FRIEDRICH-SCHILLER-UNIVERSITÄT JENA

UniDive

cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Funded by the European Union

WG1

Creation Dataset of Token Language Identification for Ukrainian-Russian Code-switching Corpus

Olha Kanishcheva[1,2], Maria Shvedova[1,3]

[1]University of Jena, [2] SET University, [3] National Technical University "KhPI

Our results:

1) A database of about 150,000 tokens containing code-switching between Ukrainian and Russian has been collected.
2) This dataset contains intra-word code-mixing, so-called Surzhik. The dataset is divided at the token level into 5 categories. The next step will be to analyze the obtained dataset and test different classification models on this data.
3) We analyzed the different types of code-switching that occur in our dataset.
4) Some metrics of code-switching have been calculated to show the complexity of the data.

Labels	Description	Tokens
UK	Ukrainian words	93 040
RU	Russian words	30 956
MIX	Ukrainian-Russian hybridized words (Surzhyk)	225
Others	Dialects, other languages, etc.	615
Punct	Punctuation	30 695