Software Development for Automated Linguistic Corpus Construction

Donjeta Bucaj1, Vanesa Gashi2

*Lorddian, Department of Research, Pristina, Kosovo, bucajdonjeta6@gmail.com1*

*University of Tirana, Tirana,Albania, vanesagashi@gmail.com2*

**Abstract:**

Building annotated linguistic corpora represents a key foundation in the universe of natural language processing (NLP), a field that rises above the demands of precision and cross-linguistic exploration. As the landscape of multilingual systems expands and the call for an unprecedented uniformity in data annotation grows stronger, the challenges that accompany building corpora become increasingly labyrinthine. This paper takes a deep dive into the world of automated tools, those ingenious creatures that facilitate the annotation and verification of linguistic data, thus revealing some of the most formidable obstacles to quality, accuracy and efficiency in this complex ecosystem.

Imagine the wonders of automation delivered automatic pre-annotation, seamless transfer of annotation from one language to another, and rigorous automatic validation of annotations transform linguists' heavy workload into a manageable task. These technological marvels not only ease the burden but also increase the coherence of data across languages. By automating these multifaceted processes, we are witnessing a tremendous acceleration in processing speed, significantly reducing the time and human resources traditionally required to build large language corpora. This automatic orchestration is essential to address the complexity of syntactic and semantic phenomena, from multi-lexical item (MWE) phrases to the complexity of relative clauses and coordination elements that are notorious for their resistance to smooth integration into different structures linguistic.

This exploration extends to examining and proposing innovative tools for curating, merging and converting corpora, creating a more coherent and centralized approach to their construction and management. Using these tools is not just a convenience; it is a necessity that guarantees the correct validation of annotated data, preserves the integrity of corpora, and prepares them for publication and further exploration by the academic avant-garde and the dynamic arena of NLP. The development of this type of software, often supported by national funding, serves as a key linchpin, advancing the construction of multilingual corpora for linguistic typology studies, machine translation, and a wide range of NLP applications.

***Keywords:*** ***Annotated Linguistic Corpora, Natural Language Processing (NLP),Automated Tools, Data Annotation, Cross-Linguistic Analysis.***