C. Bosco

October 2024

Master's Degree in Language Technologies and Digital Humanities

Evaluation *in the final exam*
for the first part of the course *Linguistic Resources for Natural Language Processing*

The **assessment for the first part** of the course will in turn be carried out in two parts:

1) the first part of the assessment (maximum 7 points) will be in itinere and will be based on a TALK *(for students who have not attended (part of) classes, a reading will be suggested, on which some questions will be asked in the oral examination described at point 2, see below)*

2) the other part of the assessment (maximum 8 points) is based on an oral **EXAMINATION** during the examination session. This oral **EXAMINATION** consists of questions on all the contents addressed in the lessons. To achieve adequate preparation students must use the materials suggested in class (slides and videos) and read the paper on language variation and multilingualism: Emily Bender - "High Resource Languages vs Low Resource Languages" available at https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/#fn15
Below is a (**NON-EXHAUSTIVE**) ***list of questions that might be asked during the exam***. The answers can be found in the materials cited above and used in the lessons.

- What is the Turing test? Why is the Turing test particularly relevant for NLP?
- What is the ALPAC report? Why is the ALPAC report particularly relevant for the history of NLP?
- In the history of NLP there are two "winters". Can you collocate them in the timeline and explain their main causes?
- Can you describe some key differences between the Penn Treebank part-of-speech tagset and the Universal Dependencies part-of-speech tagset?
- Can you give a sentence in which there is at least one case of morphological ambiguity?
- Why can punctuation be difficult for a tokenizer to dealt with?
- What are the two types of structures on which the most cases of syntactic ambiguity depend?
- What are the steps of the MATTER cycle for developing linguistic resources?
- What are the key differences between symbolic and statistical models?
- What are the main advantages of using annotated versus non-annotated sentences for training models for syntactic parsing, or, in other words, to apply supervised versus unsupervised approaches?
- What are the main features of the syntactic dependency formalisms? Can you list the most relevant formal properties of a dependency structure?
- Given a set of Part of Speech tagged sentences annotated as gold standard and with a particular model, can you explain what a true negative is for the class NOUN?

- Given a set of Part of Speech tagged sentences annotated as gold standard and with a specific model, can you explain what a false positive is for the class VERB?
- What is the formula for calculating the precision score? What is the meaning of precision?
- What is the formula for calculating the F-1 score?
- Can you give three good reasons why English can't be considered representative of natural language in general?
- What is Valico-UD?
- Can you explain the differences between parallel corpora, aligned corpora and comparable corpora?
- Can you explain the difference between corpus-based and rule-based approaches to deal with morphology and syntax?
- What are the key differences in the development of a gold standard dataset versus a silver standard dataset?
- Can you explain how training set data and test set data are used within the evaluation of a model?