Subject: Speech and Language Processing   **Course Code**: COMP 473
Type: Elective   **FM**: 100 (50 internal + 50 final)

Course Description:
An Introduction to Natural Language Processing, Computational Linguistics, and Speech
Recognition: Morphology, Syntax, Semantics, Discourse, Speech.

Prerequisites:
It is expected that students have taken prior courses like Statistics and Probability, Discrete
Mathematics, Artificial Intelligence, Programming and Data Structures. For the understanding and
implementation of the algorithms, it is essential that the students have a fairly good command of
some of the high level programming languages like C, C++, Python or Java.

Course Objectives:
To provide the students a general overview of the basics as well as the advanced concepts of Speech
and Language Processing (SLP). Upon the completion of the course, students are expected to be
able to apply the different concepts of SLP in different scenarios, both theoretical and practical.

Grading policy:

Final exam: 50

Internals:

Lab and practical assignments: 10
Final Lab exam and viva: 10
First internal: 10
Second internal: 10
Mini-Project and term report on the project: 10

Text book:

1. Speech and Language Processing – Jurafsky and Martin, Second Edition. Pearson Education.
2. Guide to localization of Open Source Software – Nepalinux team, Madan Purasakar Pustakalaya.
Center for Research in Urdu Language Processing. National University of Computer and Emerging
Sciences. Lahore, Pakistan. Edited by Bal Krishna Bal, 2007.

Reference books:

1. Unicode Demystified. A Practical Programmer's Guide to the Encoding Standard. Richard
Gillam. Addison-Wesley, 2003.
2. Natural Language Processing with Python. Stephen Bird, Ewan Klein, and Edward Loper.
O'Reilly Media, 2009. http://www.nltk.org/book/

COMP 473 – Speech and Language Processing

Lesson Plan

Unit	Topic	Assignments due	Suggested Readings
1	Introduction to NLP Origins and importance of NLP Challenges in NLP (Difficulties, Ambiguities and Evolution) Language and Knowledge (Phonetics, Syntax, Semantics, Pragmatics and Discourse) A multi-disciplinary field (Psychology, IR) Applications of NLP	Usually, assignments are due one week after they are announced.	Chapter 1 – Jurafsky and Martin. (**3 hrs**)
2.	Localization and Language Computing Introduction to Localization Key Concepts of Localization Unicode and Multilingual Computing	-	Guide to Localization of Open Source Software (**4 hrs**)
3.	Finite State Machines (FSM) and Morphology Introduction to FSM and FST Morphological Processes Principles of Word Construction (Suffix, Prefix, Stem, Affixes) English, Nepali, Turkish Morphological Representation and FSM Lexicon, Morphotactics and Orthographic rules Morphological Parsing and FST Mealy Machines FST Operations	-	Chapters 2-3, Jurafsky and Martin (**6 hrs**)
4.	Parts of Speech (PoS) Tagging and Hidden Markov Models(HMM) PoS Tagsets Rule-based PoS tagging Stochastic PoS tagging Transformation based tagging	-	Chapter 5 – Jurafsky and Martin (**4 hrs**)
5.	Syntactic Analysis Context Free Grammar (CFG) & Probabilistic CFG Word's Constituency (Phrase – Level, Sentence Level) Parsing (Top-Down and Bottom-Up) CYK Parser Probabilistic Parsing	-	Chapters 12-14, Jurafsky and Martin (**7 hrs**)

6.	Lexical Semantics Lexeme, Lexicon, Senses, Lexical relations, WordNet(Lexical Database), Word Sense Disambiguation (WSD), Word Similarity	-	Chapters 19-20, Jurafsky and Martin (**7 hrs**)
7.	Pragmatics & Discourse Analysis Monologue and Dialogue Reference Resolution Coherence and Cohesion Discourse Structure	-	Chapter 21 – Jurafsky and Martin (**7 hrs**)
8.	Applications of SLP Information Extraction Information Retrieval Machine Translation Wordnet and Ontologies Opinion Mining and Sentiment Analysis Text-to-Speech Automatic Speech Recognition	-	Chapters 9, 22-25, Jurafsky and Martin (**7 hrs**)

Lab and Practical Assignments

Note: Students are expected to do adequate readings of the concerned Chapters from the book "Natural Language Processing with Python" before every lab/practical class. A lot of these would be hands-on exercises in the classroom hours and take-away-home assignments.

Tasks	Timelines
Lab 1: Language Processing and Python Computing with Language : Text and Words A closer look at Python: Texts as Lists of Words Computing with Language: Simple Statistics Back to Python: Making decisions and taking control Automatic Language Understanding	Week 1
Lab 2: Accessing Text Corpora and Lexical Resources Accessing Text Corpora Conditional Frequency Distributions More Python: Reusing code Lexical Resources WordNet	Week 2
Lab 3: Processing Raw Text Accessing Text from the Web and from Disk Strings: Text Processing at the Lowest Level Text Processing with Unicode Regular Expressions for Detecting Word Patterns	

Useful Applications for Regular Expressions Normalizing Text Regular Expressions for Tokenizing Text Segmentation Formatting: From Lists to Strings	
Lab 4: Writing Structured Programs Back to basics Sequences Questions of Style Functions: The Foundation of Structured Programming Doing more with Functions Program Development Algorithm Design A Sample of Python Libraries	Week 4
Lab 5: Categorizing and Tagging Words Using a Tagger Tagged Corpora Mapping Words to Properties Using Python Dictionaries Automatic Tagging N-Gram Tagging Transformation-Based Tagging How to Determine the Category of a Word	Week 5
Lab 6: Learning to Classify Text Supervised Classification Further Examples of Supervised Classification Evaluation Decision Trees Naive Bayes Classification Maximum Entropy Classifiers Modeling Linguistic Patterns	Week 6
Lab 7: Extracting Information from Text Information Extraction Chunking Developing and Evaluating Chunkers Recursion in Linguistic Structure Named Entity Recognition Relation Extraction	Week 7
Lab 8: Analyzing Sentence Structure Some Grammatical Dilemmas What's the Use of Syntax? Context-Free Grammar Parsing with Context-Free Grammar Dependencies and Dependency Grammar Grammar Development	Week 8