Raziskovalna infrastruktura za jezikovne vire in tehnologije CLARIN.SI

Dr. Tomaž ERJAVEC, Odsek za tehnologije znanja, Institut „Jožef Stefan“,
Inštitut za slovenski jezik Frana Ramovša, ZRC SAZU

Ljubljana in online, Centralna tehniška knjižnica Univerze v Ljubljani,
Usposabljanje podatkovnih strokovnjakov, 14. – 17. 10. 2024

Financira Evropska unija NextGenerationEU
NOO NAČRT ZA OKREVANJE IN ODPORNOST
REPUBLIKA SLOVENIJA MINISTRSTVO ZA VISOKO ŠOLSTVO ZNANOST ON INOVACIJE
ctk
SPOZNAJ
PODPORA PRI UVAJANJU NAČEL ODPRTE ZNANOSTI V SLOVENIJI

Kdo potrebuje jezikovne vire, tj. digitalne jezikovne podatke?

- **Humanistične vede**
- *Empirično podprte jeziko(slo)vne raziskave:*
- temeljijo na realnih besedilih
- za učinkovito uporabo potrebujemo analitična orodja
- jezikoslovci, slovaropisci, zgodovinarji, družboslovci
- občanska znanost

- **Računalništvo**
- *Jezikovne tehnologije:*
- obdelava jezika postaja vedno bolj raziskovalno/komercialno zanimivo področje
- glavna paradigma: nadzorovano strojno učenje
- programi so večinoma jezikovno neodvisni, potrebujejo pa učne podatke za učenje modelov in testne podatke za evalvacijo

CLARIN.SI

SPOZNAJ
PODPORA PRI UVAJANJU NAČEL ODPRTE ZNANOSTI V SLOVENIJI

- Začetek dela v 2015
- CLARIN.SI je članica evropske RI CLARIN (Common Language Resources and Technology Infrastructure)
- Sedež na Institutu ”Jožef Stefan”
- CLARIN.SI je organiziran kot konzorcij 12 partnerjev:
  - univerze: Ljubljana, Maribor, Nova Gorica, Primorska
  - raziskovalni inštituti: ZRC SAZU, IJS, INZ, ZRS Koper
  - knjižnica: NUK
  - podjetji: Amebis, Alpineon
  - društvo: Slovensko društvo za jezikovne tehnologije
- Trije stebri delovanja:
  1. **repozitorij jezikovnih virov**
  2. konkordančniki in druge spletne storitve
  3. podpora digitalni humanistiki in jezikovnim tehnologijam

Repozitorij CLARIN.SI

- Arhiv trenutno ~600 jezikovnih virov in orodij (~300 (tudi) za slovenščino),
  delo ~1000 avtorjev (3.6T podatkov)
- Predvsem korpusi, besedišča, slovarji, modeli, programi
- Večina vnosov pod eno od licenc Creative Commons, obstajajo tudi bolj restriktivne licence (te zahtevajo prijavo prek EduGain)
- **Samoarhiviranje + uredniški pregled**
- Repozitorij certificiran s strani CLARIN in Core Trust Seal
- Dolgotrajno hranjenje, avtentikacija in avtorizacija, stalni identifikatorji, eksplicitni pogoji uporabe, bogat nabor licenc
- Zajem metapodatkov: CLARIN VLO, OpenAIRE, ELG
- Pomemben doprinos k odprti znanost

Primer vnosa

SPOZNAJ
PODPORA PRI UVAJANJU NAČEL ODPRTE ZNANOSTI V SLOVENIJI

https://www.clarin.si/repository/xmlui/handle/11356/1945

Search

CLARIN.SI Repozitorij O repozitoriju Kontakt CLARIN Prijava

Repozitorij CLARIN.SI / Prikaz vnosa

Išči

Corpus of daily jokes from the 24ur.com portal Šale24 1.0

Za citiranje vnosa uporabite naslednjo referenco ali jo izvozite v prednastavljeno obliko: BIBTEX CMDI

Dobranić, Filip, 2024, *Corpus of daily jokes from the 24ur.com portal Šale24 1.0*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1945.

Delite:

CLARIN.SI Data & Tools

Avtorji	Dobranić, Filip
Identifikator vnosa	http://hdl.handle.net/11356/1945
URL projekta	https://dihur.si/muki/humor
Datum objave	2024-10-03
Vrsta	corpus, text
Velikost	16658 sentences, 129063 tokens, 1915 texts

CLARIN.SI

Brskanje
› Celoten repozitorij
Moj račun
Prijava
Statistika
Statistika Piwik BETA
Splošne informacije
O vnosu v repozitorij
Citiranje
Življenjski ciklus vnosa

Jezik(i) Slovenian

Opis

This is a corpus of 1915 "jokes of the day" ("šala dneva") published by the Slovenian news portal 24ur.com. The jokes were scraped from their archive on September 18th, 2024. The initial list is lightly curated: shorter texts found in the original collection were removed from the corpus since they appear to be illustration captions without the accompanying illustrations.

Readers of the news portal vote on the jokes themselves with thumbs up and thumbs down buttons. The voting results are included as metadata with each joke. Several jokes have been published more than once. Each joke (distinguished based on exact text matches) is identified by a hash of its text and presents a list of voting results for every instance of its publication. The normalised_text field contains text with punctuation corrections. For now, this is limited to replacing '' (two consecutive apostrophes U+0027) with " (a single straight/dumb/vertical quotation mark U+0022). The former (two apostrophes) is consistently used in place of the latter in the original corpus.

Based on the name ("Šala dneva" i.e. "Joke of the day") and observed frequency of posting during September 2024 we assume each entry corresponds to a day starting from the day of data collection counting backwards. Each voting event for has an associated estimated publication date calculated with the above algorithm.

The jokes are linguistically annotated with CLASSLA-Stanza (https://github.com/clarinsi/classla), using the models for standard Slovenian. The JSONL file contains entries representing individual jokes containing:
- a hash of the original joke text used for duplicate identification (key: hash)
- original scraped text (key: original_text)
- normalised text (key: normalised_text)
- linguistically annotated normalised text in CoNLL-U format (key: processed_text)
- a list of vote objects containing joke vote metadata (key: votes)
- votes for (key: votes.for)
- votes against (key: votes.against)

Izdajatelj	Institute of Contemporary History
Ključne besede	jokes
Zbirke	CLARIN.SI data & tools

Prikaži polni zapis vnosa

Datoteke v tem vnosu

Prenesi navodila za ukazno vrstico

Ime	sale_annotated.zip
Velikost	1.7 MB
Format	application/zip
Opis	Compressed jsonl file with 1915 jokes and metadata
MD5	589039ff5ed4a68e79e2f5a42afff1f7

Prenesi datoteko | Predogled

Natančna navodila za vnos metapodatkov in podatkov

SPOZNAJ
PODPORA PRI UVAJANJU NAČEL ODPRTE ZNANOSTI V SLOVENIJI

Zaključki

- CLARIN.SI nudi možnost trajnega arhiviranja jezikovnih virov, odprt in brezplačen dostop do jezikovnih virov, orodij in storitev za (slovenske) raziskovalce in (kjer le mogoče) podjetja ter podporo pri ustvarjanju, arhiviranju in uporabi jezikovnih virov in orodij.

- Nadaljnje informacije:
- https://www.clarin.si/
- ERJAVEC, Tomaž, DOBROVOLJC, Kaja, FIŠER, Darja, JAVORŠEK, Jan Jona, KREK, Simon, KUZMAN, Taja, LASKOWSKI, Cyprian Adam, LJUBEŠIĆ, Nikola, MEDEN, Katja. **Raziskovalna infrastruktura CLARIN.S**I. *Jezikovne tehnologije in digitalna humanistika : zbornik konference. 2022*. str. 47-54. https://nl.ijs.si/jtdh22/pdf/JTDH2022_Erjavec-et-al_Raziskovalna-infrastruktura-CLARIN.SI.pdf

SPOZNAJ

PODPORA PRI UVAJANJU NAČEL ODPRTE ZNANOSTI V SLOVENIJI

Hvala za pozornost!