TEITOK

MAARTEN JANSSEN

CHARLES UNIVERSITY

27 MAY 2019

Como V.m. nunca me a respondido a mis cartas me dexado de escriuirle si bien no he dexado de seruirle en su off.º de V.m. de Duanero. Como ~~pedro~~ martin del condado no tuuo los recados. Como abian de ser co. reuoca gion de my sustituçion no obstante me dito delante del S.r regente el dicho regente declaro. Q tuuiese los poderes. co. reuocaçion mia Q co solo el acto del consiento no bastaua para mi cautela de darle posesion del dicho off.º y ansi oi dicho hoy yo lo sirue y yo le enbie a V.m. dos copias ~~de~~ del d.ho acto. El ca enero noventa y seis Q remate d.ho co. El S. Jos.
da

Letter

- Written in 1599
  - By a “normal” person
  - Merchant, writing to the customs office
  - Kept as part of a legal procedure
- Wealth of information
  - Semi-oral text
  - Contains errors, self-corrections, etc.
  - Characters often written differently from now/monks
  - Gives much insight into “normal” language from the 16th
  - Much depends on the actual text

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

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

Corpus Linguistics

- POS and lemma improve texts considerably
  - Searches become much more versatile
  - You can count in various ways
  - Statistics based linguistic analysis
- NLP tools do not do XML
  - Step 1: clean out any code
- NLP tools are meant for “normal” language
  - Step 2: normalize the text

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

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

Unlinked texts

- Typesettings only kept in TEI original
  - Not available in tagged version
  - No link between the two except the text itself
- Corrections break the link
  - Modifications need to be made in both versions
  - Failure to do so will kill the only link
  - Linguistic information unlinked to philological information
- Resulting corpus less than optimal
  - Does not keep what makes the text interesting
  - Ignores all complications and doubts

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

Verticalized

Verticalized Corpus View

XML File: Revistas/ModernizadasTeitok/neotag_ES/PSCR6140.xml

1599. Carta de Pedro Santus, mercader, para Pedro Navarro, aduanero y escribano.

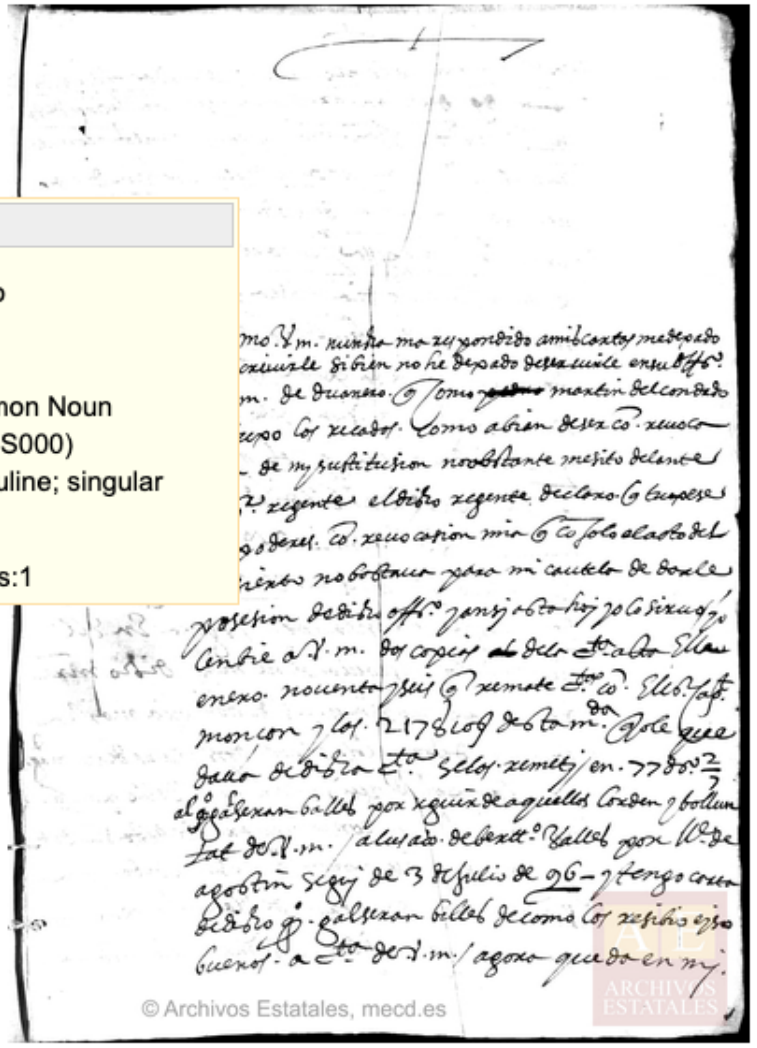
	Transcription	Expanded abbreviation	Standardization	Detailed POS	Lemma
w-1	Como	Como	Como	CS	como
w-2	Vm	Vm	VM	NP00000	VM
w-3	nunha	nunha	nunca	RN	nunca
w-4	m	m	me	PP1CS000	me
w-5	a	a	ha	VAIP3S0	haber
w-6	respondido	respondido	respondido	VMP0000	responder
w-7	a	a	a	SPS00	a
w-8	mis	mis	mis	DP1CPS	mi
w-9	cartas	cartas	cartas	NCFP000	carta
w-10	m	m	me	PP1CS000	me
w-11	e	e	he	VAIP1S0	haber
w-12	dexado	dexado	dejado	VMP0000	dejar
w-13	de	de	de	SPS00	de
w-14	escrivirle	escrivirle	escribirle	escrivirle	escrivirle
d-14-1	escribir	escribir	escribir	VMN0000	escribir
d-14-2	le	le	le	PP3CSD00	le
w-15			,	Fc	,

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>
```

Unmanagable XML

- Combined XML is barely readable
  - Need for a Graphical User Inferface (GUI)
- TEITOK is a web-based GUI
  - Corpus of TEI/XML files
  - Visualizing TEI/XML files
  - Creating and Editing TEI/XML files
  - Searching TEI/XML based corpora
- General purpose corpus tool
  - Usable for various different kinds of corpora (historic, learner)
  - Customizable (functional, visual)
  - Freely available (Git repository)

Pros and Cons

- Files get much larger
  - Files about 10x as large as in “NLP” format
  - But hard disks are cheap these days
  - Medium size corpora still fit on a pen drive
- One corpus to serve both needs
  - Unmodified TEI + linguistic annotation
  - Rich combination of information that can be exploited
- XML is flexible
  - CoNLL files have fixed numbers of columns
  - TEITOK/XML files can have any attributes you need

GUI Optimalization

- GUIs need to be optimized
  - The placement of buttons can have huge impact on efficiency
- Need for user feedback
  - 42 projects using TEITOK at this moment
  - Very different settings
  - Actively asking for feedback
- User community
  - Google group, facebook page
  - Not very active for the moment

Various projects

Historic corpus	14
Learner corpus	9
Spoken corpus	9
Reference corpus	4
Dictionary	2
Psycholinguistic	1
Less-Resourced	4
Multilingual corpus	1
Developmental	3

Self Sufficiency

- Linguistic corpora built by corpus linguists
  - Taking over from plain text
  - Too easy to mess things up in verticalized format
- Specialized corpora need specialists
  - Historical linguists (historical corpus)
  - Language aquisition specialists (learner corpus)
  - Phoneticians (spoken corpora)
- TEITOK - take care of your own corpus
  - Interface designed to be “user friendly”
  - Designed to avoid user errors

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.

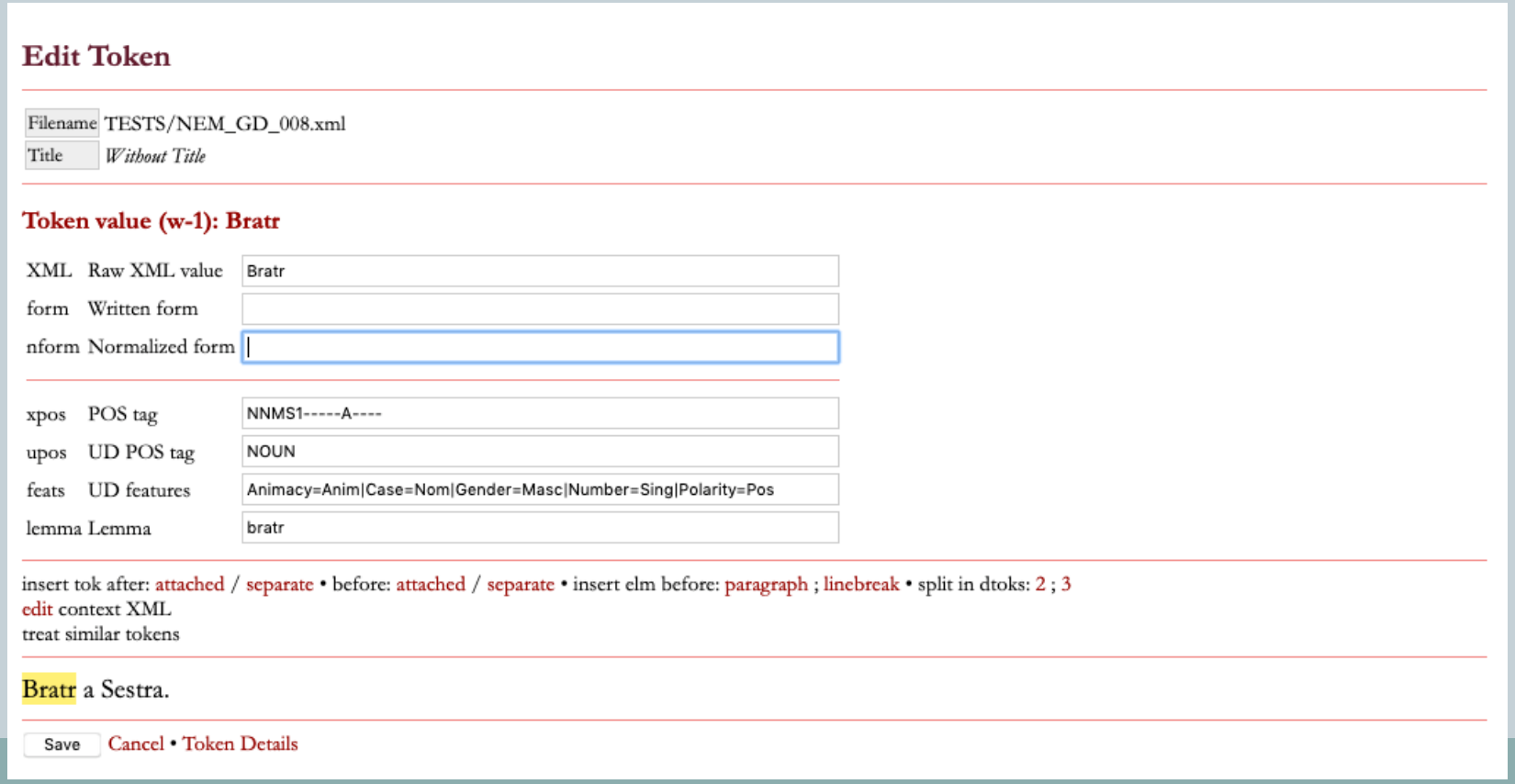
Kdo neumí nic a nechce studovat je bloubec. ~~budi~~ Bohužel, bloubec je Viktor. Ty bratr a sestra jsou moc různ~~y~~e.

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	

Multiple Orthographic Realizations

<tok lemma="Gallia" ana="NP" reg="Gallia">
<hi rend="dropcap" n="9">G</hi>ALLIA
</tok>

Normalization, Regularization, Target hypothesis, Modernization, Romanization, Diplomatic form, etc.

von 20 in Stadt X ***exestierte*** *Baugenossenschaften*

Orthographic error – TH: existierte

Morphological error – TH: existierten

Lexical error – TH: existenten

Multiple Editions from one XML

Don Afonſſo de Caſtela	Don Afonsso de Castela	Don Affonso de Castela	Dom Afonso de Castela,
de Toledo de Leon	de Toledo de Leon	de Toledo, de Leon,	de Toledo, de Leão,
Rey e ben deſ Copostela	Rey e ben des Copostela	Rey e ben des Conpost	Rei e bem de Compostela
ta o Reyno daragõ	ta o Reyno daragon	ta o reyno d'Aragon.	até o reino de Aragão.
De Cordoua• de Iahen	De Cordoua• de Iahen	De Cordova, de Jahen,	De Córdova, de Jaén,
de Seuilla outroſſi	de Seuilla outrossi	de Sevilla outrossi	de Sevilla outrossim
⁊ de Murça u gran bẽ	e de Murça u gran ben	e de Murça, u gran ben	e de Múrcia, onde grão bem
lle fez deus com aɲndi	lle fez deus com aprendi	lle fez deus com aprend	lhe fez deus como aprendi;
Do algarue q̃ gãou	Do algarve que ganou	Do Algarve, que ganou	Do Algarve, que ganhou
de mouros ⁊ noſſa fe	de mouros e nossa fe	de mouros e nossa fe	de mouros e nossa fé
meteu y• ⁊ ar poblou	meteu y• e ar poblou	meteu y, e ar poblou	meteu aí, e ar povoou
badaloz que Reyno e	badaloz que Reyno e	badaloz, que reyno e	Badajoz, que reino é

Zographensis

Zographensis

ⰵⰲⰰⱀⰳⰵⰾⰻⰵ ⱁⱅⱏ
ⰿⰰⱃⱐⰽⰰ ⸭

ⰸⰰⱍⱑⰾⱁ ⰵⰲⰰⱀⰳⰵⰾⰻⱑ · ⰻⱄ ⱈⰰ
ⱄⱀⰰ ⰱⰶⰻⱑ · ⱑⰽⱁⰶⰵ ⰵ
ⱄⱅⱏ ⱂⰺⱄⰰⱀⱁ ⰲⱏ ⱂⱃⱁⱃⱁⱌⱑⱈⱏ ·
ⱄⰵ ⰰⰸⱏ ⱂⱁⱄⱏⰾⱗ ⰰⰳⰾⱏ ⰿⱁⰺ ·
ⱂⱃⱑⰴⱏ ⰾⰻⱌⰵⰿⱏ ⰿⱁⰻⰿⱏ · ⰺⰶⰵ
ⱆⰳⱁⱅⱁⰲⰻⱅⱏ ⱂⱘⱅⱐ ⱅⰲⱁⰺ · ⱂⱃⱑ
ⰴⱏ ⱅⱁⰱⱁⱙ · ⰳⰾⰰⱄⱏ ⰲⱏⱂⰻⱙ
ⱋⰰⰳⱁ ⰲⱏ ⱂⱆⱄⱅⱏⰹⱀⰻ · ⱆⰳⱁ
ⱅⱁⰲⰰⰺⱅⰵ ⱂⱘⱅⱐ ⰳⱀⱑ · ⱂⱃⰰⰲⰹ
ⱅⰲⱁⱃⰻⱅⰵ ⱄⱅⱏⰸⰰ ⰵⰳⱁ · ⰱⰹ
ⱄⱅⱏ ⰺⱁⰰⱀⱏ ⰽⱃⱏⱄⱅⱗ ⰲⱏ ⱂⱆⱄⱅⱏⰹ
ⱀⰻ · ⰺ ⱂⱃⱁⱂⱁⰲⱑⰴⰰⱑ ⰽⱃⱏⱋⰵ
ⱀⰻⰵ ⱂⱁⰽⰰⰰⱀⰻⱙ · ⰲⱏ ⱁⱅⱏⱂⱆ
ⱋⰵⱀⰻⰵ ⰳⱃⱑⱈⱁⰿⱏ · ⰺ ⰺⱄⱈⱁ

ѥваглнѥ отъ
Марька ⸭

Зачало ѥваглнѣ · нс хва
сна бжнѣ · ѣкоже
ѥсть псано вь пророцѣхь ·
се азъ посьлѭ аглъ мои ·
прѣдь лнцемь моімь · 'іже
оуготовнтъ пѫть твоі ·
прѣдъ тобоѭ · Гласъ
въпнѭштаго въ поустынн ·
оуготова'іте пѫть гнь · правы
творнте стьза ѥго ·
Быстъ 'іоанъ кръста въ
поустынн · 'і проповѣдаа
кръштенне покаанню · въ
отъпоуштенне грѣхомъ · 'І нсхождааше

Evangeljě ot'
Mar'ka ⸭

Zachyalo evangeljě · is hva
sna bžjě · ěkože
est' psano v' proročěh' ·
se az' pos'lya aěl' moi ·
prěv' licem' moim' · iže
ougotovit' pyat' tvoi ·
prěd' toboya · Glas'
v'piyaštago v' poustyni ·
ougotovaite pyat' gn' · pravy
tvorite st'zya ego ·
byst' ioan' kr'stya v'
poysty'i · i propovědaya
kr'štenie pokaaniyu · v'
ot'pouštenie grěhom' · I ishoždaaše

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

TT-CWB-ENCODE

- Initially using verticalization + cwb-encode
  - Hard to include XML tags
- TT-CWB-ENCODE
  - Reads each XML file (C++)
  - Goes through the <tok>
  - Keeps the byte-offset of each
  - Additional nodes <p> <l> etc. by checking ID of <tok> inside
  - External stand-off annotation explicitly linked to <tok>
- .xidx
  - Indexes each pattribute and sattribute to byte offsets

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

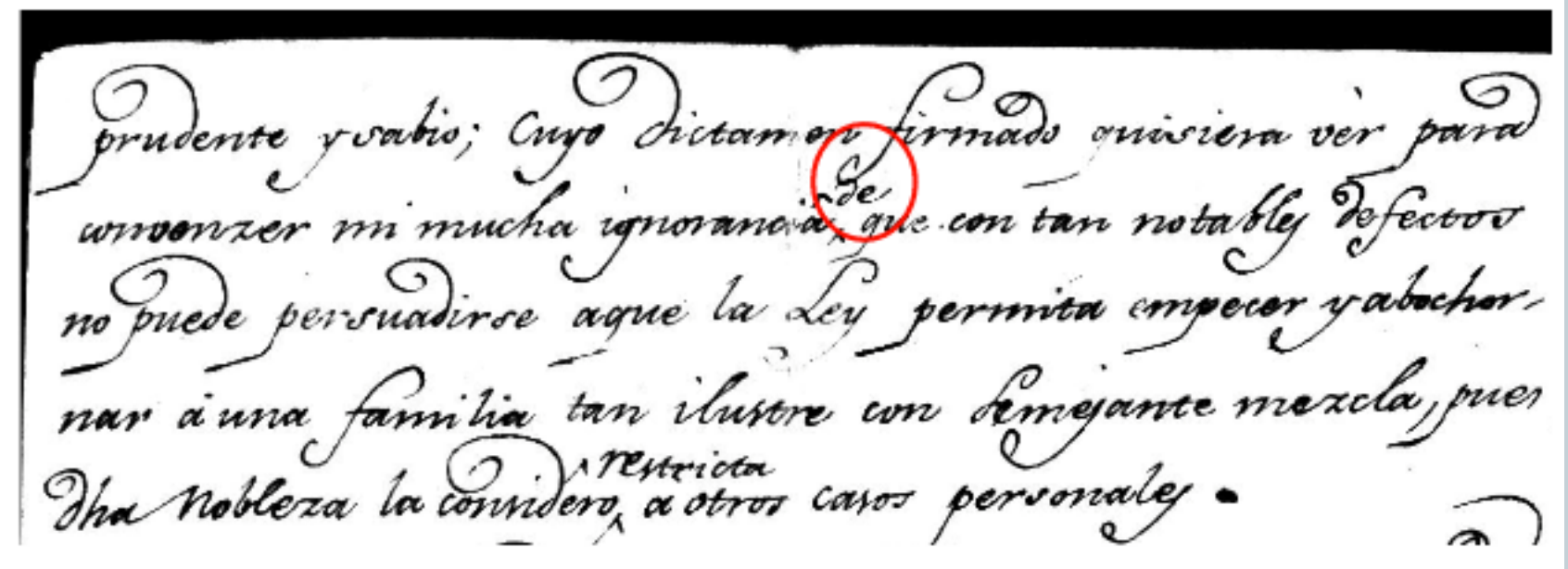
Frequency Options

Use the query above to calculate:

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

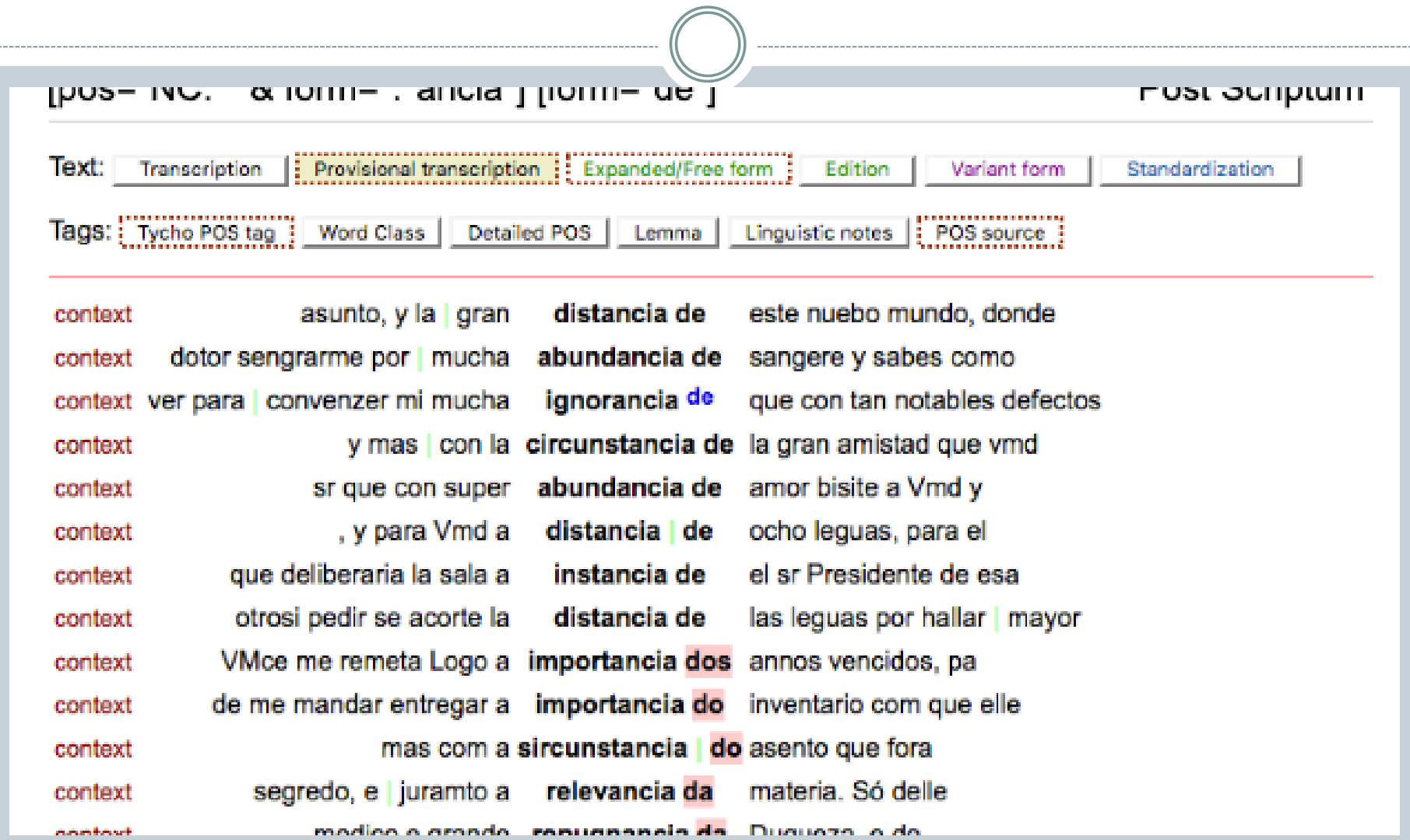
Frequency by: [select]

Source Verification


prudente y sabio; Cuyo dictamen firmado quisiera ver para
convenzer mi mucha ignorancia de que con tan notables defectos
no puede persuadirse aque la Ley permita empecer y abochor-
nar a una familia tan ilustre con semejante mezcla, pues
Dha nobleza la considero restricta a otros casos personales.

prudente y sabio; cuyo dictamen firmado quisiera ver para convencer mi mucha ignorancia [de] que con tan notables defectos no puede persuadirse a que la Ley permita empecer y abochornar a una familia tan ilustre con semejante mezcla, pues dha nobleza la considero [restricta] a otros casos personales.

Visible in Result



Clitics and Contractions

- Grammatically contracted words
  - Coq au vin => coq a le vin
  - Comedtelo => Comed te lo
  - Naňs => Na něj jsi
- Needed for grammatical searches and lemmatization
  - What is the lemma of au?
- TEITOK uses an XML trick
  - Grammatical words inside orthographic words
  - Gram words indexed and searched
  - Ort words displayed

DTOK – Grammatical Words

- Nested tokenization

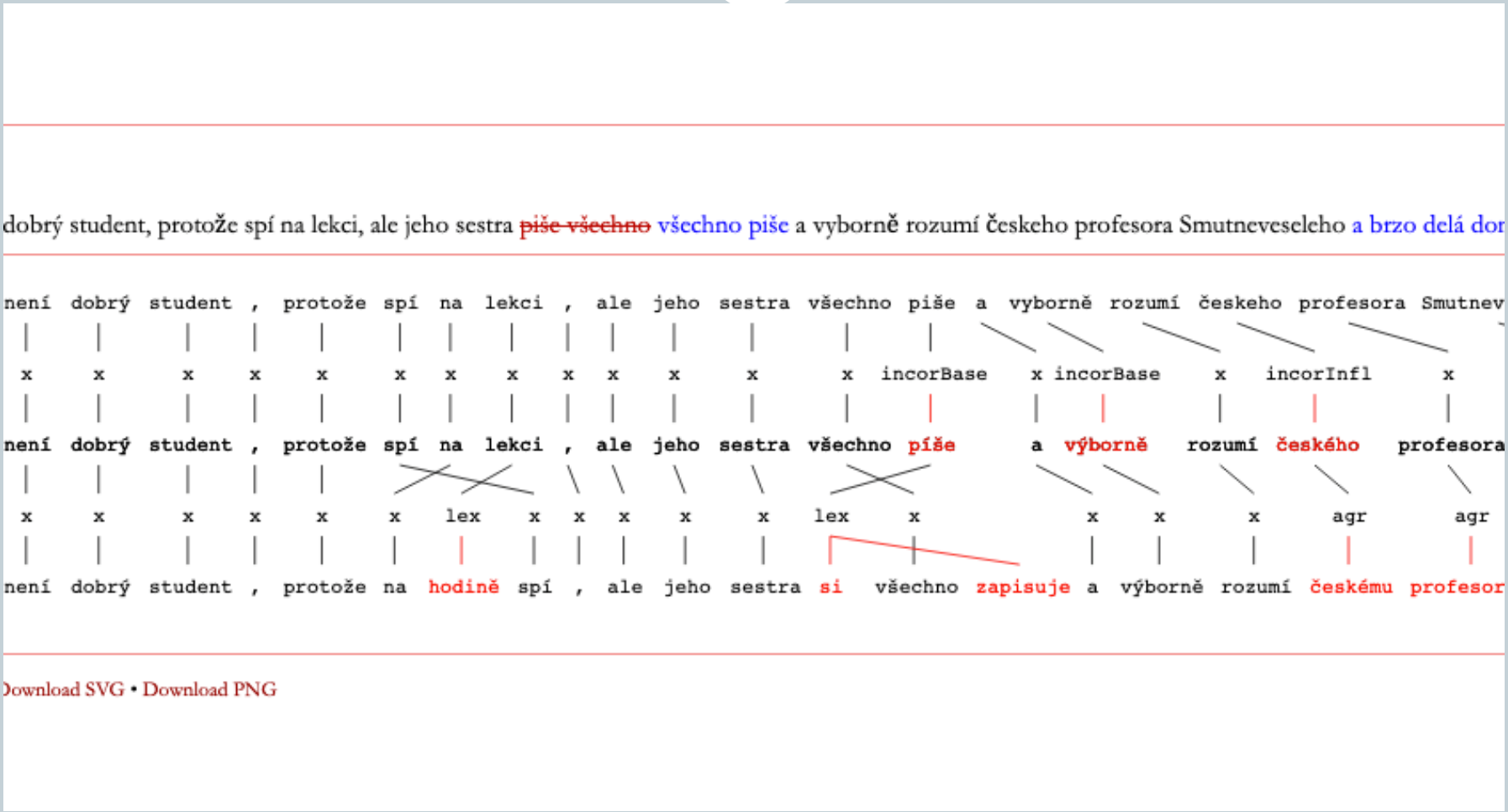
```
<tok norm="na nějs">naňs
    <dtok form="na" pos="PREP" lemma="na"/>
    <dtok form="něj" pos="PRON" lemma="on"/>
    <dtok form="jsi" pos="AUX" lemma="být"/>
</tok>
```

Modular Set-up

- TEITOK consists of many small PHP scripts
  - "All" working on the same XML file or CQP corpus
- Easy to develop new modules
  - New modules constantly being added
  - "Hello world" instructions to develop your own

```
<?php
	$raw = shell_exec("/usr/local/bin/tei2txt xmlfiles/{$_GET['cid']}");
	$maintext = "<h1>Hello World!</h1> <pre>$raw</pre>";
?>
```

CzeSL specific: feat



Search Result Visualization

Corpus Distribution

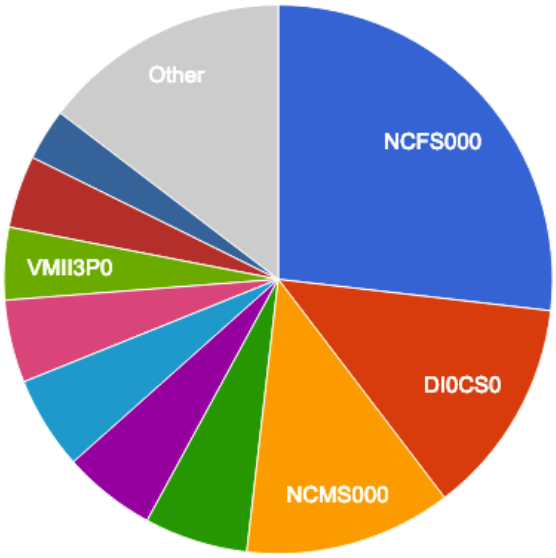
Search query: [word="ca.*"]
Group query: Group by: **POS Tag**

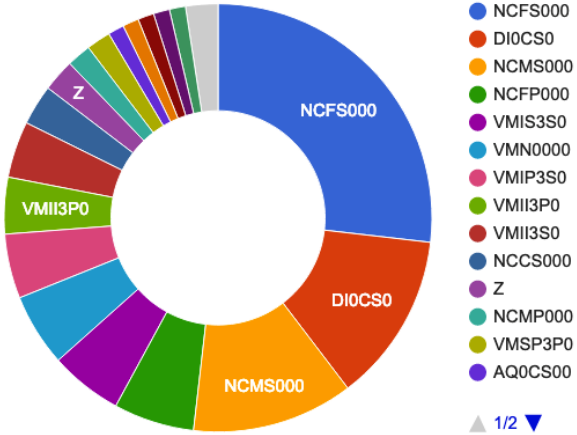
Graph: Table | Count: Count | Download: [seleccionar]

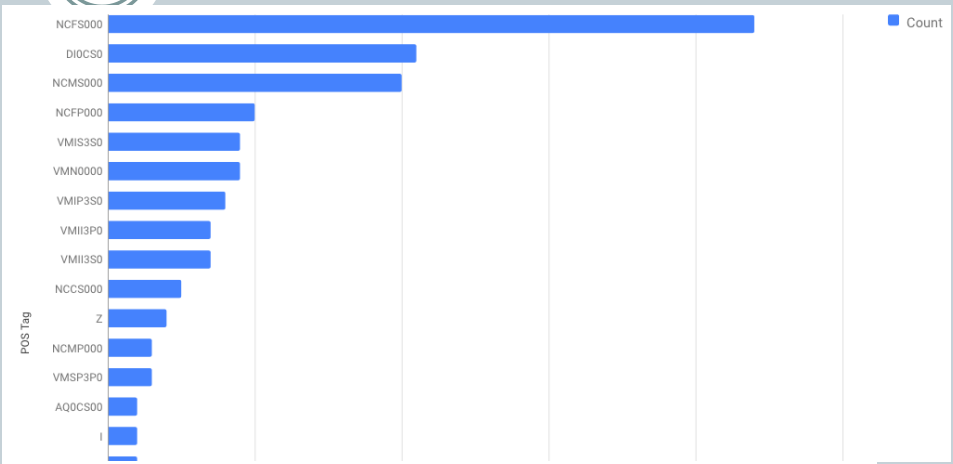
POS Tag	Count	WPM
NCFS000	44	1,314.49
DI0CS0	21	627.371
NCMS000	20	597.497
NCFP000	10	298.748
VMIS3S0	9	268.873
VMN0000	9	268.873
VMIP3S0	8	238.999
VMII3P0	7	209.124
VMII3S0	7	209.124
NCCS000	5	149.374
Z	4	119.499
NCMP000	3	89.625
VMSP3P0	3	89.625
AQ0CS00	2	59.75
I	2	59.75
VMIP3P0	2	59.75
VMIS1S0	2	59.75
VMP00SM	2	59.75
NCCN000	1	29.875
RG	1	29.875
VMII2S0	1	29.875
VMIS3P0	1	29.875

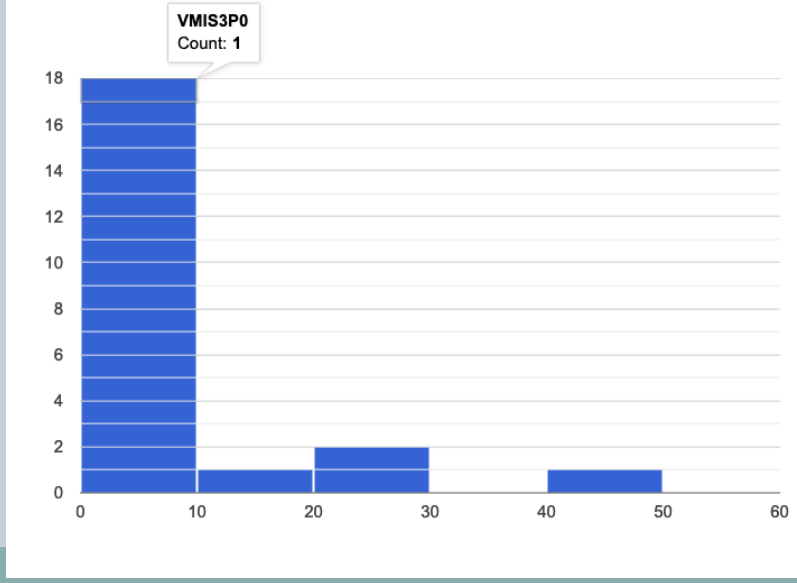
Measure	Value
Rows	22.00
Total	164.00
Minimum	1.00
Maximum	44.00
Mean	7.45
Median	3.50
Mode	2.00
Range	43.00
Standard deviation	9.68
Mean deviation	6.26
Median deviation	2.50
Mean squared error	93.70

Graphs


Other
NCFS000
VMII3P0
DI0CS0
NCMS000


NCFS000
Z
VMII3P0
DI0CS0
NCMS000
NCFS000
DI0CS0
NCMS000
NCFP000
VMIS3S0
VMN0000
VMIP3S0
VMII3P0
VMII3S0
NCCS000
Z
NCMP000
VMSP3P0
AQ0CS00
1/2


Count
POS Tag
NCFS000
DI0CS0
NCMS000
NCFP000
VMIS3S0
VMN0000
VMIP3S0
VMII3P0
VMII3S0
NCCS000
Z
NCMP000
VMSP3P0
AQ0CS00
I


VMIS3P0
Count: 1
18
16
14
12
10
8
6
4
2
0
0
10
20
30
40
50
60

POS Tag Explanation

Tagset

The PostScriptum corpus is tagged with a slightly modified version of the EAGLES tagset for Spanish, modified in such a way that it allows a conversion to the CLAWS tagset used in the Tycho Brahe project. The tagset is position-based, where the tag consists of a sequence of letters and number, where each letter or number represents a morphosyntactic feature, dependent on its position in the sequence. The positions of the sequence have a different meaning for each main POS, which is represented by the first letter. For instance, the word *bonita* has the tag AQ0FS0, where the first A indicates it is an adjective, and the S in the fifth position indicates the number, in this case singular. The full description of the main tags, with their positions, their meaning and their possible values are given below.

A	Adjective (Adj)		
1	type	Q	Qualitative
		O	Ordinal
2	degree	0	*does not apply*
		A	augmentative
		D	diminutive
		C	comparative
		S	superlative
3	gender	M	masculine
		F	feminine
		C	common
4	number	S	singular
		P	plural
		N	invariable
5	function	0	*does not apply*
		P	participle
R	Adverb (Adv)		

Inline Explanations

Meu adorado bem

Ca recebi a cua vejo o quanto me dis ademirame Vmce diser q eu onte estava alegre pareceme q alegria pa mim ja ce acabou o q lhe poco diser q nunca em minha vida não vi nem ovi o q aqui te visto e ovisto onte a noite veio pa ca pa baixo 16 molheres tam **perdidas** q era huma conciencia não ce pode durmir nada eu estive toda a noite como agora do q Vmce me dis d eu pedir ele he hum home falto de relegião como pode ele ter do de nos ele fes isto pa ce ving mce lhe dice na loje de mim porq vevia na disconfianca q eu falava aos meus parentes do carmo onte e me dice q ele hera hum home muito máo q ningem ce capacita de tal as q ele esta tão teimoso q não quer cenão q eu va pa convento e dis q la me da tudo q não quero hir pa convento esteve agora ca o irmão dele diceme q ele pos la a filha ja pa o convento q me hera milhor porq me ariscava a cahir me huma centenca eu lhe dice q agora tinha vindo abrir os olhos ja não tinha medo porq estava aqui huma q não estava com o marido pois lhe tinha fogido havia 9 meses e q veio presa e q lhe cahio hir por 5 annos pa castro marinho q eu fasia de conta o mesmo e q era o pago q ele me dava acim eu não lhe quero pedir porq não quero q diga q eu estou compeles eu ainda q estou mto ismorecida não o mostro pa as pecoas conhecidas dele porq cempre me queixo mto dele ele dise o irmão q eu o disconpunha em cartas q lhe iscrevia q em ves d eu pedir merzircordia q ele antão me pordoava q ainda o iritava eu estou a conta de Ds e de Vmce o q lhe peco he q me não iscandelise pois cabe como eu estou q istou pior do q Vmce porq Vmce he home eu sou molher porq Vmce dis q esta perdido pa amor de mim iso não he asim porq ce Vmce esta pa amor de mim eu

perdidas	
Word Class	V
Detailed POS	Verb (VMP00PF) Main; participle; plural; feminine
Lemma	perder

Manuscript Lines

Don Afonſſo de Caſtela

de Toledo de Leon

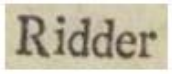
Rey e ben deſ Copostela

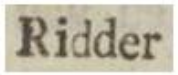
ta o Reyno daragõ

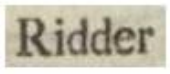
OCR -> Image search

CQP Query: [form="Ridder"] Zoeken query builder | tonen

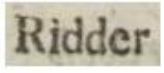
58 resultaten - getoond 1 - 50

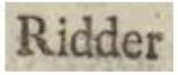
dpo3713 Ridder 

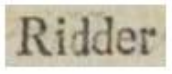
dpo3713 Ridder 

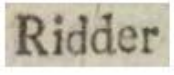
dpo3713 Ridder 

dpo3713 Ridder 

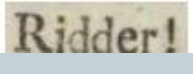
dpo3713 Ridder 

dpo3713 Ridder 

dpo3713 Ridder 

dpo3713 Ridder 

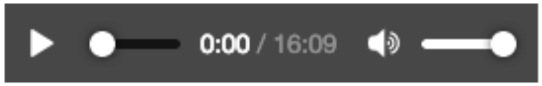
dpo3713 Ridder 

dpo3713 Ridder 

Critical Apparatus

Ds1	So befill it that seson on a day
Ht	It befell þat in þat cesoıīun on a day
Lc	Byfille þat in that seson on a day
Ha2	Byfell that . in that sesoūn on a daye .
Ha3	By felle hit in þtat saysoun · on a day
Pn	byfel in that season on a day
Base	Bifel that in that sesoun on a day
Ld2	Byfell in that sesoun vp on a day
Ad1	Be fell in that sesoñn/ on a day
Cx1	ANd fil in that seson on a day
En1	So byfyll yt þat seson on a day
Cx2	BYfyl in that seson on a day
Ma	¶ So it befell þat seson on a day
Pw	Byfille þat in that seson on aday

Oral Transcriptions



el005 - ah / o meu nome é FF //

el005 - ah / sou grega //

el005 - estou aqui há / um ano e / quatro cinco meses //

el005 - estou a trabalhar / cheguei cá como estagiária / e / acabei por ficar //

el005 - porque / fui contratada pela empresa que eu / comecei a trabalhar //

el005 - ah / era obrigatório aprender português porque o meu posto está mesmo / uma consultoria portuguesa //

el005 - e / tenho estado com pessoas / no meu escritório ah / estou a / receber muito apoio / em aprender a língua e tem sido mais fácil / porque / estou com pessoas mesmo nativas / a passar o meu dia-a-dia //

el005 - isso facilita / mais a fluência //

el005 - espero eu / hhh //

el005 - da língua //

Searching with sound

[word="casa"]

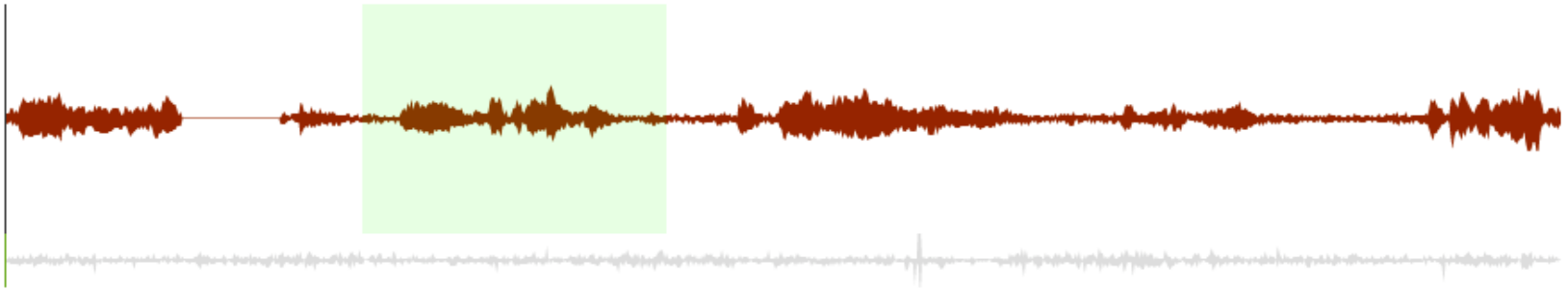
Pesquisar query builder | visualize | options

1865 resultados • Mostra 0 - 100 (seguintes)

Etiquetas: **Etiqueta POS** **Lema** **Etiqueta Sintática** **Etiqueta Morfológica** **Etiquetas Secundárias**

context	▶	no caso só o sexo | agora existe caractrísticas diferentes né |
context	▶	sabe é da pior qualidade | ninguém sabe a pessoa pra
context	▶	ruções pra gente de alta
context	▶	garantia | fica muito aquém
context	▶	necessidade começa ali | adquiria ah comprar
context	▶	com o cheque sem fundo | e fica ele apenas
context	▶	oferecer determinadas utilidades e deduzir | do salário |

Waveform View – Time-Aligned



el005 - ah / o meu nome é FF //

el005 - ah / sou grega //

el005 - estou aqui há / um ano e / quatro cinco meses //

el005 - estou a trabalhar / cheguei cá como estagiária / e / acabei por ficar //

el005 - porque / fui contratada pela empresa que eu / comecei a trabalhar //

el005 - ah / era obrigatório aprender português porque o meu posto está mesmo / uma consultoria portuguesa //

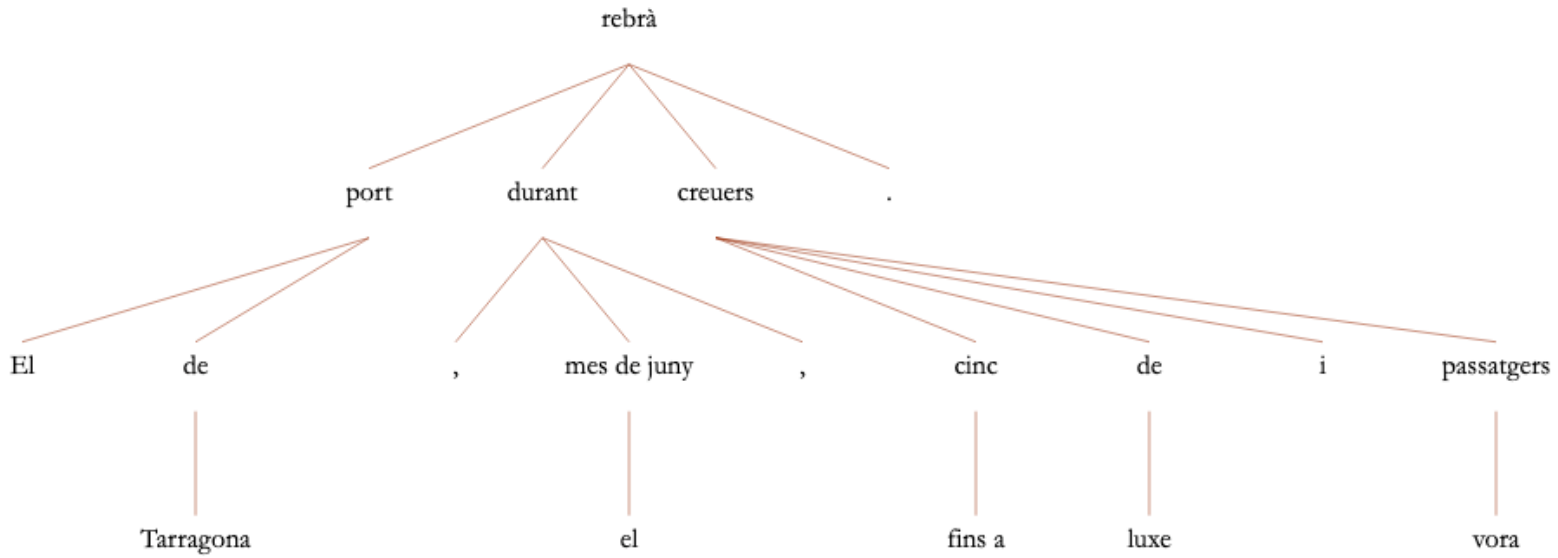
el005 - e / tenho estado com pessoas / no meu escritório a ah / estou estou a / receber muito apoio / em aprender a língua e tem sido mais fácil / porque estou a es / estou com pessoas mesmo nativas / a passar o meu dia-a-dia //

Mapping Data

Search on the Map

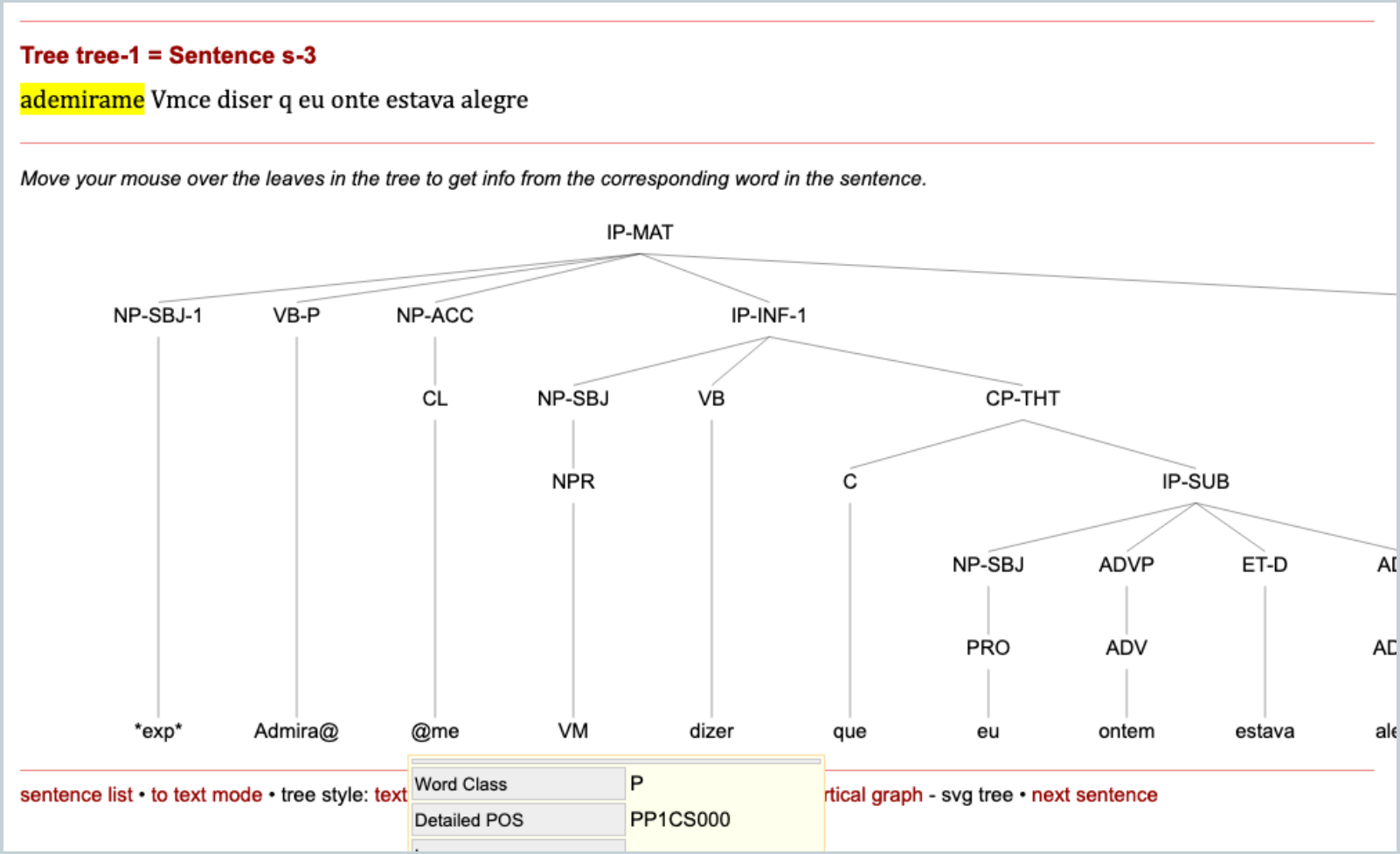
Dependency Relations

El port de Tarragona rebrà, durant tot el mes de juny, fins a cinc creuers de luxe i vora mil passatgers.



Tarragona	
POS tag	n
Morphosyntactic features	postype=proper|gen=c|num=c
Lemma	Tarragona

Constituency Trees (stand-off)



Interlinear Glossed Text (LRL)

Word	Muj	pes	leze	l
Morpheme	muj	pes	leze	l
Baseform	muj	pes	leze	l
Gloss	3SG.MASC.NOM	3SG.MASC.NOM		PAST.MASC.3SG
Meaning	my	dog	lie	
Original	Muj pes lezel			
Translation	My dog lied			

Word	Tatinek	koupi	l	kock	u
Morpheme	tatinek	koupi	l	kock	u
Baseform	tatinek	koupi	l	kock	u
Gloss	3SG.MASC.NOM		PAST.MASC.3SG		3SG.FEM.ACC
Meaning	father	buy		cat	
Original	Tatinek koupil kocku				
Translation	My father bought a cat				

Word	Krasn	a	divka	vari	la	obed
Morpheme	krasn	a	divka	vari	la	obed
Baseform	krasn		divka	vari	la	obed
Gloss		3SG.FEM.NOM	3SG.FEM.NOM		3SG.FEM.PAST	3SG.MASC.ACC
Meaning	beautiful		girl	make		lunch
Original	Krasna divka varila obed					
Translation	Beautiful girl made lunch					

More Modules

- Stand-off annotation
  - For crossing annotations (error annotation)
- HTML Editor
- Upload file
- XML Viewer/Editor
- Token editing
  - Split, merge, MWE, contractions, etc.
- User-created annotations
- Page-by-page transcription
- Collations
- Word sketches
- XDXF dictionaries

Informed NLP

- Tokens contain a lot of information

```
<tok id="w-123" form="dou" nform="dough"
     POS="NCS" lemma="dough" deprel="obj" head="w-121">
     do<unclear>u</unclear><del>w</del>
</tok>
```

- Can be used in NLP
  - Neotag POS tagger works directly on XML
  - Trained on XML, using XML parameter files
  - Can do in-context normalization
  - "He put the dou in the oven"
  - "I'd like you to dou it"

Efficient Editing

- Click-and-edit very efficient
  - For normalization – you can read the text
- Not so good for POS corrections
  - Need for much more stuctural editing
- Verticalized mode
  - View and edit in “traditional” vertical way
- Edit by Search
  - Use CQP to find and correct errors throughout the corpus

Verticalized editing

Verticalized Corpus View

XML File: TESTS/NEM_GD_008.xml

	Transcription	Normalized form	POS tag	UD POS tag	Lemma
w-1	Bratr		NNMS1-----A----	NOUN	bratr
w-2	a		J^----------	CCONJ	a
w-3	Sestra		NNFS1-----A----	NOUN	sestra
w-4	.		Z:----------	PUNCT	.
w-5	Viktor		NNMS1-----A----	PROPN	Viktor
w-6	je		VB-S---3P-AA---	AUX	být
w-7	mladý		AAMS1----1A----	ADJ	mladý
w-8	pan	pán	NNMS1-----A----	NOUN	pan
w-9	z		RR--2----------	ADP	z
w-10	~~Polska~~Ruska		NNNS2-----A----	PROPN	Rusko
w-11	.		Z:----------	PUNCT	.
w-12	Studuje		VB-S---3P-AA---	VERB	studovat
w-13	češtinu		NNFS4-----A----	NOUN	čeština
w-14	ve		RV--6----------	ADP	v
w-15	škole		NNFS6-----A----	NOUN	škola
w-16	,		Z:----------	PUNCT	,
w-17	protože		J,----------	SCONJ	protože
w-18	ne umí	neumí	TT----------+VB-S---3F	PART+VERB	neumět
w-19	psat	psát	Vf--------A----	VERB	psat
w-20	a		J^----------	CCONJ	a

CQL-driven edit

Multiple token edit via CQP Search

Define below which features you want to change in this search, and select all the tokens for which you want that change to be made. Leaving a feature empty will not eliminate it's value, but just ignore that feature in the edit.

The CQP corpus can become disaligned wrt the XML files after editing tokens.
Therefore, always regenerate the CQP corpus before using this function!

form	Written form	
nform	Normalized form	
pos	POS tag	
lemma	Lemma	

Click here to enter individual values for each result

35 resultados for [word="casa"]

File ID	Sel.	Left context	Match	Right context
xmlfiles/COSER-2910-01.xml	☐	, del castillo , esa	**casa**	grandota , la casa grande
xmlfiles/COSER-2910-01.xml	☐	esa casa grandota , la	**casa**	grande . Ahí había tres
xmlfiles/COSER-0614-01.xml	☐	pues , echarlas pienso en	**casa**	. ¿ O sea ,
xmlfiles/COSER-0614-01.xml	☐	cuando te apetece limpiar una	**casa**	. Cuando tienes ganas de
xmlfiles/COSER-0614-01.xml	☐	Pero todo era aquí en	**casa**	. Ahora ya no ,
xmlfiles/COSER-0614-01.xml	☐	. Pero ¿ si la	**casa**	estaba bien ? Entonces era
xmlfiles/COSER-0614-01.xml	☐	ropa , claro . La	**casa**	también te la daba el
xmlfiles/COSER-0614-01.xml	☐	que tenías que calentaba la	**casa**	, más o menos .
xmlfiles/COSER-0614-01.xml	☐	? Bueno , aquí en	**casa**	lo mismo , lo que
xmlfiles/COSER-0614-01.xml	☐	. Lo comenté aquí en	**casa**	, lo que este señor
xmlfiles/COSER-4102-01.xml	☐	Sí ? Sí , se	**casa**	uno . Sí , tienen

CQL-driven individual edit

Multiple token edit via CQP Search

The CQP corpus can become disaligned wrt the XML files after editing tokens.
Therefore, always regenerate the CQP corpus before using this function!

35 resultados for [word="casa"]

File ID	Left context	Match	Right context	POS tag	Lemma
context	, del castillo , esa	**casa**	grandota , la casa grande		
context	esa casa grandota , la	**casa**	grande . Ahí había tres		
context	pues , echarlas pienso en	**casa**	. ¿ O sea ,		
context	cuando te apetece limpiar una	**casa**	. Cuando tienes ganas de		
context	Pero todo era aquí en	**casa**	. Ahora ya no ,		
context	. Pero ¿ si la	**casa**	estaba bien ? Entonces era		
context	ropa , claro . La	**casa**	también te la daba el		
context	que tenías que calentaba la	**casa**	, más o menos .		
context	? Bueno , aquí en	**casa**	lo mismo , lo que		
context	. Lo comenté aquí en	**casa**	, lo que este señor		
context	Sí ? Sí , se	**casa**	uno . Sí , tienen	VMIP3S0	casar
context	y las mujeres trabajando en	**casa**	¿ no ? ¿ Y	NCFS000	casa
context	pequeño pues la cantina ,	**casa**	Zutano , casa Zutano y	NCFS000	casa
context	cantina , casa Zutano ,	**casa**	Zutano y casa Zutano ,	NCFS000	casa
context	Zutano , casa Zutano y	**casa**	Zutano , ¡ y ya	NCFS000	casa
context	... ? Los de mi	**casa**	nos fuimos a una masía	NCFS000	casa
context	, y nos volvimos a	**casa**	. Y ya no se	NCFS000	casa
context	cura ? , ¿ se	**casa**	con el cura ? ¡	NCFS000	casa
context	y el domingo ¡ a	**casa**	! asociación cultural . ¿	NCFS000	casa