# The Role of Syntactic Corpora in the Era of Large Language Models (LLMs)

**Petya Osenova**
Sofia University "St. Kl.
Ohridski" and IICT-BAS
petya@bultreebank.org

## Abstract

*Time:* This course is envisaged with duration of 2 (two) 45-minute sessions or one 90-minute session.

*Material:* The main material will be in the form of SLIDE presentations. In addition, appropriate webpages will be used as well as some demos will be shown.

*Prerequisites*: basic, but not necessarily professional, interest in and knowledge of syntactic notions, language corpora and LLMs.

*Audience*: This tutorial is aimed at the BA, MA and PhD students in Linguistics, Humanities or Computer Science who have some acquaintance with grammar and/or language corpora and/or LLMs, and would like to know more about the interaction among them. Also, they should be able to follow the presentations in English.

**Keywords:** syntactically aware language resources, LLMs, approaches to modeling syntax.

## 1 Background and Motivation

Syntactically annotated/processed data are very important for down-stream tasks such as named entity identification and linking, sequential labeling with linguistic and/or encyclopedic information, event reasoning and extraction, event perspectives handling and linking. The data can be annotated manually thus becoming gold standards, or automatically being parsed by trained models.

The resulting resources are treebanks or parsebanks. There are various approaches in modeling syntax – shallow vs. deep, constituency-based, dependency-based or mixed. The annotation schemas usually depend on the subsequent tasks, with handling the syntactic knowledge only, or with added lexical/sentence semantics, coreferences, world knowledge, etc.; with a tree-based only or with a graph-based representation.

In the era of LLMs, treebanks as well as other Language Resources (corpora and dictionaries) are NOT outdated due to a number of factors: LLMs need not only huge quantities of data but also specific linguistic knowledge about these data, LLMs can be biased, ethically compromised and can introduce unpredictable noise, etc. They are still black boxes to great extent. Therefore, syntactically aware data are needed before, at the time of and/or after the LLMs applications for the purpose of tuning, validation and evaluation of the results.

## 2 Content in brief

The tutorial will have a linguistically / language oriented focus.

The tutorial aims at introducing treebanks/parsebanks from various perspectives – theoretical orientation, granularity, formal representations, modeling linguistic knowledge with a focus on Bulgarian but with a strong multilingual and universal perspective. At the same time, the challenges of these resources are considered in the era of LLMs including chatbots. The role of the treebanks/parsebanks as well as the trained parsers or computational grammars will be discussed in relation to

the transformers (BERT, RoBERTa, etc.), ChatGPT, Bard and the like.

## 3 Timeline

**Part 1 (45 min)** will focus on the following topics:

- Introduction to the notion of syntactic corpora
- Constituency-based vs. dependency-based vs. mixed approaches
- Manually annotated vs. automatically parsed data
- Types of knowledge in syntactic data
- Transferring knowledge from one model to another: challenges

**Part 2 (45 min)** will focus on the following topics:

- Universal and multilingual annotation schemas and data: the Universal Dependencies model and its utility
- The syntactic corpora and parsers vs. LLMs: how to make the best from the two worlds.
- The following topics will be also discussed:
    - data biases,
    - domain dependance,
    - cross-lingual performance,
    - common sense and world knowledge.

## 4 About the proposer

Dr. Petya Osenova is professor in Contemporary Bulgarian Grammar (morphology, syntax and corpus linguistics) in the Faculty of Slavic Studies at Sofia University "St. Kl. Ohridski" and senior researcher in the area of Language Technologies in the Department of AI and Language Technologies at the Institute of Information and Communication Technologies, Bulgarian Academy of Sciences. Her scientific interests are in the fields of formal and computational linguistics, language resources, grammar-lexicon interface.

Petya Osenova was a key person in a number of EU projects, related to eLearning, Machine Translation, Language resources (EuroMatrixPlus, AsIsKnown, LTfLL, QTLeap, EUCases, among others). She is the responsible person for the language resources in CLaDA-BG – the CLARIN and DARIAH joint framework in Bulgaria. Petya Osenova is the Bulgarian representative at the User Involvement Committee in CLARIN-ERIC.

Petya Osenova specialized in computational linguistics as a postdoctoral fellow in Tübingen University, Germany (2003) and in Groningen University, the Netherlands (2004); as a Fulbrighter at Stanford University, the USA (2010).

In 2018 Petya Osenova received the award of Clarivate Analytics for excellence in science research in South-Eastern Europe.

## 5 References

### 5.1 Hyperlinks

Universal Dependencies website: https://universaldependencies.org/

NoSketch engine website: https://www.sketchengine.eu/nosketch-engine/

Transformer-based models for Bulgarian: https://aclanthology.org/2023.ranlp-1.77/

bgGLUE: A Bulgarian General Language Understanding Evaluation Benchmark: https://aclanthology.org/2023.acl-long.487/