Corpus linguistics. Presentation.

HAP/LAP. Corpus Linguistics.

Myself

Aitor Soroa <a.soroa@ehu.eus> (IXA group)

335 office (third plant) phone: 943 015040

Research interests:

- Lexical semantics (WSD, lexical similarity, etc)
- Knowledge acquisition
- NLP infrastructures (including annotation schemas)
- Multimodal systems
- Text generation
- Applications:
  - QA and Dialogue systems
  - Text simplification

- Google Scholar link

Course Syllabus

- Introduction to Corpus Linguistics
  - Introduction.
  - Corpus Linguistics
  - Uses of corpora
  - Corpus types
  - Corpus annotation and standards for linguistic representation.
- XML
  - XML introduction
  - XML schemas and validation
  - XPath
- Laboratories on:
  - Linux commands.
  - Word frequencies and Zipf law
  - Collocations
  - Keyword extraction
  - XML and XPath

Software we will use

- python (version 3 or higher)
  - lxml library (http://lxml.de/)
- R (https://www.r-project.org/)
- AntCoc (http://www.laurenceanthony.net/software.html)

Evaluation

- Attendance and participation: 10%
- Class assignments: 55%
- Assignments: three choices
  1. Regular assignment: 20%
  2. 'Hard' assignment: 35%
  3. Propose a subject for the final project: 35%
- Choose one of the three options above

Register to these sites

- English corpora (including BNC) http://corpus.byu.edu/
- Corpus del Español (NOW) https://www.corpusdelespanol.org/now/