DEPARTMENT OF LINGUISTICS

SCHOOL OF INDIAN LANGUAGES

UNIVERSITY OF KERALA

M.PHIL. PROGRAMME IN INTER DISCIPLINARY COMPUTATIONAL LINGUISTICS SYLLABUS

(Under Credit and Semester System w.e.f. 2016 Admissions)

Structure of the Programme

M.A.PROGRAMME IN INTER DISCIPLINARY M.PHIL. PROGRAMME IN COMPUTATIONAL LINGUISTICS

(Under Credit and Semester System w.e.f. 2016 Admissions)

Programme objectives

- To give the scientific knowledge of human Languages
- To explain the applications of computers in various Linguistics research & development
- To introduce Linguistics, Language, the nature of human language, theoretical linguistic and analytical skills for recognizing and describing the various levels of language
- To gain knowledge on computational grammar and Natural Language Processing and Speech Technology
- To know the application of Linguistics in various fields.
- To provide preliminary and analytical procedures in phonemic analysis. And also included analytical problems to be worked out which gives a practice to analyze a language and reduce to writing
- To train students about the theories and formulations of bilingualism and train the students to know more about dialectological phenomenon in these fields.
- To introduce Sociolinguistics and basic concepts and theories of both early foundational work and current issues in the field.
- To introduces the principles of linguistics and computer science and their mutual application
- To impart the knowledge on the importance of quantitative study of languages, particularly it explains the use of language data in Machine-readable form in NLP various corpus analysis tools and statistical information on language use.
- To introduces the principles of linguistics and computer science and their mutual application.

Structure of the Programme

Semester No	Course Code	Name of the Course	Number of Credits
I	LIN-711	Research Methodology	**4**
	LIN-712	Linguistic Theories	**4**
	LIN-713 (I) LIN-713 (II)	Natural Language: Understanding and Processing Current Trends in Computational Linguistics	**4** **4**
II	LIN-721	Dissertation and Viva Voce	**20**
		TOTAL CREDITS	**32**

Semester : I

Course Code : LIN-711

Course Title : RESEARCH METHODOLOGY

Credit : 4

Aim: This paper is introduced to lay strong foundation to students in Basic Research Methodology and Python Programming theories.

Objectives:

- To give the scientific knowledge of Research Methodologies
- To explain the applications of computers in various Linguistics research & development
- To introduces the principles of linguistics and computer science and their mutual application

- To impart the knowledge on the importance of quantitative study of languages, particularly it explains the use of language data in Machine-readable form in NLP various corpus analysis tools and statistical information on language use.
- To introduces the principles of linguistics and computer science and their mutual application.

Course Content

Module I: Creativity & Thinking Skills: Various views on Creativity; Stimulating Creativity; obstructions to creativity; Creativity & craft; Critical Thinking; Logical Thinking– Inductive & Deductive logic– common logical fallacies.

Module II: Research Basics: Various outlooks on research; Types of Research: pure versus applied, incremental versus innovative; Purpose and Objectives of Research: Philosophy of science; the scientific method, the research process– creative question– hypothesis Setting up of hypothesis– nature of role of hypothesis in scientific investigations– types, sources and characteristics of hypothesis- historic comparative descriptive and scientific observation and experimentation- design of experiments– critical analysis– reporting; abilities of Research Workers– Evaluation of earlier works, Identification of the research problem.

Module III: Methods of Data Collection: Primary and secondary sources– use of libraries, indexes and bibliographies. Data collection- responses– social surveys and their objectives– case study– sampling design– questionnaire; purpose, types, form and characteristics. Interview methods, objectives– types and techniques– analysis and interpretation– content analysis and classification of data. Types of statistical methods; sampling and stratification– generalization and confirmation– adoption of terminology.

Module IV: Publishing & Program: Publication process- Peer review– plagiarism –Turnitin; Open Access Publications; Evaluating quality of scientific publications– Bibliometry– Impact factors & H-index– pit falls in interpreting impact; case study of paper writing and peer review; popular journals in Computational Linguistics; An introduction to the Python programming language- Python interpreter- Python's basic data types, files, functions, and error handling, tuples, lists, dictionaries, and sets. List processing primitives, Python's object model- variables, reference counting, copying, and type checking, design functions, scoping rules, documentation strings, and anonymous functions.

Module V: Text Processing: Doctest and unittest modules, assertions. Python debugger and profiler Iterators, Generators, Text, and Binary Handling: Parsing and generating text data, string splitting, string stripping, string joining, template strings, and Unicode. Re module, regular expression pattern matching and substitution. Binary I/O and library modules for handling binary- encoded data (struct, array, etc.). Python interpreter process- Python with.

References

- Agnihotri,V (1980) Techniques of Social Research, MN Publications, New Delhi
- Anitha Rao and Bhaneji Rao, Intellectual Property rights: A primer, Eastern Book Company.
- Badckar, V H (1982) How to write assignments, research papers, Dissertation and thesis, karaka, publication, New Delhi
- Barzun, J Gruff, H E (1971) The Modern researcher, Harcount Brace Jovanovich, New York
- Glosh B N (1982) Scientific Methods and social research Sterling Publication Pvt. Ltd. New Delhi
- Gregory Bassham, William Irwin, Henry Nardone and james Wallace, Critical Thinking: A Student's Introduction, Tata McGraw Hill education Pvt. Ltd.
- Haus Raj (1987) Theory and Practice in Social Research, Surjeet Publication, New Delhi
- James Hartley, Routledge, Taylor and Francis, Academic Writing and Publishing: A Practical Hand Book.
- Jeffrey A. Lee, the scientific endeavor: a primer on scientific principles and practice, Benjamin Cummings
- Marian Petre and Gordon Rugg, The Unwritten Rules of PhD Research, Tata McGraw Hill education Pvt. Ltd.
- Marlene Caroseli, Quick Wits: 50 Activities for Developing Critical Thinking Skills, Ane Books.
- Paul Sloane and Kogan Page, The Leader's Guide to Lateral Thinking Skills, unlocking the Creativity and Innovation in You and Your Team.
- Rhonda Abrams and Julie Vallone, Winning Presentation in a Day.
- Rob Kitchin and Duncan Fuller, The Academic's Guide to Publishing, Vistaar Publications.
- Robert A. Day, How to Write and Publish a Scientific Paper.
- Rowena Murray, How to Write a Thesis, Tata McGraw Hill Education Pvt. Ltd.

- Sarah Milstein, J. D. Biersdorfer and Mathew Macdonald, Google: The Missing Manual, SHROFF Publishers.
- Tuckman, B W (1972-78) Conducting Educational research, Harcount Brace, Jovanowich, New York.

Semester : I

Course Code : LIN-712

Course Title : LINGUISTC THEORIES

Credits : 4

Aim: This paper is introduced to lay strong foundation to students in linguistic principles.

Objectives:

- To give the scientific knowledge of human Languages
- To introduce Linguistics, Language, the nature of human language, theoretical linguistic and analytical skills for recognizing and describing the various levels of language
- To know the application of Linguistics in various fields.
- To provide preliminary and analytical procedures in phonemic analysis. And also included analytical problems to be worked out which gives a practice to analyze a language and reduce to writing
- To train students about the theories and formulations of bilingualism and train the students to know more about dialectological phenomenon in these fields.

Course Content

Module I: Fundamental principles: Philological vs. linguistics approaches to language; prescriptive vs. descriptive approaches; language specific vs. cross-linguistic studies; synchronic and diachronic principles; langue and parole distinction; linguistic analysis: paradigmatic vs. syntagmatic; structural vs. functional; type and token; emic and etic; listing and generalization.

Module II: Structural principles: Principles of structuring language at the phonological, morphological, syntactical and, semantic levels; immediate constituent analysis, phrase structure grammar. Labeled bracketing, tree structuring, etc.; principles of structural semantics. Lexicology and lexicography; lexical semantics; types of dictionaries; principles and methods of dictionary making; compilation of electronic or machine readable dictionaries; compilation of dictionaries using corpus; building word nets and other lexical resources.

Module III: Generative principles: Principles of adequacy; competence vs. performance; generative principles vs. descriptive principles; transformational generative grammar; x-bar theory, government and binding theory, minimalist theory, principles and parameters.

Module IV: Typological principles: Principles of language typology; type traits in phonology, grammar and vocabulary; classification of language based on types; types based on morphological complexity: agglutinative, isolating and synthetic; types based on word order: SOV languages, SVO languages, VSO languages, etc.

Module V: Language Technology: Application of technology for the development of language; technological development of Indian Languages (TDIL); natural language processing; electronic compilation of various language tools such as dictionaries, thesaurus, encyclopedias, query systems, user interfaces etc.; computer aided language teaching and learning; machine translation; text to speech and speech text systems;

References

- Allen, J. 1995. Natural language understanding. The Benjamin, New York.
- BoSvensen 1993. Practical lexicography: Principles and methods of dictionary making.
- Cruse, A. 2000. Meaning in language: an introduction to semantics and pragmatics, Oxford University Press, Oxford.
- Crystal, D. 1971. Linguistics. Pelican.
- Fromkin, A.V. (ed.) 2001. Linguistics: an introduction to linguistic theory. Blackwell, Oxford.
- Hausser, R, 1999. Foundations of Computational linguistics: man-machine communication in natural languages. Springer.

- Helgeman, L. 2001. Introduction to Government and binding theory. Blackwell, Oxford.
- Jurafsky, D and Martin J.H. 2000. Speech and language processing: an introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Prentice Hall, New Jersey.
- Lyons J. 1968. Introduction to theoretical linguistics. Cambridge university press, Cambridge.
- Lyons J. 1982. Language and Linguistics. Cambridge University Press, Cambridge.
- Lyons, J. 1999. Linguistic Semantics. Cambridge university press, Cambridge
- Mejis, M. 1992. Computers and dictionaries. In Christopher, S (ed.) Computers and written texts. Blackwell, Oxford.
- Mejis, M. 1996. Linguistic Corpora and Lexicography. In Annual reviews of Applied Linguistics, vol. 16.
- Radford A. 1981. Transformational Syntax: A Student's Guide to Chomsky's Extended Standard theory, Cambridge university press, Cambridge.
- Radford, A. 1981. Transformational Generative Grammar Cambridge university press, Cambridge.
- Radford, A. 1983. Transformational Syntax. Cambridge University Press, Cambridge.
- Van Valin, R.D. 2000. An Introduction to Syntax. Cambridge University Press, Cambridge.
- Varma S.K. and Krishnaswamy, N. 1989. Modern Linguistics: an introduction. Oxford University Press, Delhi.
- Zgusta, L. et al. 1971. Manual of Lexicography. Mouton, The Hague.

Optional Papers:

Semester : I

Course Code : LIN 713 (I)

Course Title : NATURAL LANGUAGE: UNDERSTANDING AND PROCESSING

Credits : 4

Aim: This Paper intended to understand different NLP activities and applications for language studies.

Objectives:

- To impart the knowledge on the importance of quantitative study of languages, particularly it explains the use of language data in Machine-readable form in NLP various corpus analysis tools and statistical information on language use.
- To introduces the principles of linguistics and computer science and their mutual application.
- To give the scientific knowledge of human Languages
- To explain the applications of computers in various Linguistics research & development

Course Content

Module I: Introduction to Natural Language Processing: Introduction to Language Technology, Linguistics and Language Technology, Structure of Natural Language, Design Features of Language, Morphology, Phonology, Syntax, Semantics and Lexicography, Computational aspects of Language Technology, Computer applications of Natural Languages. Language teaching, Learning, Computer assisted Language learning and Teaching

Module II: Computational Morphology: Computational Morphology, Regular Expressions, Language automata and Computation, finite State Automata, Fundamentals of Corpus Analysis, Morphology-inflectional-derivational, Finite state Morphological Parsing with finite state Transducers, Orthographic rules, Combining FST Lexicon and Rules

Module III: Computational Language Analysis: Computational approaches to Syntax and Grammar, Computational Phonology and Text to speech, Hidden Markov Models and speech Recognition, Iturbi algorithm. Word tokenization- Pronunciation and spelling- Statistical NLP N-gram and Language Models- Word sense disambiguation and information Retrieval.

Module IV: Part-of-Speech (POS) Tagging: Morphological parsing, Transliteration-Transliteration using sequence labeling-Part-of-Speech(POS)Tagging, Tag sets, POS Tagging using SVM-Chunking-Shallow Parsing-Context-Free Grammars, Parsing with Context-Free Grammars- Probabilistic and Lexicalized Parsing, CFG Parser-Parsing Techniques-Structured Output Learning-Generalized Linear Classifiers in NLP.

Module V: Machine Translation and Applications of NLP: Machine Translation, Natural Language Generation, Transfer Metaphor, Syntactic Transformation, Lexical Transfer, Linguistic Tools and Training, Morphological Analyzer, PC KIMMO, POS Tagger, Stanford Tagger, Brills Tagger, Parser (Training), Stanford Parser, Introduction to Information Retrieval: Web Mining, Search Engines and Search Algorithms, Web Applications of Languages. Web semantics, Natural Language Tool Kit (NLTK), Word net.

References

- Akshar Bharati, Chaitanya Vineet, Sangal Rajeev. Natural Language Processing A Paninian Perspective, Prentice Hall India. 1999.
- Allen J, Natural Language Understanding, The Benajmins and Cummings Publishing Company Inc. 1994.
- Bonnie Jean Dorr, Machine Translation- A View from the Lexicon (Artificial Intelligence) by Publisher: Mit Press, 1993
- Hristo Georgiev. Language Engineering by, Continuum, 2007
- Jurafsky, D. and J. H. Martin Speech and Language Processing- Prentice-Hall. 2000.
- Martin Manning, C. D. and H. Schütze. Foundations of Statistical Language Processing, The MIT Press. 1999.
- Martin Rajman and Vincenzo Pallota. Speech and Language Engineering by, Publisher: Efpl Press, 2007

Semester : I

Course Code : LIN-713 (II)

Course Title : CURRENT TRENDS IN COMPUTATIONAL LINGUISTICS

Credits : 4

Aim: The goal of this course is to expose the student to the core techniques and applications of computational linguistics. The course will address both theoretical and applied topics in Computational Linguistics.

Objectives:

- To give the scientific knowledge of human Languages
- To explain the applications of computers in various Linguistics research & development
- To introduce Linguistics, Language, the nature of human language, theoretical linguistic and analytical skills for recognizing and describing the various levels of language
- To gain knowledge on computational grammar and Natural Language Processing and Speech Technology
- To impart the knowledge on the importance of quantitative study of languages, particularly it explains the use of language data in Machine-readable form in NLP various corpus analysis tools and statistical information on language use.
- To introduces the principles of linguistics and computer science and their mutual application.

Course Content

Module I: Introduction to Computational Linguistics and Grammar: What is Computational Linguistics? Interdisciplinary relevance: Formal Linguistics, Psycho-linguistics, Cognitive Science, Chomsky Hierarchy, **Initial Systems:** Turing Test, Dialog systems: ELIZA. Lexical Functional Grammar (LFG), Head-Driven Phrase Structure Grammar (HPSG). Context- Free grammars (CFGs), Descriptive Grammar of Malayalam.

Module II: Statistics: Probability, Joint and Conditional Probability, Bayes rule, Regression, Graph theory. **Machine Learning**: Supervised, Unsupervised and Semi-supervised learning. Decision trees (C4.5), Inductive logic programming, Naïve Bayesian Classifier, Hidden Markov Model, Singular Value Decomposition (SVD), Support Vector Machine(SVM), Conditional Random Fields(CRFs).

Module III: Computational Corpus Linguistics: Why corpus linguistics? What is a corpus? Different Corpus types, Corpora Development, World Wide Web as a corpus, British National Corpus, Speech Corpora, Multimedia corpora, Parallel Corpus, Corpus collection and design, **Font and Encoding:** Font design and development, Encoding scheme, Character encoding and decoding, UNICODE (utf8) and ASCII, ISCII **Corpus Annotation:** Tagging, Parsing, Tree-banks, Co**rpus Tools:** Dictionaries, Thesaurus creation, Tokenization, Concordance, Stemmer. **Quantitative linguistics:**Quantitative data analysis, Collocations and idioms, Text types and Genre.

Module IV: Language Modeling: Language models and their role in Text and Speech processing. Different types of Language modeling, Markov models, N-gram models, Entropy, Relative entropy, Cross entropy, Mutual information, Statistical estmation and smoothing for language models.

Module V: Text Analytics: Statistical Machine Translation (SMT), Alignment Models and Expectation Maximization (EM), EM and its use in statistical MT alignment models. Statistical phrase based systems and syntax in SMT. **Syntax and Parsing** : Context-Free Grammars (CFGs) Parsing, Top-down and bottom-up parsing, empty constituents, left recursion, Probabilistic CFGs Parsing, Dependency Parsing, Modern Statistical Parsers: Charniak Parser, The Stanford Parser, Malt Parser. **Text Mining and IE:** Document Clustering, Text Similarity, Information Extraction (IE) and Named Entity Recognition (NER), Co-reference Resolution, Statistical and Rule-based methods. **Information Retrieval:** Ranking Algorithms, Query Modification and Effectiveness, Representation of Documents, IR Models- Boolean and Vector Space Models. File Structures: Inverted Files, Signature Files. Term and Query Operations: Lexical Analysis and Stop lists, **Evaluation**: Precision, Recall and F-score: Different Evaluation metrics: BLEU, B-CUBED, IR Evaluation: Relevance Judgment, Map Score. **Open source Tool Kits for NLP applications**: GATE, WEKA, CRF++, Moses.

References:

- C. Manning, P. Raghavan, and H. Schütze: 2008, Introduction to Information Retrieval, Cambridge University Press.
- Carl Jesse Pollard, Ivan A. Sag: 1994, Head-Driven Phrase Structure Grammar, University of Chicago Press.
- Charniak, Eugene. 1993. Statistical Language Learning. The MIT Press.
- Charniak, Eugene: 1984, Introduction to artificial intelligence, Addison-Wesley.
- Christopher Manning and Hinrich Schütze: 1999, Foundations of Statistical Natural Language Processing. MIT Press. Cambridge, MA.
- Daniel Jurafsky and James H. Martin: 2000, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice-Hall.
- Ethem Alpaydin: 2005, Introduction to Machine Learning, Prentice- Hall of India. New Delhi
- Geoffrey Sampson and Diana McCarthy: 2004. Corpus Linguistics: Readings in a Widening Discipline. Continuum Press.
- http://www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm
- Jeffrey D Ullman, Rajeev Motwani and John E Hopcroft : 2000, Introduction to Automata Theory, Languages, and Computation (2nd Edition)**,** Addison Wesley.
- Martin Wynne (Ed.). 2005. Developing Linguistic Corpora: A Guide to Good Practice. Oxbow Books.
- Tom Mitchell:1997, Machine Learning, McGrow Hill
- W. B. Frakes and R. Baeza-Yates (Eds.): 1992, Information Retrieval: Data Structures & Algorithms, Prentice-Hall.

Semester : II

Course Code : LIN-721

Course Title : Dissertation and viva voce

Credits : 20

Aim: This paper is introduced to lay strong foundation to students in Language Teaching Methods.

Objectives:

- To give the scientific knowledge of human Languages
- To explain the applications of computers in various Linguistics research & development
- To introduce Linguistics, Language, the nature of human language, theoretical linguistic and analytical skills for recognizing and describing the various levels of language
- To gain knowledge on computational grammar and Natural Language Processing and Speech Technology
- To know the application of Linguistics in various fields.
- To provide preliminary and analytical procedures in phonemic analysis. And also included analytical problems to be worked out which gives a practice to analyze a language and reduce to writing
- To train students about the theories and formulations of bilingualism and train the students to know more about dialectological phenomenon in these fields.
- To introduce Sociolinguistics and basic concepts and theories of both early foundational work and current issues in the field.
- To introduces the principles of linguistics and computer science and their mutual application
- To impart the knowledge on the importance of quantitative study of languages, particularly it explains the use of language data in Machine-readable form in NLP various corpus analysis tools and statistical information on language use.
- To introduces the principles of linguistics and computer science and their mutual application.