 Learning the regional language is more important than cultivating one's mother tongue.

11.10 Text books in English are of a better standard than books in other languages in professional fields.

11.15 You feel concerned about your children since they find it difficult to retain their mother tongue.

11.18a English is more convenient than Telugu (lang. of interaction).

11.18b Hindi is more convenient than Telugu (lang. of interaction).

Before computing the χ^2 criterion to test the hypothesis of independence of (i.e. zero association between) the responses to the two question-items, pooling of neighbouring classes had to be resorted to in order to ensure sufficient frequencies. In general, the 'undecided' category was amalgamated with the 'disagree' category and the contingency tables reduced to 2 x 2 tables. For all the 2 x 2 tables, Yate's correction for continuity was applied in computing the value of χ^2 The results are summarized in Table 2.

Table 2: Association between responses to selected pairs of question items.

Statements examined	χ^2 value	d.f.	significance	$C = \sqrt{\frac{\chi^2}{N+\chi^2}}$
1	2	3	4	5
11 1 and 11 2	67 54	1	significant at 0.1%	0.5024
11 1 and 11 3	0.10	1	non-significant	0.0223
11.4 and 11 15	4.64	4	non-significant	0.1507
11 18a and 11 18b	15.51	1	significant at 0.1% level	0.2681

Therefore the results show that in many cases the null hypothesis of complete independence was rejected by the Chi-square test indicating that the responses to the two items were far from random and unrelated. In such cases the C-coefficient of contingency would measure the extent of association between the two sets of responses. For example the degree of

association between responses to the 11 18a and 11 18b was found to be moderate, but high for the association between responses to statements 11 1 and 11.2. The C-value for statements 11 1 vs. 11.3 is low – which is obvious because the same subjects cannot be contradictory in their response to [1] favourable to English and [2] favourable to the regional language.

Language attitude of the Oriya immigrant population in Kolkata: *Duttamajumdar (2008)*

The study investigates and analyses motivational factors responsible for language choice and use of the Oriya (Indic) immigrant population in Kolkata, India. While examining the aspect of environmental languages in different domains, it tries to establish a process of acculturation on one hand and language maintenance on the other. The paper also looks at the socioeconomics of power relationships. The author uses the *Student's t-test* to establish the degree of favourable responses towards Oriya. In conclusion the author feels that there is a process of amalgamation towards Bangla, the dominant language of the region.

The adaptability of certain bilingual measurement models in the Indian bilingual context: *Choudhry (1996)*

To begin with, the author looks at the adaptability of tests of bilingual measurement used in the *west*, in the Indian context. At first he used the McCarthy (1930) and Davis (1937) method to determine the threshold of bilingualism. The two languages which were tested were Bangla and Telugu among Bangla-speaking pre-school-age children in an environment dominated by the Telugu (Dravidian) language.

Results showed that traces of bilingualism were found even amongst 2-year-olds. But the number of words per verbalisations in Bangla was better than in Telugu among pre-school-age Bangla-speaking children.

The adaptability of the Discrete point James Language Dominance Test (1975) was tried next on 30 3-5 yr.-old Bangla-speakmg children spread over five age groups to determine their dominance in either Bangla, English or Telugu, which are the languages in the verbal repertoire of the children. Procedure included 20 pictures which would evoke one-word or two-word responses; questions were mainly in the form of, "what is this?", "where is this " and "what is the.. ?" Phonological variations were overlooked if it was only one per word and for more than one, one minus mark was awarded. The maximum score possible per subject was 40 points.

Based on the results, the subjects were put into 3 categories to ascertain their language dominance and bilingual proficiency The subjects were categorised thus.

A L1 dominant
B bilingual plus L1/L2/L3
C proportionate bilingual [L1/L2] or [L1/L3]

Another test was carried out using the "The bilingual syntax measure" (Burt et al, 1981), on 12 school-age children both boys and girls in the age group 4+, 5+ and 6+ whose L1 was Bangla, L2 English and L3 Telugu. The three languages were tested separately and the test included twenty questions, not necessarily translation equivalents, that were intended to elicit particular grammatical structures about a series of seven pictures which were self-expressive. The responses were rated on a six-point scale for acceptability and point value. The discrete system of scoring was also used. Comprehension capacity and reaction time was taken into account while evaluating the sentences.

The paper looked at the use of quantitative methods by the author and other quantitative linguists in the Indian multilingual context in a few domains of linguistics like text analysis, parameters of vocabulary balance, measuring language distance using the *Mahalanobis* D^2, stylistics, quantification in language attitudinal studies and the measurement of bilingualism.

Though the paper was a modest attempt to explain the use of quantitative methods by the author and other quantitative linguists in the Indian multilingual context, one should bear in mind that recent approaches to the use of statistics in built around study design, data collection, and data analysis, and with the availability of appropriate technology, numerous question may arise as to how content should change to enhance statistical thinking and understanding of concepts over rote use of standard procedures. We need to re-emphasize that the proper interpretation of figures and transforming them into tha domain of logical reasoning is what is more important to increase the scope of proper application of quantitative methods.

Conclusion

This paper attempted to review the status of methodology in quantitative linguistics and its relevance in exploring new avenues of research The paper also looked at the use of quantitative methods by the author and other quantitative linguists in the Indian multilingual context in a few domains of linguistics like text analysis, parameters of vocabulary balance, articulatory evaluation of speech sounds using a global scale, measuring language distance using the *Mahalanobis* D^2, the process of language standardization using set theory, quantification in language attitudinal studies and the measurement of bilingualism.

References

Altmann, G. (1991). Modeling diversification phenomena in language. In. Rothe, U (ed.), *Diversification processes in language. Grammar*: Hagen, Rottman 1991 33-46.

Bagawandas, M. and G. Manimannan. (2004). Quantification of Stylistic Traits. A Statistical Approach. *Proceedings of the 7th International Conference on Textual Data Analysis*. Vol.I:71-78. Louvain-la-Neuve, Belgium: UCL Press.

Bandyopadhyay, Debaprasad and Sukesh Debnath. (1996). "Mahalanobis as language planner" *Indian Journal of Applied Linguistics*. Vol. XXII, No.1 (pp. 49-57).

Burt, M. ,K., H. C. Dulay E. Hernandez-Chavez.(1976).*The bilingualism syntax technical handbook*. New York, Harcourt Brace Jovanovich.

Choudhry, Amitav (1981). Language attitudes of a linguistic minority in a regional area. *OPIL* Vol 7· 116-130.

Choudhry, Amitav (1995). "Models of bilingual measurement and their adaptability in the Indian context", *Journal of Quantitative Linguistics* The Netherlands, Swets and Zeitlinger Pub. Vol. 2.3 258-266.

Choudhry, Amitav (1995). "The Chi-square test and its significance in studying stability in response patterns. Calcutta, *ILS Monograph Series·IV*

Choudhry, Amitav and Sukesh Debnath. 2001 "Quantitative Analysis: An Indian Experience." In *Quantitative Linguistics Text as a Linguistic Paradigm. Levels, Constituents, Constructs*. L.Uhrilova, G. Wimmer, G.Altmann & R.Kohler (eds.) Festschrift Volume in Honour of L. Hr'ebicek. Vol.60. Trier, Wissenschaftlicher Verlag.

Dattamajumdar, Satarupa. 2008. "Language attitude of the Oriya immigrant population in Kolkata" in *Readings in Quantitative Linguistics*. (Panchanan Mohanty and Reinhard Kölher eds.), Delhi, Indian Institute of Language Studies.

Davis, Edith A.(1937) *The development of linguistic skills in twins, singltons with siblings and only children from age 5 to10 years*. University Minnesota, Institute of child welfare Monograph, No.14 Minneapolis: University of Minnesota Press.

Hanken, H. (1978). *Synergetics*. Berlin-Heidelberg-New York, Springer.

James, P (1976). *James language dominance test*. Second edition. Austin Texas. Learning Concepts.

McCarthy, Dorothea. (1930). *The Language development of the pre-school child.* University of Minnesota, Institute of Child Welfare Monograph, No. 4. Minneapolis: University of Minnesota Press.

Roy, A (1986) *Word frequency count of the words of Bankimchandra Chattopadhyay and Rabindranath Tagore* Unpublished monograph. Calcutta: Linguistic Research Unit, Indian Statistical Institute.

Sircar, J (1972). *An analogue of the Waring-Herdan formulae for lexical distributions*. Tech.Rep.no.LING/1/72. Indian Statistical Institute.

Zipf, G.K. (1949). *Human behaviour and the principle of least effort*. Cambridge: A.W.Press Inc.