

Annotating corpora with WebAnno

Darja Fišer

Filozofska fakulteta Univerze v Ljubljani

Beograd, 10. 12. 2015

Outline

1. Lecture

- tool characteristics
- examples of use
- accessing the tool

2. Demo

- annotation
- curation
- project setup

LECTURE

Why annotate?

1. linguistic analysis

- analysis in context
- quantitative analysis
- theory validation
- feedback on problematic cases

2. language technologies

- testing set for annotation tools
- training set for statistical annotation tools
- automatic generation of lexicons

3. practical benefits

- searching by categories
- illustration of examples
- long-term data usability

WebAnno

<https://webanno.github.io/>

- a flexible web-based tool for (linguistic) annotation of text corpora
- developed as part of CLARIN-D (F-AG 7):
- 2012-2013: Implementation of a web-based annotation platform
- 2012-2013: Development of guidelines and best practices for annotation of non-standard varieties of German
- 2014-2016: Semantic Annotation for Digital Humanities (**WebAnno 3**)

Main features

1. Flexible

- multi-level annotation
- configurable annotation layers
- different annotation modes
- annotation of unannotated and annotated corpora

2. Team tool

- one or more annotators of the same text / corpus
- supports different roles

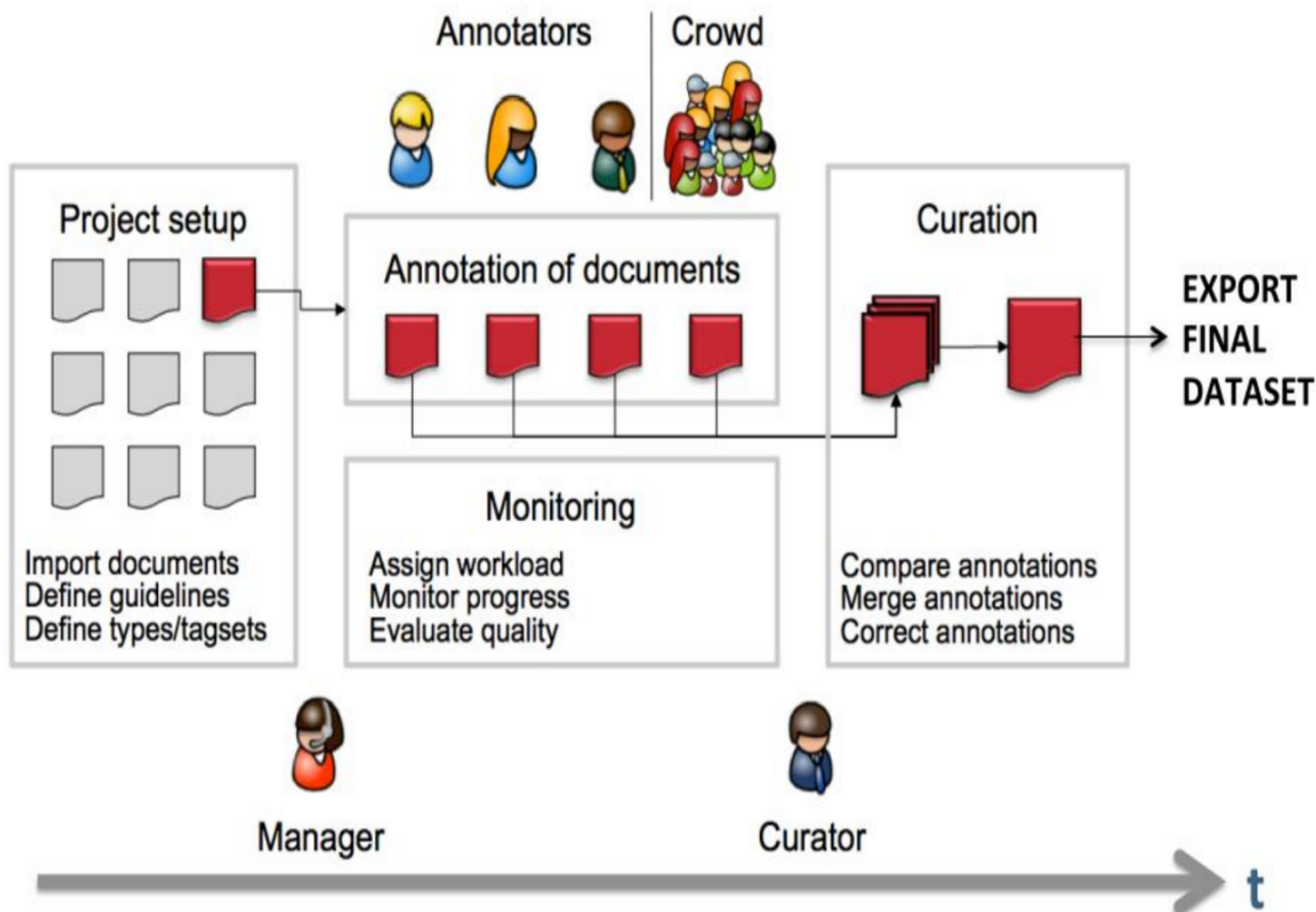
3. Web-based

- easy to access, no installation needed
- immediate storage in the backend database

4. Open source

- community participation
- good support

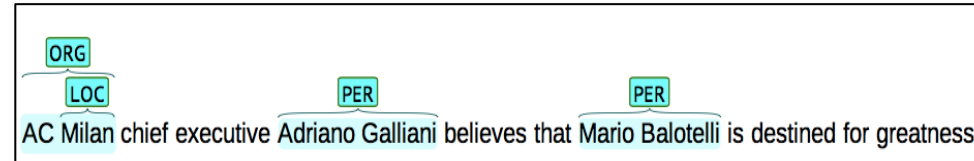
Annotation workflow



Types of annotation layers

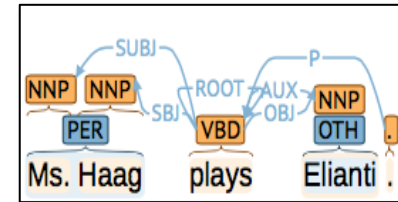
1. spans

- annotation
- strings of character or tokens
- e.g. parts-of-speech, morphological features, multi-word units, named entities



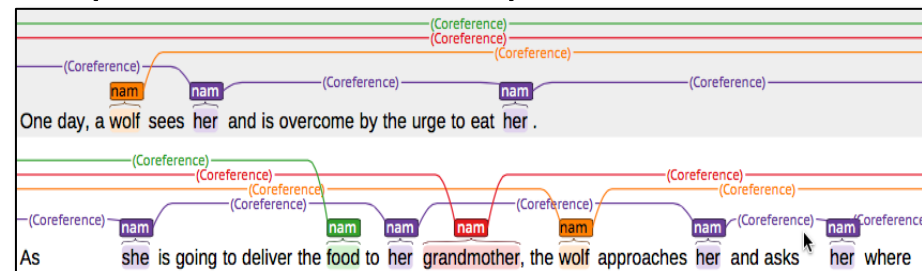
2. relations

- directed links between two spans (directed arcs)
- e.g. dependency syntax, valency, semantic roles



3. chains

- undirected links between two spans (undirected arcs)
- e.g. co-reference



Example: Named entities

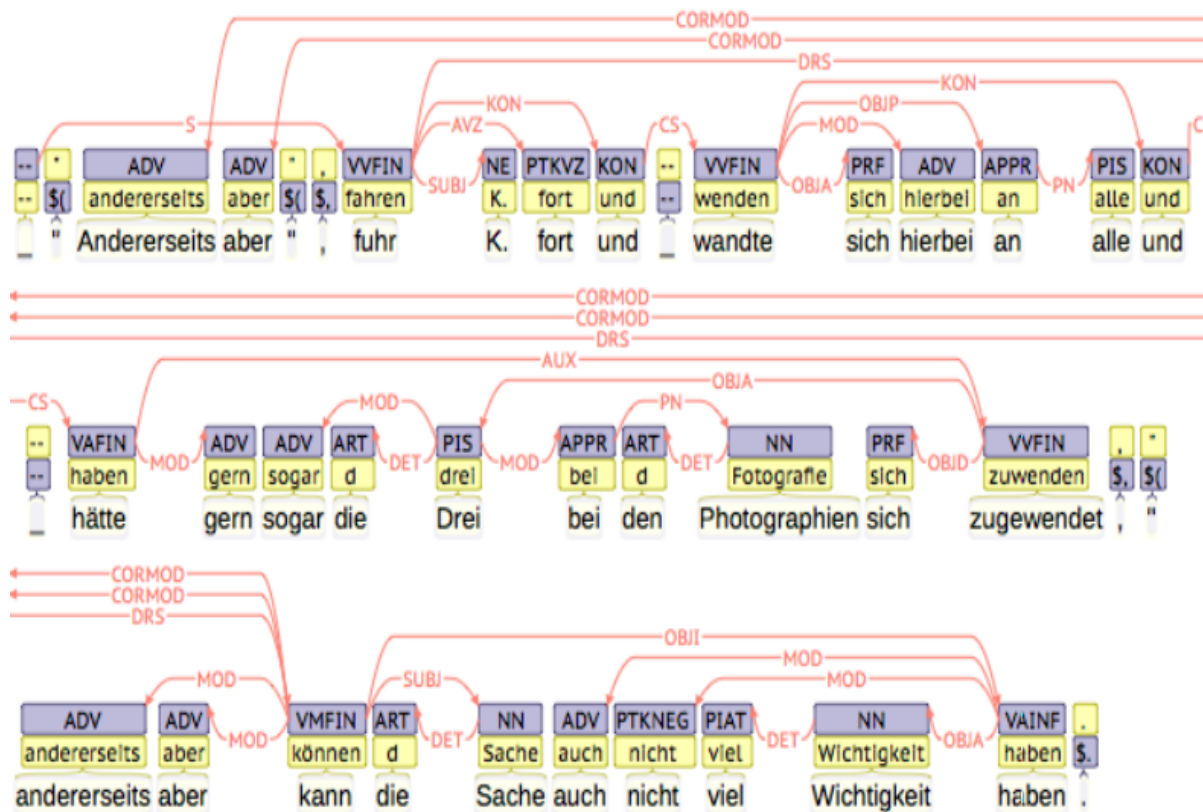
showing 6-15 of 500 sentences

Annotation

6	Am Eingang ins Areal an der <u>Freisinger Straße</u> soll ein etwa 15 Meter hoher Neubau mit drei Vollgeschossen an
7	Diese Struktur ermöglicht ein Verfeinern des reversen Adressraums in mehreren Schritten .
8	Der Taifun schwächte sich bei der Annäherung auf das <u>chinesische Festland</u> ab ; die <u>JMA</u> stufte <u>Nuri</u> aber erst tropischen Sturm zurück .
9	Bundespräsident <u>Horst Köhler</u> würdigte den Verstorbenen als " vorbildlichen Diener seiner Kirche und Christen von einzigartiger Glaub
10	Es ermöglicht das Sammeln und Einlösen von Bonuspunkten .
11	Unterschiedliche Orientierungen der Polarisationsfilter in den Brillengläsern sorgen dafür , dass jedes Auge nur sein Bild sieht und das
12	Da kamen beim Scheunenfest boshafte Charakterzüge ans Licht , wo man sie am wenigsten erwartet hätte .
13	<u>Georg Fischer</u> wird in <u>Kunshan</u> , <u>China</u> , eine Eisengiesserei errichten .
14	Als gesichert gilt wenigstens , dass sie im Consort gespielt <u>Otterstedt</u> , S. 121 bzw. um die Wende zum 17. Jahrhundert dort als beglei
15	Vor allem der bislang als beinahe sicherer Nachfolger geltende Schatzkanzler <u>Gordon Brown</u> gerät zunehmend unter Beschuss .

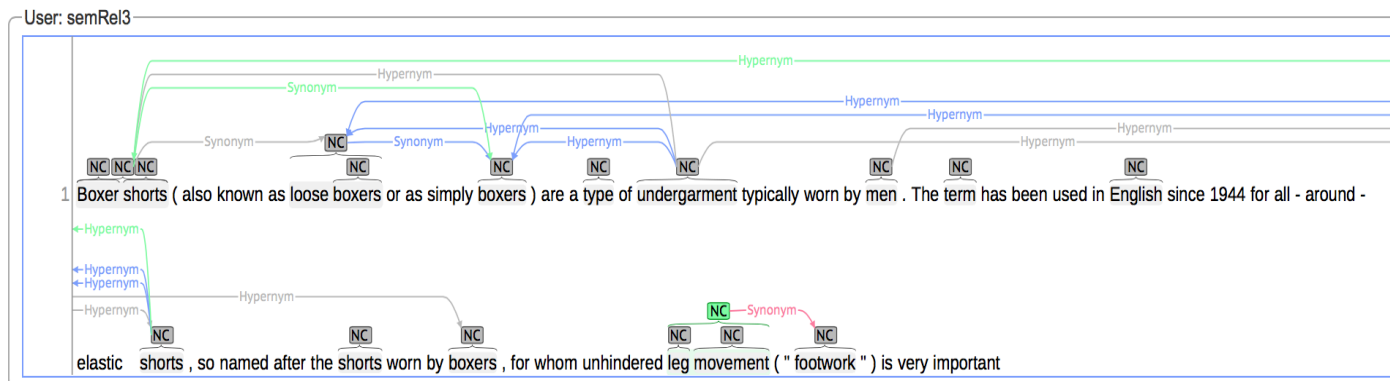
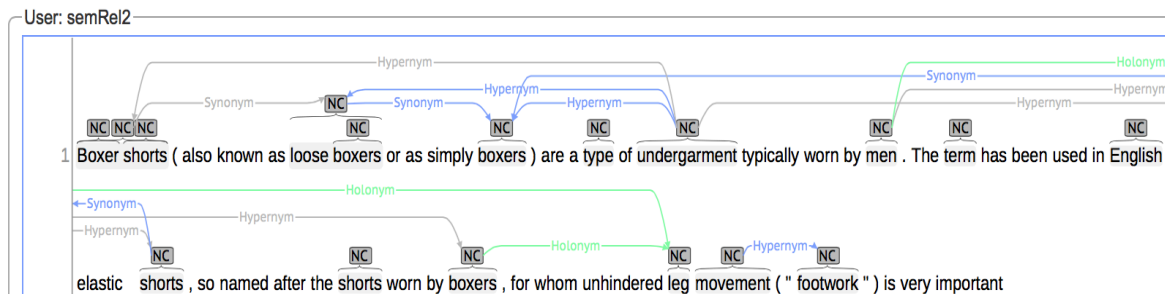
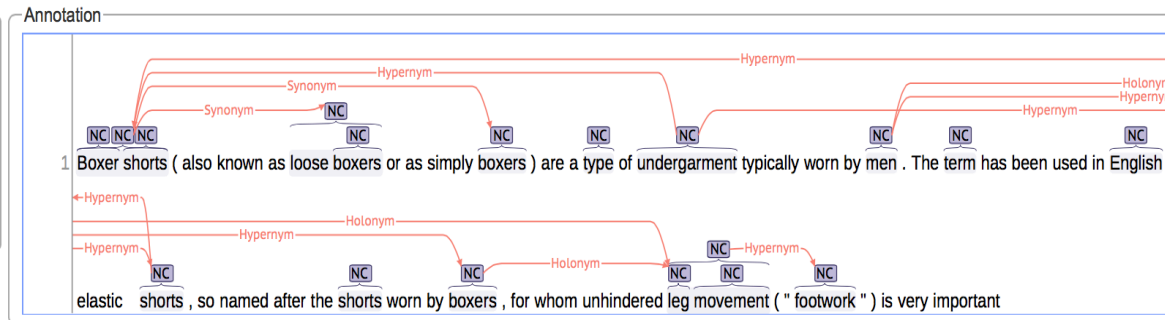
Benikova et al. (2014):
**NoSta-D Named Entity
Annotation for German:
Guidelines and Dataset.**
Proceedings LREC 2014,
Reykjavik, Iceland

Example: Nonstandard German



Dipper et al. NoSta-D: A corpus of German non-standard varieties. Non-Standard Data Sources in Corpus-Based Research (2013): 69-76.

Examples: Semantic annotation



Benikova 2015:
**Multilingual Contextual
Annotation and Analysis
of Classical Semantic
Relations between
Nominals.** MA Thesis,
Ling/LT, TU Darmstadt

Examples: Job ad parsing

showing 2-2 of 220 sentences

2 company Domsel - Consulting - Stellenangebote ad_title Software Entwickler VB (m / w) Mit neuen Ideen zum Ziel Als Personalberatung finden und vermitteln wir seit

Mitarbeiter in Festanstellung für unsere Kunden aus den Branchen Informationstechnologie und Vertrieb . Für unsere Kunden suchen wir engagierte u talent
 deren Fokus Sicherheit , Perspektiven und Innovation stehen . Unser Auftraggeber ist ein international tätiges Beratungsunternehmen , das vor allem
 Unternehmensführung und Logistik Maßstäbe setzt . Als Kompetenzunternehmen stellt es eine Schnittstelle zwischen öffentlicher Hand und Markt d
 Kunden im Vergabe - und Vertragsmanagement und beim Beschaffungsprozess von Leistungen und Waren . Durch optimale Ressourcenplanung und
 Geschäftsprozessoptimierung ist unser Auftraggeber ein Garant für den betrieblichen Erfolg seiner Kunden . Für die weitere Expansion des Unternehm

start_date zum nächstmöglichen Zeitpunkt im location_other_region Rhein - Neckar - Raum eine / n ad_title Software Entwickler (m / w) . Ihr neues Aufgabengebiet sieht wie folgt aus :

task Entwicklung einer Rich Internet Application mit Datenbankanbindung task Entwicklung von Desktop - und C / S - Anwendungen

task Entwicklung individueller Anwendungen (Point Solutions) skill_name Programmiersprachen : skill_name Visual Basic , skill_name Flex3 / skill_name Action script 3 , skill_name T - SQL

task Eigenständiges Arbeiten in einem internationalen Team mit Anwendungsberatern und Methodenspezialisten necessity Idealerweise verfügen Sie über : eine education_level qualifizierte Ausbildung im

education IT - Bereich oder eine education_level technische Ausbildung level sicherer Umgang mit skill_name Visual Basic , skill_name MS Office level gute Kenntnisse in skill_name MS VisualStudio (skill_name VB) , skill_name ASP.NET und

skill_name MS SQL Server 2005 level Praxiserfahrungen mit skill_name Adobe Flex3 necessity sind von Vorteil Unser Auftraggeber bietet Ihnen : Eine abwechslungsreiche und anspruchsvolle Tätigkeit in

position_type Festanstellung in einem interdisziplinären und internationalen Team mit Anwendungsberatern und Methodenspezialisten . Neben einer guten Dotierung bietet Ihnen unser

Auftraggeber außerdem ein flexibles position_type Arbeitszeitmodell , in dem Sie sich mit neuen Ideen talent eigenständig und talent lösungsorientiert einbringen können . Ihr Arbeitsplatz befindet sich in

zentraler Lage direkt am Hauptbahnhof location_city Mannheim . Wenn Sie eine talent Leidenschaft für IT - Themen haben , dann freuen wir uns auf Ihre vollständigen Bewerbungsunterlagen per E -
 Mail unter Angabe Ihrer Gehaltsvorstellung und Ihres möglichen Eintrittstermins an unsere E - Mail Adresse oder über unsere Homepage . Für Fragen zum Unternehmen oder zur

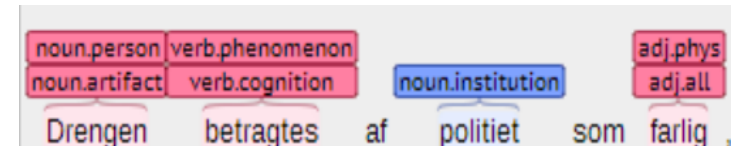
contact_person_name Aufgabe steht Ihnen contact_tel Herr Domsel contact_tel gerne unter der Telefonnummer 06241-985211 jederzeit zur Verfügung . Wir freuen uns auf Ihre Anfragen . company Domsel Consulting -

contact_tel Flatenstraße 11 – 68623 Lambertheim – 06241-985-211 – contact_url www.domsel.de company Bewerben Sie sich hier Vorname : Nachname : E - Mail : Laden Sie hier Ihren Lebenslauf hoch
 Bewerbungsanschreiben

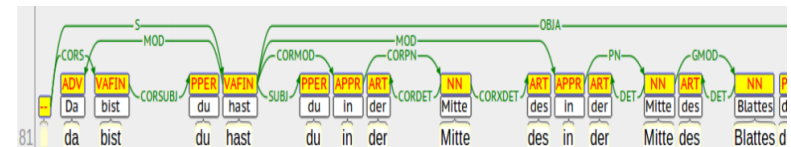
Tobias Krönke (2013): Job Ad
 Parsing for Applicant Qualification
 Prediction, MA Thesis, LT, TU
 Darmstadt, Sept. 2013

Some more examples

- **semantic** Annotation of the Danish CLARIN Reference Corpus (Pedersen et al. 2014)



- various levels of **spoken language** annotation (German, Lüdelling 2013)

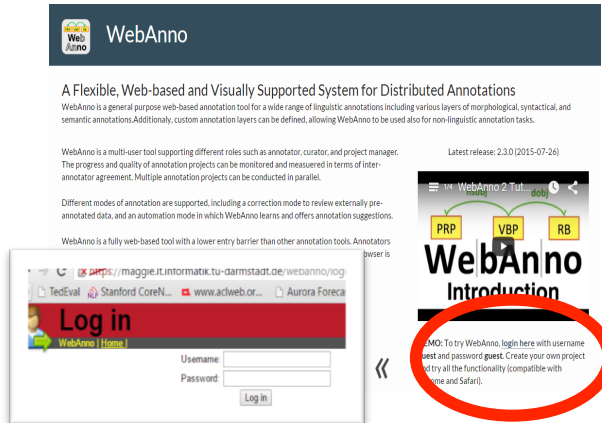


- **multi-word expressions** in online helpdesk interaction (Portuguese, Querido et al. 2015)
- **dialogue act** annotation for Twitter conversations (German, Zarisheva and Scheffer 2015)
- **biomedical entity** recognition (English, Yimam et al. 2015)

Accessing WebAnno

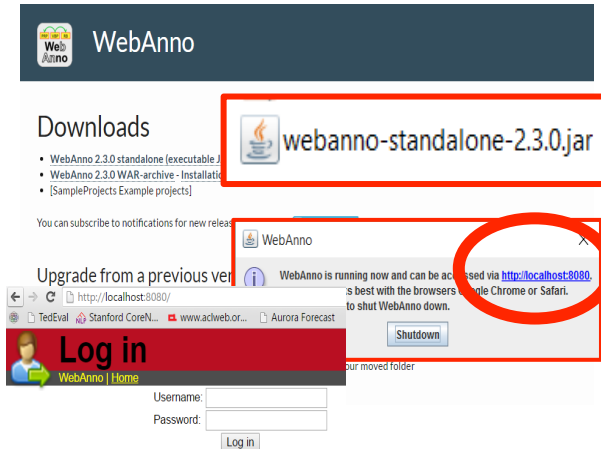
- Chrome or Safari browser
- 3 freely available installations

DEMO
(TU Darmstadt)
guest / guest



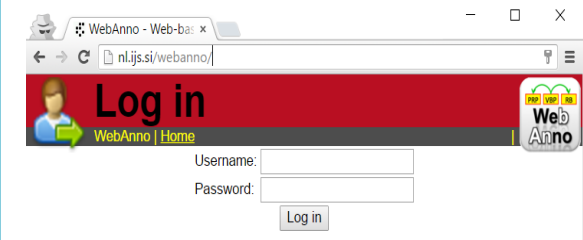
testing only

STANDALONE
(downloadable)
admin / admin



individual work

ReLDI
nljubesi@gmail.com



stable, individual
or team work,
technical support by
ReLDI

DEMO

Main menu



[Annotation](#)

- Annotate from scratch or correct previously annotated documents [mode 1]



[Curation](#)

- Compare and merge annotations from different annotators



[Correction](#)

- Check and correct previously annotated documents [mode 2]



[Automation](#)

- Annotation with simultaneous machine learning [mode 3]



[Projects](#)

- Create new projects



[Monitoring](#)

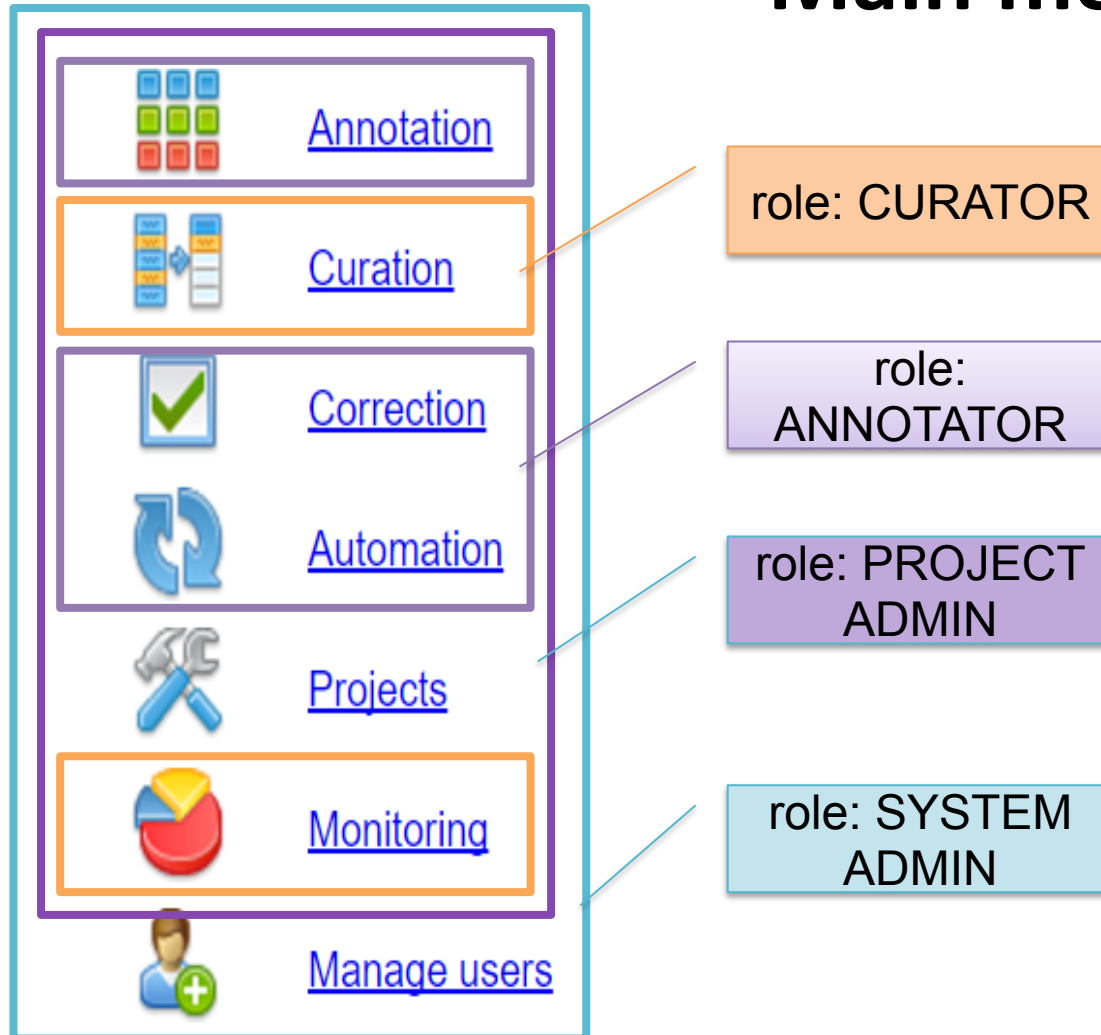
- Assign workload and monitor progress



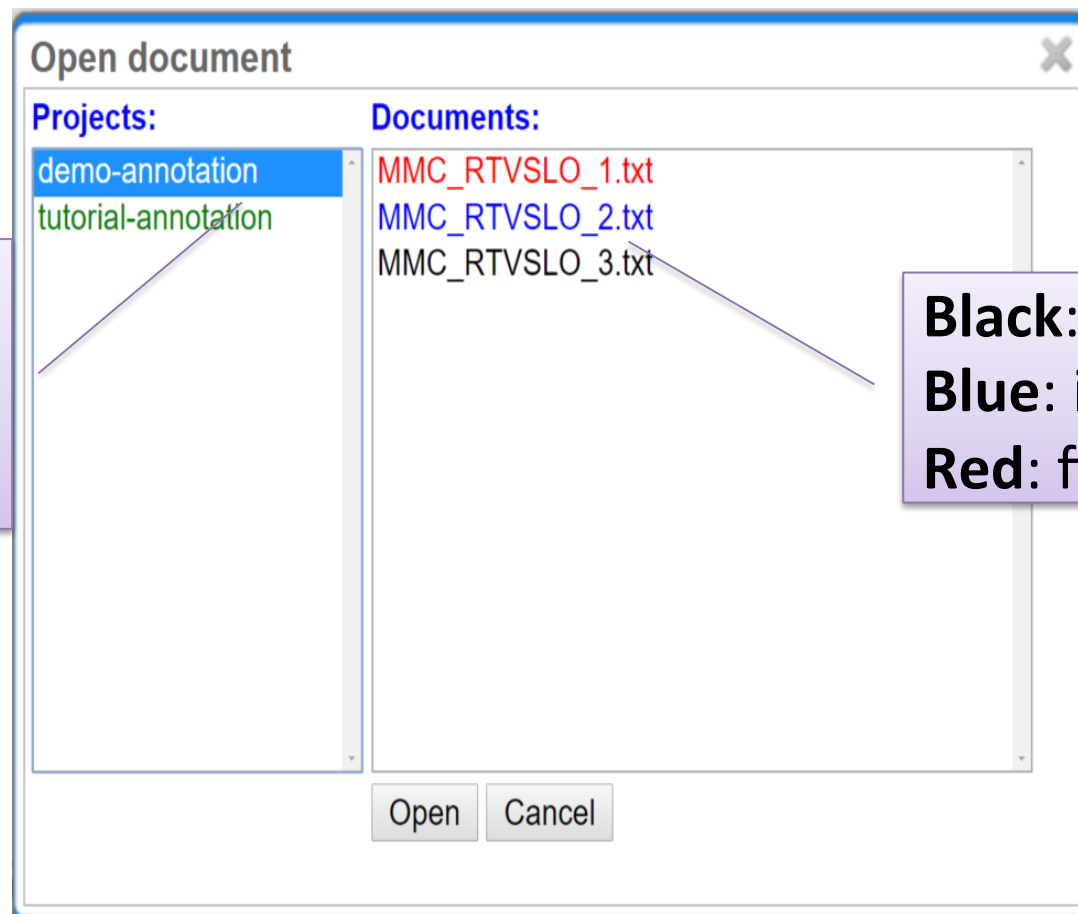
[Manage users](#)

- Create new user accounts

Main menu



Open document



Choose the project and the document for annotation.

Black: unopened
Blue: in progress
Red: finished

Navigation

Browse between assigned documents

Export current document (with annotations)

Browse current document

See guidelines for annotators



Mark **completed** document annotation

Select **annotation layers**

Specify the number of sentences displayed on one page

Moving the last annotated sentence in the middle, selecting colours

Annotation Layer and window size configuration Window

Annotation Layers:

- ☒ Besedna vrsta
- ☒ Ime
- ☒ Lema
- ☒ Odvisnostna skladnja

Number of Sentences:

Auto-scroll document while annotating : ☒

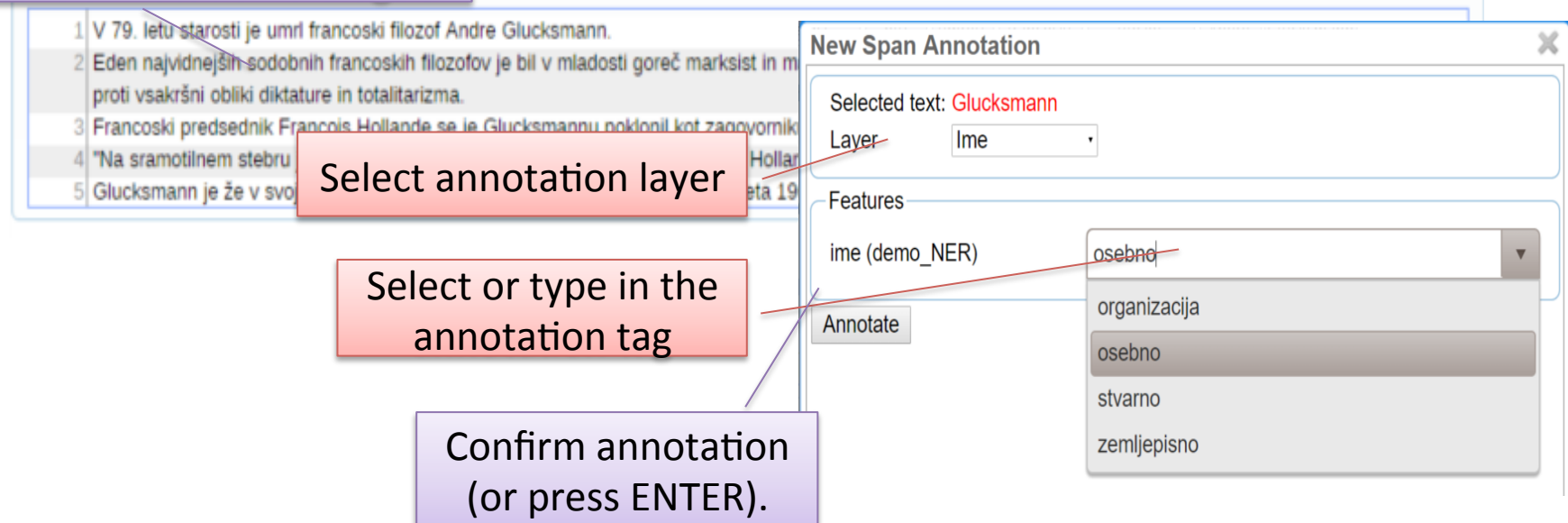
Use the same color for all tags in a layer ☐

Save Cancel

Annotation



Select a span (characters, word or words) or double-click on a word to open the annotation dialogue.



Select annotation layer

Select or type in the annotation tag

Confirm annotation (or press ENTER).

Choose one or more
annotation layers for review

Curation

Finish curating the
document

Sentences

- 1 V 79. letu starosti je umrl francoski filozof Andre Glucksmann.
Eden najvidnejših sodobnih francoskih filozofov je bil v mladosti goreč marksist in maoistični
- 2 levičar, pozneje pa se je v svojih polemičnih zapisih začel bojevati proti vsakršni obliki diktature in totalitarizma.
Francoski predsednik Francois Hollande se je
- 3 Glucksmannu poklonil kot zagovorniku zatiranih.
- 4 "Na sramotilnem stebru je izpostavljal tiste, ki so povzročali trpljenje," je tvitnil Hollande.
Glucksmann je že v svoji prvi knjigi Discours de
- 5 la guerre (Razprava o vojni) leta 1968 na muho vzel vojno in politiko jedrskega zastraševanja.

Annotation

osebno
5 Glucksmann je že v svoji prvi knjigi Discours de la guerre (Razprava o vojni) leta 1968 na muho vzel vojno in politiko jedrskega zastraševanja.

User: anno1

osebno
5 Glucksmann je že v svoji prvi knjigi Discours de la guerre (Razprava o vojni) leta 1968 na muho vzel vojno in politiko jedrskega zastraševanja.

User: kajad

osebno
5 Glucksmann je že v svoji prvi knjigi Discours de la guerre (Razprava o vojni) leta 1968 na muho vzel vojno in politiko jedrskega zastraševanja.

Frame with merged
(final) annotations

Frames with individual annotations

Grey: match

(automatically accepted to upper frame)

Blue: mismatch

(clicking the correct annotation moves it to upper frame)

Sentences in red signal a
conflict in annotation.

Curation

Sentences

1 V 79. letu starosti je umrl francoski filozof Andre Glucksmann.

Eden najvidnejših sodobnih francoskih filozofov je bil v mladosti goreč marksist in maoistični levičar, pozneje pa se je v svojih polemičnih zapisih začel bojevati proti vsakršni obliki diktature in totalitarizma.

Francoski predsednik Francois Hollande se je 3 Glucksmannu poklonil kot zagovorniku zatiranih.

4 "Na sramotilnem stebru je izpostavljal tiste, ki so povzročali trpljenje," je tvtinil Hollande.

Glucksmann je že v svoji prvi knjigi Discours de 5 la guerre (Razprava o vojni) leta 1968 na muho vzel vojno in politiko jedrskega zastraševanja.

Annotation

5 Glucksmann je že v svoji prvi knjigi Discours de la guerre vojno in politiko jedrskega zastraševanja.

New annotations can be added.

User: anno1

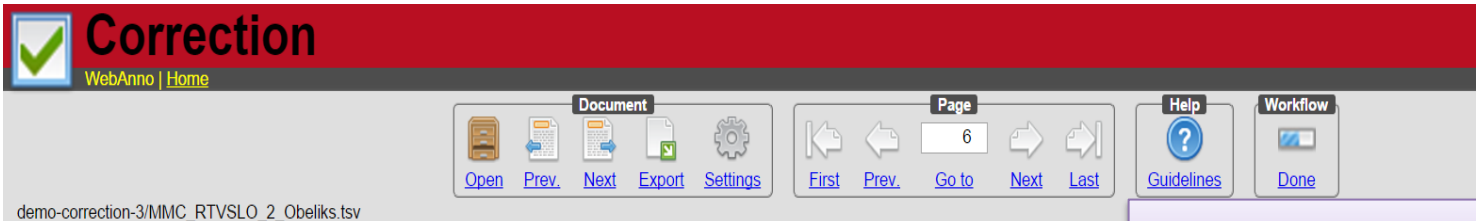
5 Glucksmann je že v svoji prvi knjigi Discours de la guerre (Razprava o vojni) leta 1968 na muho vzel vojno in politiko jedrskega zastraševanja.

User: kajad

5 Glucksmann je že v svoji prvi knjigi Discours de la guerre vojno in politiko jedrskega zastraševanja.

After selecting one of the mismatched annotation, the right one turns **green** and the wrong one turns **red**.

Correction



Annotation

Ob blesteči predstavi centrske linije so košarkarji Miamijs s 101:88 premagali Los Angeles Lakerse, ki so prvič v sezoni igrali brez Kobeja Bryanta.

Goran Dragić tudi tokrat ni blestel.

Slovenski organizator igre je v 30 minutah ob metu 2 / 10 dosegel 4 točke ter dodal še 5 skokov in 3 podaje.

Na zadnjih dveh točkah je vknjižil le 6 točk.

Rekord sezone je s 30 točkami in 11 skoki postavil Chris Bosh, četrti zaporedni dvojni dvojček (19 točk, 15 skokov) na domačem parketu pa je dosegel Hassan Whiteside.

Frame with confirmed annotations

User: Suggestion

Ob blesteči predstavi centrske linije so košarkarji Miamijs s 101:88 premagali Los Angeles Lakerse, ki so prvič v sezoni igrali brez Kobeja Bryanta.

Goran Dragić tudi tokrat ni blestel.

Slovenski organizator igre je v 30 minutah ob metu 2 / 10 dosegel 4 točke ter dodal še 5 skokov in 3 podaje.

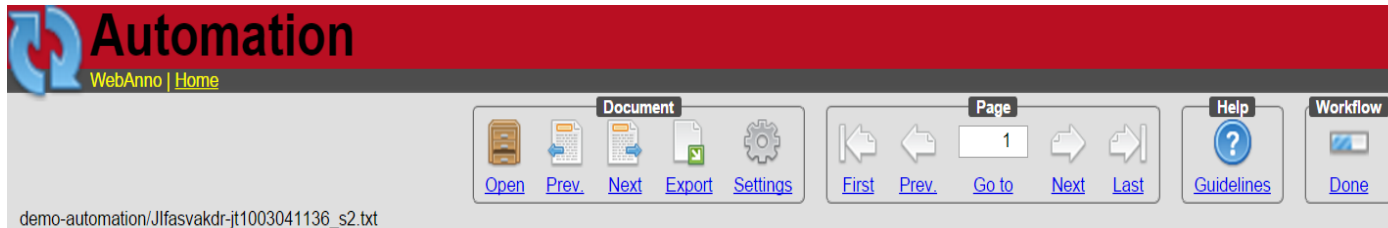
Na zadnjih dveh točkah je vknjižil le 6 točk.

Rekord sezone je s 30 točkami in 11 skoki postavil Chris Bosh, četrti zaporedni dvojni dvojček (19 točk, 15 skokov) na domačem parketu pa je dosegel Hassan Whiteside.

Frame with suggested annotations

By clicking on a suggested (blue) annotation, it is moved to upper frame and turns **grey**. If we change the suggested annotations, it turns **red**.

Automation



Annotation

		DM
1	[Pf-pred-07541:] pozdravljeni danes nas je kr mal več zbranih a ne tk da vas je prav fajn pogledat \$.	
2	[Pf-pred-07541:] en čist nekaj bi rekli en mogoče res zaključek en rezime vsega kar smo na teh preteklih treh oziroma zdaj na četrtem srečanju govorili a ne \$.	
3	[Pf-pred-07541:] dejansko zakaj je tolik pomembno da se mi sploh ukvarjamo s tem kako otroka vzgojit \$.	
4	[Pf-pred-07541:] zakaj je tolik pomembno da mi tam se dost z njim pogovarjamo pa nekaj vzpostavmo do njega odnos da spoznamo kdo ta otrok v bistvu sploh je \$.	
5	[Pf-pred-07541:] to je tist o čemer bomo dans govorili a ne \$.	
6	[Pf-pred-07541:] ker vsi vemo a ne v kakšnem svetu živimo v bistvu v svetu ki je kr zahtevn za preživetje a ne \$.	

Final annotation frame


User: Suggestion

		DM
1	[Pf-pred-07541:] pozdravljeni danes nas je kr mal več zbranih a ne tk da vas je prav fajn pogledat \$.	
2	[Pf-pred-07541:] en čist nekaj bi rekli en mogoče res zaključek en rezime vsega kar smo na teh preteklih treh oziroma zdaj na četrtem srečanju govorili a ne \$.	
3	[Pf-pred-07541:] dejansko zakaj je tolik pomembno da se mi sploh ukvarjamo s tem kako otroka vzgojit \$.	
4	[Pf-pred-07541:] zakaj je tolik pomembno da mi tam se dost z njim pogovarjamo pa nekaj vzpostavmo do njega odnos da spoznamo kdo ta otrok v bistvu sploh je \$.	
5	[Pf-pred-07541:] to je tist o čemer bomo dans govorili a ne \$.	DM
6	[Pf-pred-07541:] ker vsi vemo a ne v kakšnem svetu živimo v bistvu v svetu ki je kr zahtevn za preživetje a ne \$.	DM

Frame with dynamic annotation suggestions

Each new annotation improves the machine learning model that **suggests annotations for similar spans**. By clicking on a suggested (blue) annotation, the annotation is selected and moved to the upper frame.

Project setup


Projects Settings
[WebAnno](#) | [Home](#)

Projects

Edit existing projects

- demo-annotation
- demo-automation
- demo-correction
- goo300k
- gos-min-automation
- neoznačeno
- test-Meta
- test-auto-cyp
- test-automation
- test-bug
- test-everything
- test-hackaton-tviti
- test-multiple-users
- test-plain-text

Start a new project

- test-tviti-syntax
- testiram
- tutorial-annotation
- universal-test

Create project

Import Project

Files

Izberi datoteke

Nobena datoteka ni izbrana

☒ Create missing users

Import project

Details

General

Name: new-project

Description:

Script direction: ltr

Project Types

☒ annotation
 ☐ automation
 ☐ correction

The **annotation mode** for the project is selected and cannot be changed later on.

Import external projects

Add users

Details Users Documents Layers Tagsets Guidelines Export

Users

- jaka [user]
- kaja [admin]
- katja [user]
- spela [curator, user]

Add project collaborators from the list of all system users.

Assign collaborator roles

- user:** annotator
- curator:** reviewer
- admin:** project leader

Add Users Remove User Change Permissions

Import documents

Details Users Documents Layers Tagsets Guidelines Export

Import new documents

Format: Binary format
Files: Binary format
CoNLL 2000
CoNLL 2002
CoNLL 2006
CoNLL 2009
Plain text
WebAnno CPH TEI reader
WebAnno TSV Format
WebAnno TSV Format - Old
Weblicht TCF Format
XML format

ka ni izbrana

Import document

Documents

Select the format of one or more imported documents.

Specify layers

The screenshot shows the 'Specify layers' interface with several annotations:

- 1. define layer type (span, relation, chain)**: Points to the 'Type' dropdown in the 'Technical Properties' section, which is set to 'span'.
- 2. define layer behaviours**: Points to the 'Behaviors' section, which includes checkboxes for 'Lock to token offsets', 'Allow stacking', 'Allow crossing sentence boundary', and 'Allow multiple tokens'.
- 3. define one or more features within a layer**: Points to the 'Feature overview' section, which lists 'mwe type : [String]'.
- 4. define feature tagset (otherwise NONE)**: Points to the 'Tagset' dropdown in the 'Feature details' section, which is set to 'MWE-LexSem'.
- Edit existing layers**: Points to the 'Layers' list on the left, which includes 'Chunk', 'Coreference', 'Dependency', 'Lemma', 'Multi-word units', 'Named Entity', and 'POS'. The list is color-coded: green for default, blue for custom, and red for disabled.
- Import external layers**: Points to the 'Import layer' button in the 'Import layer' section.
- Create custom layers**: Points to the 'Create layer' button in the 'Layers' section.

Define tagsets

Edit existing tagsets

Import external tagsets

1. Create new tagset

2. Enter individual tags

Navigation: [Users](#) [Documents](#) [Layers](#) [Tagsets](#) [Guidelines](#) [Export](#)

Tagsets

- BART
- MWE-LexSem
- NER_WebAnno
- STTS
- Time

Tagset details

Name: MWE-LexSem

Language: sl

Description:

Extendible tagset?: ☒

Export as: JSON **Export tagset** **Save tagset** **Delete tagset**

Import tagsets

Format: JSON

Files: Izberi datoteke Nobena datoteka ni izbrana **Import tagset**

Tags

- PP modifier
- compound
- fixed phrase
- named entity
- proverb
- semi-fixed VP
- verb-particle
- verb-preposition

Create tag

Tag details

Name: conjunction

Description:

Save tag **Delete tag**

Export project

Details Users Documents Layers Tagsets Guidelines Export

Export

Warning: Do not leave this page or close the browser while export is running!
Leaving this page or closing the browser may lock the project settings. Unlocking them again requires a restart of the server!

Progress:

0%

Cancel

Export format:

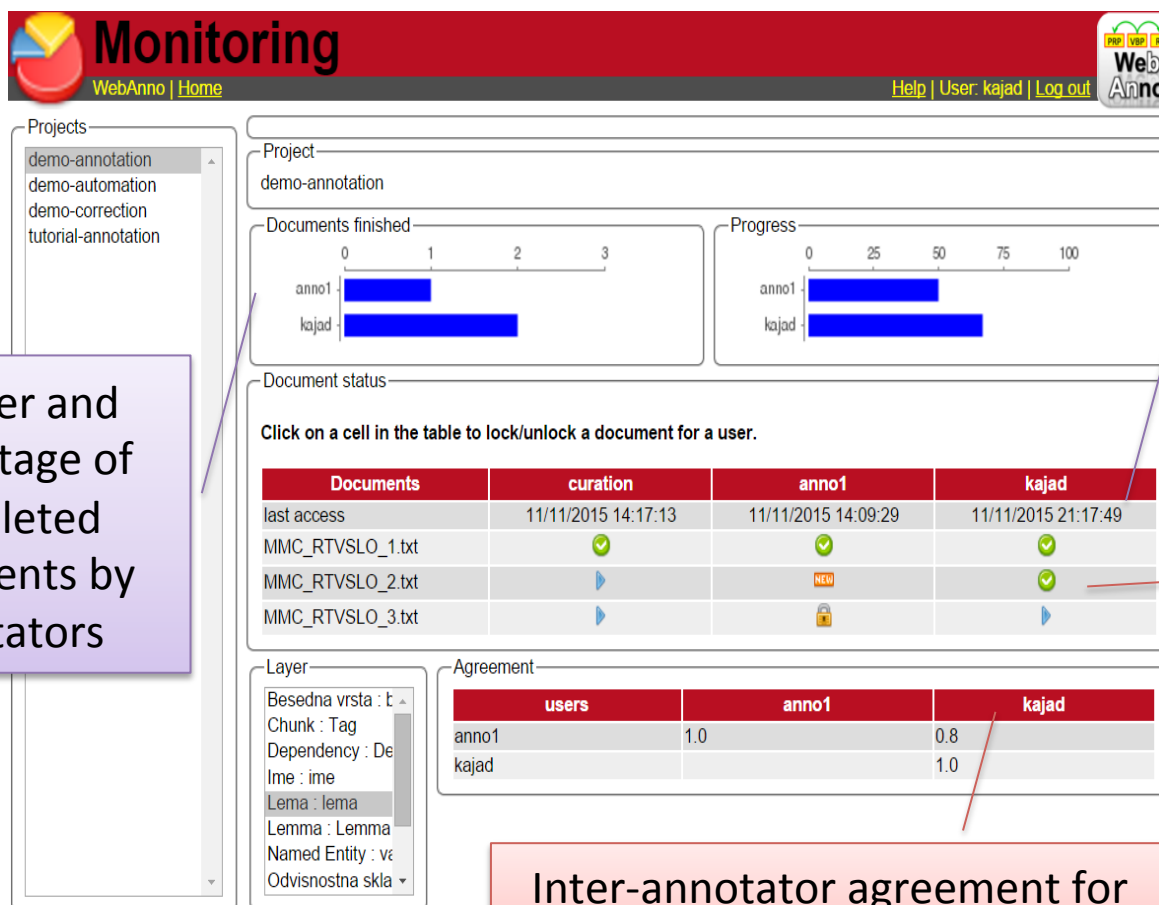
WebAnno TSV Format ▾
AUTO
Binary format
CoNLL 2000
CoNLL 2002
CoNLL 2006
CoNLL 2009
Plain text
WebAnno TSV Format
Weblicht TCF Format
XMI format

Export the whole project to the file system

Export project in preferred format

All project documents are exported, including the initial documents, annotations by individual annotators and final merged annotations. **Project settings**, such as layers and tagsets, are also exported.

Monitoring



Number and percentage of completed documents by annotators

Time of last login

Assigning workload and monitoring document status
[NEW]: unopened
arrow: in progress
check mark: completed
lock: disabled for annotation

Inter-annotator agreement for selected layer (kappa score)

Summary

- online tool
 - one or more annotation layers
 - arbitrary or predefined tags
 - different annotation modes
 - simple project setup
-
- linguistic and other types of annotation
 - simple annotation or annotation project

Help and support

- main website
<https://webanno.github.io/>
- user guide
<https://webanno.github.io/webanno/releases/2.3.0/docs/user-guide.html>
- video tutorials
<https://www.youtube.com/user/webanno>
- users' mailing list
<https://groups.google.com/forum/#!forum/webanno-user>
- github issues
<https://github.com/webanno/webanno/issues>
- publications
<https://scholar.google.si/scholar?hl=sl&q=WebAnno>
- ReLDI
nljubesi@gmail.com