Research Proposal: LLM4DH Project – Advanced Grammatical Analysis of Multilingual Corpora

Nives Hüll
University of Ljubljana
Digital Linguistics Project Seminar, University of Zagreb
nh23084@student.uni-lj.si, nhull@m.ffzg.hr

November 10, 2024

Background

This project is part of a larger initiative, contributing to the Gravitacija – LLM4DH (Large Language Models for Digital Humanities) project. I will focus on a subtask within Task T2.3, titled *Advanced Grammatical Analysis of Multilingual Corpora*, led by Kaja Dobrovoljc. This task aims to develop an approach using large language models (LLMs) for detailed grammatical analysis across multiple languages. By fine-tuning a multilingual LLM on the Universal Dependencies (UD) dataset – an extensive dataset with morphosyntactic annotations – we hope to enhance the LLM's ability to process diverse grammatical structures.

Project Scope

The overarching goal of Task T2.3 is to build a dataset and methodology that enables effective cross-linguistic grammatical analysis, data annotation, pattern extraction, and data summarization. Evaluations will focus on the model's ability to identify novel linguistic patterns, support comparative grammatical analysis across languages, and provide insights into linguistic diversity globally.

Although the original intent was to enable the model to answer comprehensive grammatical questions about world languages, this broad scope has necessitated a more focused approach. Therefore, a progressive methodology will be adopted, evaluating simpler models (e.g., using a baseline model like ChatGPT) before progressing to advanced fine-tuning or instruction-based models with retrieval-augmented generation (RAG) if needed.

Personal Research Focus

My specific task will contribute to refining the team's focus by providing targeted linguistic insights that could aid in deciding how best to fine-tune the model. To do this, I will begin by examining the *World Atlas of Language Structures* (WALS), an extensive database of structural language features. WALS documents around 190 linguistic features across languages, such as word order, and categorizes languages based on these. I will assess whether the features cataloged in WALS for languages like Slovene (and potentially Croatian) can be effectively identified within the UD system.

Research Approach and Potential Restrictions

Given the complexity of the task, it is currently difficult to predict the exact amount of time required. My initial focus will be to complete the analysis for a single linguistic level and for Slovene. If this preliminary work is successful, I plan to expand the analysis to additional levels or languages.

For data analysis, I will use tools such as *STARK*, *Q-CAT*, and possibly *Grew-match* to aid in feature extraction and assessment. This approach allows for iterative validation, enabling further expansion only if results prove feasible and meaningful in the initial Slovene-level analysis.

Research Hypothesis and Questions

Hypothesis: Most linguistic features cataloged by WALS can be identified in UD data, with possible exceptions at the phonological level. Furthermore, corpus data may provide insights into real-world language use, especially in spoken forms, that traditional grammars and WALS categorizations overlook.

Research Questions

1. **To what extent do spoken Slovene syntactic patterns align with WALS typological features?**
   Hypothesis: Spoken Slovene syntax will show deviations from WALS norms, as WALS often generalizes from written sources, whereas spoken syntax displays unique structures.

2. **Are there significant differences between spoken and written Slovene syntax in relation to WALS typological features?**
   Hypothesis: Word order and case marking in spoken Slovene will differ from WALS categories, reflecting greater variability and context sensitivity.

3. **Do WALS's binary categories sufficiently capture the syntactic nuances found in spoken Slovene?**
   Hypothesis: WALS's binary categorization lacks the nuance needed to represent complex syntactic patterns in spoken Slovene, which often display more variability than WALS's categories allow.

4. **Are spoken syntactic structures in Slovene similar to those in typologically related languages (e.g., Croatian), or do they show unique characteristics?**
   Hypothesis: Slovene spoken syntax will align more closely with other South Slavic languages, like Croatian, than with WALS generalizations, indicating regionally consistent patterns.

Challenges and Possible Solutions

One key challenge is the availability of annotated spoken and written corpora for Slovene and Croatian, both in UD format. While Slovene resources are available, I need to verify the existence of similar resources for Croatian. If spoken corpora are unavailable, I will explore segmenting written data into contexts that approximate spoken syntax. In the first phase, I may limit my analysis to Slovene only and exclude Croatian if resources for the latter are not accessible.

Additionally, answering all the research questions may prove to be too ambitious at this stage, so I may need to modify my approach based on the current progress, available knowledge, and insights gained during the project.

Potential Contribution

This project could enhance WALS by integrating corpus-based insights for Slovene, thereby updating and refining typological data. Such contributions would benefit the linguistic community, offering data that reflects current language use more accurately than traditional grammars, which may be outdated or lack a Slovene perspective.

Additionally, I aim to continue working on other parts of the project and hope that the findings from this phase will be usable and contribute to the broader goals of the project in the future.

Resources

- **Project *Gravitacija*:** `https://www.aris-rs.si/sl/medn/gravity/predstavitev.asp`

- **Universal Dependencies:** https://universaldependencies.org/
- **World Atlas of Language Structures (WALS):** https://wals.info/
- **STARK Tool:** https://github.com/clarinsi/STARK
- **Q-CAT Tool:** (https://slovnica.ijs.si/wp-content/uploads/2019/10/Q-CAT_prirocnik.pdf)
- **Grew-match Tool:** (https://match.grew.fr/)
- **Article for Reference (*Multilingual Gradient Word-Order Typology from Universal Dependencies*):** https://aclanthology.org/2024.eacl-short.6