DLA NEWS

Vol. 36 No. 12 Website: www.ijdl.org E-mail: ijdlisdl@gmail.com DECEMBER 2012

A MONTHLY OF DRAVIDIAN LINGUISTICS ASSOCIATION OF INDIA

CORPUS LINGUISTICS

When structuralism was prevalent in the 50s, language analysis and processing depended on data collected through fieldwork or from texts. Chomsky criticized this method as unreliable, as the data was not enough for any kind of linguistic exercise and he advocated for the native speaker's intuition for language analysis. He argued for data from competence rather than from performance. However, with the advent of computers which could store huge amounts of data that could be made use of for language analysis, the criticism by Chomsky was sidelined and linguistic analysis by making use of a huge corpus or corpus linguistics came into vogue. Even Chomsky himself admitted the value of performance data as a source of evidence in language acquisition and phonetics. Before the generation of electronic corpus, language databases were used in lexicography, dialectology, comparative linguistics, language teaching, child language acquisition, phonology and other fields of linguistics. Theoretically, corpus is (C)apable (O)f (R)epresenting (P)otentially (U)nlimited (S)elections of texts. It is (C)ompatible with computers, (O)perational in research and application, (R)epresentative of source language, (P)rocessable by man and machine, (U)nlimited in data and (S)ystematic in formation and representation.

Classification of the Corpus

Considering the basic traits of differences, we can divide corpus into two broad types: written corpus and speech corpus. Brown Corpus, LOB Corpus, Australian Corpus of English, etc. are examples of written corpus. London-Lund Corpus of Spoken English, Survey of Spoken English (SEU), Machine-readable Spoken English Corpus (SEC), etc. are examples for spoken corpus.

We can classify corpora in a broad manner in the following way: (1) Genre of text, (2) Nature of data, (3) Type of text, (4) Purpose of design and (5) Nature of application. Based on the genre of the text, the corpus can be classified as Written corpus, Speech corpus and Spoken corpus. Based on the nature of data, the corpus can be distinguished into the following types: General corpus, Special corpus, Sub-language corpus, Sample corpus, Literary corpus and Monitor corpus. In terms of the type of text, the corpus can be differentiated into three types. Monolingual corpus, Bilingual corpus and Multilingual corpus. In accordance with the purpose of design, the corpus can be separated into two: Unannotated corpus and Annotated corpus. Based on the nature of application, the corpus can be classified into the following types: Aligned corpus, Parallel corpus, Reference corpus, Comparable corpus (e.g. Corpus of European Union) and Opportunistic corpus.

Corpus Creation

There are various issues related with corpus design, development and management. The issues of corpus development and processing may vary depending on the type of corpus and the purpose of use. The following issues have to be addressed while creating a corpus: size of corpus, representativeness of texts, question of nativity, determination of target users, selection of time-span, selection of text type, method of data sampling, method of data input, hardware requirement, management of corpus files, method of corpus sanitation and the problem of copyright.

Text Corpus Processing

The text corpus needs to be processed. There are various corpus processing techniques, e.g. statistical analyser, concordancer, lexical collocator, key-word finder, local-word-grouper, lemmatiser, morphological processor

CONTENTS

and generator, word processor, parts-of-speech tagger, corpus annotator, parser, etc. The text processing involves the following: frequency study, word sorting, concordance, lexical collocation, key word in context (KWIC), local word grouping (LWG), word processing, tagging (part-of-speech (POS) tagging, grammatical tagging and word-sense tagging), lemmatisation, annotation (part-of-speech annotation, anaphoric annotation, prosodic annotation, semantic annotation and discoursal annotation), parsing.

Corpus in Language Technology

Corpus plays a vital role in language technology. Almost all exercises in language technology demand corpus for various reasons. Machine-learning approaches demand the use of a huge corpus. Corpus linguistics and language technology have become bedfellows. Corpus has become a knowledge resource for language technology exercises. Corpus is very much in demand for designing language technology tools. Corpus has become a source for translation support systems as well as a source for human-machine interface systems. Corpus is an inevitable resource in speech technology.

Corpus in Mainstream Linguistics

Corpora have been used in various fields of mainstream linguistic researches, analyses and applications as primary resources. Corpus is required for the following exercises in linguistics: lexicography, lexicology, coinage of technical terms, grammar-writing, semantic study, language-learning, dialect study, sociolinguistics, psycho-linguistics and stylistics.

Corpus in Machine Translation

Corpus and machine translation go hand in hand. Nowadays, machine translation is impossible without corpus. Corpus is very useful for implementing machine-learning techniques which are crucial for building machine-translation systems. Parallel corpora are vital for building statistical-based machine-translation systems. All the components of machine-translation systems including morphological analyser, syntactic parser, POS tagger, chunker, transferring system, word-sense disambiguation system, word and sentence generators, etc. require corpora for machine-learning and testing.

S. Rajendran
Amrita University, Coimbatore

SOUTHERN INDIA'S CASTE SYSTEM PREDATES ARRIVAL OF INDO-EUROPEANS

A study by the Genographic Project has given new insight into how demographic factors have shaped genetic diversity in Indian populations. Among the most surprising findings was that genetic differences between tribal and caste groups in Tamil Nadu seem to pre-date the arrival of the Indo-Europeans in the region by approximately 2,000 years.

Published in the journal *PLOS ONE*, the study was led by principal investigator Ramasamy Pitchappan of the Genographic Project's Indian Regional Centre at Chettinad Academy of Research & Education in Chennai.

Contemporary Indian populations exhibit great cultural, morphological and linguistic diversity. The study sought to answer the contentious question of whether India's contemporary genetic patterns are a result of long-term occupation, perhaps dating to just after humans left Africa about 60,000 years ago, or if they have been substantially impacted by more recent migration into the region.

"Our conclusions provide a new framework to better understand the relative impacts of demographic events and other cultural, social and economic factors that might have influenced modern genetic diversity in India", Pitchappan, the senior author, said.

Indian populations can be broadly divided into "tribal" and "non-tribal". Tribal groups constitute 8 percent of the Indian population and are characterized by traditional modes of subsistence such as hunting and gathering. In contrast, the majority of the non-tribal populations are classified as castes under the *Hindu Varna* (colour caste) system, which groups the population based primarily on occupation. The system embodies strict marital rules preventing marriage among different castes.

The study applied a novel analytical strategy to unravel the population structure and genetic history of the southernmost state of India - Tamil Nadu - which is known for its rigid caste system. One of the aims of the study was to explore whether genetic differences observed among Tamil Nadu populations could be attributed to the establishment of the *Hindu Varna* system approximately 2,000 years ago by Indo-Europeans from northern India. The genetic data was also interpreted in reference to the paleoclimatic, archaeological and historical evidences from this region.

A total of 1,680 men from 12 tribal and 19 non-tribal (caste) Tamil Nadu populations were analysed for markers on the paternally inherited Y-chromosome. Overall, the populations were characterized by Y-chromosome lineages (81 percent) that likely originated within India. The results suggest a minimal genetic influence in Tamil Nadu from

the main western Eurasian migrations reported in the last 10,000 years, including the spread of agricultural populations from the Fertile Crescent during the Neolithic period. Although non-tribal groups exhibited a slightly higher proportion of non-Indian paternal lineages than tribal populations, the common paternal lineages shared among them are likely drawn from the same ancestral genetic pool that emerged in India during the late Pleistocene and early Holocene (10,000-30,000 years ago).

The genetic data also revealed that genetic differentiation among populations in Tamil Nadu began as early as 6,000 years ago, with no significant genetic admixture among them for at least the last 3,000 years. These results indicate a minimal genetic impact from the Indo-European migrations into the region over the past 2,000 years. These results are consistent with the earliest historical records of the region that document a highly structured society prior to the establishment of the *Hindu Varna* system. Rather, the timing of the Y-chromosomal differentiation among Tamil Nadu populations seems to fit better with the emergence of agricultural technology in South India and the resulting demographic shifts during the Neolithic period.

Genographic Project manager and one of the lead authors of the study, David Soria-Hernanz explains that "the rigorous sampling and analytical approach used in the study allowed us to dissect the confounding relationships among multiple socio-cultural factors in Tamil Nadu, allowing us to further explore and test in detail the relationships between social structure and genetics".

Project director and National Geographic Explorer-in-Residence Spencer Wells noted, "This study is a wonderful example of how human culture, and particularly the shift to an agricultural mode of subsistence during the Neolithic period, has had a profound impact on modern patterns of genetic diversity".

Background: The Genographic Project seeks to chart new knowledge about the migratory history of the human species and answer age-old questions surrounding the genetic diversity of humanity. The project is a non-profit, multi-year, global research initiative. At the core of the project is a global consortium following an ethical and scientific framework and responsible for sample collection and analysis in their respective regions.

Members of the public can participate in the Genographic Project by purchasing a Genographic Participation Kit, *Geno 2.0*, from the Genographic website *www.genographic.com* where they can also choose to donate their genetic results to the research effort. A portion of the proceeds of the kits help further research and support a Legacy Fund for indigenous and traditional peoples' community-led language revitalization and cultural projects.

LITERACY RATE AMONG THE SCHEDULED TRIBES OF KERALA

No	*Name of Tribe*	*Population (KILA 2008)*	*Literate*	*Literacy Rate*
1.	Adiyan	11,218	6660	59.37
2.	Aranadan	246	102	41.46
3.	Cholanaickan	409	128	31.30
4.	Eravallan	4,408	1,984	45.00
5.	Hill Pulaya	3,415	2,019	59.12
6.	Irular (Irulan)	26,525	14,382	54.22
7.	Kadar	1,974	1,001	50.71
8.	Kanikaran (Kanikkar)	13,336	11,019	82.63
9.	Karimpalan	14,768	11,671	79.03
10.	Kattunayakan	19,995	10,332	51.67
11.	Koraga	1,644	1,162	70.68
12.	Kudiya, Melakudi	911	679	74.53
13.	Kurichiyan	35,909	27,572	76.78
14.	Kurumans	21,375	17,066	79.84
15.	Kurumbar	2,251	1,064	47.27
16.	Mahamalasar	143	54	37.76
17.	Mala Arayan	14,043	13,819	98.41
18.	Malai Pandaran	1,653	709	42.89
19.	Malavedan	4,687	3,622	77.28
20.	Mala Panickar	982	759	77.29
21.	Malasar	4,201	1,887	44.92
22.	Malavettuvan	19,728	12,008	60.87
23.	Malayan	5,550	3,113	56.09
24.	Mannan	9,229	6,120	66.31
25.	Mavilan	31,166	22,401	71.88

26. Mudugar	4,668	2,575	55.16
27. Muthuvan	19,163	11,844	61.81
28. Palleyan	1,481	1,064	71.84
29. Paniyan	92,783	53,411	57.57
30. Thachanadan Moopan	1,649	1,211	73.44
31. Ulladan	16,418	13,784	83.96
32. Uraly	4,685	3,757	80.19
33. Vetta Kuruma	6,482	3,939	60.77
34. Wayanad Kadar	673	530	78.75
TOTAL	**3,97,768**	**2,63,448**	**66.23**

[Source: KIRTADS Scheduled Tribe Population Data Bank 2012; KILA - Kerala Institute for Local Administration.]

"BALANGODA MAN" - EARLIEST H. SAPIENS IN S. INDIA?

There is evidence of *Homo erectus* population from Balangoda, Ratnapura District, Sabaragamuwa Province, Sri Lanka from about as early as 5,00,000 years Before Present. In June 2012, complete skeletons of "anatomically modern" *Homo sapiens* were discovered in Paliyangala nearby, which have been dated to about 37,000 years BP. The average height was 174 cm. for males and 166 for females. They had prominent brow ridges, depressed noses, heavy jaws and short necks. They fashioned geometric artefacts (microliths) of quartz and chert - items discovered from Batadombalena have been dated back to 31,000 years BP.

Similar objects discovered from Europe have been dated back to only about 12,000 years BP, indicating that the species had evolved faster in Balangoda than in the colder climes. They seem to have adopted a slash-and-burn technique for easier hunting but extended the technique to cultivate oats and barley from about 15,000 years BP. Skeletal remains of probably domesticated dogs have been found, dating back to about 4,500 years BP.

[Source: *Wickipedia.org/wiki/Balangoda Man*]

PAREEKSHIT THAMPURAN AWARD TO DR. N.P. UNNI

Dr. N.P. Unni, former Vice-Chancellor, Sanskrit University was awarded the *Pareekshit Thampuran Award* 2012 by the Centre for Heritage Studies, Hill Palace, Thrippunithura at a function organized on 13th November 2012. Dr. M.G.S. Narayanan presided over the function.

Dr. Unni has been nominated to the Central Sanskrit Board under the Ministry of Human Resources under Hon. Minister Pallam Raju. The nomination is for a period of 3 years from July 2012.

M. Rama

EVOLVED STRUCTURE OF LANGUAGE SHOWS LINEAGE-SPECIFIC TRENDS IN WORD-ORDER UNIVERSALS

(Continued from last issue)

To demonstrate that these correlations reflect underlying cognitive or systems biases, the languages must be sampled in a way that controls for features linked only by direct inheritance from a common ancestor[10]. However, efforts to obtain a statistically independent sample of languages confront several practical problems. First, our knowledge of language relationships is incomplete: specialists disagree about high-level groupings of languages and many languages are only tentatively assigned to language families. Second, a few large language families contain the bulk of global linguistic variation, making sampling purely from unrelated languages impractical. Some balance of related, unrelated and areally distributed languages has usually been aimed for in practice[11,12].

The approach we adopt here controls for shared inheritance by examining correlation in the evolution of traits within well-established family trees[13]. Drawing on the powerful methods developed in evolutionary biology, we can then track correlated changes during the historical processes of language evolution as languages split and diversify. Large language families, a problem for the sampling method described above, now become an essential resource because they permit the identification of coupling between character state changes over long time periods. We selected four large language families for which quantitative phylogenies are available. Austronesian (with about 1,268 languages[14] and a time depth of about 5,200[15]), Indo-European (about 449 languages[11], time depth of about 8,700 years[16]), Bantu (about 668 or 522 for Narrow Bantu[17], time depth about 4,000 years[18]) and Uto-Aztecan (about 61 languages[19], time depth about 5,000 years[20]). Between them, these languages encompass well over a third of the world's approximately 7,000 languages. We focused our analyses on the 'word-order

universals' because these are the most frequently cited exemplary candidates for strongly correlated linguistic features, with plausible motivations for interdependencies rooted in prominent formal and functional theories of grammar.

To test the extent of functional dependencies between word-order variables, we used a Bayesian phylogenetic method implemented in the software BayesTraits[21]. For eight word-order features, we compared correlated and uncorrelated evolutionary models. Thus, for each pair of features, we calculated the likelihood that the observed states of the characters were the result of the two features evolving independently, and compared this to the likelihood that the observed states were the result of coupled evolutionary change. This likelihood calculation was conducted over a posterior probability distribution of phylogenetic trees constructed using basic vocabulary data from each of the language families: 79 Indo-European languages[16,22], 130 Austronesian languages[15,23], 66 Bantu languages[24] and 26 Uto-Azteean languages (R. Ross & R.D.G., manuscript in preparation). Information of word-order typology was derived partly from the World Atlas of Language Structure database[25] and expanded with additional coding from grammatical descriptions (Supplementary Information section 1.3 and 2). The extent to which a dependent model of evaluation provides a superior explanation of the variation of word-order features to an independent model is measured using Bayes factors (BF) calculated from the marginal likelihoods over the posterior tree distribution. BF > 5 are conventionally taken as strong evidence that the dependent model is preferred over the independent model[13,26].

The results of the Bayes Traits analysis of correlated trait evolution differ considerably from the expectations derived from both universal approaches. The Greenbergian approach suggests robust tendencies towards linkages due to intrinsic system biases, while the generative approach assumes these will be 'hard' systems constraints set by discrete choices over a small innate parameter set[1,27]. Instead, our major finding is that, although there are linkages or dependencies between word-order characters within language families, these are largely lineage-specific, that is, they do not hold across language families in the way the two universals approaches predict.

1. Baker, M. *The Atoms of Language*. Basic Books. 2001.

10. Mace, R. & Pagel, M. "The Comparative Method in Anthropology". *Curr. Anthropol.* 35, 549-564. 1994.

11. Bakker, P. in *Oxford Handbook of Linguistic Typology* (Song, J.J. (Ed.)). Oxford University Press. 2010.

12. Cysouw, M. in *Quantative Linguistics: An International Handbook* (Altmann, G., Köhler, R. & Piotrowski, R. (Eds.)). 554-578. Mouton de Gruyter. 2005.

13. Pagel, M., Meade, A. & Barker, D. "Bayesian Estimation of Ancestral Character States on Phylogenies". *Syst. Biol.* 53, 673-684. 2004.

14. Gordon, R.G.J. *Ethnologue: Languages of the World* 15th Edn. SIL International. 2005.

15. Gray, R.D., Drummond A.J. & Greenhill, S.J. "Language Phylogenies reveal Expansion Pulses and Pauses in Pacific Settlement". *Science* 323, 479-483. 2009.

16. Gray, R.D. & Atkinson, Q.D. "Language-Tree Divergence Times support the Anatolian Theory of Indo-European Origin". *Nature* 48, 435-439. 2003.

17. Guthrie, M. *Comparative Bantu* Vol. 2. Gregg International. 1971.

18. Diamond, J. & Bellwood, P. "Farmers and their Languages: The First Expansions". *Science* 300, 597-603. 2003.

19. Campbell, L. *American Indian Languages: The Historical Linguistics of Native America*, 133-138. Oxford University Press. 1997.

20. Kemp, B.M. *et al.* "Evaluating the Farming/Language Dispersal Hypothesis with Genetic Variation exhibited by Populations in the Southwest and Mesoamerica". *Proc. Natl. Acad. Sci. USA* 107, 6759-6764. 2010.

21. Pagel, M. & Meade, A. "Bayesian Analysis of Correlated Evolution of Discrete Characters by Reversible-Jump Markov Chain Monte Carlo. *Am. Nat.* 167, 808-825. 2006.

22. Dyen, I., Kruskal, J.B. & Black, P. "An Indo-European Classification, A Lexicostatistical Experiment". *Trans. Am. Phil. Soc.* 82, 1-132. 1992.

23. Greenhill, S.J., Blust, R. & Gray, R.D. "The Austronesian Basic Vocabulary Database: From Bioinformatics to Lexomics". *Evol. Bioinform* 4, 271-283. 2008.

24. Holden, C.J. "Bantu Language Trees reflect the Spread of Farming across Sub-Saharan Africa: A Maximum-Parsimony Analysis". *Proc. R. Soc. Lond.* 269, 793-799. 2002.

25. Haspelmath, M., Dryer, M.S., Gil, D. & Comrie, B. *The World Atlas of Language Structure Online*. Max Planck Digital Library. 2008.

26. Raftery, A. "Approximate Bayes Factors and Accounting for Model Uncertainty in Generalised Linear Models". *Biometrika* 83, 251-266. 1996.

27. Cela-Conde, C. & Marty, G. "Noam Chomsky's Minimalist Program and the Philosophy of Mind. An Interview". *Syntax* 1, 19-36. 1998.

[To be continued]

[Courtesy: *Nature* Vol. 473, 5 May 2011]

MARXIST THINKER AND SCHOLAR P. GOVINDA PILLAI PASSED AWAY

P. Govinda Pillai, a Marxist intellectual and scholar whose reading had no particular political shades, breathed his last around 11.15 p.m. on 22nd November 2012 at Thiruvananthapuram after battling age-related ailments.

86-year old PG, as he was affectionately called, was a friend of the DLA and for some time, he had worked as a Senior Fellow in ISDL. People from different walks of life expressed condolence at the death of this Marxist thinker, literatus and philosopher.

The C.D. Meeting of the DLA on 29th November 2012 expressed sorrow at the passing away of PG. In the meeting, the members observed that PG was a multifaceted personality who excelled in his roles as a writer, philosopher and philogist. His wide and varied reading gave him a balanced perspective into language issues and he had always been a friend of the Dravidian Linguistics Association, the C.D. members remembered.

Born on 25th March 1926, PG's journey in life was from asceticism to revolutionary fervour. He joined the Communist Party in 1946, was arrested and remained behind bars till 1951. At the age of 25, he got elected to the Travancore-Cochin Legislative Assembly. During 1957-59 and 1967-69 periods, he served the country as a Member of the Legislative Assembly of Kerala. From 1964 to 1983, he was the Chief Editor of *Deśabhimani* daily and weekly, an organ of the CPI(M).

His famous books include *Isannalkkippuram* ('Literary Theories'), *Marxian Soundarya Śastram* ('Marxian Aesthetics'), *Vaigñanikaviplavamo Samskarikaviplavamo* ('Knowledge Revolution or Cultural Revolution') and *Bhakti Movement - Renaissance or Revivalism* published posthumously.

Naduvattom Gopalakrishnan

BOOKS RECEIVED

Gifted by Dr. V. Saratchandran Nair

1. *Malayalam-English-Tamil Trilingual Dictionary*. 2011. Mysore: SRLC, CIIL.
2. *Kannada-English-Malayalam Trilingual Dictionary*. 2011. Mysore: SRLC, CIIL.
3. *Telugu-English-Malayalam Trilingual Dictionary*. 2011. Mysore: SRLC, CIIL.

MALAYALAM FOR NON-MALAYALEES

ISDL has started a three-month course, *Malayalam for Non-Malayalees* in its city office near Ayurveda College Junction from 5th December 2012. Basic language-learning skills are being taught in this course.

Contribution to
Prof. V.I. Subramoniam Endowment Fund

TOTAL AS OF LAST MONTH	Rs. 3,89,120.00
CURRENT TOTAL (Including FD)	Rs. 3,89,120.00

DLA News Endowment Fund

TOTAL AS OF LAST MONTH	Rs. 1,77,377.10
CURRENT TOTAL	Rs. 1,77,377.10