Module/Course Title: Corpus Processing and Applications

- **Code number:** Y-5
- **Level of Module/Course (under-/postgraduate):** postgraduate
- **Type of Module/Course:** compulsory
- **Year of Study** 1
- **Semester** 2
- **Number of ects allocated:** 6
- **Number of teaching units:**
- **Name of lecturer / lecturers :** Katerina T. Frantzi
- **Content outline:**
  A Corpus is an organized, statistically correct and constructed using specific criteria "collection" of real language material in electronic form, representative of a language, dialect, language for specific purposes, language variety, sublanguage. Corpus-based studies deal with phenomena of real language use. Corpora play a very important role in various linguistics areas research such as phonology, morphology, syntax, semantics, pragmatics, discourse analysis, lexicography, terminology, text linguistics, stylistics, sociolinguistics, historical linguistics, applied linguistics, dialectology, forensic linguistics and more. Corpus Processing gives precision, completeness and speed to quantitative as well as qualitative analysis of the language in a way that could not be achieved using traditional techniques. Nowadays, besides Linguistics, applications of corpus processing include a great variety of disciplines that involve the study of language: Education – First and 2nd/Foreign Language Learning and Teaching, Philology, Political Sciences, Media Studies, Psychology, Sociology and more.
  During the course, students are introduced to:
  Corpora (definition, categorizations, annotation, frequency normalization, construction)
  Regular Expressions
  Finite State Automata
  Corpus Processing in the internet
  Corpus Processing in Linux environment
  Applications of corpus processing to linguistics as well as other scientific disciplines that involve language study.

- **Learning outcomes:**
  After the successful complete of the course, students should:

  - Know what a Corpus is, the categorizations of corpora, what corpus annotation is, what frequency normalization is, about corpus construction.
  - Know what Regular Expressions are and their use in natural language processing.
  - Know what Finite State Automata are and their use in natural language processing.
  - Know about Corpus Processing in the internet.

- Know about the applications of Corpus Processing to various topics.

The course Corpus Processing gives students with a theoretical background the ability to deal with issues and problems in a formal, algorithmic way, an ability acquired with the study of mathematics/computing subjects.

- **Prerequisites: –**

- **Recommended Reading:**

a) Basic Textbooks:

Γούτσος, Διονύσης και Γεωργία Φραγκάκη (2015) Εισαγωγή στη Γλωσσολογία Σωμάτων Κειμένων. Ελληνικά Ακαδημαϊκά Ηλεκτρονικά Συγγράμματα και Βοηθήματα - Αποθετήριο "Κάλλιπος".

Μικρός, Γεώργιος (2009) Ποσοτική Ανάλυση της Κοινωνιογλωσσολογικής Ποικιλίας – Θεωρητικές και Μεθοδολογικές Προσεγγίσεις. Αθήνα : Μεταίχμιο.

Φραντζή, Κατερίνα Θ. (2012) Εισαγωγή στην επεξεργασία σωμάτων κειμένων. Αθήνα: Ίων.

b) Additional References*:*

Gavrilidou, M., Karagiannis, G., Markantonatou, S. Piperidis, S. & G. Stainhaouer (2000) Proceedings of the Second International Conference on Language Resources and Evaluation, Athens, Greece.

Καπιδάκης, Σαράντος (2008) Χειρισμός πνευματικών δικαιωμάτων σε ψηφιακούς πόρους. Πρακτικά Συνεδρίου «Αρχεία, βιβλιοθήκες και δίκαιο στην κοινωνία της πληροφορίας», Αθήνα 2-3 Φεβρουαρίου 2006. Εθνική Βιβλιοθήκη της Ελλάδος, 301-307.

Κατσογιάννου, Μαριάννα & Ελένη Ευθυμίου (2004) Ελληνική Ορολογία: Έρευνα και εφαρμογές. Αθήνα: Καστανιώτης.

Kordoni, Valia, Carlos Ramisch & Aline Villavicencio (2013) Proceedings of the 9th Workshop on Multiword Expressions, NAACL HLT 2013, Atlanta, Georgia, USA, June, 2013.

Kyriacopoulou, Panagiota (2005) Analyse automatique des textes écrits: le cas du grec moderne. Θεσσαλονίκη: University Studio Press.-

Markantonatou, S., Sofianopoulos, S., Spilioti, V., Tambouratzis, G., Vasileiou, M., Yannoutsou, O. (2005) Monolingual Corpus-based MT using Chunks. In Proceedings of the Workshop on EBMT, held within the MT Summit X, Phuket, Thailand, 91-98.

Μαρκόπουλος, Γεώργιος (2006) Ζητήματα Υπολογιστικής Γλωσσολογίας: Prolog και μορφολογική ανάλυση. Αθήνα: περιοδικό Παρουσία, παράρτημα 69.

Σαριδάκης, Ιωάννης Ε. (2010) Σώματα κειμένων και μετάφραση: θεωρία και εφαρμογές, Γλώσσα και πολιτισμός. Αθήνα: Παπαζήση.

Σταματάτος, Ευστάθιος (2000) Στατιστική Αναγνώριση Είδους Κειμένου και Συγγραφέα σε Νεοελληνικά Κείμενα χωρίς Περιορισμούς, Διδακτορική Διατριβή, Τμήμα Ηλεκτρολόγων Μηχανικών
και Τεχνολογίας Υπολογιστών, Πανεπιστήμιο Πατρών.

Σταύρου, Μελίτα & Μαρία Tzevelekou (2000) Η μηχανική μετάφραση και η ελληνική γλώσσα. Αθήνα: Καστανιώτη.

- **Learning Activities and Teaching Methods:** interactive lectures

- **Assessment/Grading Methods:** assignment

- **Language of Instruction:** Greek
- **Mode of delivery (face-to-face, distance learning):** face-to-face/internet