

CLARIN: Presenting the K-centre for Ukrainian NLP and corpora

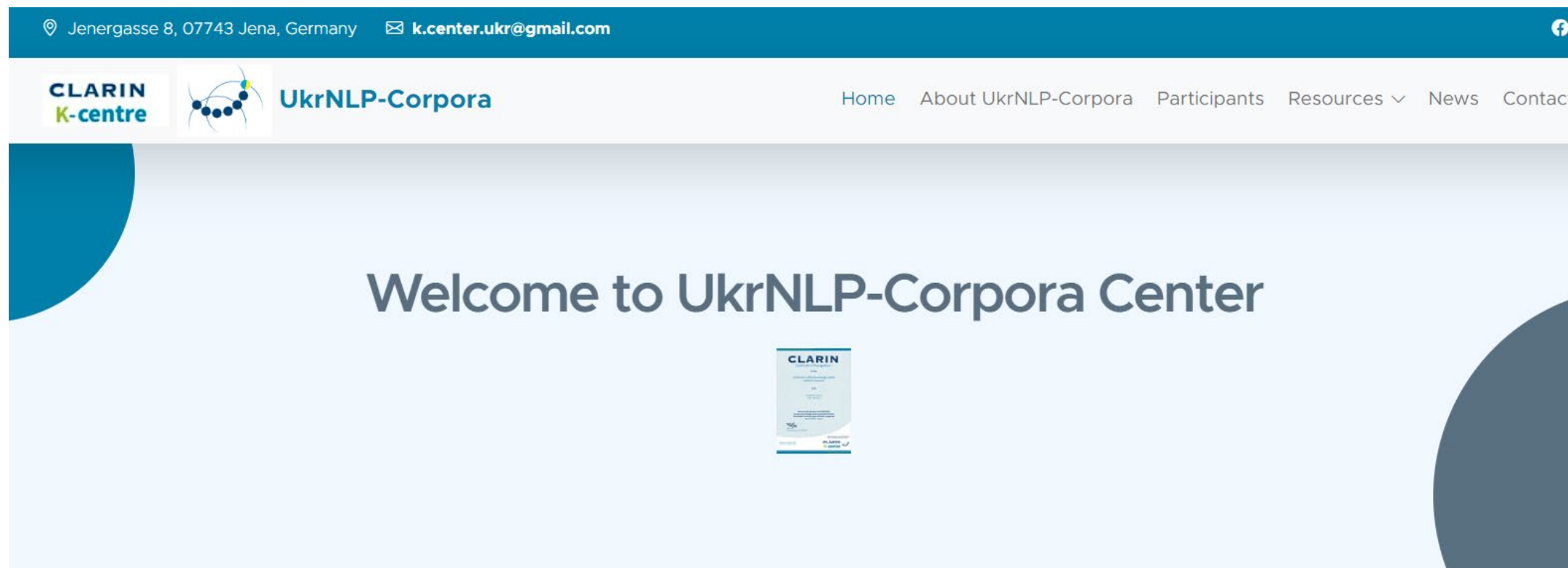
Olha Kanishcheva,
PhD, visited professor at
University of Jena

CLARIN



History of our center

(<https://uacorporus.org/k-centre/>)



Our resources (corpora)

Corpora

General Regionally Annotated Corpus of Ukrainian (GRAC) is the largest manually compiled reference corpus of Ukrainian.

Corpus Project of the Laboratory of Ukrainian contains several corpora and a dedicated morphological analyzer. The corpora include a treebank with manual disambiguation and manual tagging (140 thousand tokens), a web corpus "Zvidusil" with automatic syntactic annotation (about 3 billion tokens), parallel corpora.

Lang-uk corpus project provides collections of Ukrainian online press, fiction, and Wikipedia available for download, totaling 665 million tokens (UberText corpus), a corpus of law and legal acts counting 579 million tokens, a corpus annotated for named entities and also a build-up model for automatic annotation of named entities (people, organizations, locations, and others); different gazetteers, simple tokenizer (splitting text into paragraphs, sentences, and tokens), vector models trained on different corpora.

Ukrainian Brown corpus - open, genre-balanced and in the future annotated corpus of the modern Ukrainian language (BrUK) with a volume of 1 million word usages. The corpus is built on the basis of the well-known Brown corpus of the English language..

UA-GEC a corpus of texts with marked grammatical errors.

Our resources (tools)

Tools

Nlp-uk is an instrument based on the VESUM dictionary and the LanguageTool engine. Supports tokenization, lemmatization, POS analysis, and basic disambiguation.

Pymorphy2 — a morphological analyzer without disambiguation; the Ukrainian language is supported via the old version of VESUM.

Stanza — the Stanford library for language processing; it supports Ukrainian using the UD corpus. Features models for tokenization, lemmatization, POS and syntactic analysis.

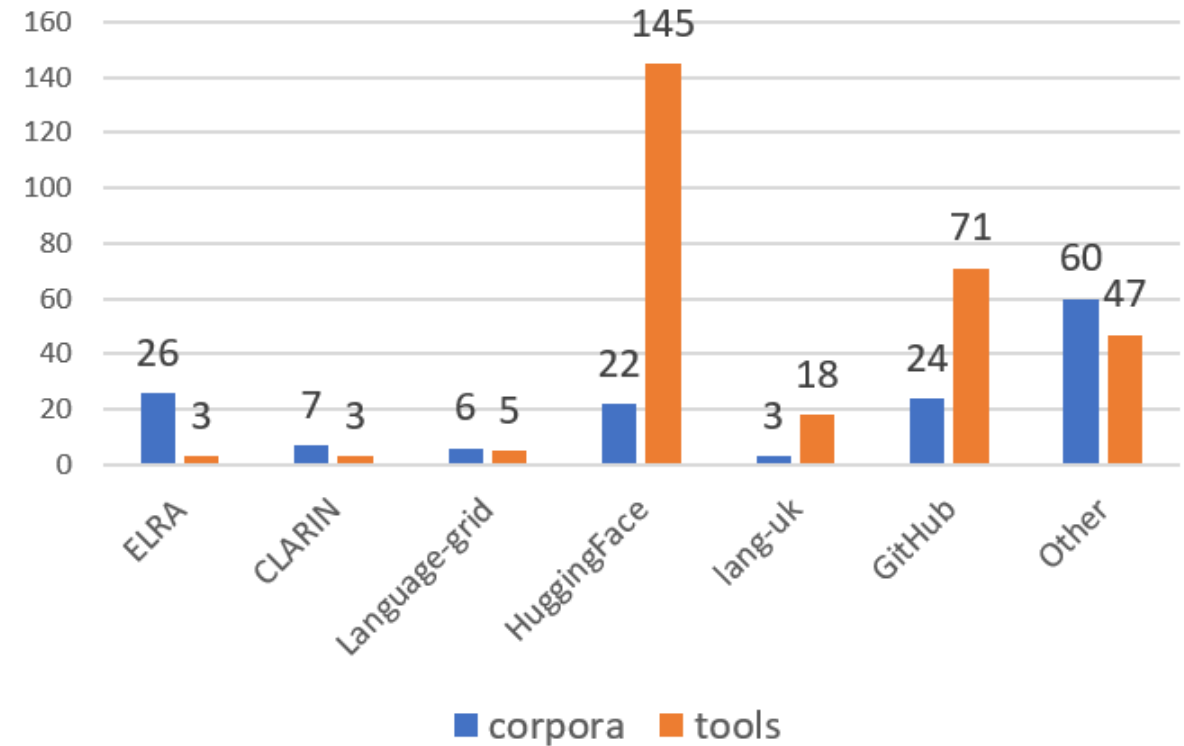
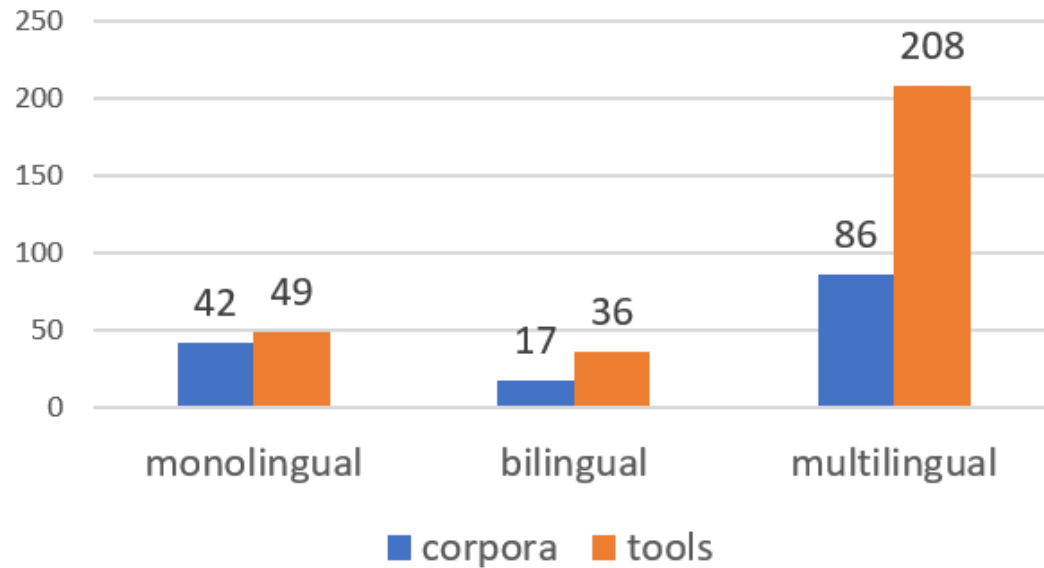
LanguageTool — spelling, stylistic, and grammar checker, which helps to correct and paraphrase texts.

Stemmer for Ukrainian language — a new stemmer for the Ukrainian language (tree_stem) created via machine learning.

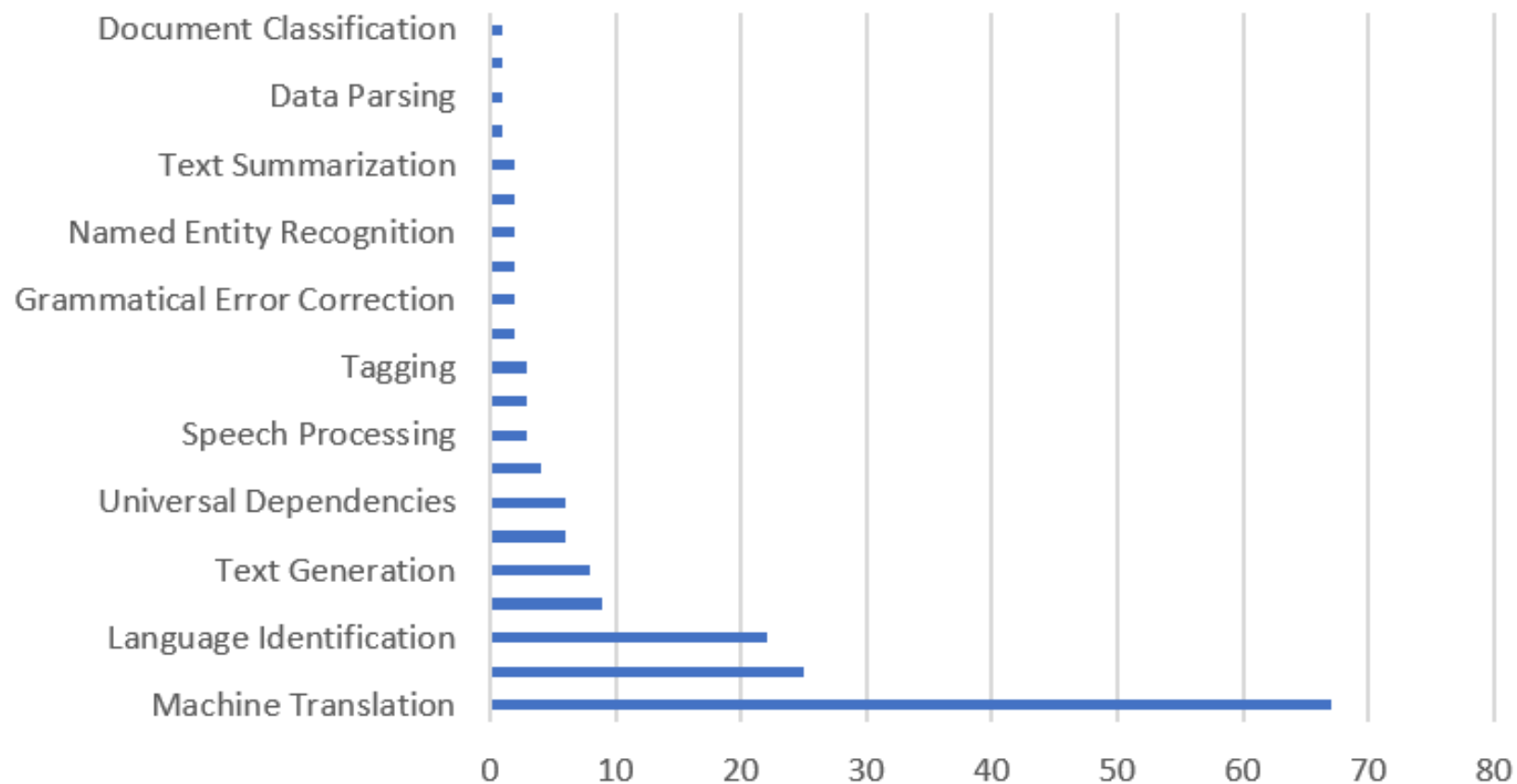
EdUKate translation software 1 — a software package that includes three tools: web frontend for machine translation featuring phonetic transcription of Ukrainian suitable for Czech speakers, API server and a tool for translation of documents with markup (html, docx, odt, pptx, odp,...).

HENSOLDT ANALYTICS services for Speech to text, Language identification, Sentiment analysis and Named entities detection, Keyword spotting, Age detection, Gender detection, Summarization.

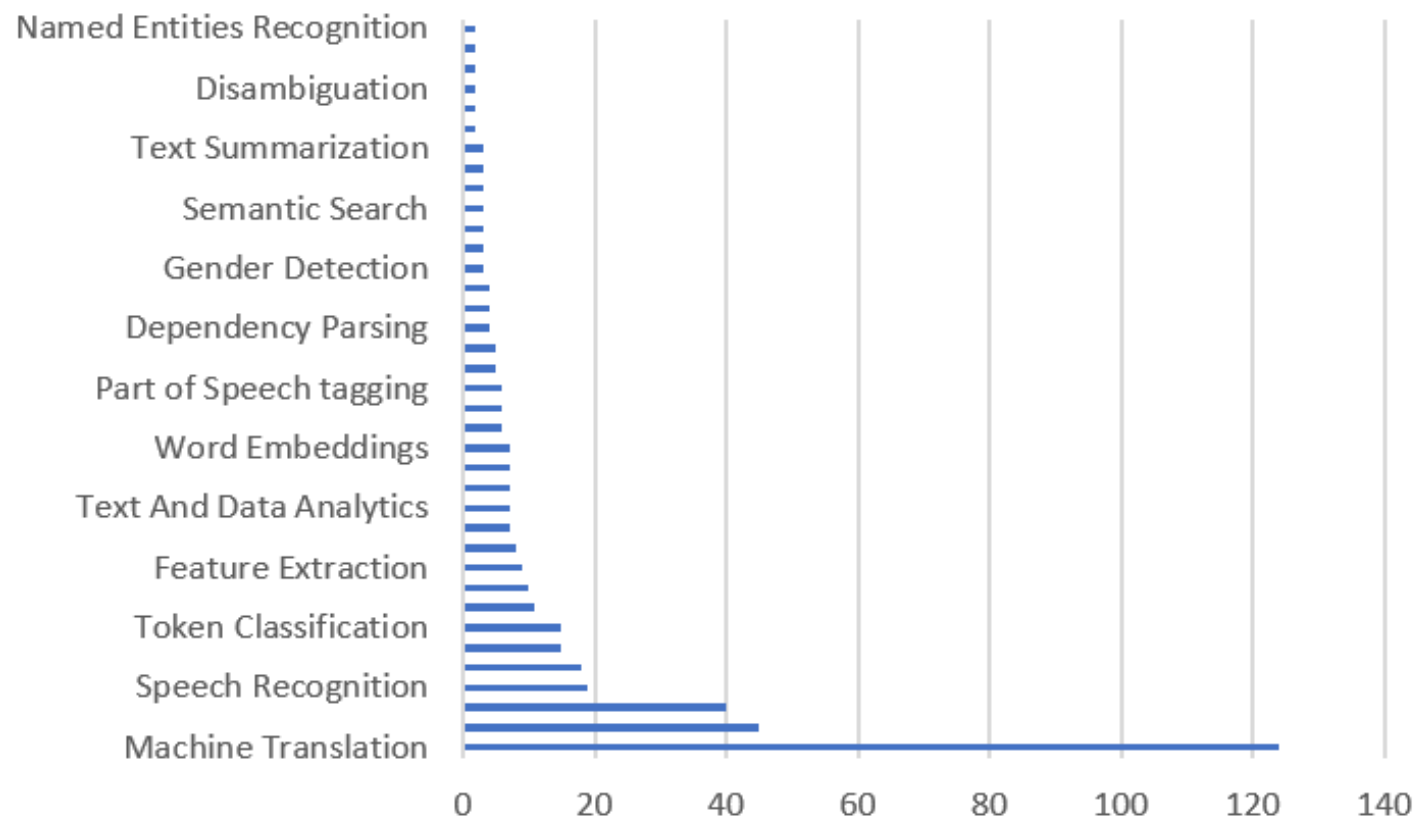
Resource statistics



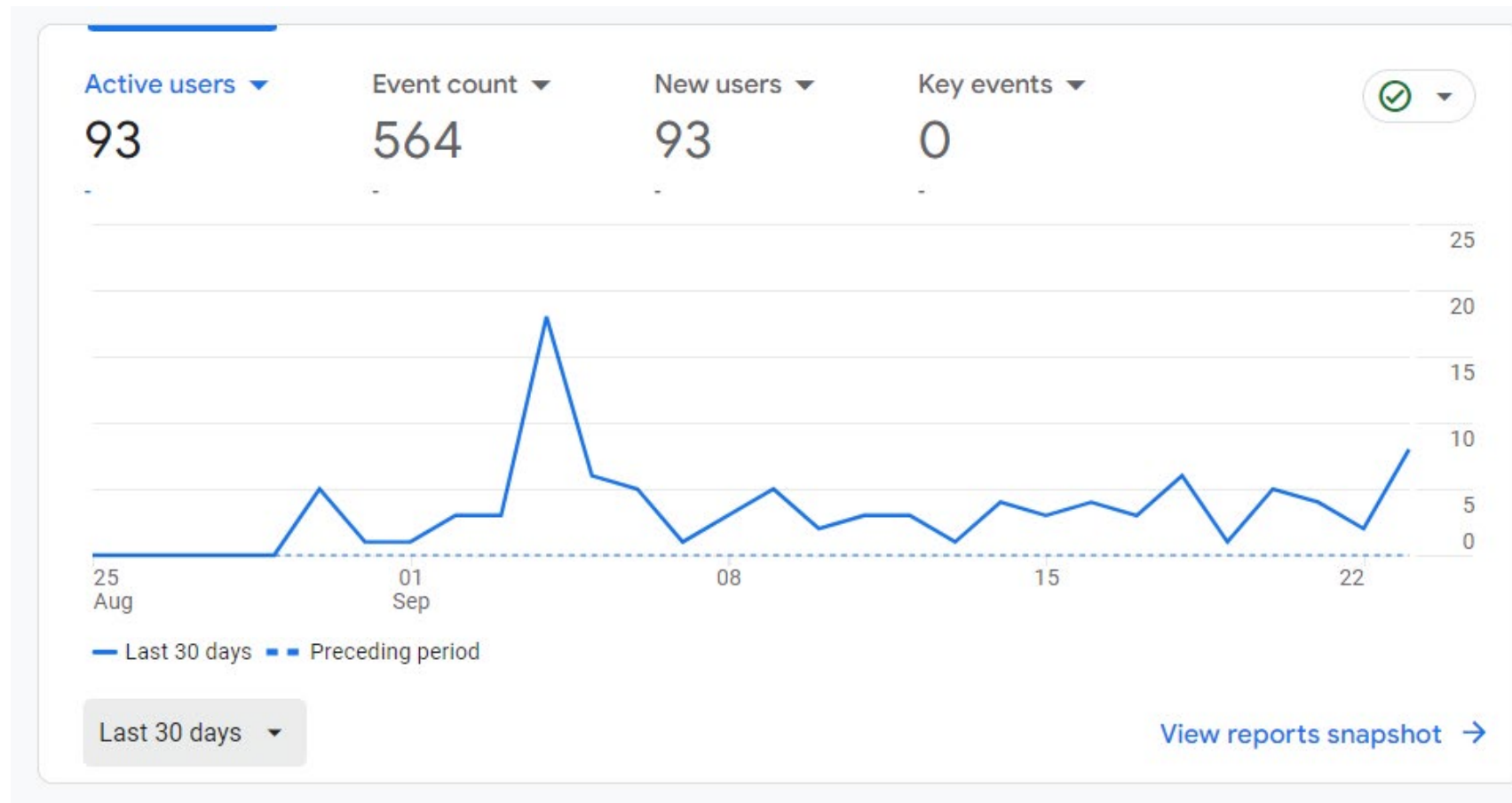
Tasks of corpora



Tasks of tools



Site analysis (users)



Site analysis (countries)

Active users▼ by Country ID▼



COUNTRY	ACTIVE USERS	
Ukraine	46	-
United States	8	-
Netherlands	7	-
Finland	5	-
Germany	3	-
Poland	3	-
Ireland	2	-
-		

Last 60 days ▼

[View countries →](#)



Thank you!

<https://uacorporus.org/k-centre/>