CLSINFRA COMPUTATIONAL LITERARY STUDIES INFRASTRUCTURE

Information extraction from the Shakespeare Drama Corpus

Silvie Cinková

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004984

Learning Goals

1. Set up a corpus
2. TEI-XML markup basics
3. Information extraction/text mining from running text
    1. Conceptualize your research question
    2. Operationalize your concepts
    3. Implement your operationalizations in corpus queries
4. Interpret your search results with elementary statistical methods

Title Sample

1. Set up a corpus
2. TEI-XML markup basics
3. Information extraction/text mining from running text
    1. Conceptualize your research question
    2. Operationalize your concepts
    3. **Implement your operationalizations in corpus queries**
4. **Interpret your search results with elementary statistical methods**

Your future: Build your own toolchain

Shakespeare Material

CLSINFRA

ABOUT CORPORA TOOLS HOW TO MERCH

Gersh DraCor	Shake DraCor	Rom DraCor
German Shakespeare Drama Corpus	Shakespeare Drama Corpus	Roman Drama Corpus
38 Number of plays	37 Number of plays	36 Number of plays
0 person + personGrp Number of characters	1,433 (M: 797, F: 116) person + personGrp Number of characters	
873,038 text Text tokens	908,286 text Text tokens	37,157 text Text tokens
31,767 (855,064) sp (Tokens)	31,066 (876,744) sp (Tokens)	18,485 (298,299) sp (Tokens)
6,883 (32,064) stage (Tokens)	10,450 (41,230) stage (Tokens)	204 (529) stage (Tokens)
Last update 3. 5. 2022 1:23:17	Last update 3. 5. 2022 0:55:50	Last update 3. 5. 2022 0:26:04

If you want to cite DraCor, please use the following reference:

BIB RIS

Fischer, Frank, et al. (2019). Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama. In *Proceedings of DH2019: "Complexities"*, Utrecht University, doi:10.5281/zenodo.4284002.

Drama Corpora Project

Unless otherwise stated, all corpora and the web design are released under Creative Commons 0 1.0

This site runs on DraCor API 0.84.0 using eXist-db 6.0.1

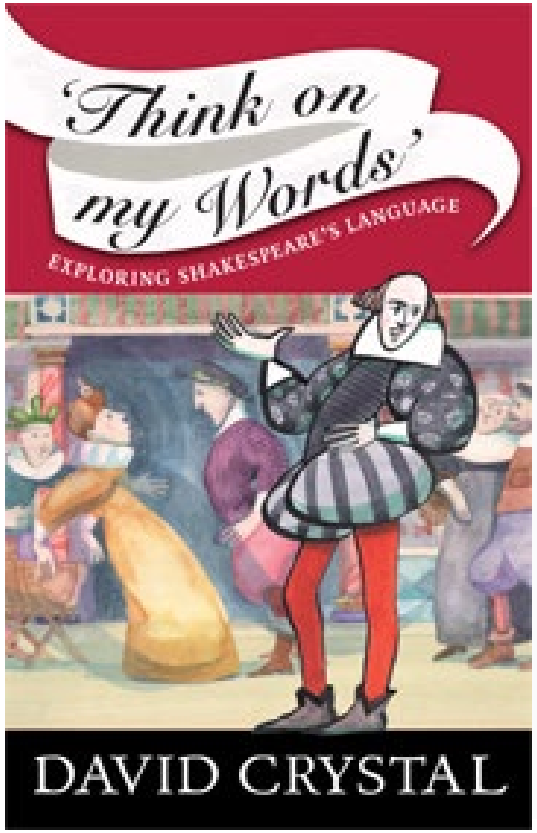
dracor.org

Research Ideas

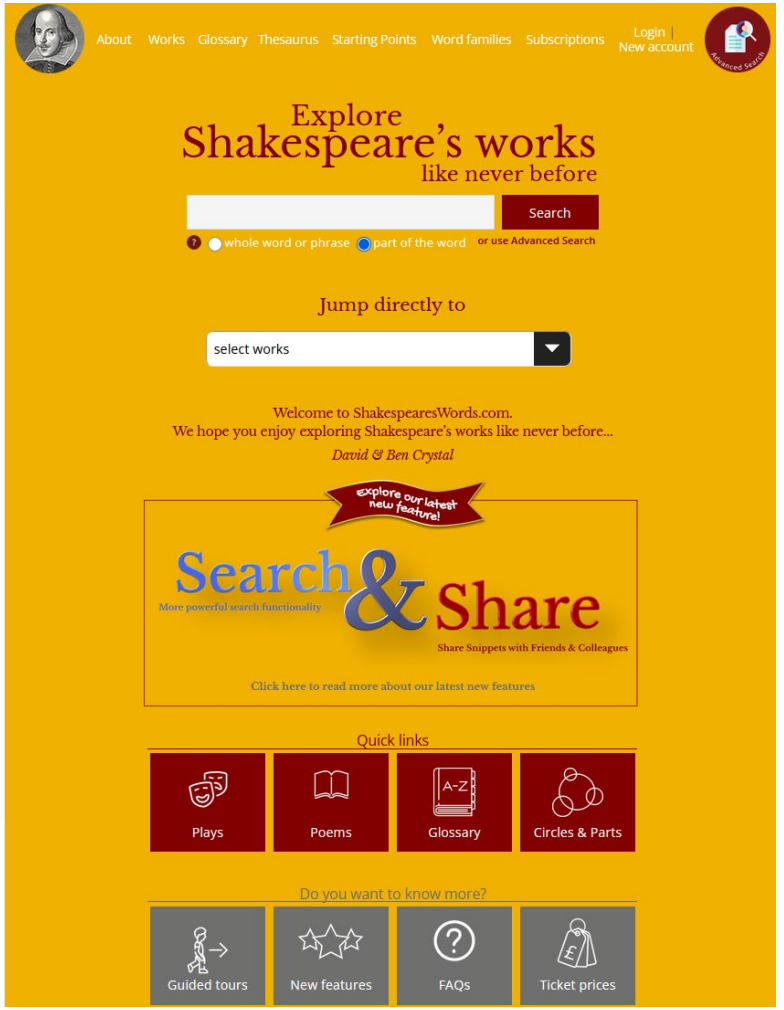
CLSINFRA

- Modern English vs. Early Modern English
  - Lexicon
  - Grammar
- Operationalized concepts of text pragmatics (“style”)
  - Content (Key Words)

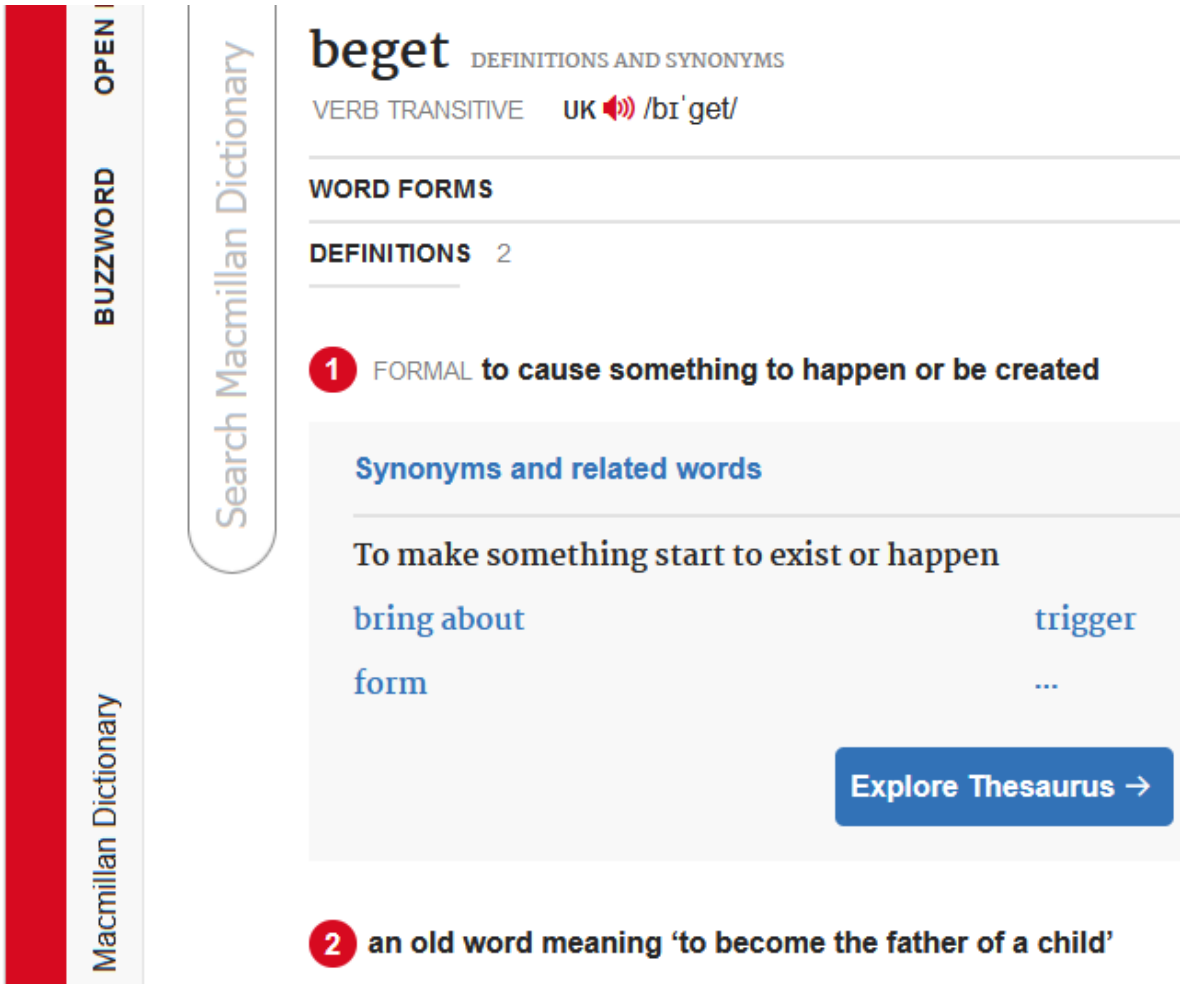
- *Words starting/ending with xxx*
- *All word forms of X*
- *X preceding Y*
- *X somewhere close around Y*
- *Y with an adjective X as attribute*
- *Passive verb X with logical subject Y*



shakespeareswords.com
50D DC7 D5B



The case of *beget*



- In which morphological forms does it occur?
- Which prefixes does it occur with?
- *Who* begets *what*?
  - Which grammatical constructions to extract?
- Shakespeare vs. *Early English Books Online* or a modern English corpus

The case of *un-* negation

- Shakespeare's creative usage
- Especially from around 1600 onwards
- With which parts of speech did he use it most?
- Compare with EEBO (Early English Books online)
  - Distribution of part of speech tags with this negation prefix

Double genitive vs. absolute genitive

CLSINFRA

- Detect all cases of genitives like this:

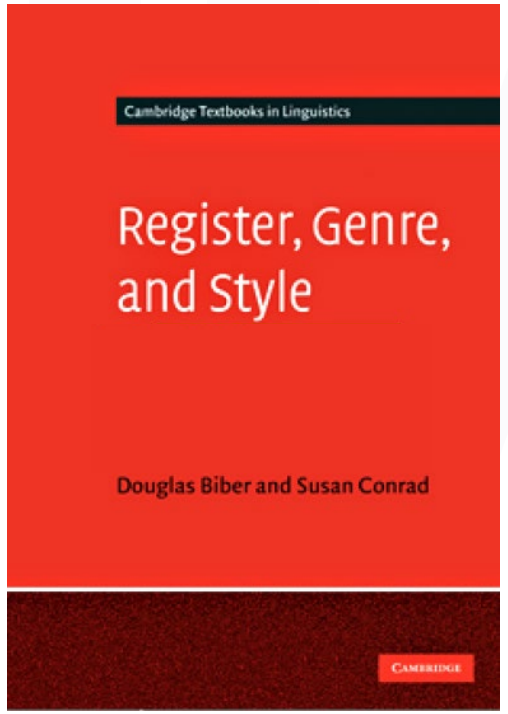
  *The young Gentleman **of** the Count Orsino**’s***

- … and this:

  For halfe thy wealth, it is Antonio’s.

Pragmatic concepts

- Social language use, Communication purpose in utterances
- Stylistic & rhetoric means
  - Described by lexical as well as grammatical features
- Genres and registers
  - Douglas Biber, since 1980s
  - Multidimensional Analysis (MDA)



Expression of stance

- Speaker reports X and indicates
  - truth estimate (true vs. false, observed vs. heard, likely vs. unlikely)

  For so I know he is, they know he is – a most arch heretic, a pestilence

  I mean that with my soul I love thy daughter

  I could find in my heart that I had not a hard heart

  I learn in this letter that Don Pedro of Aragon comes this night to Messina
  - or evaluation of X (good-bad)

  It is a problem that you don't approve of this.

Narrativity

- + simple past tense
- - 2nd person
- + past/present progressive tense
- - simple present tense
- - passive voice

Descriptivity

- + adjectives in attributive positions
- + relative clauses
- + copula predicates
- + present tense
- - progressive tense
- - modal verbs

Interactivity

- 2nd person
- questions
- vocatives
- imperatives

Uncertainty or distance

- + hedge expressions (*maybe, basically, a bit*)
- + indefinite pronouns (*some, any*)
- + some modal verbs (*can, may*)
- + conditional markers (*would, if, when, whether*)

Emotionality

CLSINFRA

- + interjections
- + exclamation marks

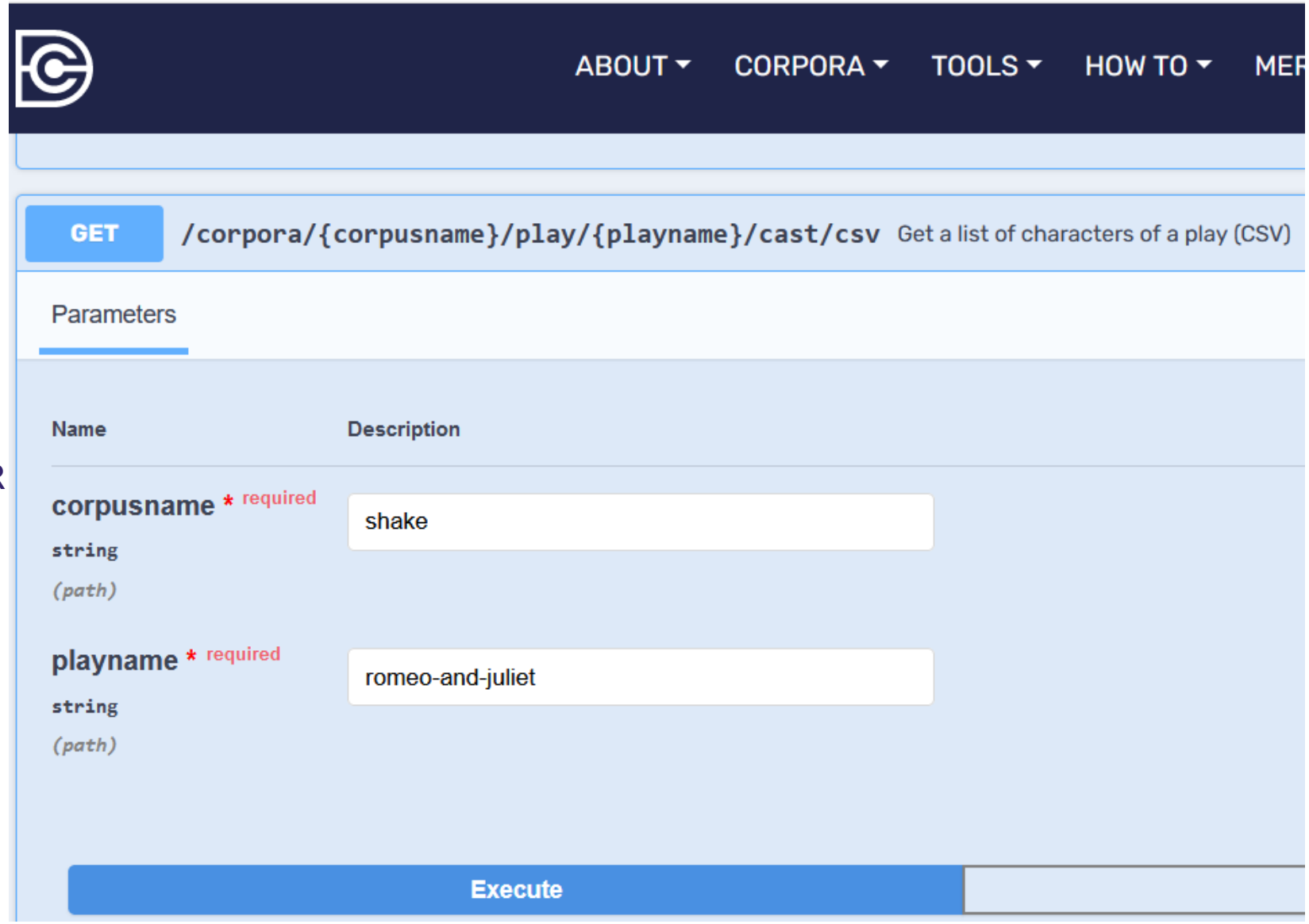
- Shakespeare: short lines by one speaker – one verse in his iambic pentameter is comprised of several speakers' lines

Programmable corpora and DraCor

DraCor “Drama Corpora”

CLSINFRA

- TEI-XML encoded drama texts in a GitHub repository
- you can slice a drama piece and download just a part
  - lines by individual speakers
  - cast list
  - stage instructions
  - acts, scenes
- use a web GUI (frontend) or an API
- even more comfortable: Python and R libraries with functions



```
curl -X 'GET' \
  'https://dracor.org/api/corpora/shake/play/romeo-and-juliet/cast/csv' \
  -H 'accept: text/csv'
```

```
https://dracor.org/api/corpora/shake/play/romeo-and-juliet/cast/csv
```

Server response

Code	Details
200	Response body

```
id,name,gender,isGroup,numOfScenes,numOfSpeechActs,numOfWords,w
"Chorus_Rom","Chorus","UNKNOWN","true","2","2","219",,"0","0","
"SERVANTS.CAPULET.Sampson_Rom","Sampson","MALE","false","1","20
"SERVANTS.CAPULET.Gregory_Rom","Gregory","MALE","false","1","15
"SERVANTS.MONTAGUE.Abram_Rom","Abram","MALE","false","1","5","2
"Benvolio_Rom","Benvolio","MALE","false","7","63","1160",,"24",
"Tybalt_Rom","Tybalt","MALE","false","3","17","263",,"21","29",
```

```
"SERVANTS.CAPULET.0.3_Rom","SERVANTS.CAPULET.0.3_Rom","MALE","false","1","2","21",,"10","10","0","0.555984555984556","0.11595979924396634"
"Cousin_Rom","Cousin_Rom","MALE","false","1","2","18",,"10","10","0","0.555984555984556","0.11595979924396634"
"SERVANTS.CAPULET.0_Rom","SERVANTS.CAPULET.0_Rom","MALE","false","1","1","4",,"10","10","0","0.555984555984556","0.11595979924396634"
```

Download

GET /corpora/{corpusname} List corpus content

GET /corpora/{corpusname}/metadata List of metadata for all plays in a corpus

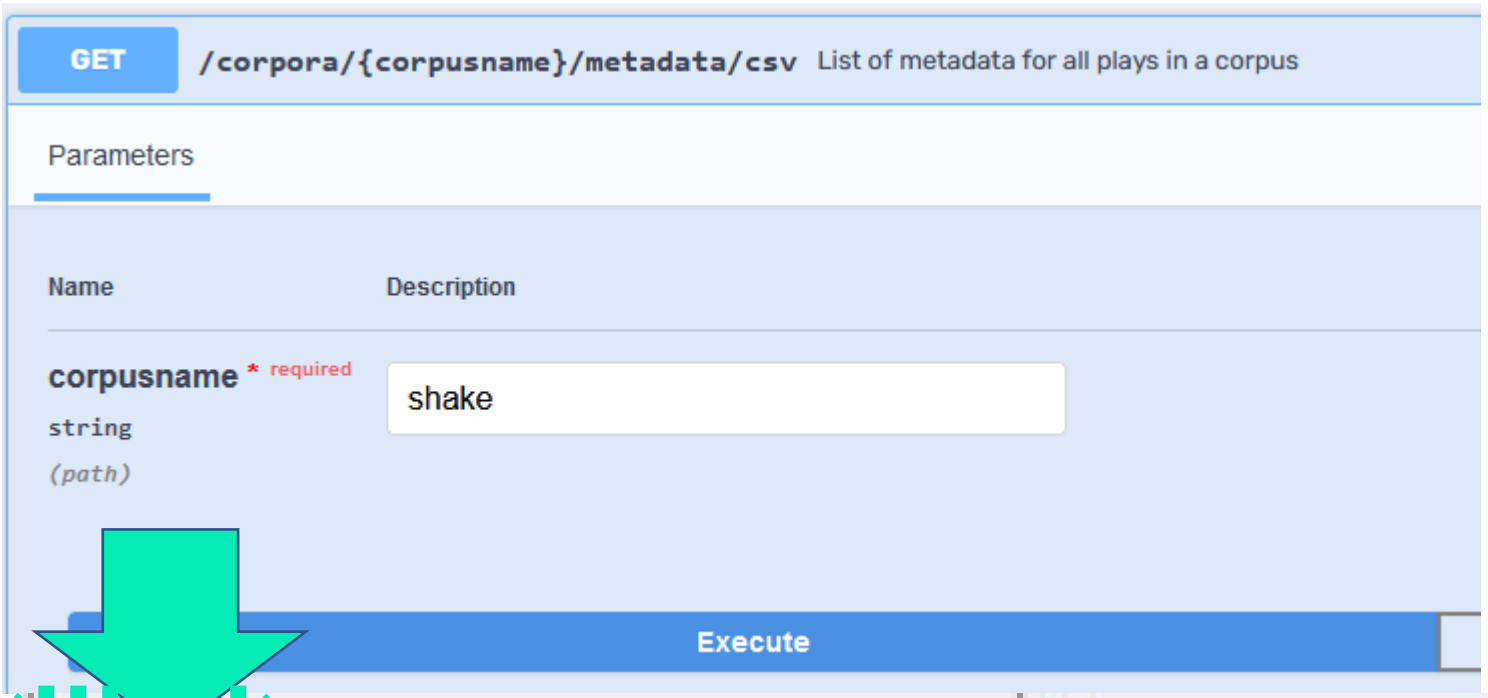
GET /corpora/{corpusname}/metadata/csv List of metadata for all plays in a corpus

~~GET /corpora/{corpusname}/metadata.csv~~ List of metadata for all plays in a corpus

GET /corpora/{corpusname}/play/{playname} Get metadata and network metrics for a single play

GET /corpora/{corpusname}/play/{playname}/metrics Get network metrics for a single play

GET /corpora/{corpusname}/play/{playname}/tei Get TEI document of a single play

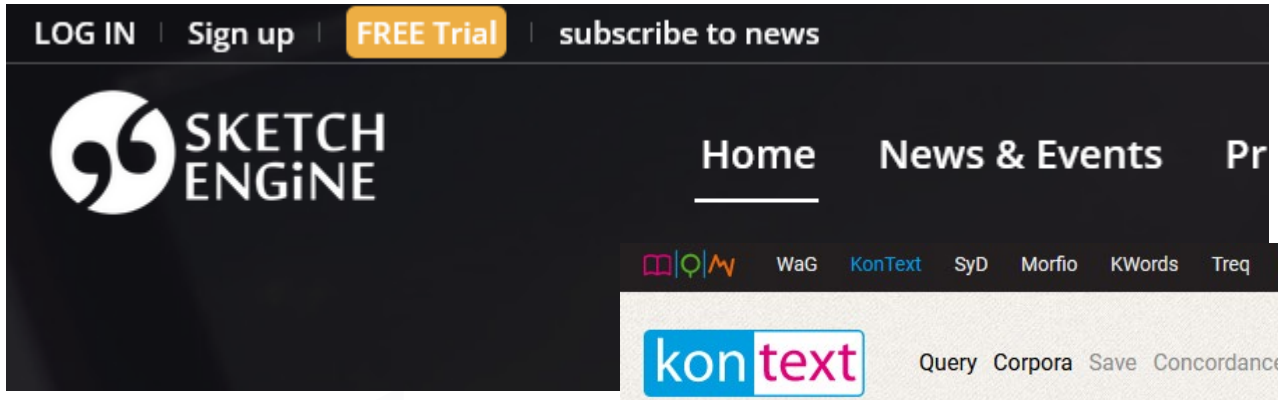


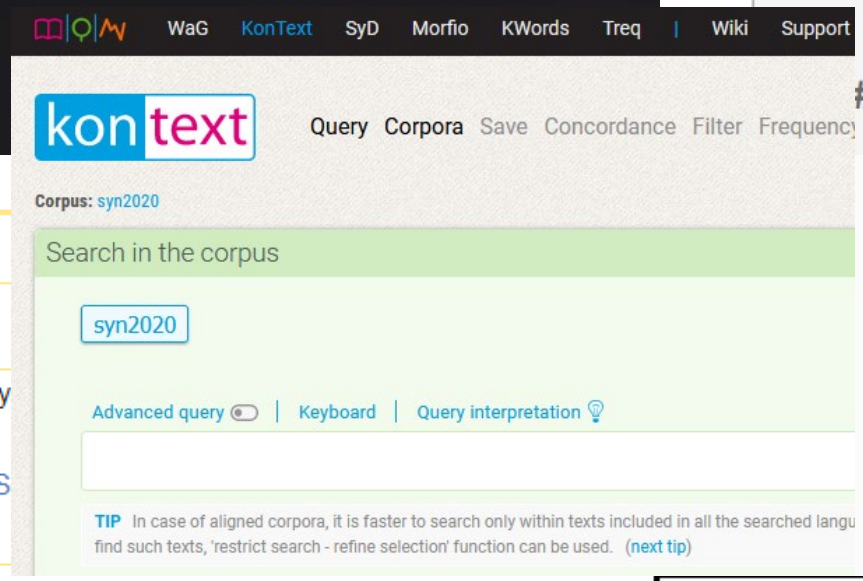
name	id	title
a-midsummer-night-s-dream	shake000008	A Midsummer Night's Dre
all-s-well-that-ends-well	shake000012	All's Well That Ends Well
antony-and-cleopatra	shake000035	Antony and Cleopatra
as-you-like-it	shake000010	As You Like It
coriolanus	shake000026	Coriolanus
cymbeline	shake000036	Cymbeline

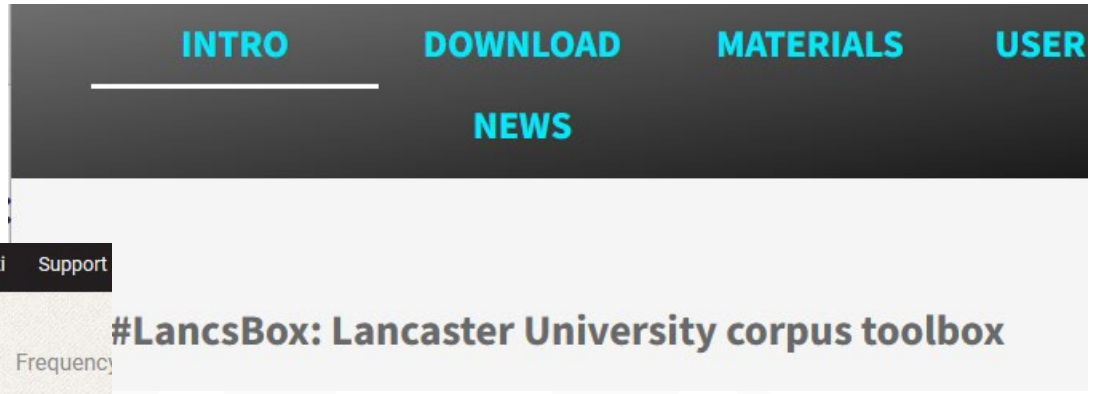
Corpus managers and query languages

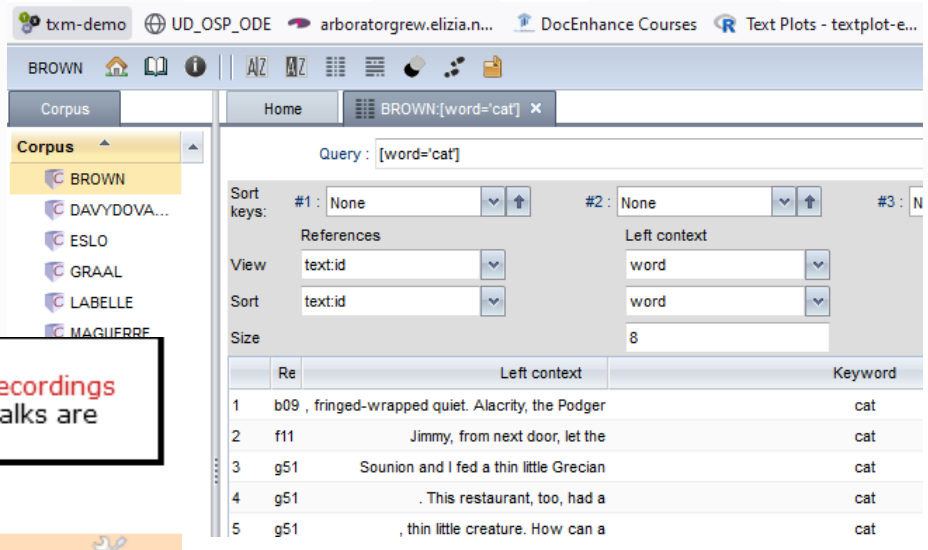
Corpus managers

CLSINFRA









CLSINFRA

TEITOK

Using TEITOK in our course

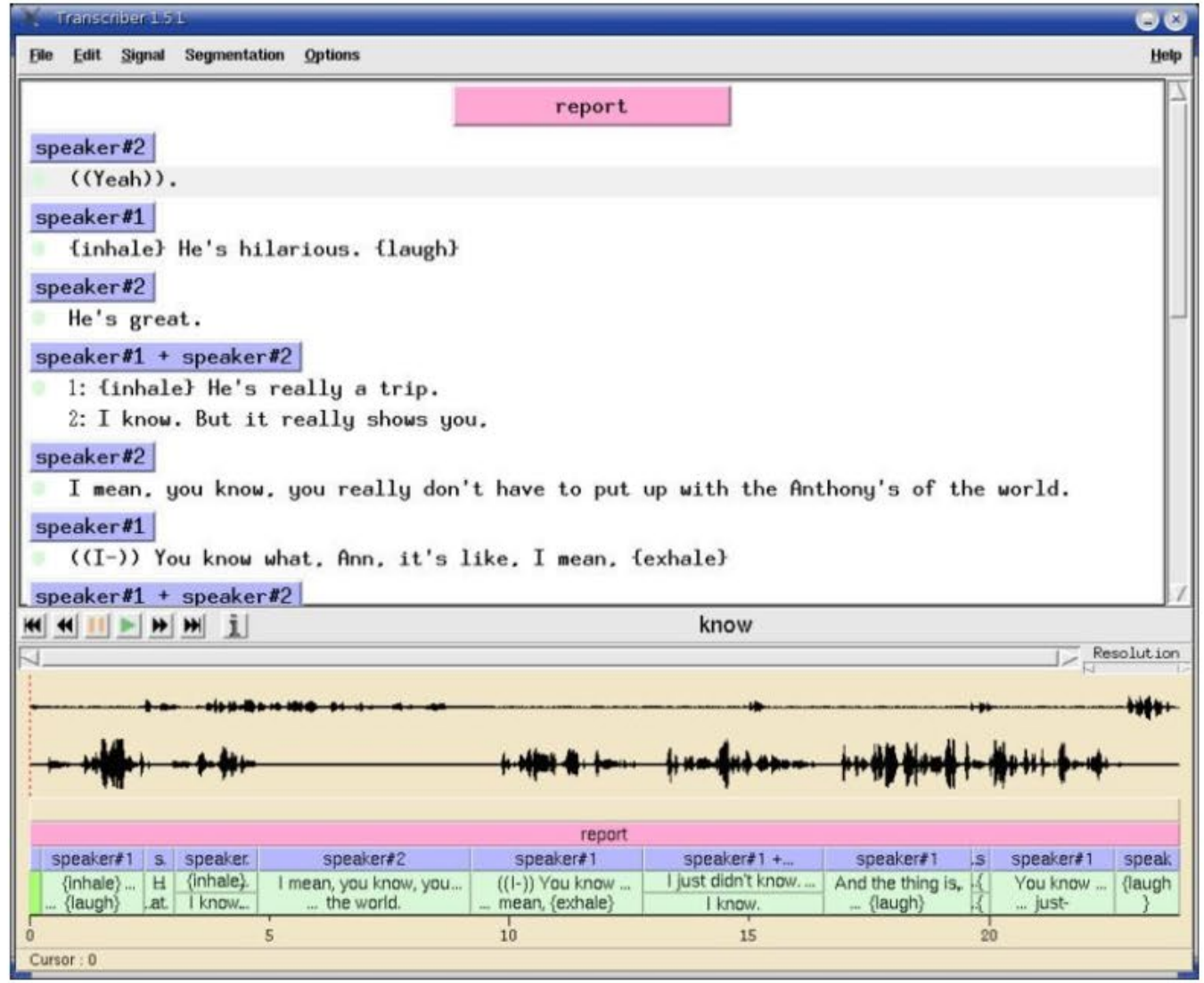
- Individual project setup by developer
- You need admin rights – user account to each project separately (as many logins as you have projects!!!)
- **dracorshake** at https://quest.ms.mff.cuni.cz/teitok-dev/teitok/teaching/dracorshake/index.php
- **For your corpus: cls** at https://quest.ms.mff.cuni.cz/teitok-dev/teitok/teaching/cls/index.php

Audio Files

I have a collection of audio files that I want to turn into a searchable corpus

Transcribe

Transcribe the audio files in transcription tools like ELAN, or Transcriber, or use speech recognition software to create an automatic transcription



speaker#2
((Yeah)).

speaker#1
{inhale} He's hilarious. {laugh}

speaker#2
He's great.

speaker#1 + speaker#2
1: {inhale} He's really a trip.
2: I know. But it really shows you.

speaker#2
I mean, you know, you really don't have to put up with the Anthony's of the world.

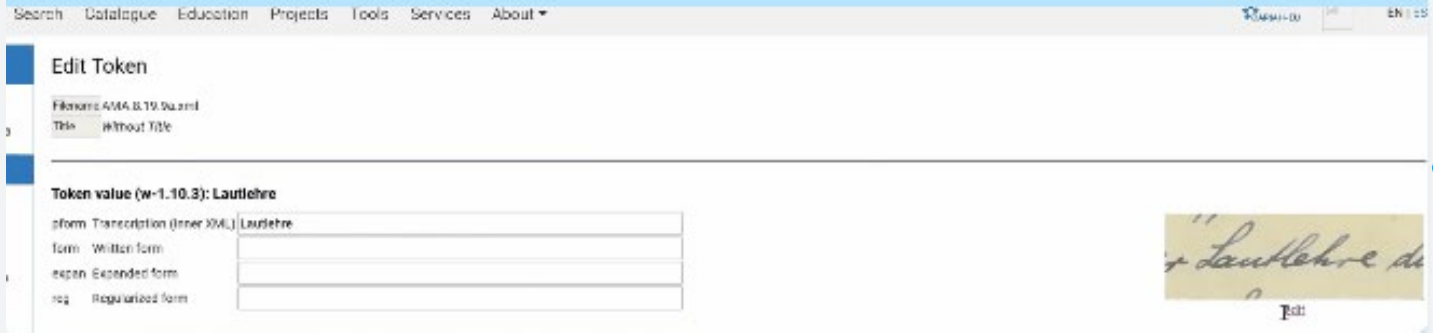
speaker#1
((I-)) You know what, Ann, it's like, I mean, {exhale}

speaker#1 + speaker#2

Import

Convert transcribed documents to TEI/XML, typically from the ELAN format, which includes not only the transcription, but also the alignment with the audio.

Image layer - graphical tokens



Facsimile Images

I have a collection of text images that I want to turn into a searchable corpus

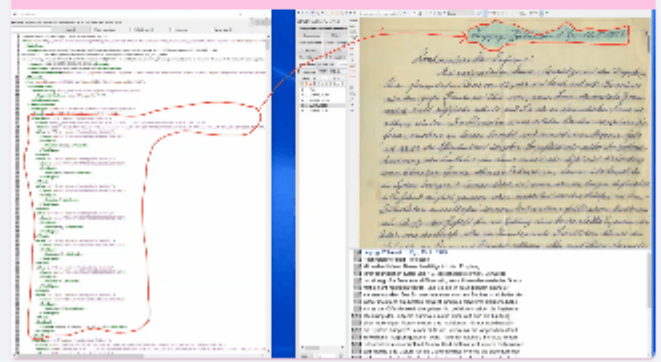
Transcribe

Transcribe in Transkribus (or similar tools) either automatically (HTR) or manually (using their transcription tools)

Import

Convert HTR documents to TEI/XML, typically from the PagesXML format, which includes not only the transcription, but also the alignment with the images.

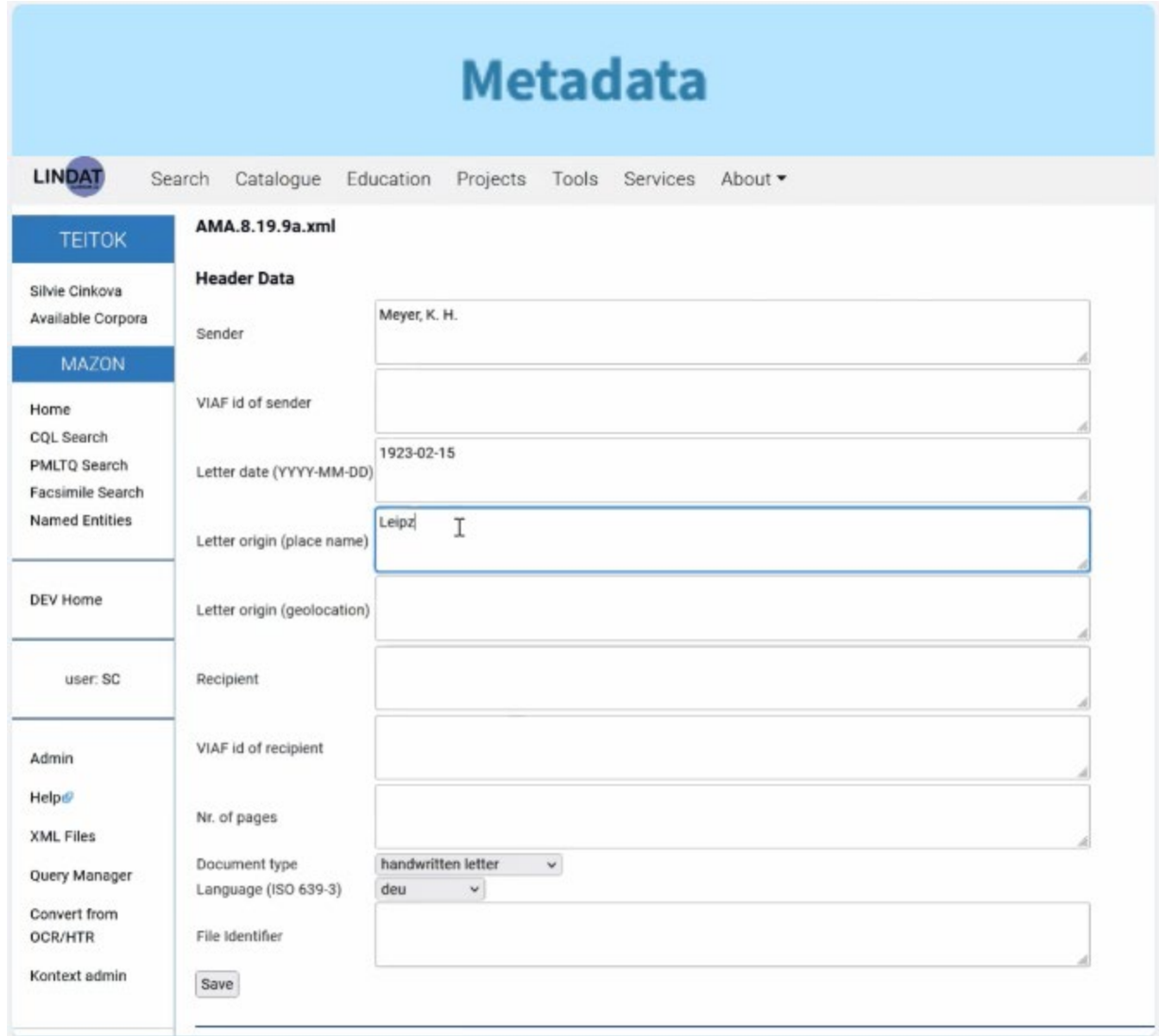
Handwritten-Text Recognition: Transkribus



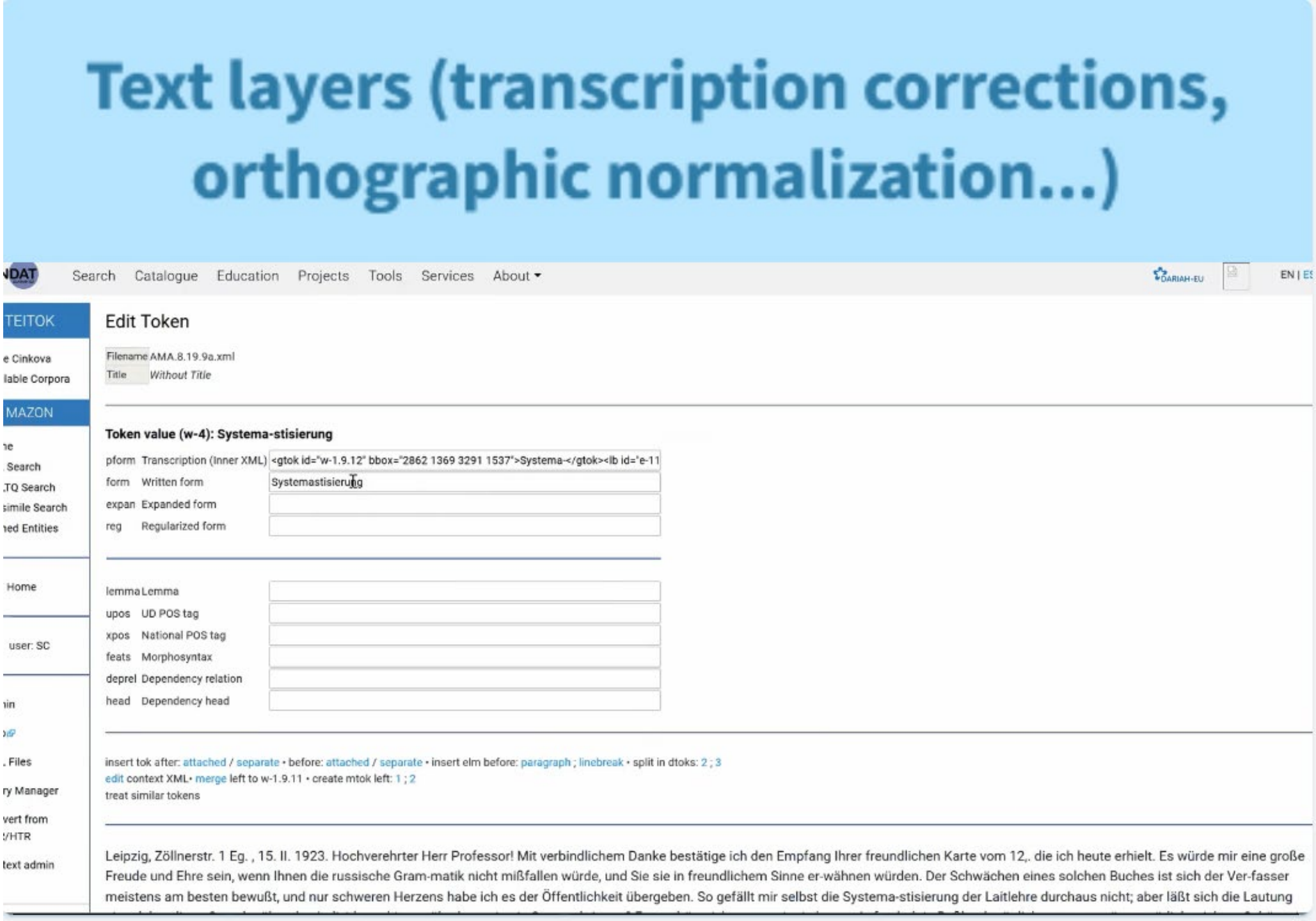
TE

A solution for TEI-
and annotation: T

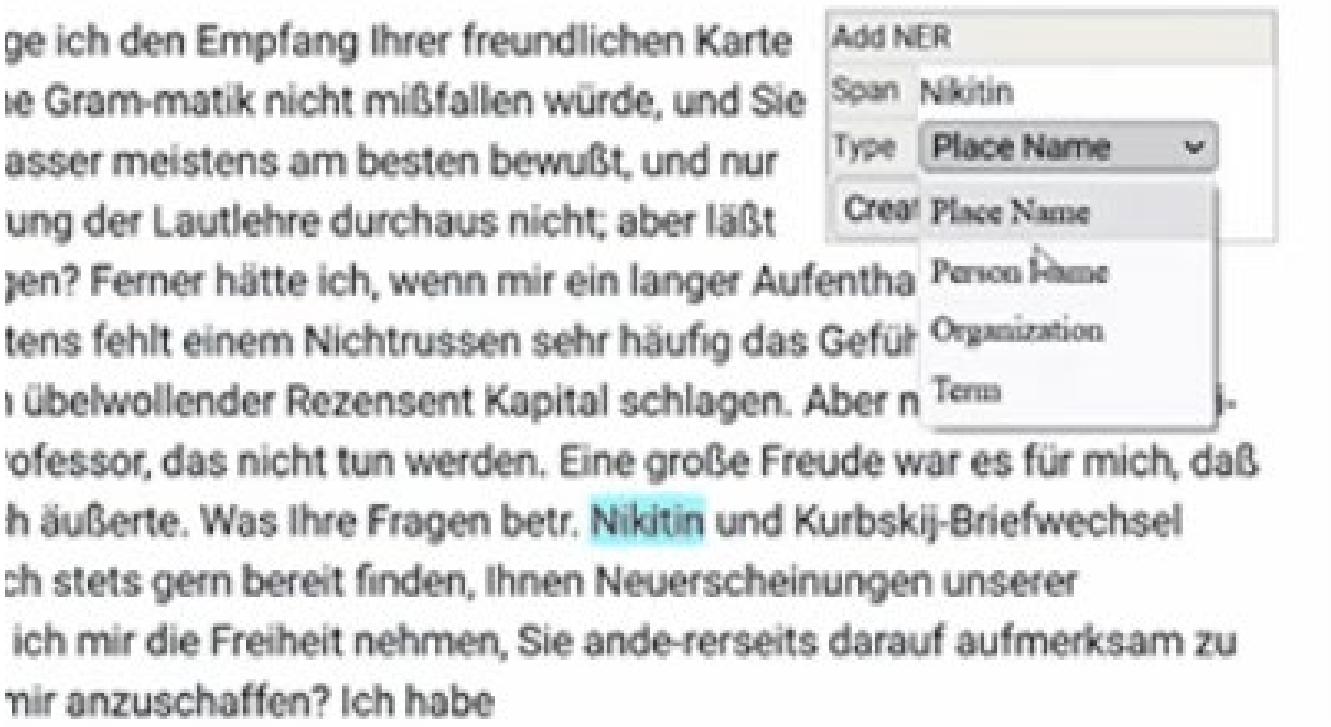
All documents in
format, which can


Metadata
LINDAT
Search
Catalogue
Education
Projects
Tools
Services
About
TEITOK
Silvie Cinkova
Available Corpora
MAZON
Home
CQL Search
PMLTQ Search
Facsimile Search
Named Entities
DEV Home
user: SC
Admin
Help
XML Files
Query Manager
Convert from OCR/HTR
Kontext admin
AMA.8.19.9a.xml
Header Data
Sender
Meyer, K. H.
VIAF id of sender
Letter date (YYYY-MM-DD)
1923-02-15
Letter origin (place name)
Leipz
Letter origin (geolocation)
Recipient
VIAF id of recipient
Nr. of pages
Document type
handwritten letter
Language (ISO 639-3)
deu
File Identifier
Save

Text layers (transcription corrections, orthographic normalization...)



Named-entity annotation in TEITOK



Corpus Search

CQL Query: [lemma="woman"] [lemma = "of"] []{0,3} within s

Search query builder

9 results • ipm: 8.44

Tags: UD POS Tag | National POS Tag | Features | Lemma | Dependency relation | Dependency head

context	her reported to be a	**woman of an invincible spirit**	. But it shall be	
context	maid's aunt , the fat	**woman of** | **Brentford , has**	a gown above . | MISTRESS	
context	He cannot abide the old	**woman of** | **Brentford .**	He swears she's a witch	
context	was 't not the wise	**woman of** | **Brentford ?**	| FALSTAFF | | Ay , marry ,	
context	gossip Report be an honest	**woman of her** | **word .**	| SOLANIO | | I would she were	
context	to desire to be a	**woman of the** | **world .**	| Enter two Pages . | | Here	
context	denied , which longs | To	**women of all fashion ;**	lastly , hurried | Here to	
context	to bear , | Making them	**women of good carriage .**	| This is she — ROMEO	
context	man . The vows of **women** | **Of no more bondage**		be to where they are	

Grew Query

Below you can type in a Grew query that will be run on all the conll-u files of this UDWiki project

```
% search for womens characteristics (or possessors)
pattern {
womannode [lemma ="woman"];
howwoman [upos = "NOUN"];
ofnode [lemma = "of"];
womannode -[nmod]-> howwoman;
howwoman -[case]->ofnode
}
```

Cluster: howwoman.lemma clear

Run Query

howwoman.lemma	Count
world	1
word	1
spirit	1
fashion	1
carriage	1
bondage	1

Export to xlsx
Export to csv
Export to txt

stored queries • store this query

Soubor Úpravy Zobrazit Historie Záložky Nástroje Nápověda

DraCor - Shakespeare Drama Corpus

https://quest.ms.mff.cuni.cz/teitok-dev/teitok/teaching/

120 %

Available Corpora

DraCor - Shakespeare Drama Corpus

Home
CQL Search
PML-TQ Search
Grew Search
Search in Kontext

DEV Home

user: SC

Below you can type in a Grew query that will be run on
project

% search for womens characteristics (or possessors)
pattern {
womannode [lemma ="woman"];
howwoman [upos = "NOUN"];
ofnode [lemma = "of"];

Cluster: howwoman.lemma clear

Run Query

howwoman.lemma = spirit (clear)

all.conllu

1 I have heard her reported to be a woman of an invincible spirit .

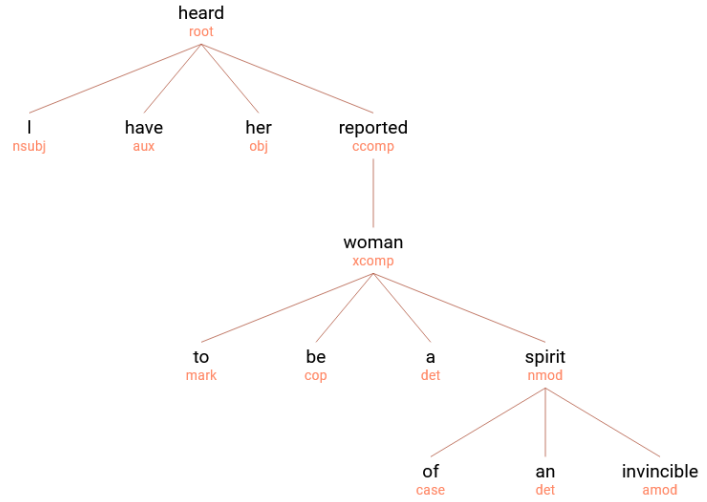
stored queries • store this query

Dependency Tree

Henry VI, Part 2

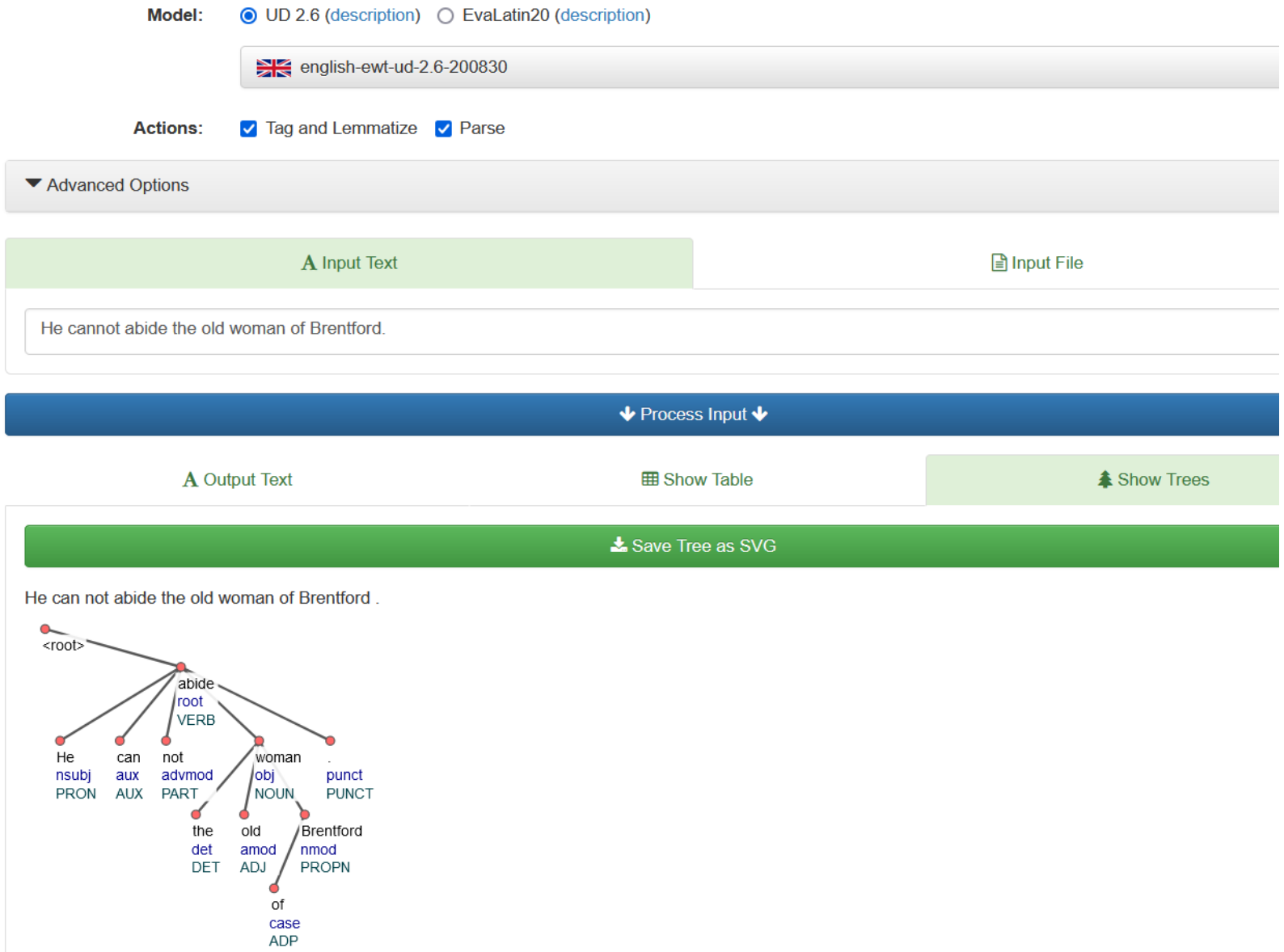
s-911 < sentence s-912

I have heard her reported to be a woman of an invincible spirit .

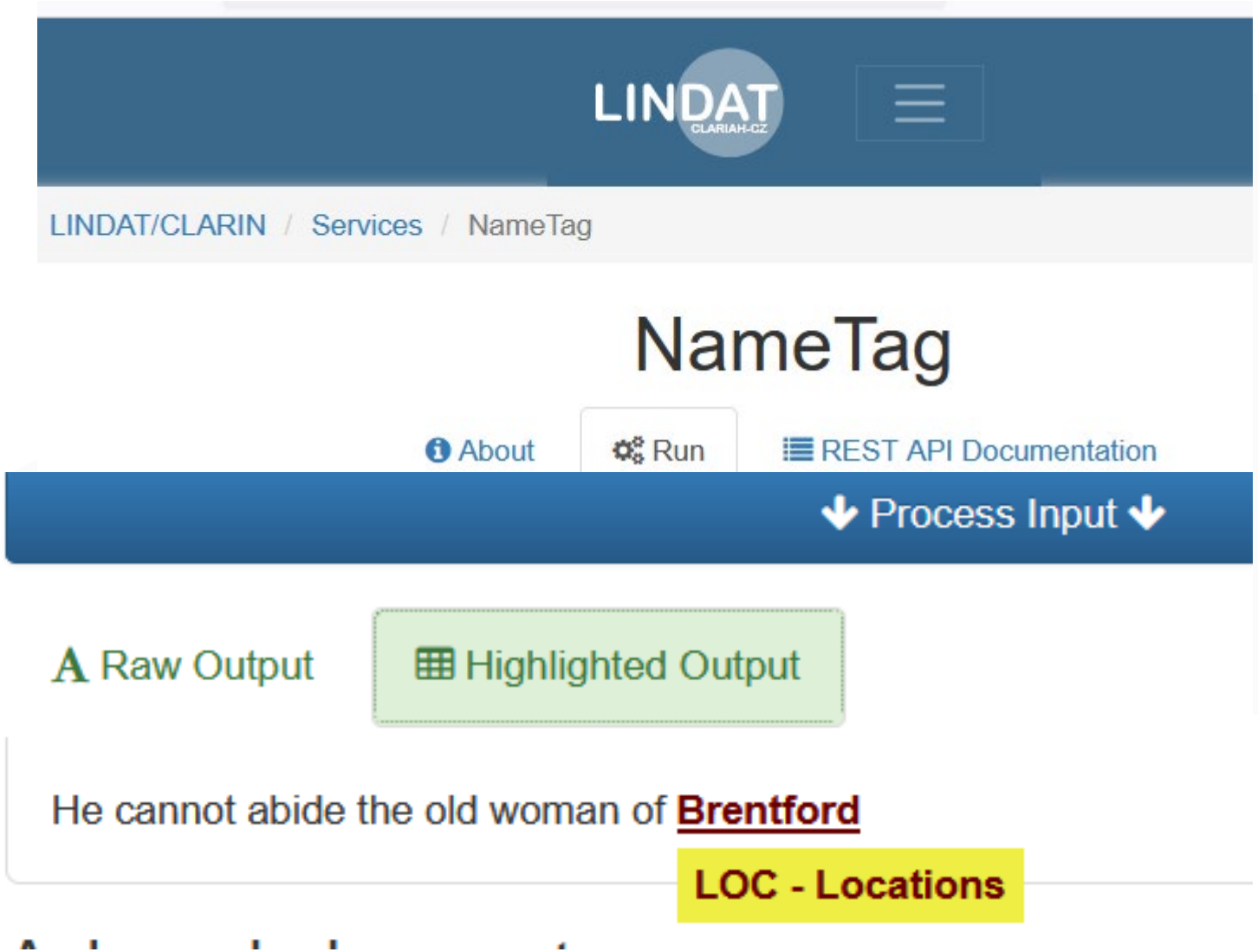


NLP tools

lemmatization, morphological tagging, syntactic parsing

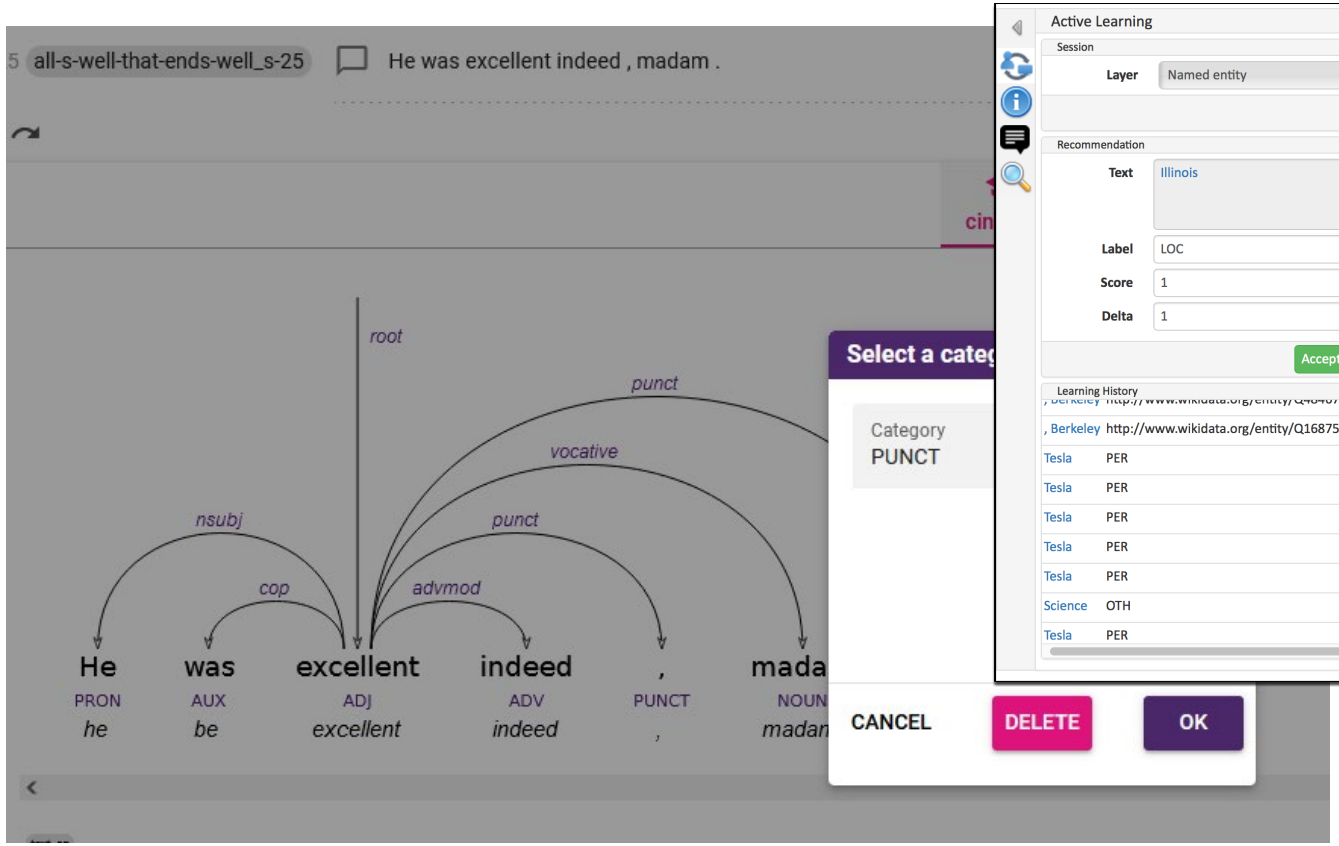


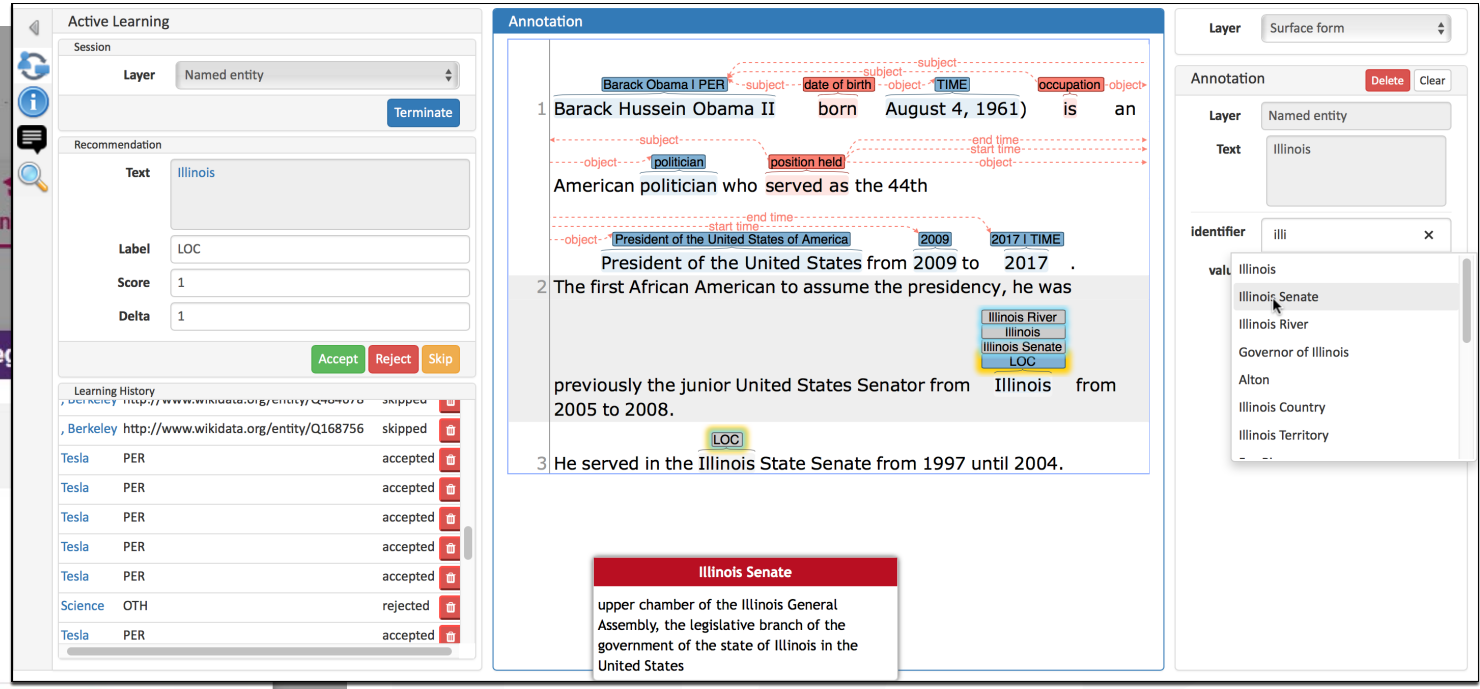
Named entity recognition



Create training data from your weird texts to get better results from the NLP tools

CLSINFRA

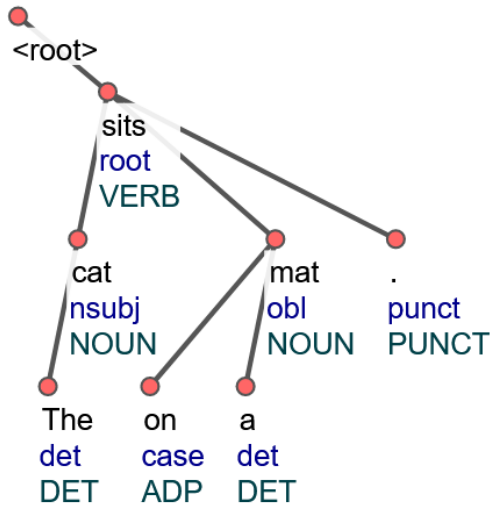


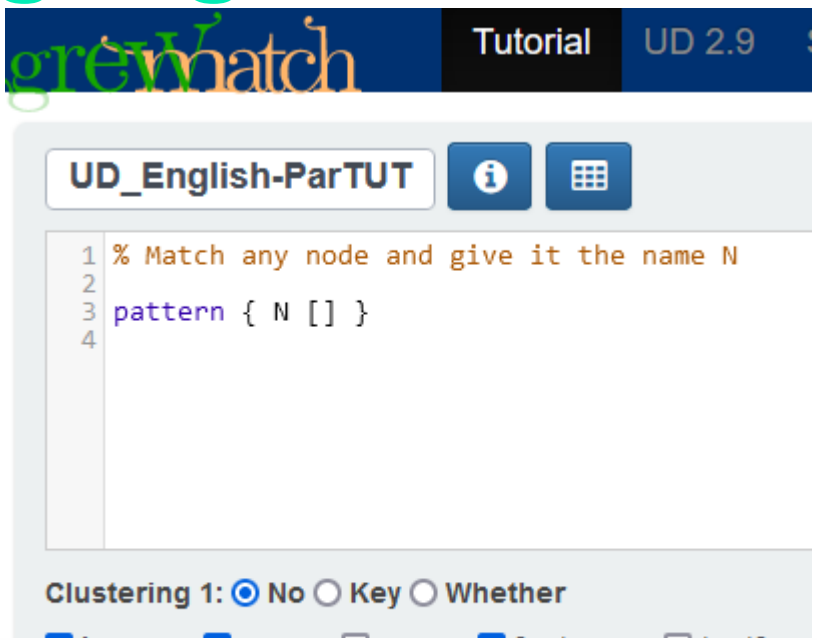


~ 5,000 tokens from your domain can do the trick!

Tree query languages

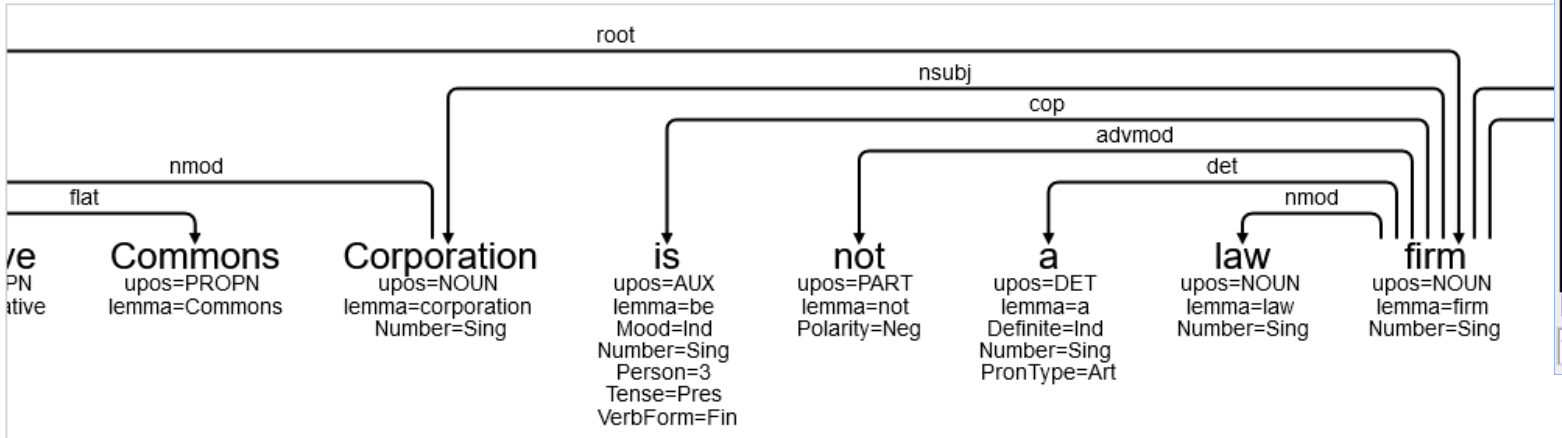
The cat sits on a mat .

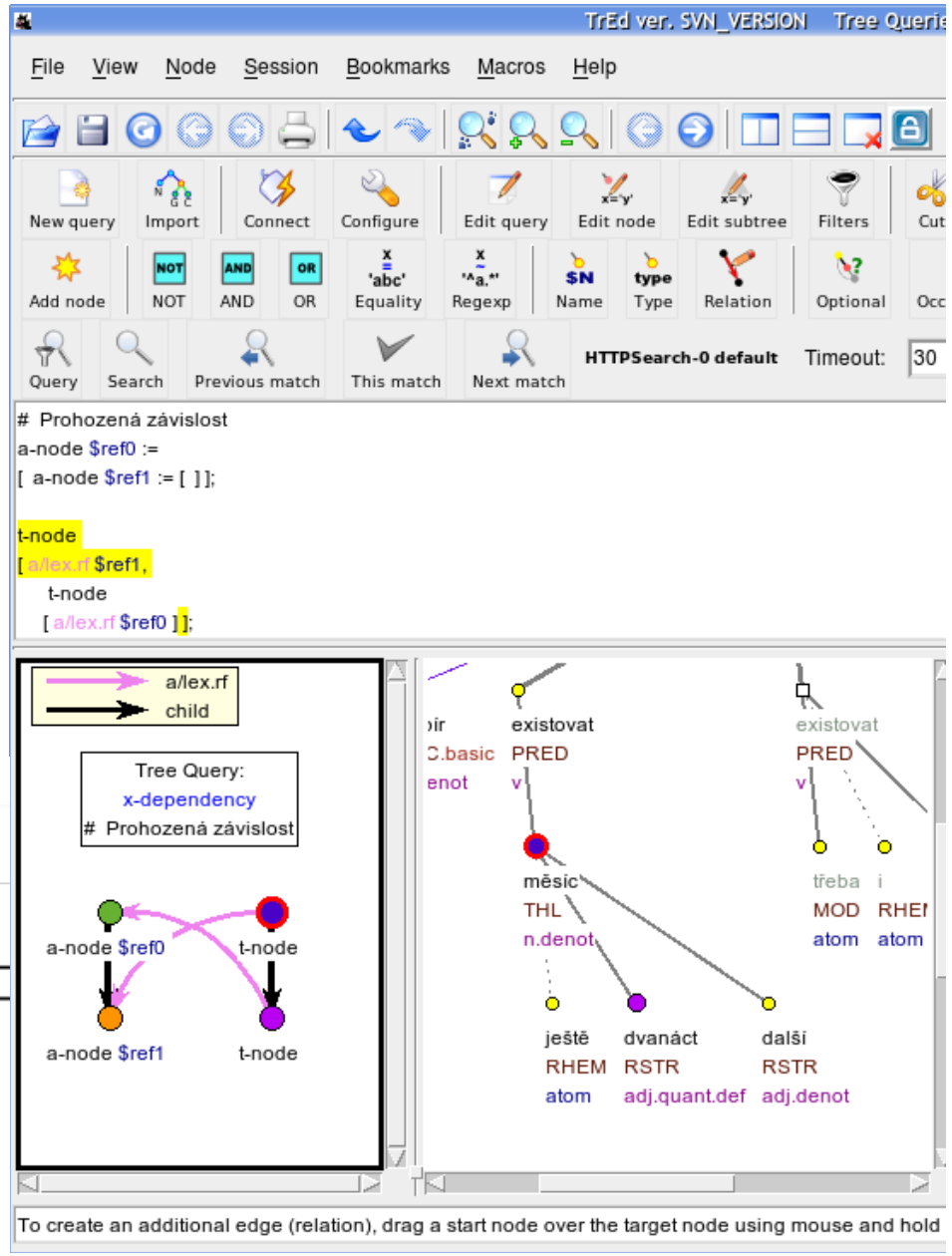




CoNLL SVG

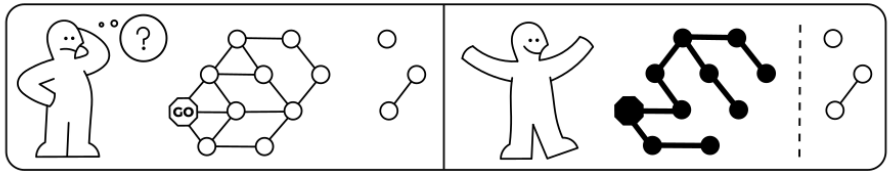
Creative Commons Corporation is not a law firm and does not provide legal services.

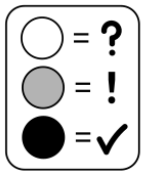


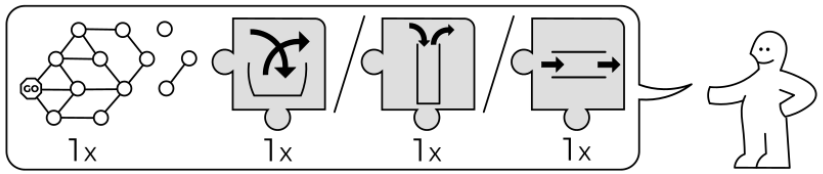


GRÅPH SCÄN

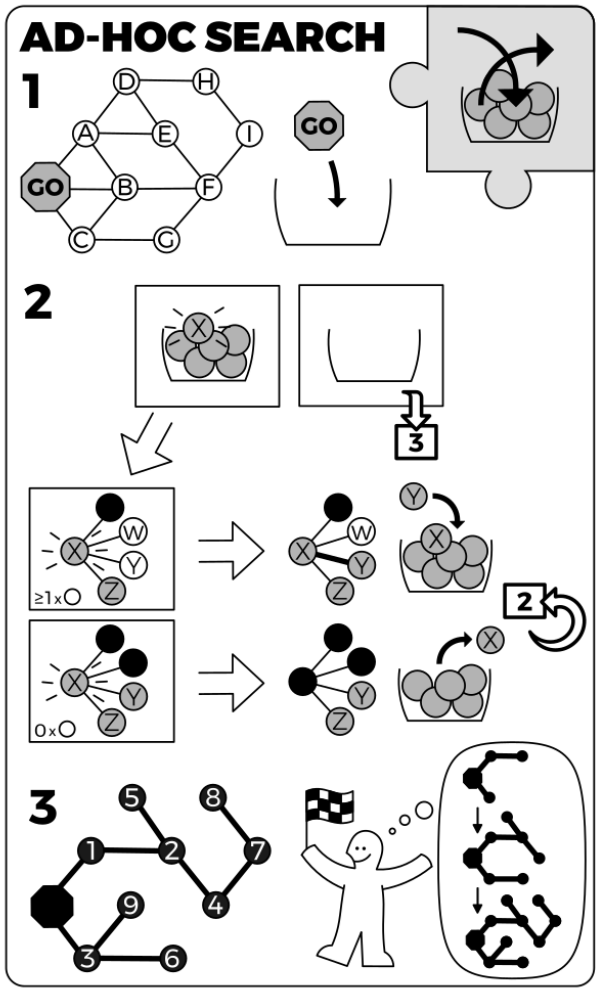
idea-instructions.com/graph-scan/
v1.1, CC by-nc-sa 4.0
IDEA



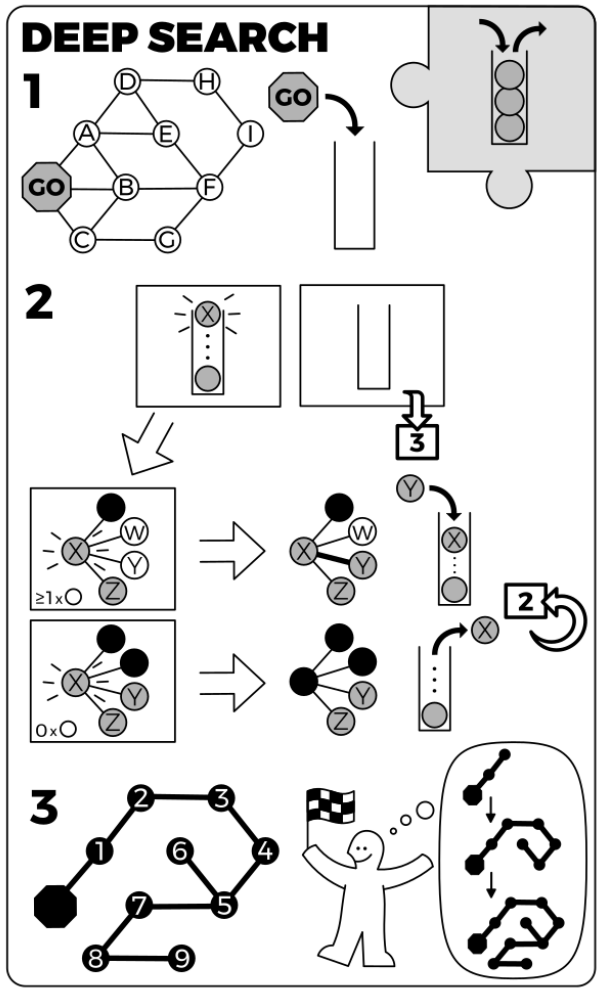




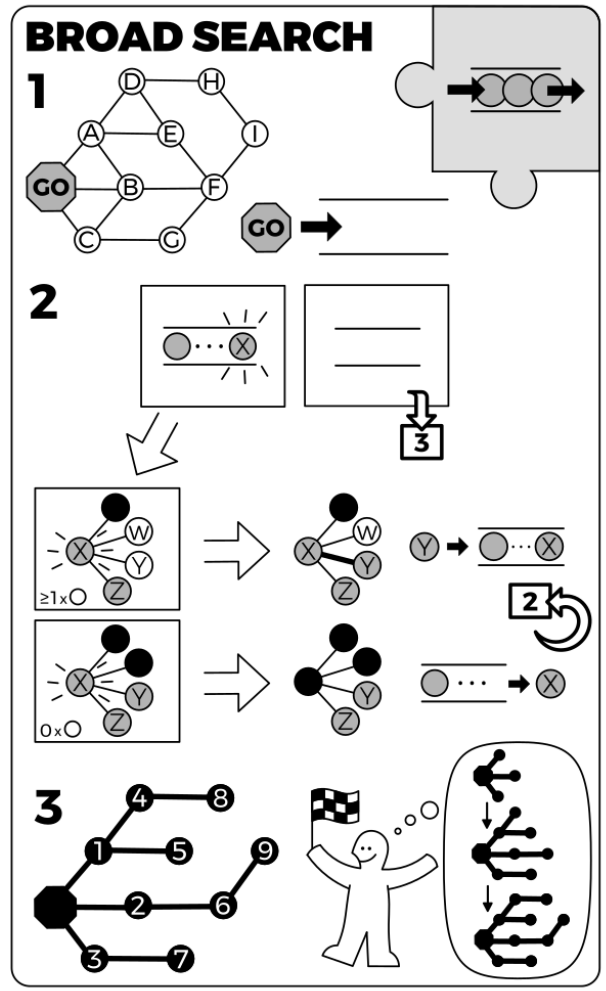
AD-HOC SEARCH



DEEP SEARCH



BROAD SEARCH



API and multi-language framework
for processing Universal Dependenci

Python Perl Java

Udapi

a framework for processing Universal Dependencies data

Links

Python: GitHub + Installation Tutorial Documentation Pap

Perl: GitHub Installation

Java: GitHub Installation

udapi.block.corefud.printentities module

class **udapi.block.corefud.printentities.PrintEntities**(*eid_re=None, min_mentions=0, print_ranges=True, mark_head=True, aggregate_mentions=True, **kwargs*) [source]

Bases: `udapi.core.block.Block`

Block corefud.PrintEntities prints all mentions of a given entity.

process_document(*doc*) [source]

Process a UD document

udapi.block.zellig_harris.enhancedeps.echildren(*node*) [source]

Return a list with node's effective children.

Parameters: **node** – An input node.

Returns: A list with node's effective children.

Return type: list

```
import re
import os.path
from udapi.core.block import Block
from collections import Counter, defaultdict

class PrintEntities(Block):                                                    [docs]
    """Block corefud.PrintEntities prints all mentions of a given entity."""

    def __init__(self, eid_re=None, min_mentions=0, print_ranges=True, mark_head=True,
                 aggregate_mentions=True, **kwargs):
        """Params:
        eid_re: regular expression constraining ID of the entities to be printed
        min_mentions: print only entities with with at least N mentions
        print_ranges: print also addressess of all mentions
                      (compactly, using the longest common prefix of sent_id)
        mark_head: mark the head (e.g. as "red **car**")
        """
        super().__init__(**kwargs)
        self.eid_re = re.compile(str(eid_re)) if eid_re else None
        self.min_mentions = min_mentions
        self.print_ranges = print_ranges
        self.mark_head = mark_head
        self.aggregate_mentions = aggregate_mentions
```