Introduction

Johann-Mattis List (University of Passau)

While the discipline of computational linguistics mostly deals with the modeling and the investigation of individual languages (often "big" languages such as English, German, Arabic, or Chinese), Multilingual Computational Linguistics focuses on the comparison of languages, trying to develop new methods and techniques by which languages can be compared automatically or in a computer-assisted manner. The comparison itself follows different perspectives (maintaining a historical, typological, or areal viewpoint). In this scientific practice course, we will take a closer look at basic theories and methods which are relevant for the discipline of Multilingual Computational Linguistics. We will look at large corpora with multiple languages of the world as well as data from individual languages and language families. If wanted, we can focus on specific language families, which are relevant for the studies of the participants. Thematically, we want to look at the inference of cognates, the detection of borrowings, the reconstruction of phylogenies, and the modeling of semantic change and sound change. If participants are specifically interested in topics that we could additionally cover, they should write a short email before the end of January, to give us time to check if we can include those topics in the course.

1 Multilingual Computational Linguistics

Given the multitude of multilingual applications in the field of computational linguistics, it may not be easy to give an exact definition of the field, since it will inevitably depend on the individual researchers' background and scientific preferences, how they fill "multilingual computational linguistics" with life. Given my specific background as a historical and comparative linguists working on computational applications that help us to increase the efficiency and accuracy of historical and typological language comparison, my major approach towards multilingual computational linguistics does not have a lot to do with the typical NLP applications that translate across a couple of well-understood languages with huge corpora. What I offer instead are a couple of novel methods and techniques by which we can compare languages synchronically and diachronically with the help of computational methods. While this may seem very narrow-minded at first sight, the scientific focus is much broader, since it touches upon a large range of topics ranging from classical linguistic typology and classical historical linguistics via more recent corpus-based approaches in linguistic typology and phylogenetic approaches in historical linguistics up to topics in psycholinguistics that try to learn more about human cognition through a close investigation of linguistic diversity.

Since the field of multilingual computational linguistics represented in this course is still in its infancy, with most major applications having only been made in the last ten years, we cannot make use of off-the-shelf tools for computational comparative linguistics but are instead in a situation where we need to design these tools and often create them from scratch. As a result, our work allows to gain concrete insights into interdisciplinary work, since we need to check with the methodology of many different disciplines (ranging from classical language comparison via bioinformatics up to computer science and digital humanities) in order to handle the problems we face in our research. As a result, we pay specific attention to *scientific problem solving*, to *open data* and *open science* in general, as well as to traditional methodologies applied in many disciplines of the humanities before the arrival of computational methods.

2 Course Organization

The course is divided into 15 sections distributed over four days. On each day, we start with a rather short morning section in which smaller topics and questions are introduced, followed by two larger sections and one practice section in the evening. On the third day, there is no practice section, and we have only one larger section before a concluding section. We use the practice sections specifically to address individual problems brought to the course by course members. Thus, although some guidelines for these topics are provided, we assume a lot of flexibility on the concrete questions here and will generally split into groups rather than working in a big group with all course members.

3 Course Plan

In the following, a detailed course plan is given.

Day 1: Introductory Topics

10–10:30: Introduction of group participants, which can be extended to the coffee break.

11–12:30: Background on Comparative Linguistics

14–15:30: Scientific Problem Solving

16–16:30: Practice Session: Discuss Unsolved Problems in Smaller Groups

Day 2: Modeling and Standards

10–10:30 Cross-Linguistic Data Formats

11–12:30 Reference Catalogs

14–15:30 Standardized Data Collections in Multilingual Computational Linguistics

16–16:30 Practice Session: Standardize and Retrostandardize Data, Parse Texts

Day 3: Inference

10-10:30 Computer-Assisted Language Comparison

11-12:30 Sequence Comparison

14-15:30 Semantic Networks

16-16:30 Practice Session: Workflow Development and Testing

Day 4: Analyzing

10-10:30 Chinese Computational Linguistics

11-12:30 Computer-Assisted Text Analysis

14-15:00 Final Discussion