TEI@LINDAT
Enriched NLP

MAARTEN JANSSEN

UFAL

24 SEPTEMBER 2020

Como V.m. nunka ma respondido amiscartas medexado
de escriuirle si bien no he dexado deseruirle en su offº.
de V.m. de duanero. Q Como ~~[illegible]~~ martin del condado
no tuvo los recados. Como abian de ser co. reuoca
sion de my sustitucion no obstante me sito delante
del Sr. regente el dicho regente declaro. Q tuviese
los poderes. co. reuocasion mia Q co solo el acto del
consiento no bastava para mi cautela de darle
posesion de dicho offº y ansi a esto hoy yo lo sirvo yo
embie a V.m. dos copias ~~[illegible]~~ del dho acto el mas
enero noventa y seis Q remate dho co. el ... [illegible]
[illegible]

Traditional transcription

Como Vm nunha m a respondido a mis cartas m e dexado | de escrivirle si bien no he dexado d escrivirle en su Off[ici]o | de Vm de duanero q[ue] como (pedro) martin del condado | no tuxo los recados. como abian de ser co[n] | revocasion de mi sustitusion no obstante me sito delante | del S[eño]r regente el dicho regente declaro q[ue] truxese | los poderes con revocasion mia q[ue] co solo el acto del | consierto no bastava para mi cautela de darle | posesion de dicho off[ici]o y ansi asta hoi io lo sirvo io | l enbie a Vm dos copias [...] de la c[uen]ta asta [...] | enero noventa y seis q[ue] remate c[uen]tas co[n] ello y ap[ar]t[e] | monçon y los 217 [...] d esta m[one]da q[ue] le | quedava de dicha c[uen]ta se los remeti en 77 [...] 2/9 | al pro[curador] galseran balles por seguir de aquellos horden y | bolluntat de Vm. acusado de bartto[lome] valles por M[erce]d de | agostin segui de 3 de julio de 96 y tengo carta | de dicho pro[curador] galseran balles de como los resibio e yso | buenos a c[uen]ta de Vm. agora queda en mi

Plain text

Como Vm nunha m a respondido a mis cartas m e dexado de escrivirle si bien no he dexado d escrivirle en su Officio de Vm de duanero que como pedro martin del condado no tuxo los recados. como abian de ser co revoca sion de mi sustitusion no obstante me sito delante del Sr regente el dicho regente declaro que truxese los poderes cn revocasion mia que co solo el acto del consierto no bastava para mi cautela de darle posesion de dicho officio y ansi asta hoi io lo sirvo io l enbie a Vm dos copias de la cuenta asta enero noventa y seis que remate cuentas co ello y aparte monçon y los 217 d esta mda que le que dava de dicha cuenta se los remeti en 77 2/9 al pro galseran balles por seguir de aquellos horden y bollun tat de Vm. acusado de bartto valles por Merced de agostin segui de 3 de julio de 96 y tengo carta de dicho pro galseran balles de como los resibio e yso buenos a cuenta de Vm. agora queda en mi

Normalized text

Como VM nunca me ha respondido a mis cartas me he dejado de escribirle, si bien no he dejado de escribirle en su oficio de VM de aduanero, que como Martín del Condado no trajo los recados como habían de ser con revocación de mi sustitución, no obstante me sito delante del señor regente. El dicho regente declaró que trajese los poderes con revocación mía, que con sólo el acto del concierto no bastaba para mi cautela de darle posesión de dicho oficio. Y así hasta hoy yo lo sirvo. Yo le envié a VM dos copias [...] de la cuenta hasta [...] enero noventa y seis, que rematé cuentas con ello, y aparte Monzón y los 217 [...] de esta moneda que le quedaba de dicha cuenta se los remití en 77 [...] 2/9 al procurador Galcerán Vallés por seguir de aquéllos orden y voluntad de VM. Acusado de Bartolomé Vallés por merced de Agustín Segui de 3 de julio de 96. Y tengo carta de dicho procurador Galcerán Vallés de como los recibió e hizo buenos a cuenta de VM. Ahora queda en mi

Tagged text

Como	CS	como
VM	NP00000	VM
nunca	RN	nunca
me	PP1CS000	me
ha	VAIP3S0	haber
respondido	VMP0000	responder
a	SPS00	a
mis	DP1CPS	mi
cartas	NCFP000	carta
me	PP1CS000	me
he	VAIP1S0	haber
dejado	VMP0000	dejar
de	SPS00	de
escribir	VMN0000	escribir
le	PP3CSD00	le
,	Fc	,
si	CS	si
bien	RG	bien

Poor representation

- Much depends on the actual text
  - “Errors” hint at pronunciation
  - Normalization kills that
- Only for statistics
  - Good for counting, bad for viewing
- Requirements
  - Searchable corpus
  - Annotated + metadata
  - Visible in orginal form
  - With digitalized images

TEI

<pb n="[16]r" facs="PSCR6140_2.JPG"/>
<lb/>Como Vm nunha m a respondido a mis cartas m e dexado
<lb/> de escrivirle si bien no he dexado d escrivirle en su Off<ex>ici</ex>o
<lb/> de Vm de duanero q<ex>ue</ex> como <del hand="PS2">pedro</del> martin del condado
<lb/> no tuxo los recados. como abian de ser co revoca
<lb subcat="false"/>sion de mi sustitusion no obstante me <unclear>sito</unclear> delante
<lb/> del Sr regente el dicho regente declaro q<ex>ue</ex> truxese
<lb/> los poderes cn revocasion mia q<ex>ue</ex> co solo el acto del
<lb/> consierto no bastava para mi cautela de darle
<lb/> posesion de dicho off<ex>ici</ex>o y ansi asta hoi io lo sirvo io
<lb/> l enbie a Vm dos copias <gap reason="cancelled" hand="PS2" extent="1 word"/> de la c<ex>uen</ex>ta asta <gap reason="illegible" extent="2 words"/>
<lb/> enero noventa y seis q<ex>ue</ex> remate c<ex>uen</ex>tas co ello y ap<ex>ar</ex>te
<lb/> monçon y los 217 <gap reason="illegible" extent="1 word"/> d esta mda q<ex>ue</ex> le que
<lb subcat="false"/>dava de dicha c<ex>uen</ex>ta se los remeti en 77 <gap reason="illegible" extent="1 word"/> 2/9
<lb/> al pro galseran balles por seguir de aquellos horden y bollun
<lb subcat="false"/>tat de Vm. acusado de bartto valles por M<ex>erce</ex>d de
<lb/> agostin segui de 3 de julio de <hi rend="underlined" subcat="annotator">96</hi> y tengo carta
<lb/> de dicho pro galseran balles de como los resibio e yso
<lb/> buenos a c<ex>uen</ex>ta de Vm. agora queda en mi

TEI

<pb n="[16]r" facs="PSCR6140_2.JPG"/>

<lb/>Como Vm nunha m a respondido a mis cartas m e dexado

<lb/> de escrivirle si bien no he dexado d escrivirle en su Off<ex>ici</ex>o

<lb/> de Vm de duanero q<ex>ue</ex> como <del hand="PS2">pedro</del> martin del condado

<lb/> no tuxo los recados. como abian de ser co revoca

TEI

<pb n="[16]r" facs="PSCR6140_2.JPG"/>

<lb/><tok>Como</tok> <tok>Vm</tok> nunha m a respondido a mis cartas m e dexado

<lb/> de escrivirle si bien no he dexado d escrivirle en su Off<ex>ici</ex>o

<lb/> de Vm de duanero <tok>q<ex>ue</ex></tok> como <del hand="PS2">pedro</del> martin del condado

<lb/> no tuxo los recados. como abian de ser co revoca

TEI

<pb n="[16]r" facs="PSCR6140_2.JPG"/>

<lb/><tok>Como</tok> <tok>Vm</tok> nunha m a respondido a mis cartas m e dexado

<lb/> de escrivirle si bien no he dexado d escrivirle en su Off<ex>ici</ex>o

<lb/> de Vm de duanero <tok form="q">q<ex>ue</ex></tok> como <del hand="PS2">pedro</del> martin del condado

<lb/> no tuxo los recados. como abian de ser co revoca

TEI

<pb n="[16]r" facs="PSCR6140_2.JPG"/>

<lb/><tok pos="ADVERB" lemma="como">Como</tok> <tok fform="Vuestra merced">Vm</tok> nunha m a respondido a mis cartas m e dexado

<lb/> de escrivirle si bien no he dexado d escrivirle en su Off<ex>ici</ex>o

<lb/> de Vm de duanero <tok form="q" fform="que">q<ex>ue</ex></tok> como <del hand="PS2">pedro</del> martin del condado

<lb/> no tuxo los recados. como abian de ser co revoca

TEI-Based NLP

- Inline tokenization inside (any) XML
  - Adding annotation inside potentially heavy code
- Very similar workflow to traditional NLP
  - while (<FILE>) vs. while ($xml->("//tok"))
- Annotations as attributes
  - Graceful degradation
  - No need to fixed columns
  - Combine CoNLL-U with audio timing and normalization

Not Pure TEI

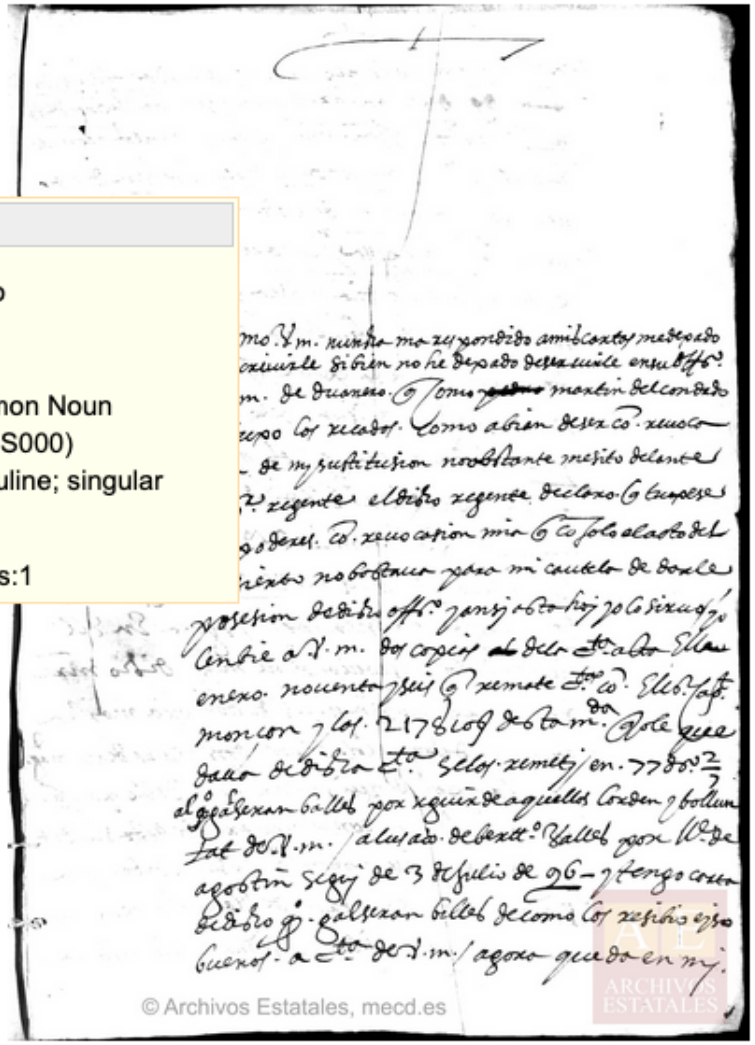
- TEI is a loose standard
  - Meaning is fixed (although often abused)
  - But many ways to encode the same information
  - TEITOK assumes a specific TEI
- Keeping information local
  - Token-based info always on the token
  - In line with traditional NLP (and more editable)
  - @bbox, @start, @end all on the element it belongs to
  - No X-Includes, X-References, etc.
- Definition file per project
  - Which features the file contains (can include non-TEI)

TEITOK text view

[fig1]

Como VM nunca me ha respondido a mis cartas me he dejado de escribirle, si bien no he dejado de escribirle en su **oficio** de VM de aduanero, que como Martín del Condado no trajo los recados como habían de ser con revocación de mi sustitución, no obstante me sito del del señor regente. El dicho regente declaró que trajes los poderes con revocación mía, que con sólo el acto concierto no bastaba para mi cautela de darle posesión de dicho oficio. Y así hasta hoy yo lo sirvo. Y le envié a VM dos copias [...] de la cuenta hasta [...] enero noventa y seis, que rematé cuentas con ello, y a Monzón y los 217 [...] de esta moneda que le quedaba de dicha cuenta se los remití en 77 [...] 2/9 al procurador Galcerán Vallés por seguir de aquéllos orden y voluntad de VM. Acusado de Bartolomé Vallés por merced de Agustín Segui de 3 de julio de 96. Y tengo carta de dicho procurador Galcerán Vallés de como los recibió e hizo buenos a cuenta de VM. Ahora queda en mi

Offo	
Expanded abbreviation	Officio
Standardization	oficio
Detailed POS	Common Noun (NCMS000) masculine; singular
Lemma	oficio
POS source	corpus:1



Full source XML


```
<p id="p-1"><seg subcat="narration"><tok id="w-1" lemma="como" mfs="CS" tagsrc="corpus:2">Como</tok> <tok id="w-2" nform="VM" lemma="VM"
mfs="NP00000" tagsrc="corpus:1">Vm</tok> <tok id="w-3" nform="nunca" lemma="nunca" mfs="RN" tagsrc="corpus:1">nunha</tok> <tok id="w-4"
nform="me" lemma="me" mfs="PP1CS000" tagsrc="corpus:1">m</tok> <tok id="w-5" nform="ha" lemma="haber" mfs="VAIP3S0"
tagsrc="corpus:2">a</tok> <tok id="w-6" lemma="responder" mfs="VMP0000" tagsrc="corpus:2">respondido</tok> <tok id="w-7" lemma="a"
mfs="SPS00" tagsrc="corpus:1">a</tok> <tok id="w-8" lemma="mi" mfs="DP1CPS" tagsrc="corpus:1">mis</tok> <tok id="w-9" lemma="carta"
mfs="NCFP000" tagsrc="corpus:1">cartas</tok> <tok nform="me" id="w-10" lemma="me" mfs="PP1CS000" tagsrc="corpus:1">m</tok> <tok id="w-11"
nform="he" lemma="haber" mfs="VAIP1S0" tagsrc="corpus:2">e</tok> <tok id="w-12" nform="dejado" lemma="dejar" mfs="VMP0000"
tagsrc="corpus:1">dexado</tok> <lb id="e-4"/> <tok id="w-13" lemma="de" mfs="SPS00" tagsrc="corpus:1">de</tok> <tok id="w-14"
nform="escribirle" tagsrc="corpus:1">escrivirle<dtok lemma="escribir" mfs="VMN0000" form="escribir" id="d-14-1"/><dtok lemma="le"
mfs="PP3CSD00" form="le" id="d-14-2"/></tok><tok id="w-15" nform="," lemma="," mfs="Fc" tagsrc="corpus:3"><ee/></tok> <tok id="w-16"
lemma="si" mfs="CS" tagsrc="corpus:2">si</tok> <tok id="w-17" lemma="bien" mfs="RG" tagsrc="corpus:4">bien</tok> <tok id="w-18" lemma="no"
mfs="RN" tagsrc="corpus:1">no</tok> <tok id="w-19" lemma="haber" mfs="VAIP1S0" tagsrc="corpus:2">he</tok> <tok id="w-20" nform="dejado"
lemma="dejar" mfs="VMP0000" tagsrc="corpus:1">dexado</tok> <tok nform="de" id="w-21" lemma="de" mfs="SPS00" tagsrc="corpus:1">d</tok> <tok
id="w-22" nform="escribirle" tagsrc="corpus:1">escrivirle<dtok lemma="escribir" mfs="VMN0000" form="escribir" id="d-22-1"/><dtok lemma="le"
mfs="PP3CSD00" form="le" id="d-22-2"/></tok> <tok id="w-23" lemma="en" mfs="SPS00" tagsrc="corpus:1">en</tok> <tok id="w-24" lemma="su"
mfs="DP3CS0" tagsrc="corpus:1">su</tok> <tok fform="Officio" id="w-25" nform="oficio" lemma="oficio" mfs="NCMS000"
tagsrc="corpus:1">Offo</tok> <lb id="e-5"/> <tok id="w-26" lemma="de" mfs="SPS00" tagsrc="corpus:1">de</tok> <tok id="w-27" nform="VM"
lemma="VM" mfs="NP00000" tagsrc="corpus:1">Vm</tok> <tok id="w-28" lemma="de" mfs="SPS00" tagsrc="corpus:1">de</tok> <tok id="w-29"
nform="aduanero" mfs="NCMS000" lemma="aduanero" tagsrc="lexicon:2">duanero</tok><tok id="w-30" nform="," lemma="," mfs="Fc"
tagsrc="corpus:3"><ee/></tok> <tok fform="que" id="w-31" lemma="que" mfs="CS" tagsrc="corpus:2">q</tok> <tok id="w-32" lemma="como"
mfs="CS" tagsrc="v">como</tok> <del hand="PS2"><tok id="w-33" form="--">pedro</tok></del> <tok id="w-34" nform="Martín" lemma="martín"
mfs="NP00000" tagsrc="corpus:1">martin</tok> <tok id="w-35" tagsrc="corpus:1">del<dtok lemma="de" mfs="SPS00" form="de" id="d-35-1"/><dtok
lemma="el" mfs="DA0MS0" form="el" id="d-35-2"/></tok> <tok id="w-36" nform="Condado" mfs="NP00000" lemma="condado" tagsrc="ending
">condado</tok> <lb id="e-6"/> <tok id="w-37" lemma="no" mfs="RN" tagsrc="corpus:1">no</tok> <tok id="w-38" dform="trujo" nform="trajo"
lemma="traer" mfs="VMIS3S0" tagsrc="corpus:1">tuxo</tok> <tok id="w-39" lemma="el" mfs="DA0MP0" tagsrc="corpus:3">los</tok> <tok id="w-40"
lemma="recado" mfs="NCMP000" tagsrc="corpus:1">recados</tok><tok id="w-41" nform="--">.</tok> <tok id="w-42" lemma="como" mfs="CS"
tagsrc="corpus:5">como</tok> <tok id="w-43" nform="habían" lemma="haber" mfs="VMII3P0" tagsrc="corpus:2">abian</tok> <tok id="w-44"
lemma="de" mfs="SPS00" tagsrc="corpus:1">de</tok> <tok id="w-45" lemma="ser" mfs="VSN0000" tagsrc="corpus:1">ser</tok> <tok fform="con"
id="w-46" lemma="con" mfs="SPS00" tagsrc="corpus:1">co</tok> <tok form="revocasion" id="w-47" nform="revocación" ltags="consonant_system"
mfs="NCFS000" lemma="revocación" tagsrc="lexicon:1">revoca<lb subcat="false" id="e-7"/>sion</tok>
```

Basic Visualization

- Directly visualizing XML document
  - Using CSS to stylize content

Bratr a Sestra.

Viktor je mladý pan z ~~Polska~~Ruska. Studuje češtinu ve škole, protože ne umí psat a čist spravně. Bydlí na koleje vedle školy, má jednu sestru Irenu, která se učí na univerzite u profesora Smutneveselého. Bohužel, Viktor není dobrý student, protože spí na lekci, ale jeho sestra ~~piše všechno~~ všechno piše a vyborně rozumí českeho profesora Smutneveseleho a brzo delá domací ukol. Večeře Irena jde na prohasku spolu z kamaradem, ale její bratr dělá nic. Jeho čeština je špatná, vím, že se vratit ve ~~Polsko~~Rusku a tam budí studovat u pomalu myt podlahy.

Kamarad Ireny je američan a chytry můž. On miluje Irenu a chce se vzít na ní. protože ona je hezká, taky chytra, rozumí ho a umí vyborně vařit.

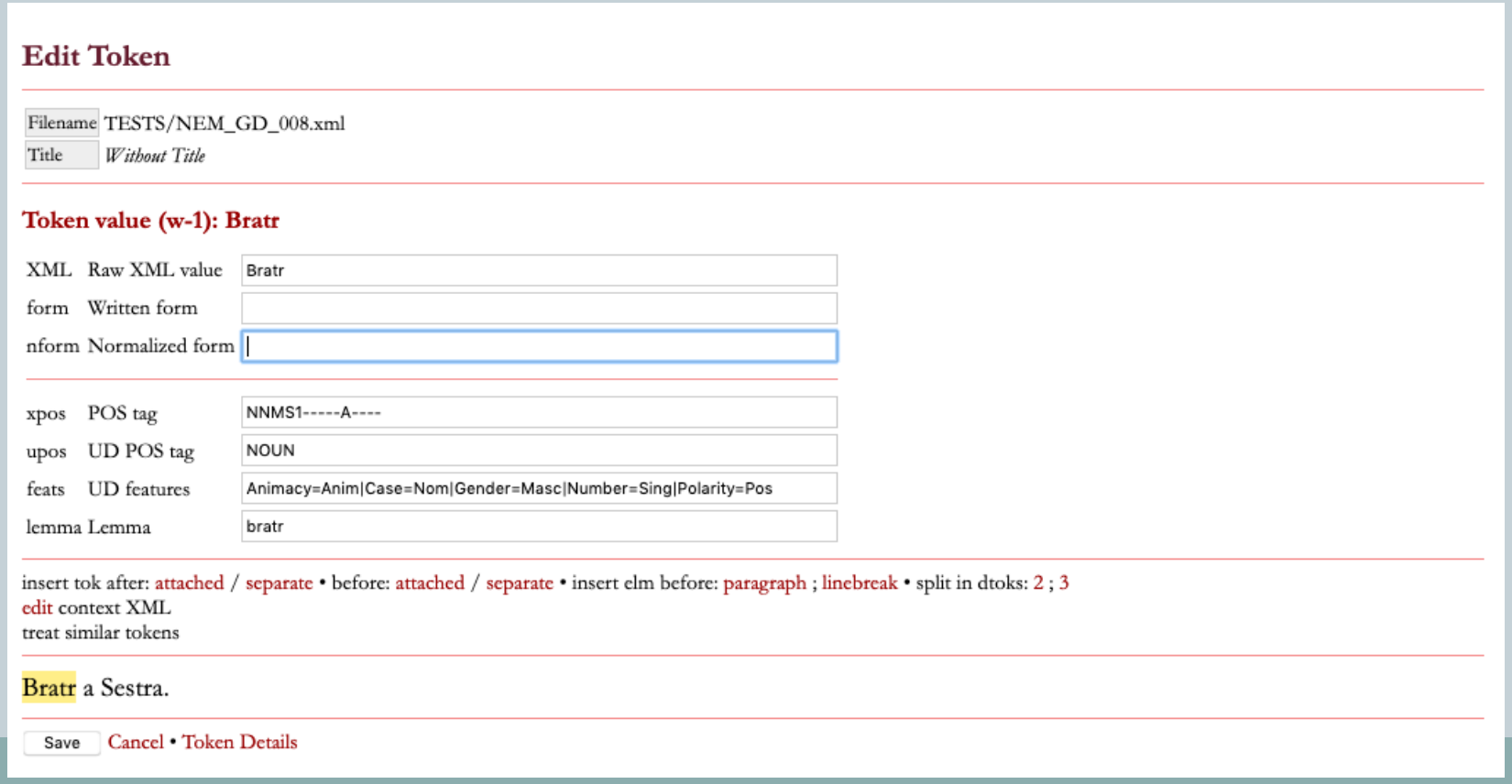
Kdo neumí nic a nechce studovat je bloubec. ~~budi~~ Bohužel, bloubec je Viktor. Ty bratr a sestra jsou moc různ~~y~~e.

To je všechno.

Konec

Edit tokens

- Edit by clicking on a word
  - Easily edit while using your corpus



Edit Metadata

Template: teiHeader-edit.tpl

Title	História da Bruxinha

Student

ID	10
Gender	masc
Age	7
Birthdate	06/07/05
Nationality	portuguesa
Lived abroad	não
Multilingual	não
Bilingual	não
Course	não
Medical assistence	sim - terapia da fala - 6 meses

Text

Type	Describe image
Task	História da Bruxinha - Chapéu
ID	BRCH

School

Searchable Index

- XML not directly searchable
  - Xquery too slow, too complex
  - CQL implementation [word="za.*"] [pos="NC"]
- Corpus WorkBench
  - Using CWB files, but with custom tools
  - Encoders writes byte-offset in XML for each corpus position
- XML indexing
  - Not directly outputting CWB results
  - CQL => positions => XML fragments

XML Search Results

Corpus Search

CQP Query: [word="Rus.*"] Search query builder | visualize | options

2 results

Text: Transcription | Written form | Normalized form

Tags: POS tag | UD POS tag | UD features | Lemma

context Viktor je mladý pan z ~~Polska~~**Ruska** . Studuje češtinu ve škole

context že se vratit ve ~~Polsko~~**Rusku** a tam | budí studovat u

Use this query for multi-token edit

Download results as TXT - Remember query

Frequency Options

Use the query above to calculate:

Collocation by: [select] | Context size: 1 | direction: Left and Right Submit Query

Frequency by: [select]

Integration with Kontext

- Kept as linked but separate tools
  - Full integration would likely break existing projects
- Two corpora created automatically from TEI
  - XML -> CWB -> vrt -> Manatee
  - For existing corpora, Manatee (or source) -> TEI/XML
- TEITOK project page
  - Description of the project/corpus + link to Kontext
- Added function in Kontext
  - Lookup word context in external source
  - TEITOK URL + text ID + token ID

TEITOK context in Kontext

	vra_ka_129_01_t_1	maminka tam má svojí kočku Zrzku . Chodíme spolu na	procházky	do lesa . A táta je doma a pracuje na
	kl9apanzuz_1	Milana2 Nováka2 a Obchodní školu Milana3	Procházky	. Ve volném čase si ráda čtu knížky ,
	vra_jt_148_01_t_1	S tátou jezdím na trénink nebo chodím s mámou na	procházky	. Mám rád oba dva rodiče . Jezdím s tátou
	AR_Mare_006_12_t_1	je velký takový zrzavý chodím sním na	procházky	a krmýho cvičím ho a mám ho rád
	kl9apanzuz_1	, které jsem dostala k Vánocům . Chodím ven na	procházky	s babičiným psem Maxem , jehož rasa je pudl a
	ho5dhajluc_1	do Anglie za taťkou . Když jsme se vrátily z	procházky	, byli puštěni pejsci Dak a Bady . Bady si
	VRA_LC_037_01_t_1	bílí . 4 . Rád si s nim chodim na	procházky	a rád si s nim hraju . Tato osoba ,
	cl8bpaledv_1			psem Argem . V těchto teplých dnech na
	vra_km_130_01_t_1			s babí a dědou k lapáku .
	cb1cchrmil_02_1			te jet na kole vykoupat do Rudy
	cb1akumzuz_01_1			uměl vyprávět tak skvělé vtipy !
	vra_lc_142_01_t_1			hrajeme . O víkendu chodíme s
	REZ_HAB_069_01_t_1			. Ještě jsem si vzpoměla že
	cl8bspipet_1			jde . Ahoj . Já jsem tvé

Default view | **TEITOK**

chodím sním na **procházky** a krmýho ~~cvičího~~ cvičím

View TEITOK document

Moving LINDAT to TEI

- Gradually move corpora to Kontext+TEITOK
  - Provide corpora in TEI/XML format
  - Attract new corpora
- Existing corpora
  - Manatee => VRT => TEITOK/XML
  - Generate TEITOK/XML from source
- New corpora
  - Make use of TEI options from the start
  - Optionally develop new visualization modules

Modular Set-up

- TEITOK uses same XML for different interfaces
  - Easy to develop new modules
- Existing alternative visualizations
  - Facsimile alignment
  - Audio alignment
  - Dependency trees
  - Geolocation mapping
- Under development
  - Aligned texts – translations and witnesses
  - Named-entity oriented modules

UD 2.6

Universal Dependencies 2.6

Home
Search
Languages

Login

Powered by TEITOK
Maarten Janssen, 2014-

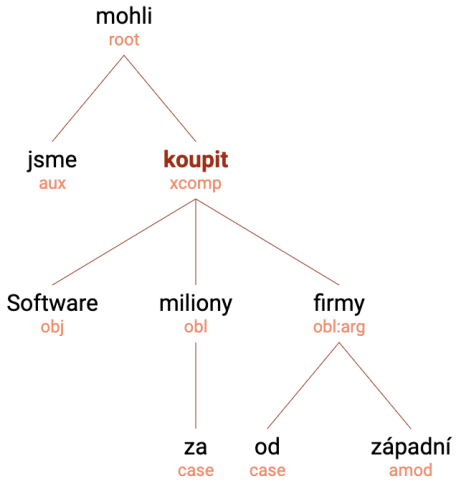
Dependency Tree

Universal Dependencies - Czech - PDT

Language	Czech
Project	PDT
Corpus Part	train
Annotation	Zeman, Daniel; Hajič, Jan

s-26 < sentence s-27 > s-28

Software jsme mohli koupit za miliony od západní firmy.



EHRI

- European Holocause Research Infrastructure
  - Recent partnership with LINDAT
- Collection(s) of holocaust related letters
  - Would be interesting as a domain-specific corpus
  - But stored as translations into German/English
- Focus on Named Entities
  - Linguistic annotations of marginal interest
- NE Module in TEITOK
  - Displays definition + corpus occurrences
  - Also useful for terminological corpora, lexicographic, etc.

Demo Page

Indexes

Bloch, Käthe

Type of index: **Person Name**

Im tschechischen Original als Blochová angegeben. Ihr weiteres Schicksal konnte von den Herausgebern bislang nicht ermittelt worden.

Occurrences

EHRI-BF-19380911_DE öffentlichen Flußghafen Prag in Ruzynĕ Käthe Bloch ein, geboren am 26

EHRI-BF-19380911_DE wollte. | Zugleich mit der Bloch flog auch der österreichische Staatsangehörige

EHRI-BF-19380911_DE längere Zeit aufhält. | Die Bloch flog um 12 Uhr ein

EHRI-BF-19380911_DE Uhr abflog und Anzenhofer die Bloch nach Prag bringen wollte,

EHRI-BF-19380911_DE bringen wollte, bat die Bloch um die Erlaubnis nach Prag

EHRI-BF-19380911_DE . 1938 fand sich die Bloch jedoch nicht zum Abflug nach

EHRI-BF-19380911_DE teilte mit, dass die Bloch dort nach einer starken nervlichen

EHRI-BF-19380911_DE Station ermittelt, dass die Bloch, obwohl sie die Karte

EHRI-BF-19380911_DE dass das gesamte Handeln der Käthe Bloch nur darauf abzielte, illegal

EHRI-BF-19380911_DE Reisepass beigefügten Valutenbestätigung hat die Bloch nur 10 RM bei sich

EHRI-BF-19380911_DE dem Schreiben den Reisepass der Bloch bei, der für das

EHRI-BF-19380911_DE hier hinterlegt wurde. | Die Bloch , die in der Zwischenzeit

EHRI-BF-19380911_DE 9. 1938 ist die Bloch nach Holland abgeflogen, nachdem

Creating New Corpora

- TEITOK initially used existing TEI/XML files
  - Created manually in Oxygen
  - Not very user friendly
- Several modules to transcribe directly in TEITOK
  - Audio, Facsimile, raw XML
- New conversion tool
  - Convert existing files to TEITOK/XML
  - Drag-and-drop or CURL
  - Word, RTF, HTML, MD, ODT, LaTeX (pandoc)
  - EXB, Praat, CHAT, Toolbox, SRT
  - FoLiA, PML, TCF
  - hOCR, ALTO
  - Brat, TMX

Annotation Pipeline

- Drag-and-drop NLP tool
  - Drop a file
  - Convert to TEITOK/XML
  - Detect language (using CWALI)
  - Ask for tool (UDPipe, Morphodita)
  - Tokenize and parse (and optionally clean first)
- Scriptable conversion from file to tagged TEI
  - Word, RTF, HTML, MD, ODT, LaTeX (pandoc)
  - EXB, Praat, CHAT, Toolbox, SRT
  - Plain text, CONLL-U
  - JPEG, GIF (tesseract)

Static Versions

- TEITOK corpora are "live"
  - Explicitly editable, reindexed frequently
- Repository requires static corpora
  - DOI assignment
  - Reproducability
- Create static dump
  - Still under discussion how
  - Copy the entire project to static number (with interface)
  - ZIP the XML files (no interface)
  - Copy the CQP corpus (searchable)
  - Automate adding to repository (create CDMI)
  - Create "groups" in Kontext

Conclusion

- TEITOK initially meant for visualization at LINDAT
  - Hidden backoffice for Kontext
  - But offers much more options
- Attractive for different corpora
  - EHRI, EEBO, Old Gospels, etc.
  - Requires user friendly conversion / creation tools
- New corpus list
  - Central index + index within each interface
  - Kontext, TEITOK, PML-TQ, Repository
  - Change corpus list in Kontext (groups)
- Design still open
  - Hiding difference between Kontext and TEITOK or keep
  - More control over TEITOK side