GIAN Course

A Practical Introduction to Corpus Linguistics

December 10-16, 2018, IIT (BHU), Varanasi

Overview

The importance of understanding and working on language is being increasingly recognized with the renewed rise of Artificial Intelligence techniques. Linguistics itself is a fascinating area of study and has repeatedly provided insights that have useful in a diverse variety of disciplines. Empirical methods have been used in Linguistics almost from the time Modern Linguistics was born. These have become more popular among researchers, both for theoretical study of language(s) and for practical applications. Corpus Linguistics is the discipline where such methods are studied. The main idea is to use statistical techniques for studying language and linguistic data, that is, corpus.

Being an empirical approach, such Linguistics requires good understanding of the principles on which statistical methods are based. As such, this is an area where statistical methods are used to form or verify hypotheses about various linguistic phenomenon. It is closely connected with other disciplines such as Applied Linguistics, Psycholinguistics, Cognitive Linguistics, Forensic Linguistics, Sociolinguistics and Computational Linguistics. Since corpora of different kind provide the primary source of evidence in empirical Linguistics, it is important for all researchers working in these areas and also in Natural Language Processing, Information Retrieval, Artificial Intelligence, Speech Processing etc. to know and understand the fundamentals of Corpus Linguistics. Otherwise, the validity of their work may not be clear. This is true of any researcher who relies on corpora for evidence or for machine learning. The number of researchers using corpora for machine learning is increasing very rapidly and language technology systems based on machine learning from corpora are being developed and deployed in real life now, even on mobile devices.

The objective of the course is to introduce the participants to Corpus Linguistics with the help of theory lectures as well as lab sessions and to make them aware of its use in various areas of research and technology development. The primary objectives of the course are as follows:

- Introducing the participants to the fundamentals of Corpus Linguistics
- A brief review of statistical techniques used for Corpus Linguistics and Quantitative Linguistics using R
- Demonstrating the use of Corpus Linguistics techniques for research questions and for practical applications
- Underlining the importance of insights from Corpus Linguistics for areas like Computational Linguistics, Natural Language Processing and Artificial Intelligence and considering case studies on this point

Stefan Thomas Gries (born 1970) is (full) professor of linguistics in the Department of Linguistics at the University of California, Santa Barbara (UCSB), Honorary Liebig-Professor of the Justus- Liebig-Universität Giessen, Germany, (since September 2011), and Visiting Chair (2013–2018) of the Centre for Corpus Approaches to Social Science at Lancaster University; in the summer semester of 2017, he is the Honorary Leibniz-Professor at the Research Academy of the University of Leipzig, Germany.

Gries earned his M.A. and Ph.D. degrees at the University of Hamburg, Germany, in 1998 and 2000. He was at the Department of Business Communication and Information Science of the University of Southern Denmark at Sønderborg (1998–2005), first as a lecturer, then as assistant professor and tenured associate professor. In 2005, he spent 10 months as a visiting scholar in the Psychology Department of the Max Planck Institute for Evolutionary Anthropology in Leipzig, Germany, before he accepted a position at UCSB, starting November 1, 2005.

Anil Kumar Singh is a researcher and a teacher who has been working in the area of NLP for the last thirteen years. He is working as an Assistant Professor in the department of Computer Science and Engineering at IIT (BHU), Varanasi, India. He has published on various topics in NLP and has organized a couple of research workshops and a couple of introductory workshops on NLP. At present, he is also involved in building machine translation systems for Bhojpuri, Maithili and Magahi, which are all less resourced languages.

Course Coordinator

Anil Kumar Singh
Phone: +91-9648277700
E-mail: aksingh.cse@iitbhu.ac.in

Address:

Associate Professor
Department of Computer Science & Engg.
IIT (BHU), Varanasi-221005 (UP), India

.....

Registration:

http://www.gian.iitkgp.ac.in/GREGN

Course Website:

https://sites.google.com/view/gian-corpus-linguistics-2018/home

Participants	**Number of participants for the course will be limited to fifty or less.**
You Should Attend If…	▪ you are an engineer or research scientist interested in language technology and want to know more about Corpus Linguistics. ▪ you are a linguist or from language studies and want to find out how your knowledge could be used for Corpus Linguistics. ▪ you are a student or faculty member from an academic institution interested in natural language processing or computational linguistics and realize the importance of Corpus Linguistics, but would like to learn more ▪ you want to get hands-on practice with Corpus Linguistics
Fees	**Participants from abroad: US $250** **Industry/Research Organizations: Rs. 10000/-** **Faculty Members from Academic Institutions: Rs. 5000/-** **Students: Rs. 1000/-** The above fees include all instructional materials, computer use for tutorials and assignments, laboratory equipment usage charges, free internet facility. The participants will be provided with accommodation on payment basis. **Please note that no lunch will be provided. Tea and snacks will be provided twice each day.**