Co-funded by the
Erasmus+ Programme
of the European Union

Development of the interdisciplinary master program on Computational Linguistics at Central Asian universities

585845-EPP-1-2017-1-ES-EPPKA2-CBHE-JP

Syllabuses

Work Package 2 / Task 2.2 / Deliverable 2.2

December 2019

Content

Formal Grammar

1. Basic Data

Subject Name	*Formal Grammar*
ECTS	5
Obligatory/Elective	Obligatory
Module	Applied Linguistics
Semester	3 (would be advisable to move it to the first or, if not possible, to the second semester)
Teaching language	English, Uzbek, Russian
Professors	D.Kselov, A.Shermatov
Department / Faculty	
University / Center	Samarkand State Institute of Foreign Languages

2. Objectives

1. Provide introduction to key ideas and issues of computational linguistics from the perspective of formal grammar.
2. Introduce the fundamental concepts of formal language, formal grammar, and automata theory.
3. Familiarize with the standard terminology and the most important theoretical tools and concepts related to formal grammar.
4. Classify machines by their power to recognize languages.
5. Employ automata to solve problems in computational linguistics.

3. Subject Content

1. Introduction to Propositional calculus.
    1.1. Basic concepts.
    1.2. Basic and derived argument forms.
    1.3. Proofs in propositional calculus.
    1.4. Soundness and completeness of the rules.
    1.5. Interpretation of a truth-functional propositional calculus.
    1.6. Alternative calculus.
    1.7. Equivalence to equational logics.
    1.8. Graphical calculi.
    1.9. Other logical calculi.
    1.10. Solvers.
2. A hierarchy of formal languages.
    2.1. Languages, grammars, automata, and inference.
    2.2. Computational definition of 'grammar'.
    2.4. Regular languages.
    2.5. Context-free languages.
    2.6. Mildly Context-Sensitive languages.
    2.7. Context-sensitive languages.
    2.8. Description of natural language phenomena.
3. Automata models.
    3.1. Definitions and concepts.
    3.2. Finite-state automata and regular languages.
    3.3. Push-down automata and Context-free languages.
    3.4. Automata models for Mildly Context-Sensitive languages.

4. Probabilistic grammars.

4.1. Definitions and concepts.

4.2. Probabilistic Regular grammars.

4.3. Probabilistic Context-free grammars.

4.4. Probabilistic Mildly Context-Sensitive Grammars.

---

4. Recommended Bibliography

Rozenberg, G. & Salomaa, A. (eds), 1997. Handbook of formal grammars. 3 Vols. New York: Springer.

Hopcroft, J.E., Motwani, R. & Ullman, J.D., 2006. Introduction to Automata Theory, Languages, and Computation. Addison Wesley.

Levelt W.J.M., 2008. An Introduction to the Theory of Formal Languages and Automata. John Benjamins Publishing Company. Amsterdam / Philadelphia.

Jeffrey Heinz, Colin de la Higuera, Menno van Zaanen, 2016. Grammatical inference for Computational Linguistics. Morgan & Claypool Publishers.

Christopher D. Manning and Hinrich Schütze, 1999. Foundations of Statistical Natural Language Processing. MIT Press.

Laura Kallmeyer, 2010. Parsing Beyond Context-Free Grammars. Springer-Verlag, Berlin & Heidelberg.

Пулатов А.К., 2011. «Компьютер лингвистикаси». Учеб.на узб.яз. (Pulatov A.K. 2011. ‘Computational Linguistics’. Textbook in Uzbek)

Recommended software:

JFLAP (http://www.jflap.org/). If chosen, see also: Susan Rodger and Thomas Finley, 2006. JFLAP - An Interactive Formal Languages and Automata Package, ISBN 0763738344, Jones and Bartlett. Available online at: https://www2.cs.duke.edu/csed/jflap/jflapbook/ (visited on August 2019)

Nooj (http://www.nooj-association.org/)

5. Competences

Upon completion of this course, the student will be able to:

1. Understand the concept of abstract machines and their power to recognize languages.
2. To know the adequacy and limitations of different models of automata for modeling and solving computing problems.
3. Design grammars for formal languages

6. Methodology

36 hours of theoretical instruction, 36 hours of practical courses, and 48 hours for self-study.

7. Assignments and Evaluation

10% attendance and active participation, current tasks performance;

40% midterm control (results of tests);

50% final examination (the research and final written work).

8. Requirements

- Students should attend theoretical and practical classes, do home assignments, study additional materials given for independent work.
- Students will be asked to demonstrate their understanding of the content on a weekly basis through discussions and assignments.
- Students are expected to post questions to the Class MOODLE Site to get help with the material.

Introduction to Programming for NLP (Python)

1. Basic Data

Subject Name	Introduction to Programming for NLP(Python)
ECTS	5
Obligatory/Elective	Obligatory
Module	Computational Technologies
Semester	1
Teaching language	English
Professors	Mersaid Aripov, Sanatbek Matlatipov and Muftakh Khakimov
Department / Faculty	Applied mathematics and computer analysis
University / Center	National University of Uzbekistan

2. Objectives

This course is a **practical** introduction to NLP. In the course, students will learn by example, write real programs, and grasp the value of being able to test an idea through implementation. The course is also suitable for non-computer science background students as the course will teach them in **programming.** The course provides extensive illustrations and exercises from NLP. The course will also try to be pragmatic in striking a balance between theory and application, identifying the connections and the tensions.

3. Subject Content

1. Language Processing and Python
   - 1.1 Computing with Language: Texts and Words
   - 1.2 A Closer Look at Python: Texts as Lists of Words
   - 1.3 Computing with Language: Simple Statistics
   - 1.4 Back to Python: Making Decisions and Taking Control
   - 1.5 Automatic Natural Language Understanding
2. Accessing Text Corpora and Lexical Resources
   - 2.1 Accessing Text Corpora
   - 2.2 Conditional Frequency Distributions
   - 2.3 More Python: Reusing Code

8. Analyzing Sentence Structure
   8.1 Parsing with Context-Free Grammar
   8.2 Dependencies and Dependency Grammar
   8.3 Grammar Development
9. Building Feature-Based Grammars
   9.1 Grammatical Features
   9.2 Processing Feature Structures
   9.3 Extending a Feature-Based Gramma
10. Analyzing the Meaning of Sentences
   10.1 Natural Language Understanding
   10.2 Propositional Logic
   10.3 First-Order Logic
   10.4 Discourse Semantics

---

4. Main and Recommended Bibliography

- Bird, Steven, Ewan Klein, and Edward Loper (2009), Natural Language Processing with Python, O'Reilly Media.
-Speech and Language Processing, by Jurafsky and Martin (Prentice Hall, 2008).
-Kübler, McDonald and Nivre's "Dependency Parsing"

---

5. Competences

In this course, student will learn:

• How simple programs can help you manipulate and analyze language data, and how to write these programs
• How key concepts from NLP and linguistics are used to describe and analyze language
• How data structures and algorithms are used in NLP
• How language data is stored in standard formats, and how data can be used to evaluate the performance of NLP techniques

---

6. Methodology

As the basic literature the book «Natural Language Processing with Python» authors Steven Bird, Ewan Klein, and Edward Loper will be used. More low stated methodology is based on heads of this book. Mark Luttsa's book «Learning Python» forth edition, and also the book on Python in

the Uzbek language which will be specially published within the limits of the present project will be in addition used.

As independent (house) works exercises which are set in the end of corresponding heads of the basic book will be set. Identical examples on processing of the Uzbek language will be besides, set.

Chapter	Lecture	Independent (house) works
Chapter 1. Language Processing and Python	2	6
Chapter 2. Accessing to Text Cases and Lexical Resources	2	6
Chapter 3. Processing of the Crude Text	2	6
Chapter 4. Processing the Structured Programs	2	6
Chapter 5. Categorizing and Marks of Words	4	12
Chapter 6. Studying Text Classification	4	12
Chapter 7. Extraction Information from Text	4	12
Chapter 8. Analyzing sentence structure	4	12
Chapter 9. Building Feature-based Grammar	3	9
Chapter 10. Analyzing the Meaning of Sentences	3	9
Total of hours	**30**	**90**

PRACTICE AND EXERCISES

1. Installation and interpreter Python adjustment – 2 hours
2. Start of programs on Python – 2 hours
3. Ways of start of the program – 2 hours
4. Installation and acquaintance with Natural language Tooling (NLTK) – 4 hours
5. Exercises on types of data and operations with components of computer linguistics – 4 hours
6. Exercises on giving, expression and a conclusion with components of computer linguistics – 3 hours
7. Exercises on trains and files with components of computer linguistics – 3 hours
8. Exercises under conditional instructions with components of computer linguistics – 4 hours
9. Exercises to cycles with components of computer linguistics – 3 hours
10. Exercises managing Linguistic Data – 3 hours

Total of hours – 30 hours

7. Assignments and Evaluation

5% - active attendance and participation – At the analysis of syntax of the Uzbek language
20% - Oral examination – Words and phrases of the Uzbek language
75% - Final work – Drawing up of various programs on the Python for the analysis of structure of the Uzbek offers

8. Requirements

Master students must bring their laptops to all sessions of the course, install a Windows distribution with the Python interpreter, including the NLTK (Natural language toolkit) module.

Master students will study a course of Python for processing of the Uzbek and English language. On studying of the given course magistrant

should have representations:

- about ways of processing of words;
- about ways of processing of phrases;
- about ways of static processing of offers;
- about ways of processing of the case;
- about possibilities WordNet;

should know and be able to use:

- processing of any text;
- to make the structured programs;
- functions and libraries of the python for natural language processing;
- to classify texts;
- extraction of the information from the text;
- to carry out the analysis of structure of the text;
- context-free grammar;

and also should have skills on:

- to the analysis of value of offers;
- to semantics of the Uzbek language offers;
- to semantics of informal conversation of the Uzbek language;
- to management of linguistic data;
- to application of meta languages for modeling of natural languages

Language Analysis

1. Basic Data

Subject Name	*Language Analysis*
Instruction level	Master, part of Master in Computational Linguistics Program
ECTS	5
Obligatory/Elective	Obligatory
Module	Applied Linguistics
Admission	a Bachelor Degree in some of the following subjects: - linguistics - computer science, - computational linguistics or language technology,
Prerequisites	English language competence B2 level, Linguistics or Computer Science bachelor degree
Semester	1
Teaching language	English
Teaching competence of the instructor	Knowledge of Linguistics including phonetics, phonology, morphology, syntax, discourse analysis, semantic; approaches to language analysis Skills of language analysis Ability to organize collaborative learning environment
Department / Faculty	Foreign Philology
University / Center	A.Baitursynov Kostanay State University, Kostanay, Kazakhstan; email: ksuintrel@gmail.com
Course description	The course explores the structure of a language; covers different approaches to language analysis; considers the elements of a language analysis; covers concepts and practices for analysis of English by computer with emphasis on the applications of computational analysis to problems in applied linguistics
Teaching activity type	5 hours of Lectures; 40 hours of practical classes including seminars; 90 hours of independent learning; 15 hours preparation for the exam 8 homework assignments; final exam

2. Objectives

1. Understand the purpose of a language analysis.
2. Explore the structure of the language (language family, type, etc.).
3. Understand the methodology of a language analysis.
4. Learn elements in language analysis (author, text type, audience, etc.).
5. Identify language analysis features (vocabulary, syntax, perspective, grammar, imagery).
6. Explore Analysis techniques.

3. Subject Content and Teaching plan

1. The notion of a language sign and language system.
2. A framework for analysis. Levels of language analysis.
3. Level of phonemes.
   3.1 The units of the text.
   3.2 Phonemes and symbols.
   3.3Pronunciation (contractions, question forms, tag questions).
   3.4 The written form and the spoken form.
4. Morphological level.
   4.1 Types of morphemes.
   4.2 Parts of speech.
5. Syntactical level.
   5.1 Syntactical relations.
   5.2 The notion of syntactical predicate.
   5.3 Types of sentences.
   5.4 Word order and structures that follow (transitive verbs, verb plus infinitive or gerund).
   5.5 Pragmatics of a sentence.
6. Semantics.
   6.1 Word forms.
   6.2 Function and meaning.
   6.3 Synonymy. Homonymy.
   6.4 Word combinations. Types of word combinations.
   6.5 Words and Concepts.
7. Applications of computational analysis to problems in applied linguistics

Teaching plan

Week	Topic	Lectures**	Readings*	Assignments***
1	The notion of a language sign and language system.	The notion of a language sign and language system.	https://language-teaching-learning.com/about/language-as-a-system-of-sign/ https://www.gelbukh.com/clbook/Computational-Linguistics.htm#_Toc86751630	http://www.fon.hum.uva.nl/praat/ 1) Describe the English/ Kazakh/ Uzbek/Russian Language as a system of signs 2) Study Praat tool and do a phonetic analysis of a given text
2	A framework for analysis. Levels of language analysis.	A framework for analysis. Levels of language analysis.	http://www.eltconcourse.com/training/courses/lacourse/language_analysis_course_index.html	
3	Level of phonemes.	Level of phonemes.	http://www.eltconcourse.com/training/courses/lacourse/whatisaphoneme_la_course.html	
4	Level of phonemes.			
5	Level of phonemes.			
6	Morphological level.	Morphological level of analysis	http://www.eltconcourse.com/training/courses/lacourse/morphemes_la_course.html https://www.cs.bham.ac.uk/~pjh/sem1a5/pt2/pt2_intro_morphology.html	3) Do a morphological analysis of the given 10 words per language: English/Kazakh/ Russian 4) Choose 5 long words and do their morphological analysis
7	Morphological level.			
8	Syntactical level.	Syntactical level of analysis	http://www.eltconcourse.com/training/courses/lacourse/phrase_structure_la_course.html http://www.eltconcourse.com/training/courses/lacourse/sentences_la_course.html https://www.slideshare.net/dr.shadiabanjar/syntactic-analysispptx	5) Do a syntactical analysis of the given 5 sentences per language: English/Kazakh/ Russian 6) Study the text and present its analysis at syntactical level
9	Syntactical level.			
10	Syntactical level.			
11	Semantics	Semantics	http://www.eltconcourse.com/training/courses/lacourse/word_class_la_course.html https://zetaglobal.com/blog-posts/understanding-semantic-analysis-title-totally-meta/ https://www.gelbukh.com/clbook/Computational-Linguistics.htm#_Toc86751665 https://www.slideshare.net/qiuyuel2/what-is-semantic-analysis	7) Do the semantic analysis of the texts in English/Kazakh/Russian
12	Semantics			
	Semantic analysis			

13	Language analysis tools	Language analysis tools and computational analysis to problems in applied linguistics	https://www.gelbukh.com/clbook/Computational-Linguistics.htm#_Toc86751649 https://ox.libguides.com/c.php?g=422982&p=2888571 https://libguides.lib.rochester.edu/LIN/tools	8) Analyze the given text, using minimum three tools, explain/compare the findings within each tool
14	Applications of computational analysis to problems in applied linguistics			
15	Complex Problem solving with language analysis tools			

*the recommendations for readings are to be edited

** The number of lectures is given in accordance to A.Baitursynov KSU demands

*** Obligatory assignment for each class is to participate in class discussion on the topic and demonstrate theoretical knowledge

4. Recommended Bibliography

The Oxford Handbook of Linguistic Analysis. Second edition.Edited by Bernd Heine, Heiko Narrog. Oxford University Press.2015.1180pp.

http://www.eltconcourse.com/training/courses/lacourse/language_analysis_course_index.html
online tools to analyze the texts:
http://www.fon.hum.uva.nl/praat/
https://www.online-utility.org/text/analyzer.jsp (Non-English language texts are supported)
https://ox.libguides.com/c.php?g=422982&p=2888571
https://libguides.lib.rochester.edu/LIN/tools

5. Course outcomes: Competences

Upon completion of this course, the student will be able to:

- analyze sentences, texts and make synthesis.
- apply existing tools for the processing of different languages (morphological, syntactic and semantic analysis)
- identify the most relevant symbolic methods for language technology research.
- investigate the design of language processing systems.
- use existing applications in the field of language technology.
- use massive linguistic resources for different languages.
- work in an interdisciplinary team.
- improve oral and written communication in English.

6. Methodology

5 hours of Lectures; 40 hours of practical classes including seminars; 90 hours of independent learning; 15 hours preparation for the exam

8 homework assignments; final exam

In the course study blended learning\teaching is applied: the course is studied within Moodle e-learning platform

7. Assignments and Evaluation

60% active attendance and participation, task performance (8 homework assignments, in-class activity of the students will be graded).
40% Oral examination.

8. Requirements

Learners should attend theoretical and practical classes, do home assignments, study additional materials given for self-study.

Plagiarism: When doing written assignments and producing presentations, please, document all of your source material. If any text is taken from somebody else, quotation must be identified with the reference to its source. Any sources from which you obtain numbers, ideas, or other material from must be cited. In case plagiarism is identified (7% and less of authenticity) the work of a student maybe disqualified; 70% of authenticity is acceptable.

Language Analysis

1. Basic Data

Subject Name	Language Analysis
ECTS	5
Obligatory/Elective	Obligatory
Module	Linguistic analysis
Semester	1
Teaching language	Uzbek
Professor	Nilufar Abdurakhmonova
Department / Faculty	Information and contemporary technology department in Uzbek language and literature faculty
University / Center	Tashkent State university of Uzbek language and literature

2. Objectives

1. To introduce students to the different types of language analysis
This course serves as an introduction to language analysis, presented from a functional typological perspective. Students will learn how languages around the world construct words and sentences. We will explore the similarities and differences in the ways languages do this, and we will consider how language is shaped by human cognition,

culture, and speakers' communicative goals. This class will focus on the hands-on analysis of language data. Students will gain linguistic and analytic problem-solving skills, an appreciation for the diversity of the world's languages, and the foundations necessary to understand the complexity of language and the problems involved in the computational processing of language.

2. To explain the different areas of linguistics

- Phonetics and phonology: The sounds of human languages, their acoustic and physiological properties, how they are classified, produced, and perceived, as well as
  how speech sounds are organized into systems of contrast.
- Morphology: The abstract rules/constraints governing the internal structure of words, how they are formed, their categories, and how they are related to other words in the speaker's 'mental lexicon'.
- Syntax: Principles governing how words are combined to form phrases and sentences.
- Semantics, pragmatics: The relationship between linguistic form and linguistic meaning/use. How words are interpreted, how the meanings of phrases and sentences are analysed on the basis of the meanings of their parts, and how speakers employ linguistic expressions to perform communicative tasks (making assertions, asking questions, issuing commands, etc.).

3. To provide a broad overview of the field, and to acquaint you with some of the research questions and debates with which linguists are currently engaged.

Throughout the course, we focus on doing analysis—developing sound argumentation and problem-solving skills, learning to identify and analyze data in order to construct productive, testable hypotheses about language. We hope to make you aware of the complexity and sophistication of your own (largely unconscious) linguistic knowledge, and in so doing, inspire you to question some of your own preconceptions about how language works.

3. Subject content

General introduction to language analysis

1. Theoretical and applied linguistics
   - 1.1. Language typology
   - 1.2. Theoretical approaches to language analysis
   - 1.3. Linguistic units in language acquisition

2. Phonetics & phonology
   - 2.1. Phonological analysis
   - 2.2. Experimental phonetics.
   - 2.3. Phonology and graphemes

3. Morphology – comparison of word structure and parts of speech in different languages

3.1. Lemmatization, tokenization and stemming in different languages (e.g. Uzbek, Russian and English)
3.2. Uzbek morphology and parts of speech

4. Syntax and parsing
   4.1. Syntactic structures and models used by different languages (e.g. Uzbek, Russian and English)
   4.2. Uzbek syntactic structures and parsing
   4.3. Syntactic Relation of constituencies.

5. Semantic analysis
   5.1. The Lexicon
   5.2. Experimental semantics

6. Language as communication
   6.1. Text linguistics
   6.2. Discourse and discourse analysis

4. Recommended Bibliography

1. The Oxford Handbook of Linguistic Analysis. Second edition. Edited by Bernd Heine, Heiko Narrog. Oxford University Press.2015.1180pp.

2. Abdurahmonova Nilufar Mashina tarjimasining lingvistik ta'minoti (monograph-Linguistic database of machine translation), 2018

3. Abdurahmonova Nilufar Mashina tarjimasining lingvistik asoslari (manual-linguistic foundation of machine translation)

And internet resources

5. Competences

Upon completion of this course, the student will be able to:

1. Apply linguistic theories to the problems of natural language processing
2. Understand the different levels of linguistic analysis
3. Use online and traditional paper language resources competently

6. Methodology

There will be theoretical and practical classes, but students are expected to study the bibliography and teaching materials provided autonomously and prepare work assigned by the teacher.

25 hours of theoretical instruction
25 hours of practical classes
40 hours of autonomous study
40 hours of preparation of assignments
15 hours of presentation and evaluation of assignments
5 hours of tests

7. Assignments and Evaluation

10% active attendance and participation
60% project work
30% test

There will be two types of assignments: (a) analytic assignments, which involve you working out the solution to a problem set independently, and (b), written assignments, which involve you describing a (pre-determined) analysis in written form. • Assignments are due (a) in class if in hard copy or (b) by a time specified online if electronic. • Late assignments receive a 10% penalty for each subsequent weekday, counting 24-hour periods. Late assignments are not accepted once the assignment has been handed back. • The midterm and final exams will include written short-answer and multiple choice components.

8. Requirements

Students must bring their laptops to all sessions of the course. Auditory should be provided with interactive whiteboard or projector and computer

Language Resources

1. Basic Data

Subject Name	Language Resources
ECTS	5
Obligatory/Elective	Obligatory
Module	NLP
Semester	1
Teaching language	English
Professors	Albina Dossanova
Department / Faculty	Philology and World Languages
University / Center	Al-Farabi Kazakh National University

2. Objective

The objective of the course is to enable students to improve both their in-depth understanding of the structure, purpose and functions of different language resources and their ability to apply existing language resources and to create new ones for natural language processing.

3. Subject Content

1. Introduction to language resources (LR).
   1.1 Basic concepts of LR. Overview of software tools for the preparation, collection, management and use of language resources. Grammar/ language models.
   1.2 Basic principles for LR infrastructure: documentation, development, and availability.
2. Corpus as a LR. Basic concepts of a corpus.
   2.1 Overview of national text corpora.
   2.2 Corpus management software. Corpus structures: files, paragraphs and sentences.
   2.3 Corpus architecture: components of data collection and storage, preprocessing, analysis and visualization.
   2.4 Corpora and NLP applications.

3. The lexicon as a LR. Introduction to lexicography.
   3.1 Monolingual, bilingual and multilingual dictionaries.
   3.2 Dictionary structure and content.
   3.3 Ethical and legal aspects of reusing of dictionary material.
   3.4 General language and Wordnet.
4. The database as a LR. Databases for written and spoken language data.
   4.1 Terminology database as a type of the LR database.
   4.2 Design of a terminology database.
   4.3 Implementation and use a terminology database.
5. Ontology as a LR. Basic concepts of an ontology.
   5.1 Architecture of an ontology. Types, properties, and relationship types.
   5.2 Integrated development environments for ontology building.
   5.3 Ontology-based NLP applications.

Laboratory work

1. A review of language resources in the domain area, identification of needs.
2. Development of the corpus structure.
3. Design of database for the corpus: a conceptual model.
4. Design of database for the corpus: a logical model.
5. Database development: creating tables.
6. Setting up the tool for collecting URLs from websites in a given domain area.
7. Setting up a tool for collecting news. Gathering news and comments. Storing data in the database.
8. Expansion of the initial dictionary of stop words with stop words from a given domain area. Data cleaning.
9. Application of the tokenization tool. Storing sentences in the database.
10. Application of the lemmatization tool. Storing lemmas in the database.
11. Application of the morphological analyzer. Storing results in the database.
12. Application of the indexing tool. Saving indexed data in the database.
13. Application of the frequency analyzer.
14. Application of the keyword extraction tool. Storing keywords in the database.
15. Creating a terminological dictionary. Storing terminological dictionary in the database.

4. Recommended Bibliography

Allan J. 1999. Understanding Natural Language. Cambridge (UK),

Baroni, Marco and Silvia Bernardini, 2006. (eds.) Wacky! Working papers on the Web as Corpus. GEDIT, Bologna.

Baroni, Marco and Silvia Bernardini. 2004. 'Bootcat: Bootstrapping corpora and terms from the web'. In Proceedings of LREC 2004, pages 1313–1316.

Bejoint, H. 2000. Modern Lexicography, Oxford University press.

Bolshakov I.A., Gelbukh A.F. Computational Linguistics: Models, Resources, Applications. Mexico, 2004.

Ciamarita, Massimiliano and Marco Baroni. 2006. Measuring web-corpus randomness: A progress report. In Baroni, Marco and Silvia Bernardini, 2006. (eds.) Wacky! Working papers on the Web as Corpus. GEDIT, Bologna.

Fellbaum, Christiane (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Jurafsky, D. & J.H. Martin. Speech and Language Processing (3rd ed. Draft version). Available at: https://web.stanford.edu/~jurafsky/slp3/

Sharoff, Serge. 2006. 'Creating general-purpose corpora using automated search engine queries'. In Baroni, Marco and Silvia Bernardini, 2006. (eds.) Wacky! Working papers on the Web as Corpus. GEDIT, Bologna.

Weisser, Martin. 2016. Practical corpus linguistics: an introduction to corpus-based language analysis. Wiley-Blackwell.

Websites

ELRA – European Language Resources Association - http://www.elra.info/en/

LDC – Linguistic Data Consortium - https://www.ldc.upenn.edu

Natural Language Processing – Articles on Natural Language Processing – Article on 'What is Corpus'.http://language.worldofcomputing.net/linguistics/introduction/what-is-corpus.html

Natural Language Processing for Hackers - https://nlpforhackers.io/corpora/ - last updated March 2019

Wordnet: An Electronic Lexical Database https://wordnet.princeton.edu

5. Competences

Upon completion of this course, the students will be able to:

1. Describe the theoretical and organizational methodological issues of construction and functioning of the subsystem of linguistic support.
2. Develop new tools for data collection, storing, and managing.
3. Reuse the existing and develop the new language resources.
4. Adapt the language resources for certain NLP applications.
5. Plan and carry out group practical and research projects to develop of language resource prototype.

6. Methodology

Course consists of:

30 hours of lectures (during 15 weeks)
30 hours of lab classes (during 15 weeks)
30 hours - preparation of individual assignments
30 hours – preparation of group assignments
30 hours – reading and research

Most of the theoretical information needed for the course will be explained during lectures. During lectures, students will be given reading and research assignments. They are expected to prepare them in time. During lab classes, we will cover practical aspects of course language resources. There will be assignments involving studying of documentation, working with modern development tools, as well as discussions.

7. Assignments and Evaluation

Participation in lecture activities – 21%
Lab assignments – 56%
Independent work assignments – 23%
TOTAL – 100%

There will be 2 intermediate controls (IC), midterm exam (MT) and final exam (FC). Each one will be graded with maximum 100%.
Final grade will be calculated according to following:

$$\text{Finalgrade} = \frac{\text{IC1} + \text{IC2}}{2} \cdot 0{,}6 + 0{,}1\text{MT} + 0{,}3\text{FC}$$

8. Requirements

1. For each class students have to prepare in advance, according to the class schedule. Assignments should be completed by the class, after the topic has been discussed.

2. Homework assignments will be given throughout the semester according to the class schedule.

3. Most homework assignments will include reading and a few questions that can be answered by simple research. For the research students might need to use relevant training resources.

When doing homework students have to keep in mind following rules:

- Homework must be carried out within a specified time.

- Students can work on homework together with other students, providing each one of them works on a specific problem or task.

Machine learning in Natural Language Processing (NLP)

1. Basic Data

Subject Name	Machine learning in Natural Language Processing (NLP)
ECTS	5
Obligatory/Elective	Obligatory
Module	Applied Linguistics
Semester	1
Teaching language	English/Uzbek
Professors	Gayrat Matlatipov, Khabibulla Madatov
Department / Faculty	Information Technology
University / Center	Urgench State University

2. Objectives

This course will introduce a number of concepts and techniques developed in the field of Machine Learning and review their applications to Natural Language Processing (NLP) applications, and, to a lesser extent, to issues in Computational Linguistics. Natural Language Processing is a sub-discipline of Artificial Intelligence which studies algorithms and methods for building systems (or, more commonly, components of larger systems) to deal with linguistic input. Sometimes a distinction is drawn between NLP and Computational Linguistics whereby the latter is viewed as the study of linguistic ability viewed as a computational process, whereas the former is viewed more from an engineering (application directed) perspective1 . Although the boundaries between these fields are not always clear, we will accept the distinction and focus on applications. In order to do so, we will introduce a number of tools developed in the related field of Machine Learning, and illustrate their use in NLP tasks ranging from text classification to document analysis and clustering. These tools and techniques will be illustrated through

a series of case studies designed to help you understand the main concepts behind Machine Learning while getting acquainted with modern NLP applications.

---

3. Subject *Content*

1. Introduction to Machine learning.
    - 1.1. Rationalist and Empiricist Approaches to language
    - 1.2. Scientific content
        - 1.2.1 Questions that linguistics should answer
        - 1.2.2 Non-categorical phenomena in language
        - 1.2.3 Language and cognition probabilistic phenomena
    - 1.3. The Ambiguity of Language: Why NLP Is Difficult
    - 1.4. Dirty Hands
        - 1.4.1 Lexical resources
        - 1.4.2 Word counts
        - 1.4.3 Zipf's laws
        - 1.4.4 Collocations
        - 1.4.5 Concordances
2. Mathematical Foundation
    - 2.1 Elementary Probability theory
    - 2.2 Essential Information theory
3. Corpus-Based Work
    - 3.1 Getting Set Up
        - 3.1.1 Computers
        - 3.1.2 Corpora
        - 3.1.3 Software
    - 3.2 Looking at Text
        - 3.2.1 Low-level formatting issues
        - 3.2.2 Tokenization: What is a word?
        - 3.2.3 Morphology
        - 3.2.4 Sentences
    - 3.3 Marked-up Data
    - 3.4 Frequency
    - 3.5 Mean and Variance
    - 3.6 Hypothesis Testing
4. Vector Semantics
    - 4.1 Lexical Semantics,Vector Semantics
    - 4.2 Words and Vectors, Cosine for measuring similarity
    - 4.3 TF-IDF: Weighing terms in the vector
    - 4.4 Word embeddings
5. Neural Networks and Neural Language Models
    - 5.1. Neural Networks, Deep Learning
    - 5.2 Neural Language Model
6. Word Sense Disambiguation

6.1 Methodological Preliminaries
6.2 Supervised Disambiguation
6.3 Dictionary-Based Disambiguation
6.4 Unsupervised Disambiguation

7. Lexical Acquisition
   7.1 Evaluation Measures
   7.2 Verb Subcatigoryzation
   7.3 Attachment ambiguity
   7.4 Semantic similarity

8. Markov model
   8.1 Markov Model
   8.2 Hidden Markov Model(HMM)
   8.3 The tree Fundamental Questions for HMMs
   8.4 HMMs: Implementations, Properties, and Variants

9. Part of speech Tagging
   9.1 The introduction Sources in tagging
   9.2 Markov model Taggers
   9.3 Hidden Markov model Taggers
   9.4 Transformation –Based Learning of Tags
   9.5 Tagging accuracy and Uses of Taggers

4. Recommended Bibliography

- Christopher D.Mnning and Hinrich Shcutze, Foundations of Statistical Natural Language Processing , 2nd Edition, 1999.
- Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition by Daniel Jurafsky and James Martin, Pearson Prentice Hall, 2nd Edition 2008
- Ian Goodfellow and Yoshua Bengio and Aaron Courville, Deep Learning, 2016
- Yoav Goldberg, Neural Network Methods in Natural Language Processing, 2017

5. Competences

On completion of this course, the student will be able to:

1. Understanding the basic consept of Machine Learning and current ML techniques available for NLP tasks;
2. Proper knowledge about Corpora and ways to build it as well as NLP tasks they are used for;
3. Learn what is vector semantics and weighing terms in vectors, how to create word-embeddings, also students will learn popular representations (GloVe, FastText, ElMO) and different types of embeddings(word, character, sentence);
4. Basic knowledge about neural Networks and their usage in NLP, Deep Learning methods Language Modeling tasks;

5. Students should be able to use and explain appropriate state-of-the-art symbolic parsing techniques, and, where a labelled corpus is available, statistical parsing techniques (generative and discriminative)
6. Given an NLU system, students should be able to choose appropriate evaluation metrics for the system, use error analysis to propose improvements, and relate it to features of human models of language interpretation at various levels of processing

6. Methodology

36 hours of theoretical instruction, 36 hours of exercises and practical courses, 48 hours independent work.

7. Assignments and Evaluation

5% active attendance and participation.
20% Oral examination.
75% Final work.

8. Requirements

Students must bring their laptops to all sessions of the course;

Basic knowledge of any programming language (Python is highly prioritized).

Machine Translation Technologies

1. Basic Data

Subject Name	Machine Translation Technologies
ECTS	5
Obligatory/Elective	Obligatory
Module	NLP module including applications
Semester	3
Teaching language	English
Professors	Zhandos Zhumanov, Ualsher Tukeyev
Department / Faculty	Information Systems Department, Faculty of Information Technologies
University / Center	Al-Farabi Kazakh National University

2. Objectives

1. To study approaches to solving the problem of machine translation.
2. To understand advantages and disadvantages of different approaches to machine translation.
3. To learn several machine translation tools that use different approaches.

3. Subject Content

1. Overview of machine translation problems
2. Rule-based machine translation
    - 2.1 Basic concepts of RBMT
    - 2.2 RBMT models
    - 2.3 Overview of resources needed for RBMT
    - 2.4 Apertium machine translation platform
3. Statistical machine translation
    - 3.1 Basic concepts of SMT
    - 3.2 SMT models
    - 3.3 Overview of resources needed for SMT
    - 3.4 Moses statistical machine translation system
4. Neural machine translation
    - 4.1 Basic concepts of NMT
    - 4.2 NMT models
    - 4.3 Overview of resources needed for NMT
    - 4.4 TensorFlow Neural Machine Translation Tutorial

5. Hybrid machine translation. Applications of Machine Translation.

4. Recommended Bibliography

Apertium 2.0: Official documentation. Latest update November 2018. Available at: http://wiki.apertium.org/wiki/Documentation.

Jurafsky, D. & J.H. Martin. (2008) Speech and Language Processing, 2nd Edition. Chapter 25. Prentice Hall.

Jurafsky, D. & J.H. Martin. Speech and Language Processing (3rd ed. Draft version). Available at: https://web.stanford.edu/~jurafsky/slp3/

Koehn, P. (2010) Statistical Machine Translation, 1st Edition. Cambridge University Press.

Koehn, P. (2019) Moses – Statistical Machine Translation System - User Manual and Code Guide.. Available at http://www.statmt.org/moses/manual/manual.pdf

Lamb, S.L. (1962/2003) 'On the Mechanization of syntactic analysis'. In Nirenburg, S. Somers, H. & Wilks, Y. Eds. (2003).

Nirenburg, S. Somers, H. & Wilks, Y. Eds. (2003) Readings in machine translation. MIT Press.

TensorFlow tutorials – Available at: https://www.tensorflow.org/tutorials/

Wilks, Y. (2009). Machine Translation - Its Scope and Limits. Springer.

Websites of interest

Andy Way's webpage - https://www.computing.dcu.ie/~away/ latest update August 2019

John Hutchins - Publications on Machine Translation, Computer-based Translation Technologies, linguistics and other topics. http://www.hutchinsweb.me.uk. Latest update May 2014.

Mikel L. Forcado's webpage - https://www.dlsi.ua.es//~mlf/

5. Competences

Upon completion of this course, the student will be able to:

1. Understand and solve different challenges that may occur in connection to machine translation tasks.
2. Understand various machine translation approaches.
3. Understand positive and negative aspects of various machine translation approaches.
4. Know how to apply machine translation tools to computational linguistics tasks.
5. Apply knowledge and skills in machine translation to real life problems.

6. Methodology

Course consists of:

15 hours of lectures (during 15 weeks)

30 hours of lab classes (during 15 weeks)

65 hours homework and assignments

40 hours autonomous study and research

Most of the theoretical information needed for the course will be explained during lectures. During lectures, students will be given reading and research assignments. They are expected to prepare them in time. During lab classes, we will cover practical aspects of machine translation technologies. There will be assignments involving studying of documentation, working with modern development tools, as well as discussions.

7. Assignments and Evaluation

Participation in lecture activities – 21%
Lab assignments – 56%
Independent work assignments – 23%
TOTAL – 100%

There will be 3 intermediate controls (IC) and final exam (FC). Each one will be graded with maximum 100%.

Final grade will be calculated according to following:

$$\text{Finalgrade} = \frac{\text{IC1} + \text{IC2} + \text{IC3}}{3} \cdot 0{,}6 + 0{,}4\text{FC}$$

8. Requirements

1. For each class students have to prepare in advance, according to the class schedule. Assignments should be completed by the class after the topic has been discussed.

2. Homework assignments will be given throughout the semester according to the class schedule.

3. Most homework assignments will include reading and a few questions that can be answered by simple research. For the research, students need to use relevant training resources.

4. Students should bring their laptops for every class.

When doing homework students have to keep in mind the following rules:

- Homework must be carried out within a specified time.

- Students can work on homework together with other students, providing each of them work on a specific problem or task.

Natural Language Understanding

1. Basic Data

Subject Name	Natural Language Understanding
ECTS	5
Obligatory/Elective	Obligatory
Module	Applied Linguistics
Semester	1
Teaching language	English/Uzbek
Professors	Gayrat Matlatipov
Department / Faculty	Information Technology
University / Center	Urgench State University

2. Objectives

The objective of the course is to learn the basic concepts in the statistical processing of natural languages.

This course presents an introduction to general topics and techniques used in natural language processing today, primarily focusing on statistical approaches. The course provides an overview of the primary areas of research in language processing as well as a detailed exploration of the models and techniques used both in research and in commercial natural language systems.

3. Subject Content

1. Course overview
    a. Goals of NLU
    b. Course Set-up, Jupyter Notebook, NumPy, PyTorch

2. Distributed word representations
    a. High-level goals and guiding hypotheses
    b. Matrix designs
    c. Vector comparison
    d. Basic reweighting
    e. Subword information
    f. Visualization

g. Dimensionality reduction
h. Retrofitting

3. Supervised sentiment analysis
    a. Sentiment as a deep and important NLU problem
    b. General practical tips for sentiment analysis
    c. The Stanford Sentiment Treebank (SST)
    d. Methods: hyperparameters and classifier comparison
    e. Feature representation
    f. RNN classifiers
    g. Tree-structured networks

4. Relation extraction with distant supervision
    a. Task definition
    b. Experiments

5. Natural language inference
    a. Overview of NLI
    b. SNLI and MultiNLI
    c. Hand-built features and experiments
    d. Sentence-encoding models
    e. Chained models
    f. Attention (Global, Word-by word)
    g. Error analysis

6. Grounded language understanding

7. Evaluation metrics
    a. Evaluation metrics
    b. Evaluation methods
    c. Evaluating NLU models with harder generalization tasks

8. Contextual word representations
    a. Representing long texts for NLU

---

4. Recommended Bibliography

- D. Manning, Christopher. (2015). Computational Linguistics and Deep Learning. Computational Linguistics. 41. 699-705. 10.1162/COLI_a_00239.
- Alessandro Lenci. Distributional Models of Word Meaning. 2018. Annual Review of Linguistics. P 151-171. V 4.

- Smith, Noah A.. "Contextual Word Representations: A Contextual Introduction." ArXivabs/1902.06006 (2019): n. pag. https://arxiv.org/pdf/1902.06006.pdf
- Percy Liang,Christopher Potts, Bringing Machine Learning and Compositional Semantics Together, 2015, Annual Review of Linguistics, P 355-376, V 1, 10.1146/annurev-linguist-030514-125312
- Hector J. Levesque. On our best behaviour. P.27-35. 2014. V.212. Artifitial Intelligence. https://doi.org/10.1016/j.artint.2014.03.007
- Potts, Christopher. "A case for deep learning in semantics: Response to Pater." Language, 2019. Project MUSE, doi:10.1353/lan.2019.0003
- Karlgren, J., & Sahlgren, M. (2001). From words to understanding. In Uesaka, Y., Kanerva, P., & Asoh, H. (Eds.), Foundations of Real-World Intelligence, pp. 294–308. CSLI Publications.
- Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th International Conference on Machine Learning (ICML-08), pp. 160–167
- James Allen, Natural Language Understanding, 2nd Edition, 1999.
- Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition by Daniel Jurafsky and James Martin, Pearson Prentice Hall, 2nd Edition 2008
- Macherey, Klaus & Josef Och, Franz & Ney, Hermann. (2001). Natural Language Understanding Using Statistical Machine Translation.
- Kim, Yoon, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science, pages 61–65, Baltimore, MD.

5. Competences

On completion of this course, the student will be able to:

1. Students should be able to use and explain appropriate state-of-the-art symbolic parsing techniques, and, where a labelled corpus is available, statistical parsing techniques (generative and discriminative)
2. Given an NLU system, students should be able to choose appropriate evaluation metrics for the system, use error analysis to propose improvements, and relate it to features of human models of language interpretation at various levels of processing
3. Given an example of a problem in coreference resolution, discourse segmentation, and discourse parsing, students should be able to provide a written description of how current symbolic and statistical techniques help solve the problem
4. Given a model and a labelled corpus, students should be able to employ existing ML software packages to train the model on the corpus in order to perform a lexical semantic task

5. Given an open-ended problem of choosing informative features for a particular NLP task and a description of the available training resources, the student should be able to give a well-justified, written and/or practical, selection of such informative features

6. Methodology

36 hours of theoretical instruction, 36 hours of exercises and practical courses, 48 hours independent work.

7. Assignments and Evaluation

5% active attendance and participation.
20% Oral examination.
75% Final work.

8. Requirements

Students must bring their laptops to all sessions of the course; programming skills at least at the level of Computer Programming for Speech and Language Processing are also required.

Ontology design tools

1. Basic Data

Subject Name	Ontology design tools
ECTS	3
Obligatory/Elective	Elective
Module	Applied Linguistics
Semester	3
Teaching language	Kazakh / Russian
Professors	Razakhova B., Niyazova R.
Department / Faculty	Computer Science and Information Security / Information Technology
University / Center	Eurasian National University

2. Objectives

1. Formation of practical skills in the design and application of ontologies in the processing of natural languages
2. Investigate and use ontological domain model for the development of components of intelligent systems

---

3. Subject Content

1. Technology of development of domain ontology
    - 1.1 Methodology for constructing a product ontology.
    - 1.2 Designing an ontology based on conceptual domain model.
2. Main editors of ontology.
    - 2.1 Editors ontology.
    - 2.2 Main editors of ontology.
    - 2.3 Editor Protégé.
3. Description logic
    - 3.1 General information.
    - 3.2 Syntax DL.
    - 3.3 The syntax of the ALC logic.
    - 3.4 DL semantics and ALC logic.
    - 3.4 Relationship with predicate logic.
    - 3.5 Knowledge base.
4. Algorithm for ALC logic
    - 4.1 Logical analysis.
    - 4.2 Properties DL.
    - 4.3 Solvability of ALC logic.
    - 4.4 The concept of the resolving algorithm.
    - 4.5 ALC for logic without terminology and with terminology
5. Practical application of ontologies.
    - 5.1 Semantic web.
    - 5.2 Development and management of terminology.
    - 5.3 Conceptual modeling.
    - 5.4 Knowledge management systems.
    - 5.5 Integration of heterogeneous data sources.
    - 5.6 Content specification of heterogeneous data sources.
    - 5.7 Information search.
    - 5.8 Semantic search.
    - 5.9 Ontologies in e-commerce.

4. Recommended Bibliography

7. Цуканова Н. И. Онтологическая модель представления и организации знаний. Учебное пособие для вузов. – М.: Горячая линия – Телеком, 2015. – 272 с.: ил.
8. Соловьев В.Д., Добров Б.В. и др. Онтологии и тезаурусы, учебное пособие, Москва, -2006
9. A Practical Guide To Building OWL Ontologies Using Prot'eg'e 4 and CO-ODE Tools Edition 1.3.: – The University Of Manchester Copyrightc The University Of Manchester. –2011.
10. Matthew Horridge et al. 2011. A Practical Guide To Building OWL Ontologies Using Protégé 4 and CO-ODE Tools, Edition 1.3, published by the University of Manchester, 24 Mar. - 2011, 108 pp.
11. SPARQL Query Language for RDF: http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/.
12. Description logic (lectures) of Evgeny Zolin: http://lpcs.math.msu.su/~zolin/dl/

5. Competences

Upon completion of this course, the student will be able to:

1. Apply independently develop ontologies for a certain subject area, solve problems using ontology-based systems
2. Know basic concepts of the field of knowledge representation, examples of best practices in the development of systems based on ontologies, a set of design tools and ontologies presentation...
3. Use solve the problems of automatic text processing, intelligent search

6. Methodology

30 hours of theoretical instruction, 30 hours of practical courses, 120 hours of independent work

7. Assignments and Evaluation

Current evaluation: during the classes

Interim: According to the performance of students during 1-7 and 8-15 weeks.

Final exam: testing

Forms of control

Current evaluation - 15%

SSL - 15%

Interim control:

Colloquium - 10%

Testing - 10%

Course work – 10%

Current and interim control at least 60%

Final control of at least 40 %

The policy of subject

Requirements of discipline: mandatory attendance of classes, active participation in discussions of preliminary preparation for lectures and seminars on teaching aids and basic literature , quality and timely fulfillment of the SSL, participation in all kinds of control (current control, SSL control, final test).

8. Requirements

Students must bring their laptops to all sessions of the course, install Protégé

Ontology and semantic technology

1. Basic Data

Subject Name	Ontology and semantic technology
ECTS	5
Obligatory/Elective	Obligatory
Module	Applied Linguistics
Semester	2
Teaching language	Uzbek/ Kazakh / Russian
Professors	Abdurakhmonova N., Razakhova B., Niyazova R.
Department / Faculty	Information technology department in Uzbek language and literature in TSUULL Computer Science and Information Security / Information Technology in ENU

2. Objectives

1. Formation of basic knowledge about ontologies
2. Formation of practical skills in the design and application of ontologies in the processing of natural languages
3. Investigate and use ontological domain model for the development of components of intelligent systems

3. Subject Content

I. Introduction
1.1. Definitions of ontology.
1.2. Types of ontologies.
1.3. Two main approaches to the construction of ontologies.
1.4. The principle of ontology independence from natural language.
1.5. Linguistic ontology. Ontologies and automatic text processing.

II. Definition of basic concepts.
2.1. Semiotic systems.
2.2. The general scheme of an intelligent system that understands the text.
2.3. Semantic roles of Fillmore, Schenk's semantic scenarios, Relational situational models.
III. Models of knowledge representation
3.1. Data and knowledge.
3.2. Production models.
3.3. Semantic networks
3.4. Thesauri: WordNet.
3.5. Frames: FrameNet. PropBank.
3.6. Formal logical models.
3.7. The acquisition of knowledge in the processing of text. Use of knowledge

IV. Languages of knowledge representation and request.
4.1. SPARQL query language.
4.2. OWL - basic concepts.
4.3. Constructions language OWL.
4.4. Ontology management

V. Semantic models for Turkic languages
5.1. Toolkit for filling the thesaurus
5.2. Multilingual ontological grammar for Turkic languages.
5.3. Multi-user multilingual multi-functional online tools for filling the structural-parametric model of the Turkic morpheme.

4. Recommended Bibliography

1. Цуканова Н. И. Онтологическая модель представления и организации знаний. Учебное пособие для вузов. – М.: Горячая линия – Телеком, 2015. – 272 с.: ил.
2. Соловьев В.Д., Добров Б.В. и др. Онтологии и тезаурусы, учебное пособие, Москва, -2006
3. A Practical Guide To Building OWL Ontologies Using Prot'eg'e 4 and CO-ODE Tools Edition 1.3.: – The University Of Manchester Copyrightc The University Of Manchester. –2011.
4. Matthew Horridge et al. 2011. A Practical Guide To Building OWL Ontologies Using Protégé 4 and CO-ODE Tools, Edition 1.3, published by the University of Manchester, 24 Mar. - 2011, 108 pp.
5. SPARQL Query Language for RDF: http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/.
6. Description logic (lectures) of Evgeny Zolin: http://lpcs.math.msu.su/~zolin/dl/

5. Competences

Upon completion of this course, the student will be able to:

1. Know semantic technology problems.
2. Understand different approaches to solving the problem of semantic analysis.
3. Use different computational tools for semantic analysis.
5. Apply understanding of problems and different approaches to real life problems.

Tools

1. The Protégé Ontology Editor (version 4.1 beta) can be downloaded from http://protege.stanford.edu/
2. SWOOP is an open-source editor with built-in access to the Pellet reasoner
3. Jena, a Java framework for RDF and OWL. Includes RDF and OWL APIs, and the ability to read/write RDF/XML into these APIs

6. Methodology

30 hours of theoretical instruction, 30 hours of practical courses, 120 hours of independent work

7. Assignments and Evaluation

10% attendance and active participation, current tasks performance;

40% midterm control (results of tests);
50% final examination (the research and final written work).

8. Requirements

- Students should attend theoretical and practical classes, do home assignments, study additional materials given for independent work.
- Students will be asked to demonstrate their understanding of the content on a weekly basis through discussions and assignments.
- Students are expected to post questions to the Class MOODLE Site to get help with the material.

Speech Processing

1. Basic Data

Subject Name	Speech Processing
ECTS	5
Obligatory/Elective	Obligatory
Module	NLP module
Semester	3
Teaching language	Kazakh, English, Russian
Professors	Gulmira Bekmanova
Department / Faculty	Computer Science and Information Security
University / Center	L.N. Gumilyov ENU

2. Objectives

1. Formation of basic knowledge about speech technologies
2. Formation of practical skills in the design and application of speech processing tools

3. Subject Content

1. Introduction
2. Speech Fundamentals:
2.1 Articulatory Phonetics
2.2 Production and Classification of Speech Sounds; Acoustic Phonetics
2.3 Acoustics of Speech Production;
3. Digital Signal Processing Elements
3.1 Discrete signals (sequences)
3.2 Physical feasibility, sustainability
3.3 Representation of discrete signals and systems in the frequency domain
3.4 Z-Transformation
3.5 Inverse Z-Transformation
3.6 Non-recursive filters
4. Speech signal processing
4.1. "H – L" - processing a numeric array. Signal smoothing
4.2. Pre-recording speech signal
4.3. Quasiperiod Sequence Calculation
4.4 Check recorded speech
4.5 Clarification of the boundaries of the end of the speech
4.6 Time Alignment and Normalization
4.7 Dynamic Time Warping
5 Speech Recognition
5.1 Large Vocabulary Continuous Speech Recognition: Architecture Of A
5.2 Large Vocabulary Continuous Speech Recognition System.
5.3 Algorithms for the recognition of separate words.
5.4 Algorithms of continuous speech recognition.
5.5 Algorithms for ponephone recognition.
5.6 Recognition using hidden Markov models.
5.7 Recognition using neural networks.
5.8 Diphon recognition.
6 Speech Synthesis
6.1Text-To-Speech Synthesis: Concatenative and Waveform Synthesis Methods,
6.2 Diphone synthesis

4. Recommended Bibliography

Shelepov V. Yu. ''Speech recognition lectures'', Nauka Svita, 2012

Bekmanova G. "Some methods of Kazakh language computer processing", Master PO, Astana 2015

Armstrong, Susan, *Natural language processing using very large corpora,* Dordrecht : Kluwer Academic Publishers, 1999.

Lawrence Rabiner And Biing-Hwang Juang, "Fundamentals Of Speech Recognition", Pearson Education, 2003.

Daniel Jurafsky And James H Martin, “Speech And Language Processing – An Introduction To Natural Language Processing, Computational Linguistics, And Speech Recognition”, Pearson Education, 2002.
Frederick Jelinek, “Statistical Methods Of Speech Reco gnition”, MIT Press, 1997.

5. Competences

Upon completion of this course, the student will be able to:
1. Express the speech signal in terms of its time domain and frequency domain representations and the different ways in which it can be modelled;
2. Derive expressions for simple features used in speech classification applications;
3. Explain the operation of example algorithms covered in lectures;
4. Synthesis block diagrams for speech applications, explain the purpose of the various blocks, and describe in detail algorithms that could be used to implement them;
5. Implement components of speech processing systems.

6. Methodology

30 hours of theoretical instruction, 30 hours of practical courses, 120 hours of independent work

7. Assignments and Evaluation

5% active attendance and participation.
20% Oral examination.
75% Final examination (the research and final written work)

8. Requirements

Students must bring their laptops to all sessions of the course, Mat Lab and using programming environment

Types of control (current, mid-term)

Current evaluation: during the classes

Interim: According to the performance of students during 1-7 and 8-15 weeks.

Final exam: testing

Forms of control

Current evaluation - 15%

SSL - 15%

Interim control:

Colloquium - 10%

Testing - 10%

Course work – 10%

Current and interim control at least 60%

Final control of at least 40 %

The policy of subject

Requirements of discipline: mandatory attendance of classes, active participation in discussions of preliminary preparation for lectures and seminars on teaching aids and basic literature, quality and timely fulfillment of the SSL, participation in all kinds of control (current control, SSL control, final test).

Statistical methods in Natural Language Processing

1. Basic Data

Subject Name	Statistical methods in Natural Language Processing
ECTS	5
Obligatory/Elective	Obligatory
Module	Statistical methods in Natural Language Processing
Semester	1
Teaching language	English
Professors	Кудубаева С.А.,Туребаева Р.Д., Сауханова Ж.С.
Department / Faculty	
University / Center	

2. Objectives

The goal of mastering the discipline “Statistical methods in linguistics” is to develop in undergraduates theoretical knowledge of mathematical and statistical methods, practical skills of applying a theoretical basis for calculating and interpreting data from a linguistic experiment carried out at any language level, as well as the ability to form scientific conclusions based on the results.

The program of study of the discipline sets as its task:

- to study in detail the methods of collecting, organizing and processing statistical linguistic data to identify existing patterns;
- to study in detail random variables according to the results of observation of

experiments at various language levels;
- to know the mathematical methods for natural language experiments and to be able to describe the statistical data obtained from the point of view of a scientific approach;
- to form a body of knowledge about the statistical properties of natural language;

3. Subject Content

1. Introduction.

    1.1. Quantitative analysis of texts.

    1.2. Statistical approach to the research of language structures.

2.Organization of the samples for statistical analysis.

    2.1. The rules of the organization of the sample.

    2.2. Organization of mechanical sampling.

    2.3. Organization of random sampling.

    2.4. Determination of the required sample volume.

    2.5. Statistical characteristics of the sample.

3. Comparison of statistical characteristics of different samples.

    3.1. Comparison of samples on the average frequency oscillation band.

    3.2. Test samples for statistical homogeneity.

    3.3. t-distribution criterion.

    3.4. Determining the materiality of differences of percentages.

    3.5. The pooling of samples.

    3.6. Comparison of variational series.

4. Establishing dependencies between different phenomena in the sample.

    4.1. Conjugate features.

    4.2. Rank correlation.

    4.3. The calculation of the correlation coefficient.

5. Nonparametric criteria of difference.

    5.1. Method of expert assessments.

    5.2. Similarity factor.

5.3. Criterion signs.

5.4. Serial criterion.

5.5. Wilcoxon test. .

6. Primary processing of experimental results.

6.1. Statistical observations and their results.

6.2. Grouping and tabulating data.

6.3. Graphic representation of data.

6.4. Generalizing the numerical characteristics of a series of distribution.

6.5. Variation and characteristics (degree of dispersion) of a random variable.

7. Errors made when testing statistical hypotheses.

7.1. The level of significance of statistical hypotheses.

7.2. The concept of degrees of freedom.

7.3. Unbiased estimates.

8. Similarity measures.

8.1. Similarity in mathematical terms.

8.2. Vector-space models.

8.3. Latent Semantic Analysis.

8.4. Bayesian models.

9. Statistics and natural language.

9.1. Markov models.

9.2. Zipf-Mandelbrot Law and Zipf's law.

9.3. Psycholinguistic issues.

10. Applications to Natural Language Processing.

10.1. Part-of-speech tagging.

10.2. Word sense disambiguation.

10.3. Text mining.

10.4. Statistical Machine Translation.

4. Recommended Bibliography

Перебенос В.И., 2016. Статистические методы в лингвистике / под ред. М.Р. Кауль. – М.: РГГУ.

Зубов А.В., Зубова И.И., 2004. Информационные технологии в лингвистике: учеб. Пособие для студентов линг. фак. вузов. – М.: Изд. центр «Академия»

Леонтьева Н.Н., 2006. Автоматическое понимание текстов: системы, модели, ресурсы: учеб. Пособие для студентов линг. фак. вузов. – М.: Изд. центр «Академия».

Соснина Е.П., 2012. Введение в прикладную лингвистику: учебное пособие / Е. П. Соснина. – 2-е изд., исп. и доп. – Ульяновск: УлГТУ.

Баранов, А.Н., 2001. Введение в прикладную лингвистику: учеб. пособие / А.Н. Баранов. - М. : Эди-ториал УРСС

Степанов Ю.С. Методы и принципы современной лингвистики. 7-е изд. – М. : ЛИБРОКОМ, 2009 (М. : ООО "ЛЕНАНД"). – 310 с.

Холлендер М., Вулф Д.А., 1983. Непараметрические методы статистики. М..

Шаховский В.И., Шейгал Е.И., 2008. Методика лингвистических исследований. – Волгоград.

Гинзбург А.И., 2003. Статистика. - СПб: Питер

Oakes M., 1998. Statistics for Corpus Linguistics. Edinburgh University Press

C. Manning & H. Schütze, 1999. Foundations of Statistical Natural Language Processing. MIT Press.

F. Jelinek, 1998. Statistical Methods for Speech Recognition. MIT Press.

D.Jurafsky and J.H.Martin, 2009. Speech and Language Processing (2nd edition). Prentice Hall.

Philipp Koehn, 2012. Statistical Machine Translation. Cambridge University Press.

5. Competences

Upon completion of this course, the student will

know:

- the scope of statistical laws in language and speech;

- the technique of statistical calculations in relation to the phenomena of language and speech;

- the basic mathematical-statistical concepts, methods and techniques for processing linguistic information;

be able to:

- correctly organize the selection of language material for statistical analysis;

- calculate the main statistical characteristics of the sample;

- compare the statistical characteristics of different samples;

- establish the relationship between the various phenomena in the sample;

- use statistical methods in their research;

- analyze different types of texts in their historical and theoretical aspects;

own:

- methods and procedures adopted in statistical studies, in the analysis of a large amount of language material;
- elements of mathematical statistical processing of linguistic information

6. Methodology

15 hours of theoretical instruction, 30 hours of excercises and practical courses, homework,...

7. Assignments and Evaluation

Current evaluation: during the classes

Interim: According to the performance of students during 1-7 and 8-15 weeks.

Final exam: testing

Forms of control

Current evaluation - 15%

SSL - 15%

Interim control:

Colloquium - 10%

Testing - 10%

Course work – 10%

Current and interim control at least 60%

Final control of at least 40 %

The policy of subject

Requirements of discipline: mandatory attendance of classes, active participation in discussions of preliminary preparation for lectures and seminars on teaching aids and basic literature , quality and timely fulfillment of the SSL, participation in all kinds of control (current control, SSL control, final test).

8. Requirements

- Students are expected to have background knowledge that could be found in standard textbooks on calculus and probability
- Students must bring their laptops to all sessions of the course, install the Python interpreter and Matlab