C. Bosco
October 2024
Master's Degree in Language Technologies and Digital Humanities

# Evaluation *in itinere* for the first part of the course *Linguistic Resources for Natural Language Processing*

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

- the first part of the assessment (maximum 7 points) is done *in itinere* (during the classes) and is based on a **TALK (whose features are described below)**[1]

- the other part of the assessment (maximum 8 points) is based on an oral examination during the examination session. The features of the examination are described in the file "*Indications for the first part of the exam: the oral examination*", that will be published on Moodle before the end of the classes of the first part of the course.

The characteristics of the **TALK** are as follows.

## How?

- Students must prepare a presentation (using PowerPoint or similar programs), i.e. the TALK, to show and discuss with the teacher and other students
- They can work individually, but they are strongly encouraged to work in teams of at least 2 to a maximum of 4 students. ***The main criterion for forming a team should be that its members have the same knowledge of a particular language on which the group's work can focus.***
- The TALK should last about 5 minutes for each member of the group. ***The time limit is the same for all students in order to encourage the ability to synthesize and select the most important aspects (as in a talk for a conference), but also to allow the teacher to apply a comparative evaluation of their talks.***

## With what content?

- For each team (or individual for students working individually, e.g. because they do not take the talk during the classes), the content of the talk must be the application of morphological and syntactic analysis to a text corpus in a particular selected language. ***Work on corpora containing data on a specific phenomenon is explicitly encouraged, e.g. a corpus of sentences in which relative clauses occur, or a corpus of sentences in which multi-word expressions occur, or a corpus of particularly ambiguous sentences ...***
- Each team must focus on a specific language and may only consist of members who have a good knowledge of that language (native speakers or advanced learners). ***Work on languages other than English is strongly encouraged in order to give all participants the opportunity to be exposed to examples and resources for as wide a variety of languages as possible***.
- In practice, each team has to collect a corpus of sentences in the specific selected language. The corpus must contain only texts in the chosen language and at least 10 sentences for each member of the group. ***Working on larger data sets is strongly encouraged to provide the opportunity to observe a greater variety of phenomena or a greater number of examples of a particular phenomenon.***

[1] **Students who did not take the TALK during the course** (e.g. because they could not attend classes), **can make it up during the final exam**; they must write an email to cristina.bosco@unito.it to arrange the topic of the TALK.

- Analysis can be done by applying **UDpipe** to the corpus (or using other available tools) and one or more models available for the selected language. In their talks the students must present interesting cases, such as errors they discovered in the results of the automatic analysis at the level of PoS tagging and parsing. They can also use different formats for the visualization of the syntactic trees.
- Considering that the language discussed in the talk can be unknown for the most of students, for each **example** given in the slides of the talk **two translations** in English must be provided: one translation in fluent English is needed to show the meaning of the sentence, while a word-by-word translation help the audience to understand the linguistic phenomena discussed.

**<u>How must the TALK be structured internally?</u>**

- The TALK must begin with a description of the corpus, based on the following: the language, the size of the corpus (number of sentences and number of tokens), the source from which the sentences were extracted.
- The TALK must include the discussion of some cases, phenomena and features of the observed language. Students are not asked to provide comments about all the sentences of the corpus they collected and analyzed, but to focus on the specific and interesting linguistic aspects they find in the data they collected in their corpora.
- The TALK must conclude with some lessons learned from the analysis of the corpus and observations on how the analysis can be extended in the future.

**<u>When?</u>**

- The team must be built and registered on Moodle (in the section about lesson 9 – October 2 you can find the link to "***Choice your group!***"). **The deadline for registration is October 9**.
- The talks will take place in the second hour of classes in the second half of October, starting on the 14th (15th, 16th, 21th, 22th, 23th); a calendar will be drawn up and communicated to the students after the teams have been formed.
- The **slides** of the talk must be **sent to cristina.bosco@unito.it at least 3 days before** the date on which the talk is scheduled (better if more in advance).

**Example:**

How?

the talk of a group of 3 members focusing on Russian must include three Russian speakers (or advanced Russian learners); it will be about 15 minutes long and based on a corpus of at least 30 Russian sentences.

With what content?

The group will collect a corpus (of at least 30 sentences) in which all sentences contain at least one transitive verb, the talk of this group will be based on the observation of the behavior of transitive verbs in the collected corpus and will answer questions such as: Are all complements of transitive verbs lexically realized? Is it the subject always a noun or it can be a verb? In how many cases are the verbs also connected to some modifiers? …