MHLT H6015: Digital Corpus Linguistics

Short Title:	Digital Corpus Linguistics APPROVED
Full Title:	Digital Corpus Linguistics

Module Code:	MHLT H6015

ECTS credits:	10

NFQ Level:	9

Module Delivered in	1 programme(s)

Module Contributor:	Irene Murtagh

Module Description:	The aim of this module is: To provide students with in-depth knowledge and skills of the role of digital corpora in natural language processing as a human language technology, in particular their design, analysis and correctness. To give students the skills necessary a wide range of techniques for digital corpus development and natural language applications. To provide students with the necessary theoretical and practical applications framework for research using a digital corpus for natural language applications. To instil an in-depth appreciation of the importance of quality in the digital corpus development.

Learning Outcomes:

On successful completion of this module the learner will be able to

1. Articulate the role and place of the digital corpus in an NLP strategy
2. Apply best practice research methods in using a digital corpus of natural language through appropriate corpus markup and annotation
3. Undertake research-motivated analysis of text using a corpus
4. Use and design text-oriented software programs
5. Critically differentiate current issues in digital corpora creation and use

Module Content & Assessment

Indicative Content
The Role of the Digital Corpus in NLP (30%) The use of corpus access techniques; Interpretation of corpus data; Corpus-based theories of language
Research Methods in using a Digital Corpus of Natural Language (30%) Corpus markup and annotation, Analysis of text using the corpus, Use and design of software programs ELAN software, AntConc software, Text oriented software, The R programming language, Corpus collection / corpora balance, Corpora gold standard, Research project design.
Advances in Applications of the Digital Corpus (30%) Corpora in translation, Supporting cultural studies, Historical studies, Digital humanities, Forensic linguistics using text analysis, Language teaching, Language processing.
Current Issues in Digital Corpora Creation and Use (10%) This address emerging issues in the creation of digital corpora and treebanks, Linguistics corpus-based methods, Methods of corpus linguistics (frequency lists, concordances, collocations), and the discipline directions in a digital world.

Indicative Assessment Breakdown	%
Course Work Assessment %	100.00%

Course Work Assessment %

Assessment Type	*Assessment Description*	*Outcome addressed*	*% of total*	*Assessment Date*
Practical/Skills Evaluation	Weekly practical work based on lecture material	1,2,3,4,5	20.00	Every Week
Project	The student will typically be expected to carry out a small-scale corpus-based project of their own design, using monolingual or bilingual corpora. The content is flexible to cater for the needs and interests of individual students.	4,5	40.00	n/a
Project	The student will typically undertake empirical work in applications of the digital corpus to corpora in translation, digital humanities, forensic linguistics using text analysis or language processing or current advancements in the creation of digital corpora and treebanks.	2,3	40.00	n/a

No Final Exam Assessment %

Indicative Reassessment Requirement
Coursework Only *This module is reassessed solely on the basis of re-submitted coursework. There is no repeat written examination.*

ITB reserves the right to alter the nature and timings of assessment

Indicative Module Workload & Resources

Indicative Workload: Full Time	
Frequency	**Indicative Average Weekly Learner Workload**
Every Week	24.00
Every Week	24.00
Every Week	152.00

Indicative Workload: Part Time	
Frequency	**Indicative Average Weekly Learner Workload**
Every Week	24.00
Every Week	24.00
Every Week	152.00

Resources

Recommended Book Resources

Malcolm Coulthard and Alison Johnson, *An introduction to forensic linguistics*, London ; Routledge, 2007. [ISBN: 0415320232]

Roger Garside, Geoffrey Leech, Tony McEnery, editors, *Corpus annotation*, London ; Longman, 1997. [ISBN: 0582298377]

Gries, Stefan Th. 2009, *What is Corpus Linguistics?*, Language and Linguistics Compass 3 http://www.linguistics.ucsb.edu/faculty/stgries/research/2009_STG_CorpLing_LangLingCompass.pdf

Grant S. Ingersoll, Thomas S. Morton, Andrew L. Farris, *Taming Text*, Manning Publications [ISBN: 193398838X]

Tony McEnery and Andrew Wilson 2001, *Corpus linguistics*, Edinburgh University Press Edinburgh [ISBN: 0748611657]

Supplementary Book Resources

Steven Bird, Ewan Klein, and Edward Loper 2009, *Natural language processing with Python*, O'Reilly Sebastopol, Calif. [ISBN: 0596516495]

Nancy Ide and James Pustejovsky 2014, *The Handbook of Linguistic Annotation.*, Developing linguistic theories using annotated corpora, http://web.stanford.edu/~cgpotts/papers/demarneffe-potts-lingann.pdf

Jockers, Matthew 2014, *Quantitative Methods in the Humanities and Social Sciences*, Springer; 2014 edition (June 11, 2014) [ISBN: 3319031635]

Jacob Perkins, *Python Text Processing with NLTK 2.0 Cookbook*, Packt Publishing [ISBN: 1849513600]

Matthew Russell 2013, *Mining The Social Web*, O'Reilly Media

This module does not have any article/paper resources

This module does not have any other resources

Module Delivered in

Programme Code	Programme	Semester	Delivery
BN_KMHLT_R	Master of Science in Computing in Multimodal Human Language Technology	2	Elective