INSTITUT FÜR ÜBERSETZEN UND DOLMETSCHEN

UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

Collaboration between Tbilisi University and IÜD, Heidelberg University

Project participants: Oleg Kapanadze, Bogdan Babych, Vahram Atayan

Heidelberg, 3 December 2021

Computational analysis for low-resourced and morphosyntactically complex languages

Main areas of the project:

Finite state transducers (FST) for morphologically complex languages:

- running FST for annotating Georgian (KA) corpora
    - producing Part-of-Speech PoS, Morphological information and lemmas
    - collection of relevant corpora
    - evaluating coverage and quality
- automating extension of transducers from partially parsed data; human development workflow

Developing other core linguistic technologies:

- extending tagging/lemmatization coverage for Georgian (KA) | Armenian (HY) | Ukrainian (UK) Nouns, Adjectives, Adverbs, Verbs *(applications: terminology extraction, word vectors, translation equivalents / lexicography, parsing)*
    - finite-state or rule-based methods
    - corpus-based methods
    - machine learning methods (use of token-, context- and syntax-based features)
    - evaluation / development of evaluation sets
    - Linguist's Workbench - interactive annotation environment for linguistics
        for extending the coverage
        for evaluation of tagging & lemmatization in corpora

- identification and lemmatization of KA | HY | UK Multiword Expressions *(applications: Multiword translation equivalents, next-generation linguistically-aware Translation Memories, improvement and evaluation of Machine Translation)*
    - treebank-based methods (syntactic nodes and types of syntactic relations)
    - chunking / shallow parsing / context+part-of-speech-based methods

- morphological disambiguation: choosing correct analysis + lemma for ambiguous tokens in corpus *(applications: corpus annotation, word vectors)*

- word sense disambiguation, e.g., for light verb construction 'take part' vs. 'participate' *(applications: corpus annotation of word senses; word vectors)*

- treebanks and parsing for KA | HY | UK *(applications: argumentation mining; discovering translation equivalents, next-generation linguistically-aware Translation Memories, improving / evaluating MT)*
    - NLTK,
    - training Stanford parser...
    - dependency vs. constituency representations

- WordNet and ontologies for the general lexicon of KA | HY | UK and terminological domains
available linguistic resources, e.g., EN | DE Word Net; FrameNet
extensions with word vectors