Click here to be redirected to the virtual room of the Project Expo.

Project abbreviation: MIC 21

Project name: Multilingual Image Corpus

Project coordinator: Prof. Svetla Koeva

Project consortium: Institute for Bulgarian Language “Prof. Lyubomir Andreychin”

Funding: The Multilingual Image Corpus (MIC 21) project was supported by the European Language Grid project through its open call for pilot projects. The European Language Grid project has received funding from the European Union’s Horizon 2020 Research and Innovation programme under Grant Agreement no. 825627 (ELG).

Project duration: 01.03.2021 - 28.02.2022

Main key words: image dataset, ontology of visual objects

Background of the research topic: One of the processing tasks for large multimodal data streams is automatic image description (image classification, object segmentation and classification). Although the number and the diversity of image datasets is constantly expanding still there is a huge demand for more datasets in terms of variety of domains and object classes covered.

Goal of the project: The goal of the project Multilingual Image Corpus (MIC 21) is to provide a large image dataset with annotated objects and object descriptions in (at least) 20 European languages.

Project abstract: The Multilingual Image Corpus consists of an Ontology of visual objects (based on WordNet) and a collection of thematically related images whose objects are annotated with segmentation masks and labels describing the ontology classes. The dataset is designed both for image classification and object detection and for semantic segmentation. The main contributions of our work are: a) the provision of large collection of high-quality copyright free images; b) the formulation of the Ontology of visual objects based on WordNet noun hierarchies; c) the precise manual correction of automatic object segmentation within the images and the annotation of object classes; and d) the association of objects and images with extended multilingual descriptions based on WordNet inner- and interlingual relations.

Publications:

Svetla Koeva. Multilingual Image Corpus: Annotation protocol. In: Galia Angelova, Maria Kunilovskaya, Ruslan Mitkov, Ivelina Nikolova-Koleva (Eds). RANLP 2021. Deep Learning for Natural Language Processing Methods and Applications. Proceedings, 2021, 701-708.