The CLARIN.SI repository

Tomaž Erjavec
Department of Knowledge Technologies, Jožef Stefan Institute
Repository

- Currently the most important service of CLARIN.SI
- Long-term and safe archiving of language resources (https, Nagios)
- Explicit terms of use (terms of service, licence)
- Ethical codex (Code of conduct)
- Most resources in standard encoding (Unicode, XML)
- Certified as CLARIN Centre B
  - Digital Seal of Approval, DSA
  - Currently being (re)certified for Core Trust Seal, CTS
Repository platform

- Based on the DSpace platform, developed for open digital repositories
- DSpace modified for use by CLARIN repositories: CLARIN/DSpace developed by Czech CLARIN
- Used by Czech, Italian, Norwegian, Polish, and Slovenian CLARIN
- Maintenance on GitHub, Slovenian fork on GitLab
Metadata

- Standard encoding of metadata:
  - Component Metadata Infrastructure (CMDI)
  - Dublin Core (DC)
- Metadata always CC0
- Meta-data harvesting:
  - CLARIN Virtual Language Observatory (VLO)
  - also:
Permanent identifiers

- Each repository entry is assigned a PID
- So, even if the repository platform is changed, the PIDs stay the same (but need to be reconfigured)
- Best known solution: DOI
- CLARIN used the Handle system
- http://hdl.handle.net/11356/1044 → https://www.clarin.si/repository/xmlui/handle/11356/1044
- Important for citation of repository items:

Please use the following text to cite this item or export to a predefined format:

Krsnik, Luka; Dobrovoljc, Kaja and Robnik-Šikonja, Marko, 2019, Dependency tree extraction tool STARK 1.0, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1284.
Top level page
Top level page

Description:
The siParl corpus contains minutes of the Assembly of the Republic of Slovenia for 11th legislative period 1990-1992, minutes of the National Assembly of the Republic of Slovenia from the 1st to the 7th legislative period ...

This item contains 6 files (11.43 GB).

Consonant-vowel structures in the GOS 1.0 corpus

Author(s):
Čibej, Jaka ; Arhar Holdt, Špela ; Dobrovoljc, Kaja ; Krek, Simon

Description:
The lists contain consonant-vowel structures of all lemmas, word forms, and normalized word forms in the GOS 1.0 Corpus of Spoken Slovene (http://hdl.handle.net/11356/1040). In each unit, its characters were converted as ...

This item contains 7 files (3.6 MB).

Consonant-vowel structures in the Gigafida 2.0 corpus

Author(s):
Čibej, Jaka ; Arhar Holdt, Špela ; Dobrovoljc, Kaja ; Krek, Simon

Description:
The lists contain consonant-vowel structures of all lemmas and word forms in the Gigafida 2.0 corpus. In each unit, its characters were converted as follows: C - consonant (in lists with finegrained character categorizations, ...
Log-in

- Necessary for making a new repository item, for accessing non-CC items and for using some advanced functions
- For those without EduGain, CLARIN ERIC also gives accounts
How to find interesting resources

- Browsing by language, type of resource, keywords, author, etc.
- Search (fuzzy matching)
- Advanced facet search
- Currently 163 items (without prior versions)

<table>
<thead>
<tr>
<th>Type</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>corpus</td>
<td>78</td>
</tr>
<tr>
<td>languageDescription</td>
<td>2</td>
</tr>
<tr>
<td>lexicalConceptualResource</td>
<td>66</td>
</tr>
<tr>
<td>toolService</td>
<td>16</td>
</tr>
</tbody>
</table>
Landing page of a repository item 1.

Training corpus ssj500k 2.2

Please use the following text to cite this item or export to a predefined format:


This resource is also integrated in the following services:

Authors
Krek, Simon; Dobrovoljc, Kaja; Erjavec, Tomaž; Može, Sara; Ledinek, Nina; Holz, Nanika; Zupan, Katja; Gantar, Polona; Kuzman, Taja; Čibaj, Jaka; Arhar Holdt, Špola; Kavčič, Teja; Škrjanec, Iza; Marko, Dafne; Jezeršek, Lucija; Zajc, Anja

Item identifier
http://hdl.handle.net/11356/1210

Project URL
http://eng.slovenscina.eu/tehnologije/ucni-korpus

Referenced by
http://dx.doi.org/10.18653/v1/W17-1406

Date issued
2019-01-26

Type
corpus

Size
586248 tokens, 27829 sentences, 500295 words

Language(s)
Slovenian
Landing page of a repository item 2

Description
The sij500k training corpus contains about 500,000 tokens manually annotated on the levels of tokenisation, sentence segmentation, morphosyntactic tagging, and lemmatisation. About half of the corpus is also manually annotated with syntactic dependencies, named entities, and verbal multiword expressions. About a quarter of the corpus is annotated with semantic role labels. The morphosyntactic tags and syntactic dependencies are included both in the JOS/MULTEXT-East framework, as well as in the framework of Universal Dependencies.

The vocabulary of (1) and (2) is provided in the back element and (3), (4), and (5) in the teiHeader of the TEI encoded corpus. The semantic role labels are also documented in the teiHeader.

In contrast to the previous version 2.1, this version corrects various errata in spacing and text metadata and adds UD morphological and (where it was possible to do so automatically) dependency annotations to the corpus. Note that the UD annotations are not included in the vertical file.

Publisher
Centre for Language Resources and Technologies, University of Ljubljana

Acknowledgement
Ministry of Education, Science and Sport 3311-08-986003 "Communication in Slovene"
ARRS (Slovenian Research Agency) P2-103 "Knowledge Technologies"
ARRS (Slovenian Research Agency) J6-8256 "New grammar of contemporary standard Slovene: sources and methods"
ARRS (Slovenian Research Agency) MR-37487 "Young Researcher Programme"
ARRS (Slovenian Research Agency) P6-0411 "Language Resources and Technologies for Slovene"

Subject(s) part-of-speech tagging  dependency treebank  parsing  named entities  tokenisation  manual annotation  TEI  verbal multiword expressions  semantic role labelling  CONLL-U
Landing page of a repository item 3

Show full item record

Files in this item

- Download instructions for command line
- Download all files in item (40.95 MB)

This item is Publicly Available and licensed under:
Creative Commons - Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

<table>
<thead>
<tr>
<th>Name</th>
<th>ssj500k.conllu.zip</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>10 MB</td>
</tr>
<tr>
<td>Format</td>
<td>application/zip</td>
</tr>
<tr>
<td>Description</td>
<td>Corpus in CONLL-U format, complete corpus with UD morphology and separately the UD syntactically annotated part, also split into train/dev/test.</td>
</tr>
<tr>
<td>MD5</td>
<td>f65ae2995a2a7acfe43b1a5aa3140dca</td>
</tr>
</tbody>
</table>

Download file   Preview
<table>
<thead>
<tr>
<th>dc.contributor.author</th>
<th>Holz, Nanika</th>
</tr>
</thead>
<tbody>
<tr>
<td>dc.contributor.author</td>
<td>Zupan, Katja</td>
</tr>
<tr>
<td>dc.contributor.author</td>
<td>Gantar, Polona</td>
</tr>
<tr>
<td>dc.contributor.author</td>
<td>Kuzman, Taja</td>
</tr>
<tr>
<td>dc.contributor.author</td>
<td>Čibej, Jaka</td>
</tr>
<tr>
<td>dc.contributor.author</td>
<td>Arhar Holdt, Špela</td>
</tr>
<tr>
<td>dc.contributor.author</td>
<td>Kavčič, Teja</td>
</tr>
<tr>
<td>dc.contributor.author</td>
<td>Škrjanec, Iza</td>
</tr>
<tr>
<td>dc.contributor.author</td>
<td>Marko, Dafne</td>
</tr>
<tr>
<td>dc.contributor.author</td>
<td>Jezeršek, Lucija</td>
</tr>
<tr>
<td>dc.contributor.author</td>
<td>Zajc, Anja</td>
</tr>
<tr>
<td>dc.date.accessioned</td>
<td>2019-01-26T20:37:28Z</td>
</tr>
<tr>
<td>dc.date.available</td>
<td>2019-01-26T20:37:28Z</td>
</tr>
<tr>
<td>dc.date.issued</td>
<td>2019-01-26</td>
</tr>
<tr>
<td>dc.identifier.uri</td>
<td><a href="http://hdl.handle.net/11356/1210">http://hdl.handle.net/11356/1210</a></td>
</tr>
<tr>
<td>dc.description</td>
<td>The ssj500k training corpus contains about 500,000 tokens manually annotated on the levels of tokenisation, sentence segmentation, morphosyntactic tagging, and lemmatisation. About half of the corpus...</td>
</tr>
</tbody>
</table>
Piwik

Get a monthly report for this item via email

Click on a data point to summarize by year / month.

Views **1275**

Downloads **18297**

Statistics for the year 2018
Repository: download

- Most entries available under Creative Commons licences
- Others: AAI login, agree to licence conditions: CLARIN.SI Licence ACA ID-BY-NC-INF-NORED
- Cite the resource!
- Always use the handle, not the URL!

Emoji Sentiment Ranking 1.0

Please use the following text to cite this item or export to a predefined format:

Repository: deposit

- AAI log-in
- Fairly simple workflow:

  Item submission

  1. Basic Info
  2. Who's involved
  3. Describe
  4. Upload
  5. License
  6. Note
  7. Review
  8. Complete

- However: follow best practices, look at other entries!
- Choosing the licence
- Possibility to embargo the data
- Can enter a new version of a resource
- Option to share item editing
- The entry is checked by one of the CLARIN.SI editors
Data Formats

- Media files: wav, avi, jpg, png…
- Data with tabular structure: TSV, CSV, XML
- Hierarchical data: XML
  - using one of the „standard“ schemas, e.g. TEI, …, TRS, ELAN
  - using your own schema
  - in all cases: schema + documentation must be part of the entry
- + Derived formats + Documentation (can be PDF)
- Deposit as ZIP
What is in the repository?

- Language models, training data, lexicons for tagging, lemmatisation, syntax of Slovene, Croatian, Serbian, for standard and non-standard language
- Various types of word embeddings
- Large annotated corpora of various text types
- Speech corpora
- Multilingual corpora
- Machine readable dictionaries
- Some software
Tools

• “Language models, training data, lexicons for tagging, lemmatisation, syntax of Slovene, Croatian, Serbian, for standard and non-standard language”
• tokenisers, morphosyntactic taggers for sl, hbs
• parsers (also UD-PIPE)
• Viewers and editors for linguistic annotation
CLARINSI@GitHub

clarin-dspace
Forked from ufal/clarin-dspace
LINDAT/CLARIN digital repository based on DSpace
• Java 942  ★ 0  ⌂ 8  Ⓛ 0  Updated 6 days ago

mte-msd
MULTTEXT-East morphosyntactic specifications
• HTML 0  ★ 1  ⌂ 0  Ⓛ 0  Updated 21 days ago

babushka-bench
Benchmarking NLP tools on Slovene, Croatian and Serbian
• Python 1  ★ 2  ⌂ 1  Ⓛ 0  Updated on Mar 19

classia-stanfordnlp
Forked from stanfordnlp/stanza
CLASSIA Fork of the Official Stanford NLP Python Library for Many Human Languages
• Python 505  ★ 2  ⌂ 0  Ⓛ 0  Updated on Mar 19

reldi-tokeniser