Department of Computational Linguistics
Institute for Bulgarian Language
Bulgarian Academy of Sciences

52 Shipchenski Prohod Blvd.
Building 17
Sofia 1113, Bulgaria

UniDive

COST Action CA21167: Universality, diversity and idiosyncrasy in language technology

Proposal to host short-term scientific mission (STSM)

1. Name and address of the hosting institution

Host:

Department of Computational Linguistics
Institute for Bulgarian Language
Bulgarian Academy of Sciences.

Postal Address:

52 Shipchenski Prohod Blvd.
Building 17
Sofia 1113, Bulgaria

Webpage: https://dcl.bas.bg/en/

2. Contact information

Information about the STSM:

Svetla Koeva svetla@dcl.bas.bg, Ivelina Stoyanova iva@dcl.bas.bg

3. Approximate duration and dates of the STSM

The duration of the proposed STSM would ideally be 1 – 2 weeks. There is flexibility with regards to the time of the year when the STSM can take place and it needs to be agreed with the host institution in advance.

It can also be combined with attendance to the International Conference Computational Linguistics in Bulgaria (9 – 10 September 2024) in Sofia.

4. Research topic and description

Research topic: **Developing monolingual and multilingual language resources for low-resource languages**

Description: The proposed topic aims to address the challenges of low-resource languages and the compilation of language resources for them. Our experience in developing language resources for Bulgarian can be helpful for researchers who also deal with the challenges of collecting data, processing and annotation, alignment, developing lexical and semantic resources, theoretical and applied studies on morphologically rich and/or low-resource languages.

We are interested in collaborative research in order to explore the notions of universality and diversity, in particular in developing language resources and technologies.

We share experience on the following resources:

(1) Bulgarian National Corpus – structure, compilation, metadata, subcorpora, search engine, collecting special purpose large corpora, processing and annotation.

(2) Specialised corpora – Bulgarian-English Sentence- and Clause-Aligned Corpus, Bulgarian Semantically Annotated Corpus, other parallel or comparable corpora. These will be analysed in comparison with other similar resources and how they can be used in combination.

(3) BulNet – the Bulgarian WordNet, would be analysed as connected to the Princeton WordNet (and hence, any other wordnet based on it). This turns the resource into a multilingual lexical-semantic resource.

(4) FrameNet – a large conceptual resource and our work on its applications for Bulgarian, in particular assigning FrameNet frames onto WordNet synsets. We also present the system BulFrame facilitating the conceptual description of lexical entries.

(5) Various MWE-related resources – dictionaries of MWEs, aligned Bulgarian-Romanian MWE vocabulary, morphological description of MWEs, Bulgarian PARSEME corpus annotated with verbal MWEs, etc.

(6) Verb semantics resources and theoretical studies – we have worked extensively on classes of predicates, for example statives, in comparison between Bulgarian, Russian, English. We welcome researchers working in this field to expand the research.

(7) Multimodal resources – Multilingual Image Corpus (MIC) and the ontology of visual objects.

The following objectives of the COST Action are addressed:

- This will contribute directly to the objectives of UniDive WG1 (Corpus Annotation) and WG2 (Lexicon-Corpus Interface) by exploring the capabilities of various corpora and resources. Expanding the resources for different languages, and in particular for low-resource languages aligns with WG3's focus on cross-lingual and multilingual language technology.
- These resources and the experience gained will facilitate ongoing efforts to develop language technologies which perform better on low-resourced languages. Novice methods for transfer of knowledge from well-resourced languages to less-resourced languages based on universality principles is one way to achieve better quality of resources.
- The STSM also can help fostering fruitful long-term collaborations both in theoretical research and in developing multilingual language resources and technologies.