



<||/> 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



The screenshot shows the DataCamp website with a dark blue background. At the top left is the DataCamp logo and a 'WE'RE HIRING' badge. To the right are navigation links for 'Products', 'For Business', and 'Pricing'. The main heading is 'Build data skills online' in white. Below it is a sub-headline: 'Data drives everything. Get the skills you need for the future of work.' There are two prominent buttons: a green one that says 'Start Learning For Free' and a white one with a black border that says 'DataCamp For Business'. At the bottom, there is a row of logos for various data science and business intelligence tools: Python, R, SQL, Tableau, Power BI, Excel, and Oracle.

datacamp **WE'RE HIRING** Products ▾ For Business Pricing

Build data skills online

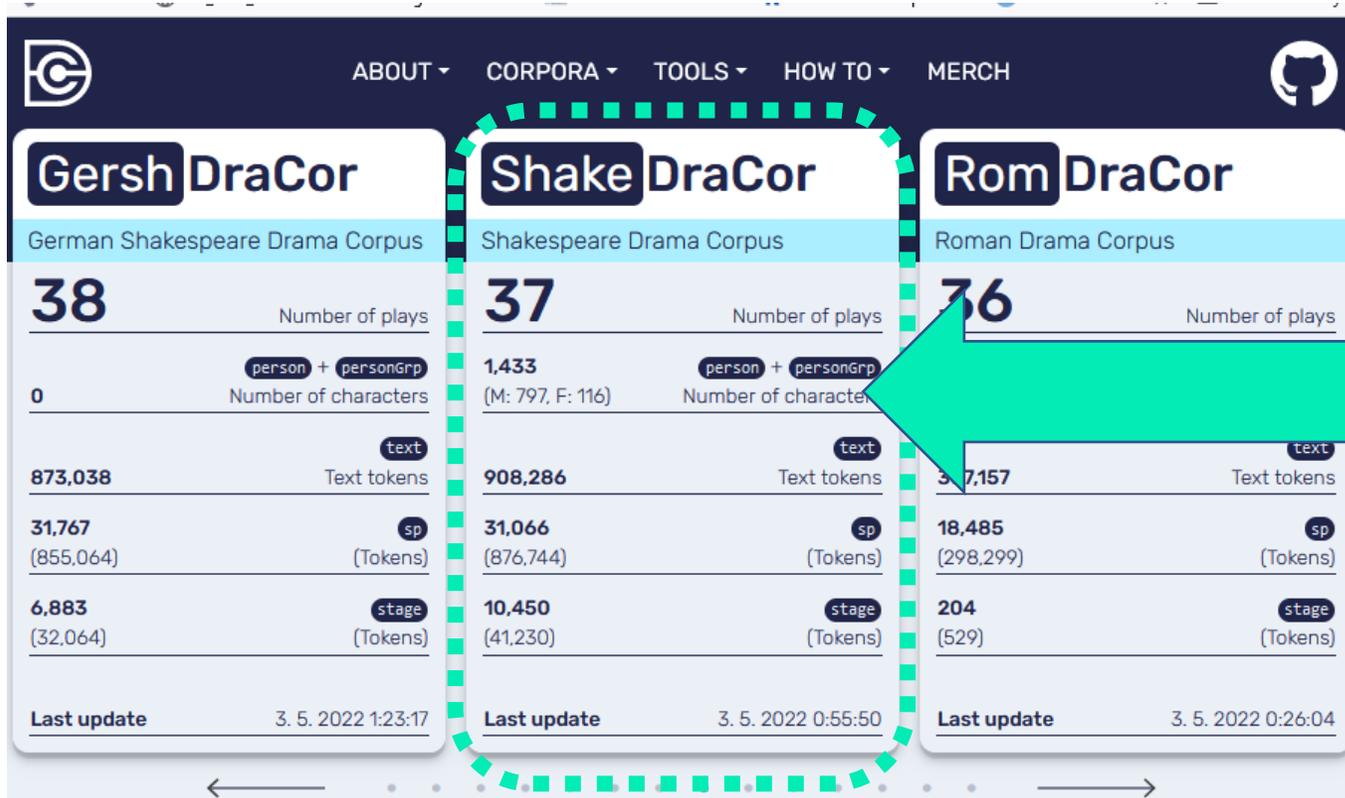
Data drives everything. Get the skills you need for the future of work.

[Start Learning For Free](#)

[DataCamp For Business](#)

python R SQL + a b l e a u Power BI Excel ORACLE

Shakespeare Material



Gersh DraCor	Shake DraCor	Rom DraCor
German Shakespeare Drama Corpus	Shakespeare Drama Corpus	Roman Drama Corpus
38	37	36
Number of plays	Number of plays	Number of plays
0	1,433 (M: 797, F: 116)	
Number of characters	Number of characters	
873,038	908,286	5,7157
Text tokens	Text tokens	Text tokens
31,767 (855,064)	31,066 (876,744)	18,485 (298,299)
(Tokens)	(Tokens)	(Tokens)
6,883	10,450	204
(Tokens)	(Tokens)	(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



The World's Largest Shakespeare Collection

The Folger Shakespeare Library, established on Capitol Hill in 1932 as a gift to the American people, is home to the world's largest collection of First Folios, the

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



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 CC BY

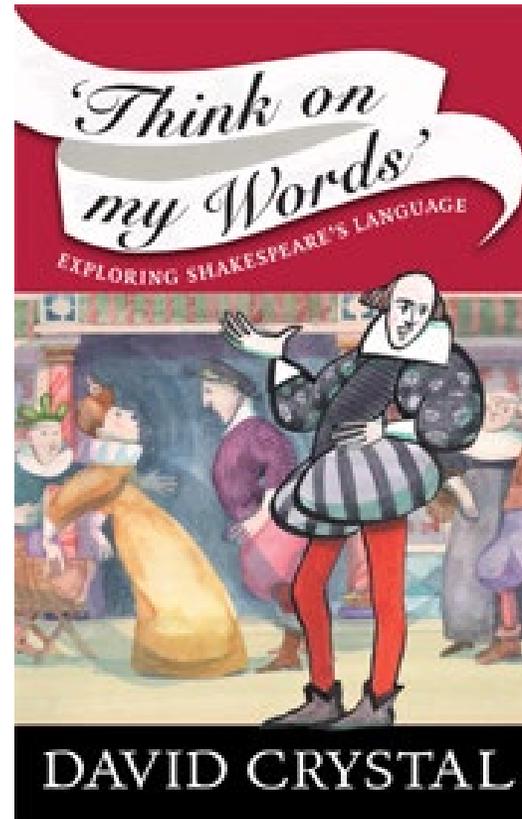
This site runs on DraCor API 0.84.0 using eXist-db 6.0.1

dracor.org

Research Ideas

- 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*

Macmillan Dictionary

OPEN | BUZZWORD

Search Macmillan Dictionary

beget DEFINITIONS AND SYNONYMS
VERB TRANSITIVE UK  /bi'get/

WORD FORMS

DEFINITIONS 2

1 FORMAL **to cause something to happen or be created**

Synonyms and related words

To make something start to exist or happen

bring about	trigger
form	...

[Explore Thesaurus →](#)

2 an old word meaning 'to become the father of a child'

- 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

- Detect all cases of genitives like this:

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

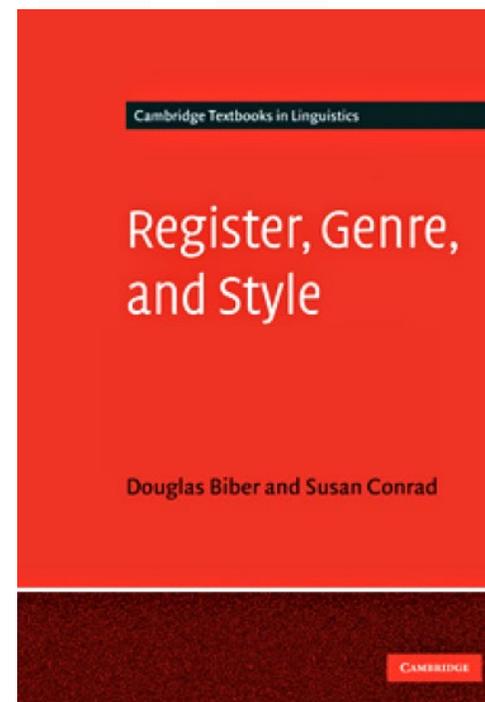
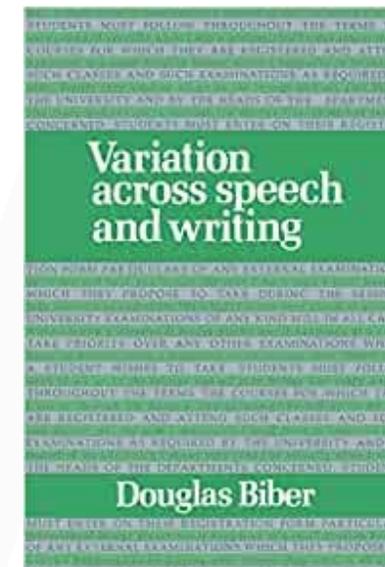
- ... and this:

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

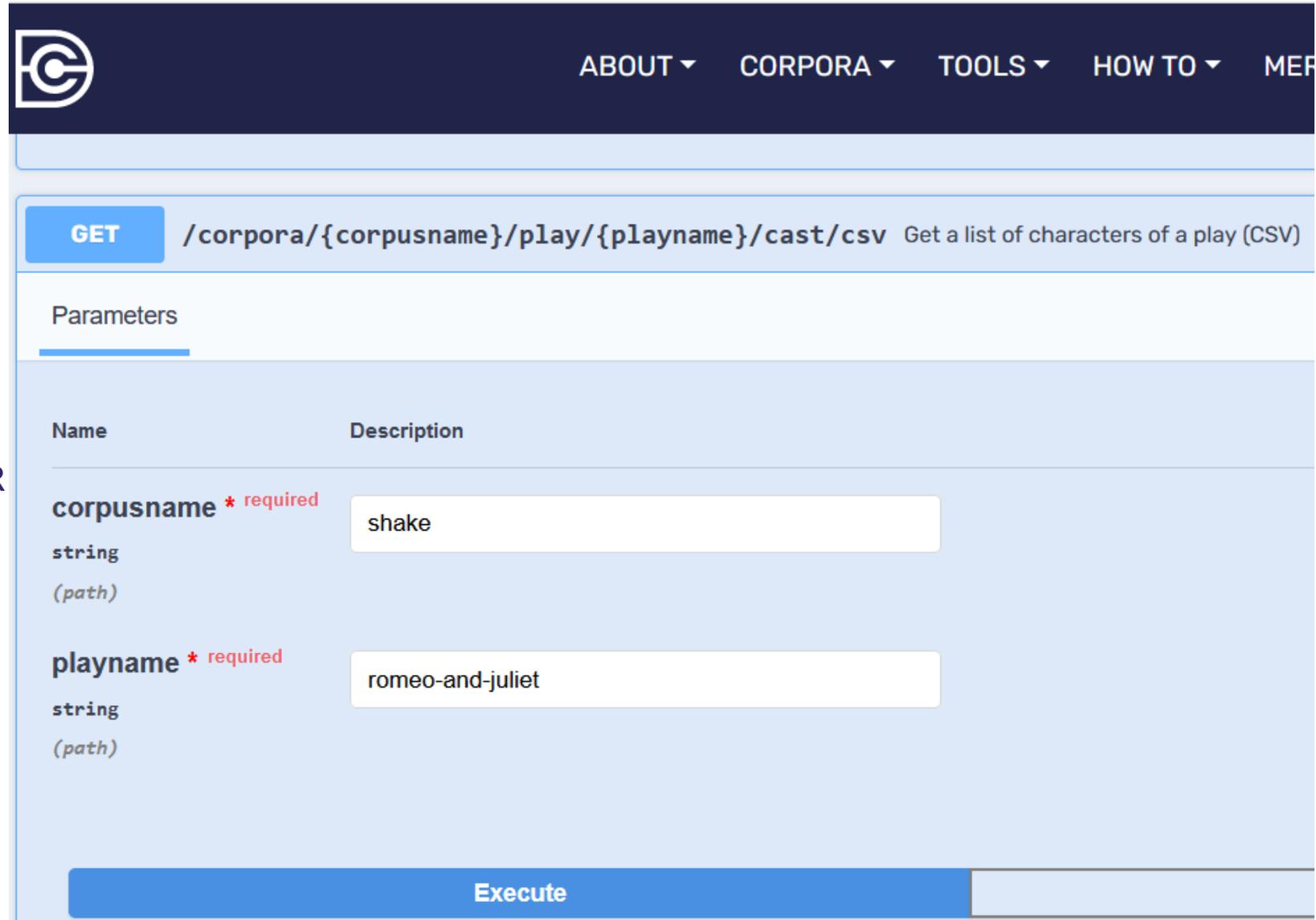
- + interjections
- + exclamation marks
- Shakespeare: short lines by one speaker – one verse in his iambic pentameter is comprised of several speakers' lines



Programmable corpora and DraCor

DraCor “Drama Corpora”

- 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



The screenshot shows the DraCor web interface. At the top, there is a dark blue navigation bar with the DraCor logo on the left and menu items: ABOUT, CORPORA, TOOLS, HOW TO, and MENU. Below the navigation bar, there is a light blue header area with a 'GET' button and the API endpoint: `/corpora/{corpusname}/play/{playname}/cast/csv`. To the right of the endpoint is the description: 'Get a list of characters of a play (CSV)'. Below this is a 'Parameters' section with a table:

Name	Description
corpusname * required string (path)	<input type="text" value="shake"/>
playname * required string (path)	<input type="text" value="romeo-and-juliet"/>

At the bottom of the interface, there is a large blue 'Execute' button.

```
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

```
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



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

Parameters

Name	Description
corpusname * required string (path)	<input type="text" value="shake"/>

Execute



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



LOG IN | Sign up | **FREE Trial** | subscribe to news

SKETCH ENGINE

Home News & Events Pr

WaG KonText SyD Morfio KWords Treq Wiki Support

INTRO DOWNLOAD MATERIALS USER

NEWS

#LancsBox: Lancaster University corpus toolbox

kon text Query Corpora Save Concordance Filter Frequency

Corpus: syn2020

Search in the corpus

syn2020

Advanced query | Keyboard | Query interpretation

TIP In case of aligned corpora, it is faster to search only within texts included in all the searched langu find such texts, 'restrict search - refine selection' function can be used. (next tip)

txm-demo UD_OSP_ODE arboratorgrew.elizia.n... DocEnhance Courses Text Plots - textplot-e...

BROWN

Corpus Home BROWN:[word='cat']

Query : [word='cat']

Sort keys: #1 : None #2 : None #3 : N

References Left context

View text.id word

Sort text.id word

Size 8

Re	Left context	Keyword
1	b09 , fringed-wrapped quiet. Alacrity, the Podger	cat
2	f11 Jimmy, from next door, let the	cat
3	g51 Sounion and I fed a thin little Grecian	cat
4	g51 . This restaurant, too, had a	cat
5	g51 , thin little creature. How can a	cat

Latest Release



AntConc

A freeware corpus analy and text analysis.

[AntConc Homepage] [S

Downloads:



Latest news
Slides and video recordings from our CL2021 talks are available now.



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>

Transcribe

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

The screenshot shows the Transcriber 1.5.1 application window. The menu bar includes File, Edit, Signal, Segmentation, Options, and Help. A pink 'report' button is at the top. The main text area contains a transcription with speaker labels and phonetic annotations. Below the text is a control bar with playback icons and a 'know' label. Underneath is an audio waveform. At the bottom is a table with columns for speaker, segment, and time.

report

speaker#2
{inhale} ((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
{inhale} ((I-)) You know what, Ann, it's like, I mean, {exhale}

speaker#1 + speaker#2

know

speaker#1	s.	speaker.	speaker#2	speaker#1	speaker#1 +...	speaker#1	s	speaker#1	speak
{inhale} ...	H	{inhale}.	I mean, you know, you...	((I-)) You know ...	I just didn't know. ...	And the thing is, {	{	You know ...	{laugh
... {laugh}	.at.	I know...	... the world.	... mean, {exhale}	I know.	... {laugh}	}	... just-	}

Cursor : 0

Audio Files

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

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

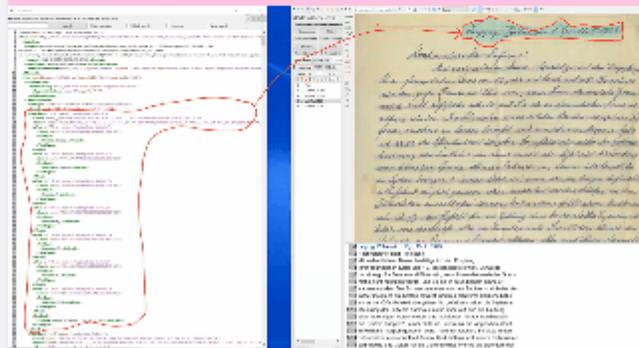
Import

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

Transcribe

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

Handwritten-Text Recognition: Transkribus



TEI

A solution for TEI- and annotation: T
All documents in
format, which can





Metadata



[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	<input type="text" value="Meyer, K. H."/>
VIAF id of sender	<input type="text"/>
Letter date (YYYY-MM-DD)	<input type="text" value="1923-02-15"/>
Letter origin (place name)	<input type="text" value="Leipz I"/>
Letter origin (geolocation)	<input type="text"/>
Recipient	<input type="text"/>
VIAF id of recipient	<input type="text"/>
Nr. of pages	<input type="text"/>
Document type	<input type="text" value="handwritten letter"/>
Language (ISO 639-3)	<input type="text" value="deu"/>
File identifier	<input type="text"/>

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



TEITOK

e Cinkova
lable Corpora

MAZON

te
Search
TQ Search
simile Search
ted Entities

Home

user: SC

in

)/P

. Files

ry Manager

vert from

/HTR

text admin

Edit Token

Filename AMA.8.19.9a.xml

Title Without Title

Token value (w-4): Systema-stisierung

pform Transcription (Inner XML) <gtok id="w-1.9.12" bbox="2862 1369 3291 1537">Systema-</gtok><lb id="e-11

wform Written form Systemastisierug

expan Expanded form

reg Regularized form

lemma Lemma

upos UD POS tag

xpos National POS tag

feats Morphosyntax

deprel Dependency relation

head Dependency head

insert tok after: [attached](#) / [separate](#) • before: [attached](#) / [separate](#) • insert elm before: [paragraph](#) ; [linebreak](#) • split in dtoks: 2 ; 3

[edit context XML](#) • [merge](#) left to w-1.9.11 • create mtok left: 1 ; 2

[treat similar tokens](#)

Leipzig, Zöllnerstr. 1 Eg., 15. II. 1923. Hochverehrter Herr Professor! Mit verbindlichem Danke bestätige ich den Empfang Ihrer freundlichen Karte vom 12., die ich heute erhielt. Es würde mir eine große Freude und Ehre sein, wenn Ihnen die russische Gram-matik nicht mißfallen würde, und Sie sie in freundlichem Sinne er-wähnen würden. Der Schwächen eines solchen Buches ist sich der Ver-fasser meistens am besten bewußt, und nur schweren Herzens habe ich es der Öffentlichkeit übergeben. So gefällt mir selbst die Systema-stisierung der Laitleure durchaus nicht; aber läßt sich die Lautung



Named-entity annotation in TEITOK

ge ich den Empfang Ihrer freundlichen Karte
ie Gram-matik nicht mißfallen würde, und Sie
asser meistens am besten bewußt, und nur
ung der Lautlehre durchaus nicht; aber läßt
jen? Ferner hätte ich, wenn mir ein langer Aufentha
tens fehlt einem Nichtrussen sehr häufig das Gefüt
i übelwollender Rezensent Kapital schlagen. Aber n
ofessor, das nicht tun werden. Eine große Freude war es für mich, daß
h äußerte. Was Ihre Fragen betr. Nikitin und Kurbskij-Briefwechsel
ch stets gern bereit finden, Ihnen Neuerscheinungen unserer
ich mir die Freiheit nehmen, Sie ande-rerseits darauf aufmerksam zu
mir anzuschaffen? Ich habe

Add NER	
Span	Nikitin
Type	Place Name
Create	Place Name
	Person Name
	Organization
	Term



Corpus Search

CQL Query:

[query builder](#)

9 results • ipm: 8.44

Tags:

[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: 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

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: 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](#)

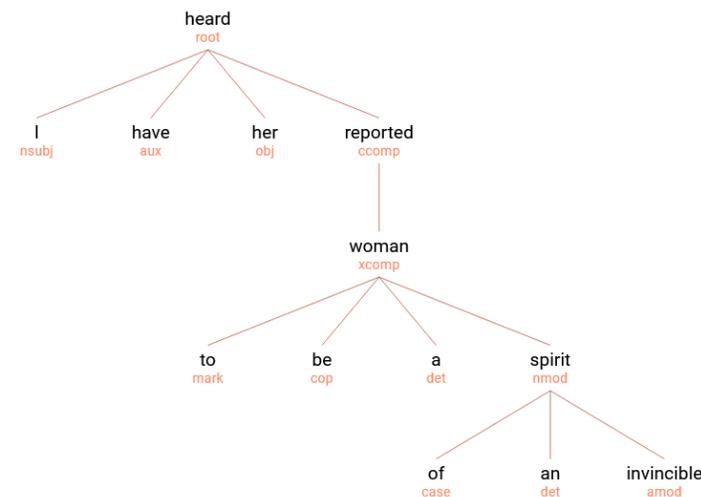
Dependency Tree

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

Model: UD 2.6 (description) EvaLatin20 (description)

 english-ewt-ud-2.6-200830

Actions: Tag and Lemmatize Parse

▼ Advanced Options

▲ Input Text

📄 Input File

He cannot abide the old woman of Brentford.

↓ Process Input ↓

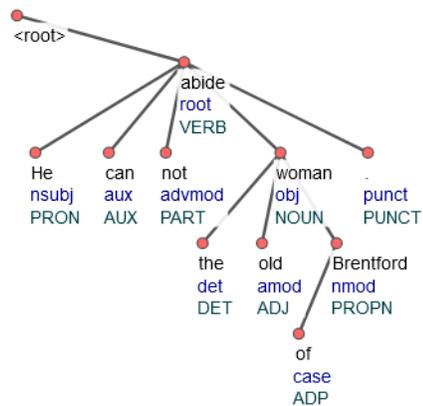
▲ Output Text

📄 Show Table

🌲 Show Trees

📄 Save Tree as SVG

He can not abide the old woman of Brentford .



LINDAT
CLARIAH-CZ



LINDAT/CLARIN / Services / UDPipe

UDPipe

[About](#)

[Run](#)

[REST API Documentation](#)



Named entity recognition

LINDAT
CLARIAH-CZ

LINDAT/CLARIN / Services / NameTag

NameTag

[About](#) [Run](#) [REST API Documentation](#)

↓ Process Input ↓

[Raw Output](#) [Highlighted Output](#)

He cannot abide the old woman of Brentford

LOC - Locations



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



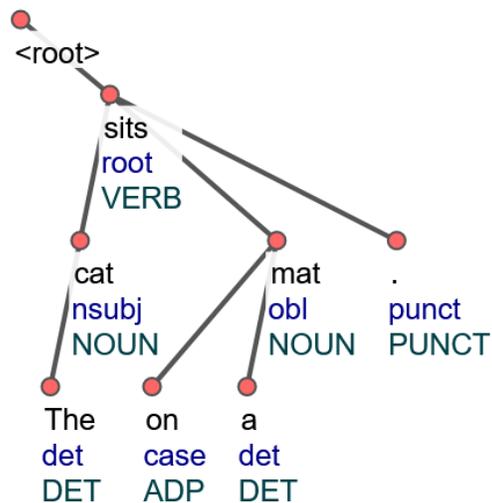
The screenshot displays a complex NLP tool interface with several key components:

- Text Input:** "He was excellent indeed, madam."
- Active Learning Panel:** Shows a "Layer" set to "Named entity" and a "Recommendation" for the text "Illinois" with a "Label" of "LOC" and a "Score" of 1. It includes "Accept", "Reject", and "Skip" buttons.
- Annotation Panel:** Displays a list of text segments with semantic annotations. For example, "Barack Hussein Obama II" is annotated as a "PERSON" (PER), "born August 4, 1961" as a "DATE" (TIME), and "President of the United States from 2009 to 2017" as a "POSITION" (TIME). A dropdown menu for "Illinois" is open, showing options like "Illinois Senate", "Illinois River", "Illinois", "Governor of Illinois", "Alton", "Illinois Country", and "Illinois Territory".
- Text Analysis Diagram:** A dependency parse tree for the sentence "He was excellent indeed, madam." showing relationships like "nsubj", "cop", "advmod", "punct", and "vocative".
- Category Selection:** A "Select a category" dialog box is open, showing "PUNCT" as the selected category.
- Learning History:** A table showing the tool's learning process, including entries for "Tesla" (PER) which were "accepted" and "Science" (OTH) which was "rejected".

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

Tree query languages

The cat sits on a mat .



UD_English-ParTUT

```

1 % Match any node and give it the name N
2
3 pattern { N [] }
4
  
```

Clustering 1: No Key Whether

TrEd ver. SVN_VERSION Tree Query

File View Node Session Bookmarks Macros Help

New query Import Connect Configure Edit query Edit node Edit subtree Filters Cut

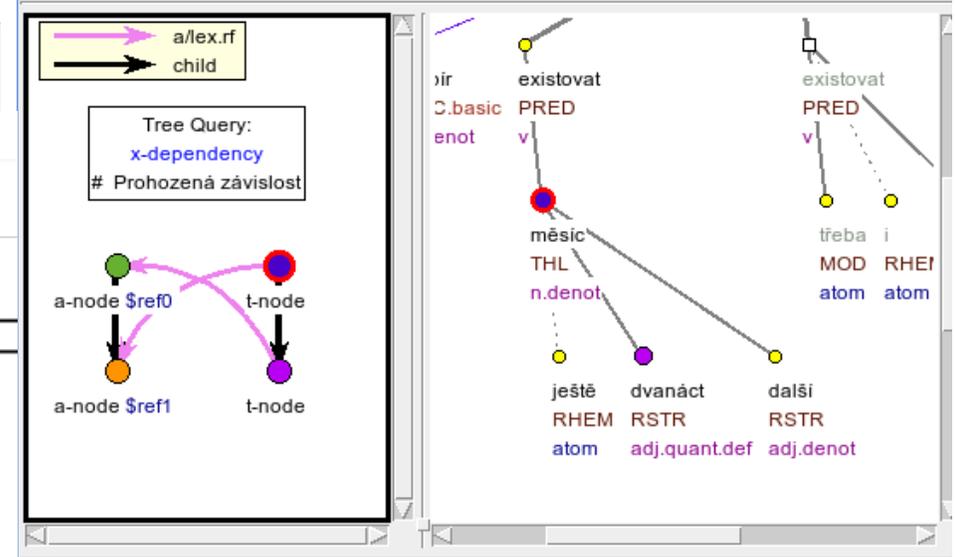
Add node NOT AND OR Equality Regexp Name Type Relation Optional Occ

Query Search Previous match This match Next match HTTPSearch-0 default Timeout: 30

```

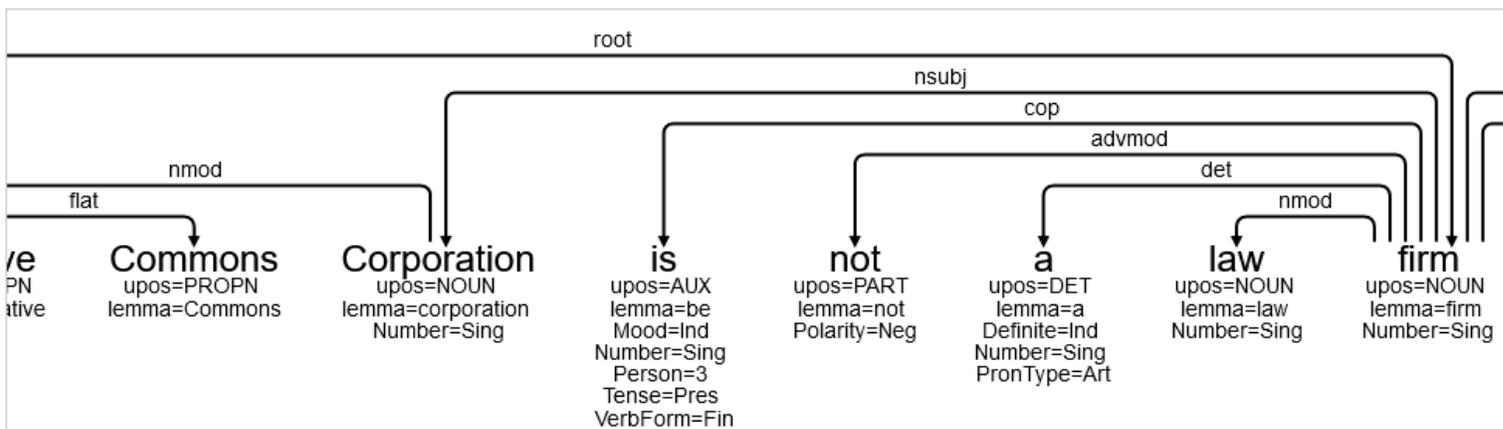
# Prohozená závislost
a-node $ref0 :=
[ a-node $ref1 := [ ] ];

t-node
[ a/lex.rf $ref1,
  t-node
  [ a/lex.rf $ref0 ]];
  
```

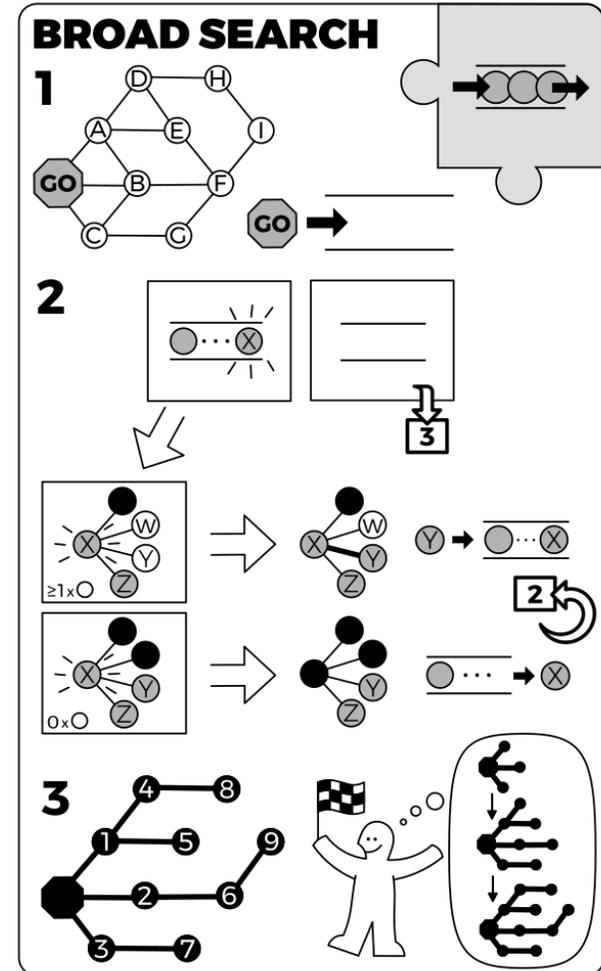
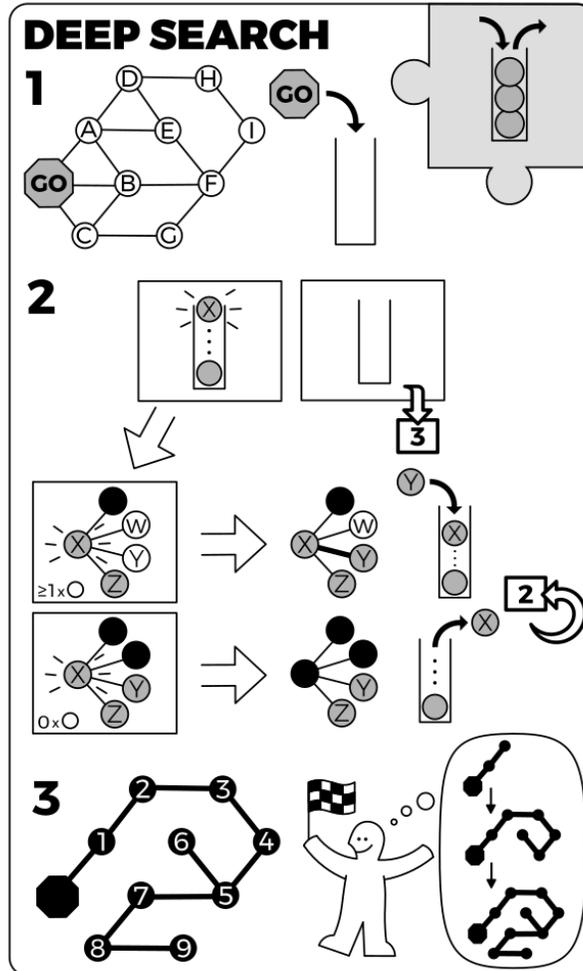
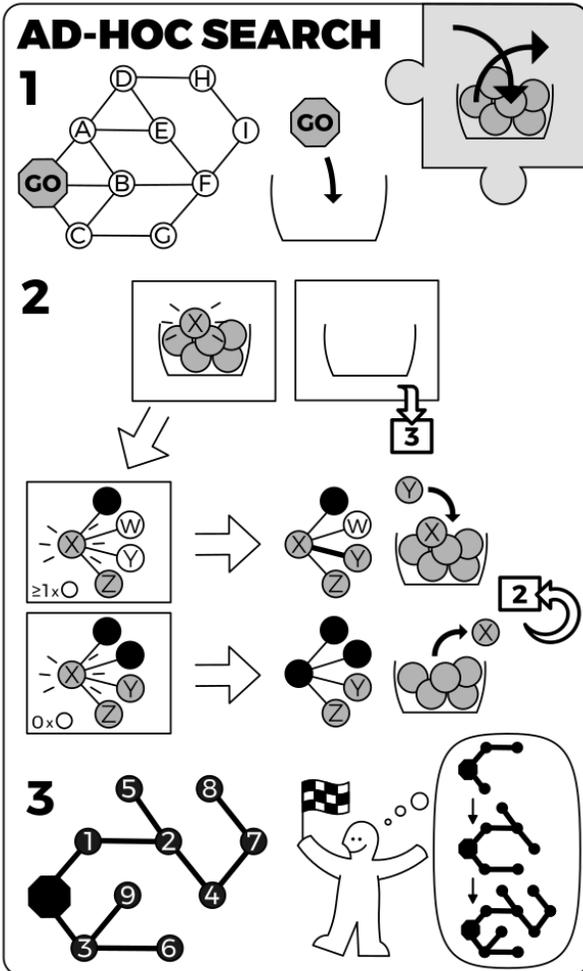
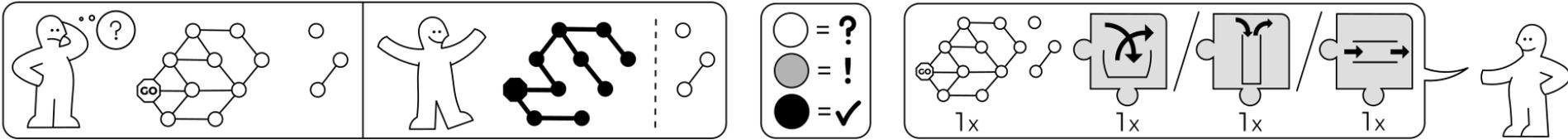


To create an additional edge (relation), drag a start node over the target node using mouse and hold

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



GRÅPH SCÄN





API and multi-language framework
for processing Universal Dependencies

Python

Perl

Java

Udapi

a framework for processing [Universal Dependencies](#) data

Links

Python: [GitHub + Installation Tutorial Documentation Page](#)

Perl: [GitHub Installation](#)

Java: [GitHub Installation](#)

`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

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

```
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 addresses 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
```