

Research lifecycle in the EOSC cloud using CLARIN technology / Ikerketa prozesua EOSC lainoan CLARIN teknologiak erabilita.

Proposer(s) / Proposatzailea(k):

Mikel Iruskietea

Contact / Kontaktua:

mikel.iruskietea@ehu.eus

Description / Deskribapena

The student can choose a hot topic in Digital Humanities and follow the following use-case with her/his research topic:

Searching for tools & processing the dataset

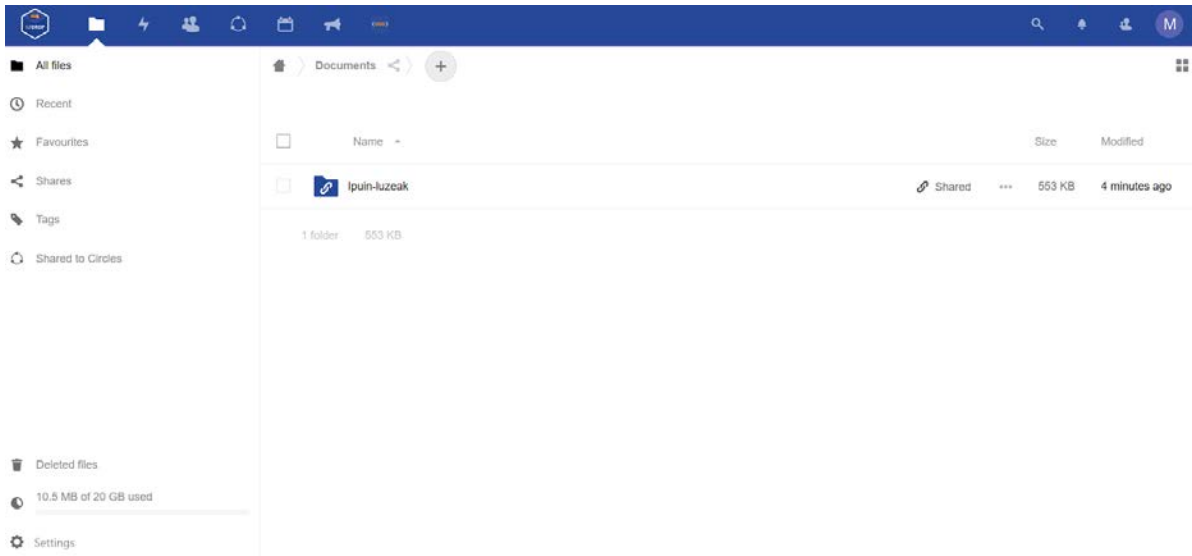
1. EOSC portal search

- At the [EOSC Portal market place](#), the researcher searches for "[language analysis](#)". This leads to 1 result: the Language Resource Switchboard.
- In the description of the [Language Resource Switchboard](#), several relevant analysis methods are listed (e.g. Topic Modelling and Stylometry). It also states it can be invoked from [B2DROP](#).
- Researcher goes to the [B2DROP](#) landing page in the EOSC-portal market place and goes to the [actual application](#).

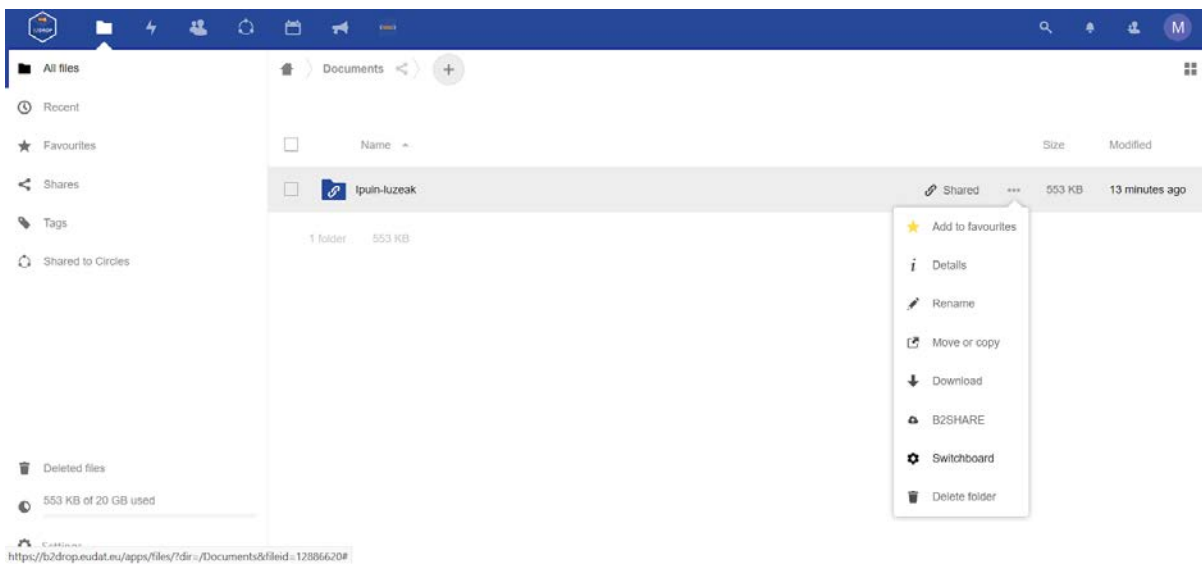


2. B2DROP file upload

- The researcher uses single-sign-on (B2ACCESS) to login to his B2DROP workspace.
- There she uploads the dataset and shares the file with a **Share Link**.
 - Basque tales (from 3 to 6 years): <https://b2drop.eudat.eu/s/naW8Hcaadmaf8ds>




- Now he clicks on the ... icon next to the file and selects **Switchboard**.



3. Language Resource Switchboard

- After being redirected to the Language Resource Switchboard, she indicates that the input file contains BASQUE data.


Language Resource Switchboard Upload Tool Inventory Help CLARIN 

Resource: Mediatype: Language:

Matching Tools

Group by task


▼ Distant Reading

>


Start Tool

Voyant Tools [↗](#)

▼ Text Analytics


>


Start Tool

WebLicht Advanced Mode [↗](#)


About v2.1.2
Service provided by CLARIN
Contact


- Then he clicks on **Show Tools**.
- This combination results in 1 available tool, WEBLIHGT. She clicks on this tool to see more details.
- Now he invokes WEBLIHGT via the **Click to start tool** button.

Sign in via the CLARIN Service Provider Federation CLARIN 


Select your home organisation below. This is usually the organisation where you work or study. Signing in here will allow you to access certain CLARIN resources and services which are only available to users who have logged in. If you cannot find your organisation in the list below, please select the clarin.eu website account and use your CLARIN website credentials. If you don't have such credentials you can register an account here. For questions please contact spl@clarin.eu.


Previously chosen home organisation



clarin.eu website account




Home organisation list


clarin.eu website account




AAI@EduHr Single Sign-On Service



Aalborg University

UDPipe

About Run REST API Documentation

UDPipe is a trainable pipeline for tokenization, tagging, lemmatization and dependency parsing of CoNLL-U files. UDPipe is language-agnostic and can be trained given annotated data in CoNLL-U format. Trained models are provided for nearly all UD treebanks. UDPipe is available as a binary for Linux/Windows/OS X, as a library for C++, Python, Perl, Java, C#, and as a web service. Third-party R CRAN package also exists.

UDPipe is a free software distributed under the Mozilla Public License 2.0 and the linguistic models are free for non-commercial use and distributed under the CC BY-NC-SA license, although for some models the original data used to create the model may impose additional licensing conditions. UDPipe is versioned using Semantic Versioning.

Copyright 2017 by Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Czech Republic.

Description of the available methods is available in the API Documentation and the models are described in the UDPipe User's Manual.

Service

The service is freely available for testing. Respect the CC BY-NC-SA licence of the models – **explicit written permission of the authors is required for any commercial exploitation of the system.** If you use the service, you agree that data obtained by us during such use can be used for further improvements of the systems at UFAL. All comments and reactions are welcome.

Model: UD 2.5 (description) UD 2.4 (description) UD 2.0 (description) UD 1.2 (description)

basque-bdt-ud-2.5-191206

Actions: Tag and Lemmatize Parse

Process Input

Output Text

Show Table

Show Trees

Save Output File

```
# newdoc
# newpar
# sent_id = 1
# text = BABARRUN ALE MAGIKOAK
1 BABARRUN Babarrun PROPN _ _ 3 nmod _ SpacesBefore='\n'
2 ALE ale PROPN _ _ 1 flat _ _
3 MAGIKOAK MAGIKOAK PROPN _ _ Case=Erg(Definite=Def)Number=Sing 0 root _ SpacesAfter='\n\n'

# newpar
# sent_id = 2
# text = Andoni izeneko mutiko bat baserrian bizi zen bere amarekin bakar-bakarrik.
1 Andoni Andoni PROPN _ Case=Dat(Definite=Def)Number=Sing 6 iobj _ _
2 izeneko izen NOUN _ _ 3 nmod _ _
3 mutiko mutiko NOUN _ _ 6 nsubj _ _
4 bat bat NUM _ NumType=Card 3 nummod _ _
5 baserrian baserri NOUN _ Animacy=Inan(Case=Ine)Definite=Def)Number=Sing 6 obl _ _
6 bizi bizi ADJ _ Case=Abs(Definite=Ind 0 root _ _
7 zen izan AUX _ Aspect=Prog(Mood=Ind)Number(abs)=Sing(Person(abs)=3 6 aux _ _
8 bere bera DET _ Case=Gen)Number=Sing 9 nmod _ _
9 amarekin ama NOUN _ Case=Com(Definite=Def)Number=Sing 6 obl _ _
10 bakar-bakarrik bakar-bakarrik ADV _ _ 6 advmod _ SpacesAfter=No
11 . . PUNCT _ _ 6 punct _ _
```

Process Input

Output Text

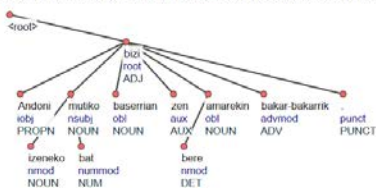
Show Table

Show Trees

Save Tree as SVG

Previous 1 2 3 4 5 6 7 8 9 10 11 12 ... Next

Andoni izeneko mutiko bat baserrian bizi zen bere amarekin bakar-bakarrik .



View Tool List **HELP/FAQ**

Main Page Chain 1 x + New Chain

Show tools with status: development production superseded withdrawn

Next Choices (Double-click on an icon to add it to the chain)

Input and Chain Selection

Run Tools Clear Results Download chain

Title (Plain Text)	Char: UDPipe tokenizer	Char: UDPipe tagger	Char: UDPipe parser
BASARRUN ALE MAGKOKAK Andoni izeneko mutiko bat baserrian bizi zen bere	Document Type: CONLL-U conllu format conllu minor Language: Basque	conllu lemmas conllu upostags conllu xpostags conllu feats	conllu heads conllu displs

Calling UDPipe tagger ...

TÜNDRA FileBank_ee3a1b14-d284-4579-994a-b965c61aabe7 Treebanks Tutorial About Old TÜNDRA v1 CLARIN-D **HELP/FAQ**

Query

Enter either a TIGERSearch query, or simply a word in quotation marks.

History Query Language Help

Sentence 2 out of 289

Andoni izeneko mutiko bat baserrian bizi zen bere amarekin bakar-bakarrik .

Visualization

The visualization shows a dependency parse tree for the sentence. The root node is 'root', which connects to 'bizi' (ADJ). Other nodes include 'Andoni' (PROPN), 'izeneko' (NOUN), 'mutiko' (NOUN), 'bat' (NUM), 'baserrian' (NOUN), 'zen' (AUX), 'bere' (DET), 'amarekin' (NOUN), and 'bakar-bakarrik' (ADV). Relations shown include iobj, nsubj, nmod, nummod, obl, aux, advmod, and punct.

5. Publishing the results in B2SHARE



- At the [EOSC-portal market place](#), the researcher searches for “Store and publish research data”. Then he finds [B2SHARE](#) as potential publication platform.
- From there he navigates to [B2SHARE](#) and uses single-sign on to authenticate. Since he authenticated to B2DROP before, it is not necessary anymore to enter a username and password.

- He clicks on the **Create a new record** button, enters a title, selects the CLARIN community and finally clicks on **Create Draft Record**.
- After entering the necessary metadata, he checks the **Submit draft for publication** checkbox and clicks on the **Save and Publish** button. This will make the dataset [available via a persistent identifier](#).
- Her submissions will be findable in the [Virtual Language Observatory](#) and [B2FIND](#) within a day.

The screenshot shows the EUDAT website interface. At the top, there is a navigation bar with 'GO TO EUDAT WEBSITE', 'EUDAT', and 'B2SHARE' logos, a search bar, and a user profile for 'mikel.iruskietia@gmail.com'. Below the navigation, the record title 'Haur Hezkuntzako ipuin-bilduma' is displayed, along with the author 'Iruskietia, Mikel' and the date 'Mar 14, 2020'. The description, disciplines, and keywords are listed. A 'Files' section shows a file named '01-3U-1T-Arroxali-LUZEa.txt' with a size of 1.36KB. The 'Basic metadata' section indicates 'Open Access' is 'True' and the license is 'Creative Commons Attribution-ShareAlike (CC-BY-SA)'.

Alternative data publication option

Many CLARIN centres are providing [depositing services](#).

Goals / Helburuak

Study and carry out how CLARIN tools can be used to perform the complete research lifecycle in the cloud using EOSC services.

Requirements / Betebeharrak

Computer scientists
or linguist

Framework / Esparrua

In collaboration with CLARIN and IXA-CLARIN-k.

Tasks and plan / Atazak eta plana

1. Study and understand how EOSC and CLARIN works:
2. Design a small research lifecycle using CLARIN data and technologies.
3. Compare the results with other methods and tools.

References:

- Krauwier, S., & Hinrichs, E. (2014). The CLARIN research infrastructure: resources and tools for e-humanities scholars. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014) (pp. 1525-1531). European Language Resources Association (ELRA).
- Bassett, S., Wessels, L., Krauwier, S., Maegaard, B., Hollander, H., Admiraal, F., ... & Uiterwaal, F. (2019, March). Connecting the Humanities through Research Infrastructures.
<https://www.clarin.eu/eosc>