

# CLARIN.SI: The Slovenian node of CLARIN

<https://www.clarin.si/>

**CLARIN.SI**

Common Language Resources and  
Technology Infrastructure, Slovenia



The screenshot shows a detailed view of a repository item. At the top, there's a header with the title 'Ukrainian web corpus MaCoCu-uk 1.0'. Below it, a note says 'Please use the following text to cite this item or export to a predefined format: Bañón, Marta; et al., 2023, Ukrainian web corpus MaCoCu-uk 1.0, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, <http://hdl.handle.net/11356/1838>'. The main content area includes sections for Authors (Bañón, Marta; et al.), Item identifier (<http://hdl.handle.net/11356/1838>), Project URL (<https://macocu.eu/>), Referenced by (<https://hdl.handle.net/11370/685514a8-947e-44f9-83cf-90356c5f1684>), Date issued (2023-05-24), Type (corpus, text), and Size (21471613 texts, 6181945683 words). On the right, there's a sidebar with links for BIBTEX, CMDI, CLARIN.SI Data & Tools, and various repository statistics.

## Repository

- ⇒ CTS and CLARIN certified repository of language data and software
- ⇒ 562 entries (3.4T): Slovenian (278), Croatian (74), Serbian (66), Bulgarian (27), Macedonian (17), Montenegrin (9)
- ⇒ Large annotated corpora, training corpora, lexical resources, language models, ELEXIS catalogue of dictionaries

## Concordancers

The screenshot shows the ParlaMint-AT 3.0 interface. At the top, there's a search bar with 'ParlaMint-AT 3.0 (Austrian parliament)'. Below it, a section titled 'SINGLE-WORDS' shows a list of words from a reference corpus: ParlaMint-AT 3.0 (Austrian parliament) (items: 14,950). The list is organized into three columns: Word, Word, and Word. The words listed include 'teuerung', 'gasheizung', 'ukraininer', 'putin', 'totschnig', 'gas', 'putins', 'ukrainisch', 'ukraine', 'hospiz-', 'gasreserve', 'linhart', 'gutschein', 'speicher', 'tursky', 'seelsorge', 'erdgas', 'wasserstoff', 'geosphere', 'orratung', 'russisch', 'vladimir', 'meteorologie', 'ivermectin', 'mindestpensionistin', 'kraus-winkler', and 'elli'. On the left, there's a sidebar with various icons for filtering and searching.

### On-line corpus analysis:

- ⇒ 150 corpora, 35 languages, 28 billion words
- ⇒ Annotations: lemmas, PoS, NER, syntactic dependencies
- ⇒ Several concordancers: KonText, noSketch Engine with(out) log-in
- ⇒ Openly accessible (API!)
- ⇒ Other CLARIN.SI tools: GitLab, WebAnno

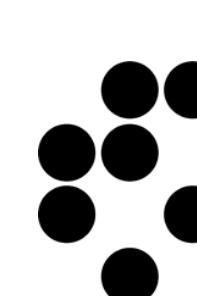
## Engagement

### Community:

- CLARIN.SI consortium: 13 main Slovenian language-related organisations
- Co-organiser of biennial LT & DH conferences, support for other events
- Talks, tutorials, LR development support

### Projects:

- Annual CLARIN.SI supported projects: 24 completed (2018–)
- Implementation of European Cohesion: 500k EUR for equipment (2019–2021)
- RDA Node Slovenia (2020–2021)

 **Jožef Stefan Institute**

*University of Ljubljana*



**ZRC SAZU**



**alpineon))**



inštitut za novejšo zgodovino



**AMEBIS**



**trojína**

