TEITOK

COMBINING LANGUAGE AND LINGUISTIC INFORMATION WITHOUT COMPROMISE

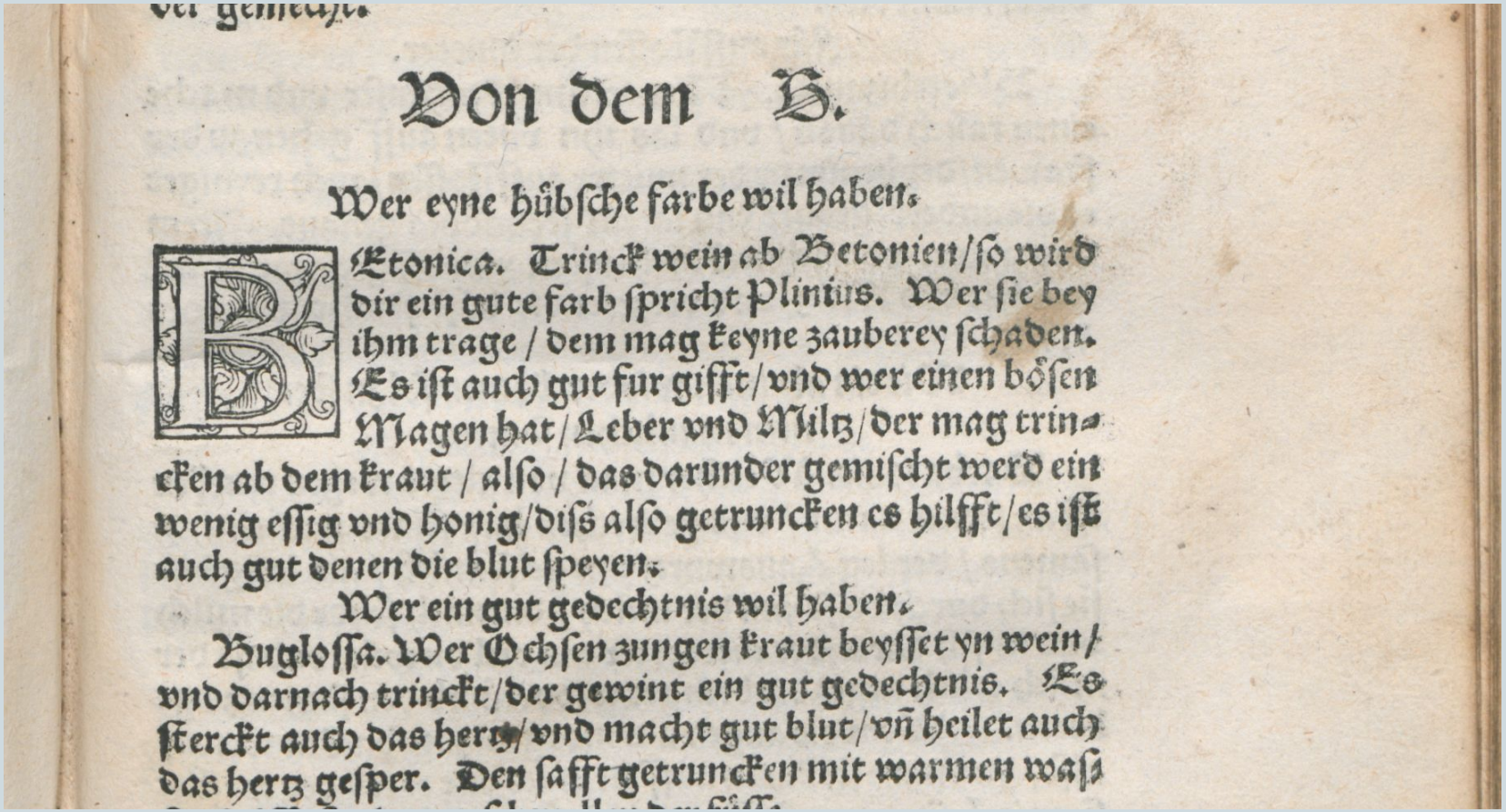
Language vs. Linguistics

- Linguistic information (ANNIS)
  - Tokenization, POS, lemma, dependencies, syntax, semantics
  - Normalizations, sentences
- Language information (TEI)
  - Lines, pages, paragraphs
  - Bold, italics, colour, typesettings
  - Hands, additions, deletions
- Document information (Dublin Core)
  - Facsimile images
  - Year, library, author, etc.

Traditional Method

- Annotated Corpora are text based
- Step 1 is to clean the text
  - Throw away any non-linguistic information
- Step 2 is to tokenize
  - Verticalizing text one word per line
  - Removes information about inter-word spacing
  - Often also splitting contractions

Facsimile Image



Von dem B.

Wer eyne hübsche farbe wil haben.

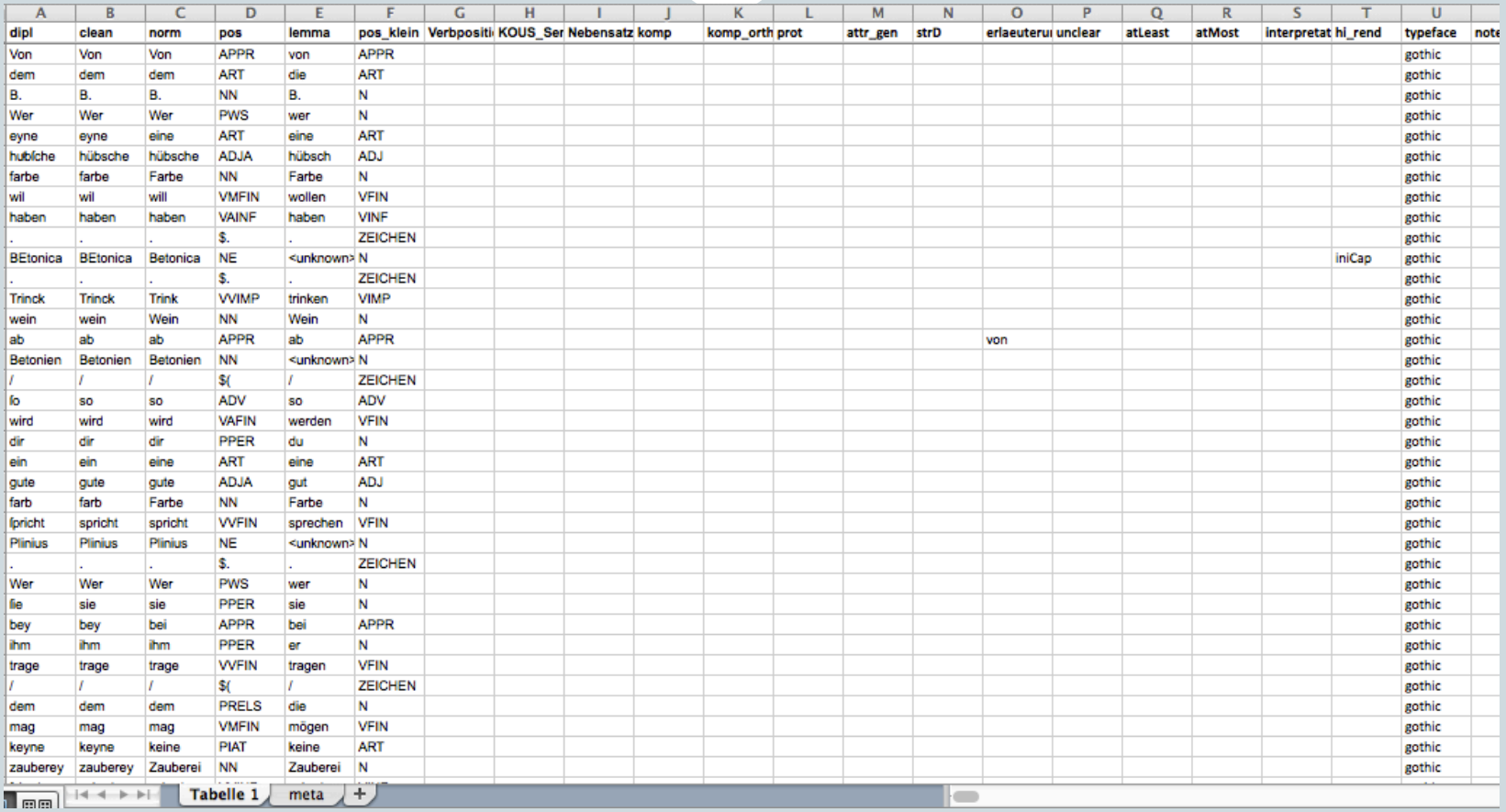
BEtonica. Trinck wein ab Betonien/so wird dir ein gute farb spricht Plinius. Wer sie bey ihm trage / dem mag keyne zauberey schaden. Es ist auch gut fur gifft/vnd wer einen bösen Magen hat/Leber vnd Miltz/der mag trincken ab dem kraut / also / das darunder gemischt werd ein wenig essig vnd honig/diss also getruncken es hilfft/es ist auch gut denen die blut speyen.

Wer ein gut gedechtnis wil haben.

Buglossa. Wer Ochsen zungen kraut beysset yn wein/ vnd darnach trinckt/der gewint ein gut gedechtnis. Es sterckt auch das hertz/vnd macht gut blut/vñ heilet auch das hertz gesper. Den safft getruncken mit warmen was

Example from the RIDGES corpus

Verticalized text

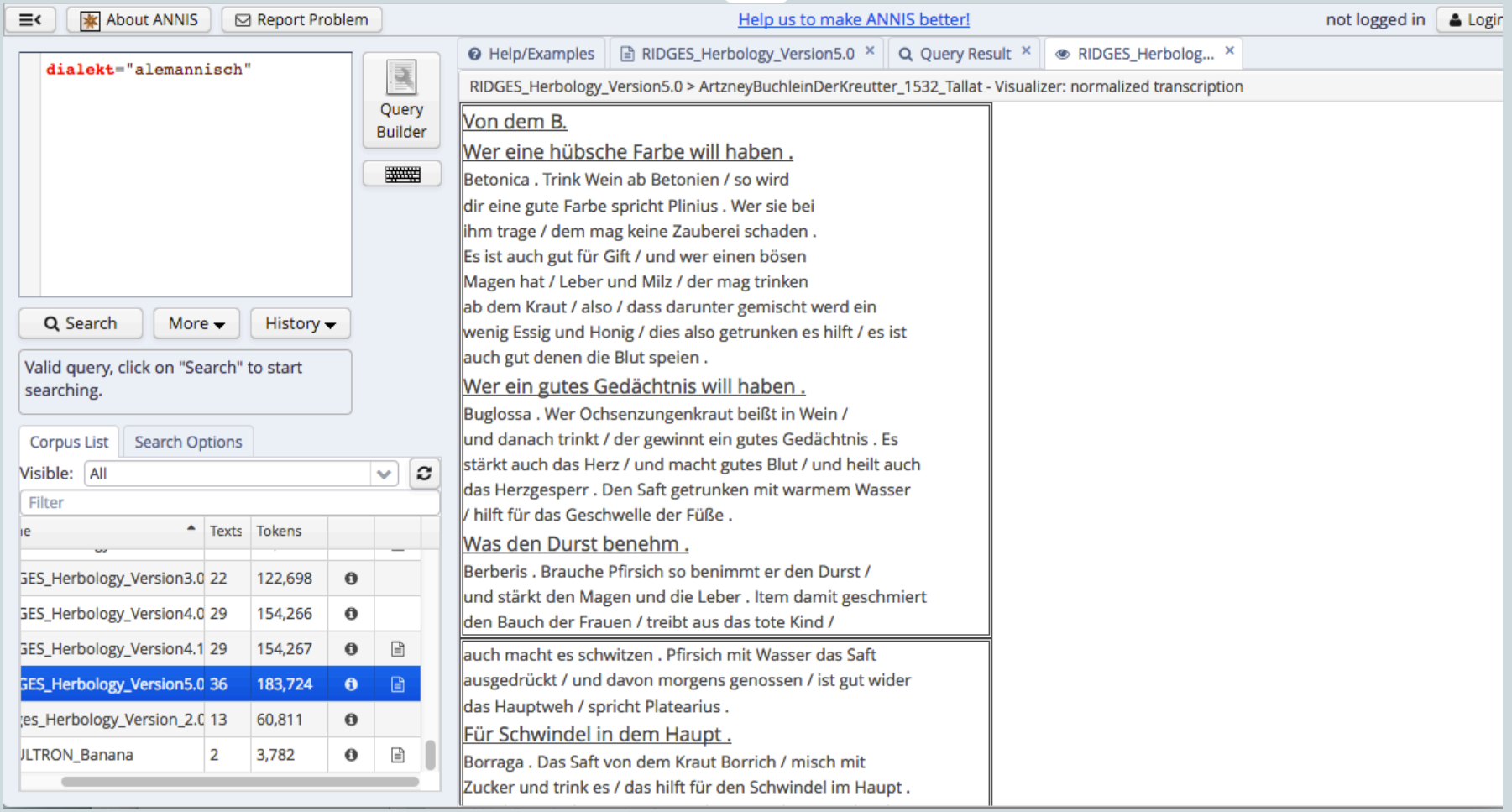


A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	
dipl	clean	norm	pos	lemma	pos_klein	Verbpositi	KOUS_Ser	Nebensatz	komp	komp_orth	prot	attr_gen	strD	erlaeterui	unclear	atLeast	atMost	interpretat	hi_rend	typeface	note
Von	Von	Von	APPR	von	APPR															gothic	
dem	dem	dem	ART	die	ART															gothic	
B.	B.	B.	NN	B.	N															gothic	
Wer	Wer	Wer	PWS	wer	N															gothic	
eyne	eyne	eine	ART	eine	ART															gothic	
hubſche	hübsche	hübsche	ADJA	hübsch	ADJ															gothic	
farbe	farbe	Farbe	NN	Farbe	N															gothic	
wil	wil	will	VMFIN	wollen	VFIN															gothic	
haben	haben	haben	VAINF	haben	VINF															gothic	
.	.	.	$.	.	ZEICHEN															gothic	
BEtonica	BEtonica	Betonica	NE	<unknown>	N														iniCap	gothic	
.	.	.	$.	.	ZEICHEN															gothic	
Trinck	Trinck	Trink	VVIMP	trinken	VIMP															gothic	
wein	wein	Wein	NN	Wein	N															gothic	
ab	ab	ab	APPR	ab	APPR									von						gothic	
Betonien	Betonien	Betonien	NN	<unknown>	N															gothic	
/	/	/	$(	/	ZEICHEN															gothic	
ſo	so	so	ADV	so	ADV															gothic	
wird	wird	wird	VAFIN	werden	VFIN															gothic	
dir	dir	dir	PPER	du	N															gothic	
ein	ein	eine	ART	eine	ART															gothic	
gute	gute	gute	ADJA	gut	ADJ															gothic	
farb	farb	Farbe	NN	Farbe	N															gothic	
ſpricht	spricht	spricht	VVFIN	sprechen	VFIN															gothic	
Plinius	Plinius	Plinius	NE	<unknown>	N															gothic	
.	.	.	$.	.	ZEICHEN															gothic	
Wer	Wer	Wer	PWS	wer	N															gothic	
ſie	sie	sie	PPER	sie	N															gothic	
bey	bey	bei	APPR	bei	APPR															gothic	
ihm	ihm	ihm	PPER	er	N															gothic	
trage	trage	trage	VVFIN	tragen	VFIN															gothic	
/	/	/	$(	/	ZEICHEN															gothic	
dem	dem	dem	PRELS	die	N															gothic	
mag	mag	mag	VMFIN	mögen	VFIN															gothic	
keyne	keyne	keine	PIAT	keine	ART															gothic	
zauberey	zauberey	Zauberei	NN	Zauberei	N															gothic	

Tabelle 1 | meta | +

Final version (Excel) of the verticalized text with columns

ANNIS - Text


About ANNIS
Report Problem
Help us to make ANNIS better!
not logged in
Logir
dialekt="alemannisch"
Query Builder
Search
More
History
Valid query, click on "Search" to start searching.
Corpus List
Search Options
Visible: All
Filter
Texts
Tokens
GES_Herbology_Version3.0 22 122,698
GES_Herbology_Version4.0 29 154,266
GES_Herbology_Version4.1 29 154,267
GES_Herbology_Version5.0 36 183,724
es_Herbology_Version_2.0 13 60,811
ULTRON_Banana 2 3,782
Help/Examples
RIDGES_Herbology_Version5.0
Query Result
RIDGES_Herbolog...
RIDGES_Herbology_Version5.0 > ArtzneyBuchleinDerKreutter_1532_Tallat - Visualizer: normalized transcription
Von dem B.
Wer eine hübsche Farbe will haben .
Betonica . Trink Wein ab Betonien / so wird
dir eine gute Farbe spricht Plinius . Wer sie bei
ihm trage / dem mag keine Zauberei schaden .
Es ist auch gut für Gift / und wer einen bösen
Magen hat / Leber und Milz / der mag trinken
ab dem Kraut / also / dass darunter gemischt werd ein
wenig Essig und Honig / dies also getrunken es hilft / es ist
auch gut denen die Blut speien .
Wer ein gutes Gedächtnis will haben .
Buglossa . Wer Ochsenzungenkraut beißt in Wein /
und danach trinkt / der gewinnt ein gutes Gedächtnis . Es
stärkt auch das Herz / und macht gutes Blut / und heilt auch
das Herzgesperr . Den Saft getrunken mit warmem Wasser
/ hilft für das Geschwelle der Füße .
Was den Durst benehm .
Berberis . Brauche Pfirsich so benimmt er den Durst /
und stärkt den Magen und die Leber . Item damit geschmiert
den Bauch der Frauen / treibt aus das tote Kind /
auch macht es schwitzen . Pfirsich mit Wasser das Saft
ausgedrückt / und davon morgens genossen / ist gut wider
das Hauptweh / spricht Platearius .
Für Schwindel in dem Haupt .
Borraga . Das Saft von dem Kraut Borrich / misch mit
Zucker und trink es / das hilft für den Schwindel im Haupt .

Linguistics Only

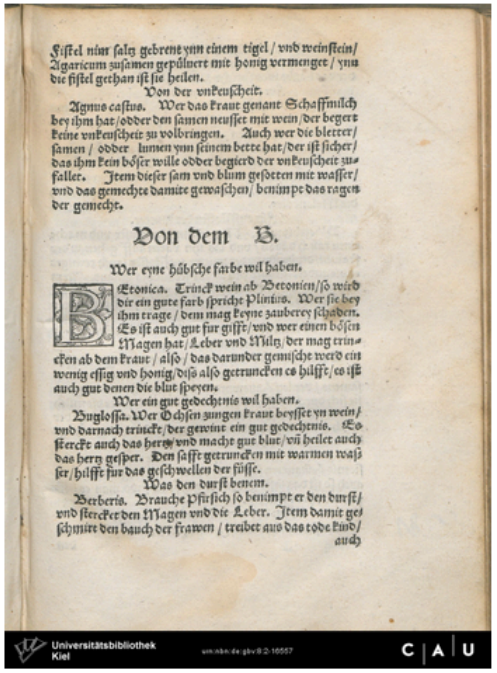
- Corpora mostly cannot be rendered as “text”
  - Example before still has linebreaks, making it better to read
  - Full-text representation often not even offered
- Good for linguistics
  - Statistical analysis
  - Complex queries
  - Grammatical analysis
- Bad for most other things
  - Not friendly (or even possible) to read
  - No link to facsimile images

TEITOK

- TEITOK starts from TEI/XML
  - Each file is in full TEI/XML format
- Annotation is added inline
  - Tokens are added non-destructively to the XML
  - Annotations are attributes over nodes
- Extralinguistic rendering
  - TEI encodings (such as dropcaps) rendered as CSS
  - Images rendered next to document (from <pb/>)
- Graphical display in browser
  - Examples to follow

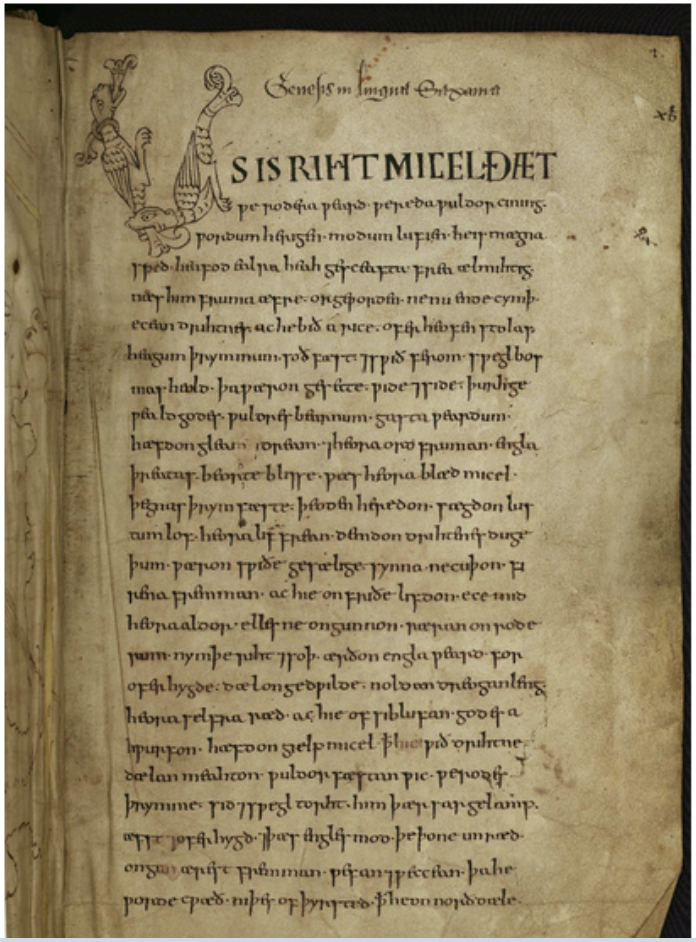
RIDGES

Von dem B.
Wer eyne hübſche farbe wil haben.
BEtonica. Trinck wein ab Betonien / ſo wird
dir ein gute farb ſpricht Plinius. Wer ſie bey
ihm trage / dem mag keyne zauberey ſchaden.
Es iſt auch gut fur gifft / vnd wer einen boſſen
Magen hat / Leber vnd Miltz / der mag trin
cken ab dem kraut / alſo / das darunder gemiſcht werd ein
wenig eſſig vnd honig / diſs alſo getruncken es hilfft / es iſt
auch gut denen die blut ſpeyen.
Wer ein gut gedechtnis wil haben.
Bugloſſa. Wer Ochſen zungen kraut beyſſet yn wein /
vnd darnach trinckt / der gewint ein gut gedechtnis. Es
ſterckt auch das hertz / vnd macht gut blut / v heilet auch
das hertz geſper. Den ſafft getruncken mit warmen wa
ſer / hilfft fur das geſchwellen der fuſſſe.
Was den durſt benem.
Berberis. Brauche Pfirſich ſo benimpt er den durſt /
vnd ſtercket den Magen vnd die Leber. Item damit ge
ſchmirt den bauch der frawen / treibet aus das tode kind /


Fiſtel nim ſalz gebrent ynn einem tigel / vnd weinſtein /
Agaricum zuſamen gepülwert mit honig vermenget / ynn
die fiſtel gethan iſt ſie heilen.
Von der vnkeuſcheit.
Agnus caſtus. Wer das kraut genant Schaffmilch
bey ihm hat / odder den ſamen neuſſet mit wein / der begert
keine vnkeuſcheit zu volbringen. Auch wer die bletter /
ſamen / odder lumen ynn ſeinem bette hat / der iſt ſicher /
das ihm kein böſer wille odder begierd der vnkeuſcheit zu
falle. Item dieſer ſam vnd blum geſotten mit waſſer /
vnd das gemechte damite gewaſchen / benimpt das ragen
der gemechte.
Von dem B.
Wer eyne hübſche farbe wil haben.
BEtonica. Trinck wein ab Betonien / ſo wird
dir ein gute farb ſpriche Plinius. Wer ſie bey
ihm trage / dem mag keyne zauberey ſchaden.
Es iſt auch gut fur gifft / vnd wer einen böſen
Magen hat / Leber vnd Miltz / der mag trin-
cken ab dem kraut / alſo / das darunder gemiſcht werd ein
wenig eſſig vnd honig / diſs alſo getruncken es hilfft / es iſt
auch gut denen die blut ſpeyen.
Wer ein gut gedechtnis wil haben.
Bugloſſa. Wer Ochſen zungen kraut beyſſet yn wein /
vnd darnach trincke / der gewint ein gut gedechtnis. Es
ſtercke auch das hertz vnd macht gut blut / vñ heilet auch
das hertz geſper. Den ſafft getruncken mit warmen waſ-
ſer / hilfft fur das geſchwellen der füſſe.
Was den durſt benem.
Berberis. Brauche Pfirſich ſo benimpt er den durſt /
vnd ſtercket den Magen vnd die Leber. Item damit ge
ſchmiret den bauch der frawen / treibet aus das tode kind /
auch

Junius 11

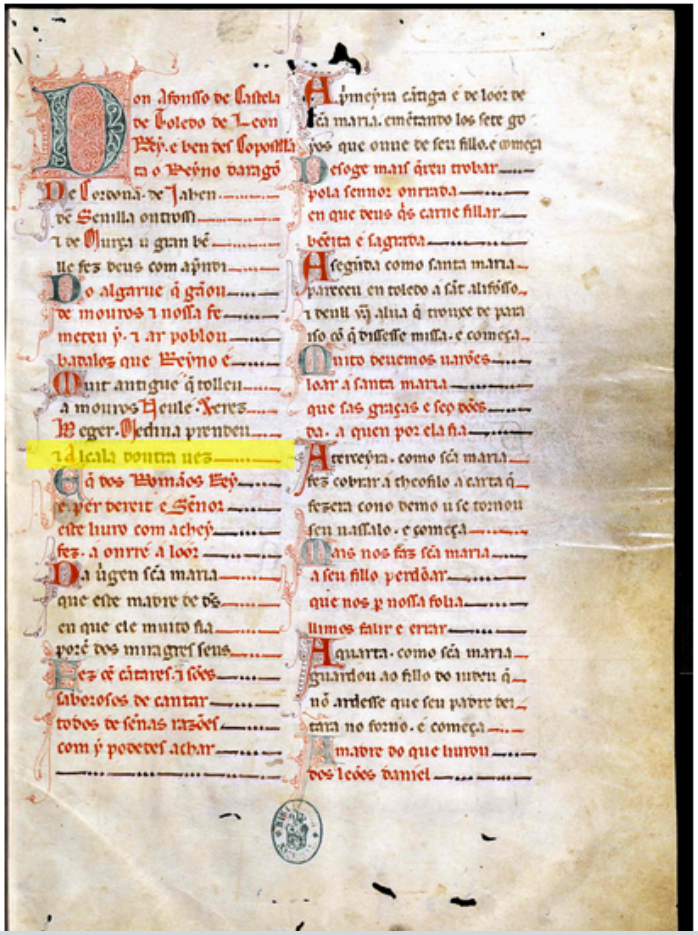
US IS RIHT MICEL ÐÆT
þe rodera þeard · þereda þuldorcining ·
þordum herigen · modum lufien · he iſ mægna
ſþed · heafod ealra heah geſceafta · frea ælmihtig ·
næſ him fruma æfre · or geporden · ne nu ende cymþ ·
ecean drihtneſ · ac he bið a rice · ofer heofen ſtolaſ ·
heagum þrymmum · ſoðfæſt ⁊ ſpið ferom ſpeglboſ
maſ heold · þa pæron geſette · þide ⁊ ſide · þurh ge
þeald godeſ · þuldreſ bearnum · gaſta þeardum ·
hæfdon gleam ⁊ dream · ⁊ heora ord fruman · engla
þreataſ · beorhte bliſſe · þæſ heora blæd micel ·
þegnaſ þrymfæſte · þeoden heredon · ſægdon luſ
tum lof · heora liffrean · demdon drihteneſ duge
þum · pæron ſpiðe geſælige · ſynna ne cuþon · fi
rena fremman · ac hie on friðe lifdon · ece mid
heora aldor · elleſ ne ongunnon · ræran on rode
rum · nymþe riht ⁊ ſoþ · ærðon engla þeard · for
oferhygde · dæl on gedpilde · noldan dreogan leng ·
heora ſelfra ræd · ac hie of ſiblufan · godeſ a
hpurfon · hæfdon gielp micel · þ hie pið drihtne ·



http://alfclul.clul.ul.pt/teitok/junius11/index.php?action=file&cid=pag1

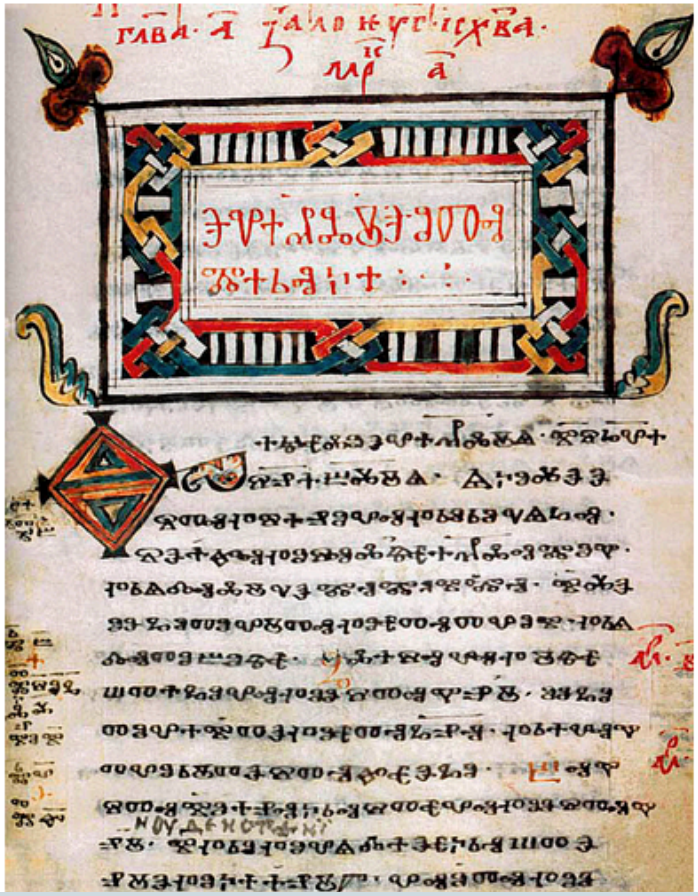
Cantigas de Santa Maria

Don Afonſſo de Caſtela
de Toledo de Leon
Rey e ben deſ Copostela
ta o Reyno daragõ
De Cordoua• de Iahen
de Seuilla outroſſi
ꝛ de Murça u gran bẽ
lle fez deus com aɲndi
Do algarue q̃ gãou
de mouros ꝛ noſſa fe
meteu y• ꝛ ar poblou
badaloz que Reyno e
Muit antigue q̃ tolleu
a mouros Neule • Xerez
Beger • Medina prendeu
e Alcala doutra uez
E q̃ dos Romãos Rey
e per dereit e Señor
este liuro com achey
fez a onrre a loor
Da uirgen ſſa maria



http://alfclul.clul.ul.pt/teitok/cantigas/index.php?action=file&cid=cantigas-00

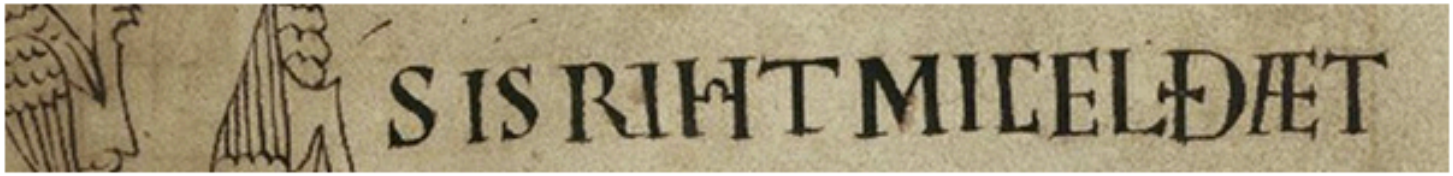
Zographensis



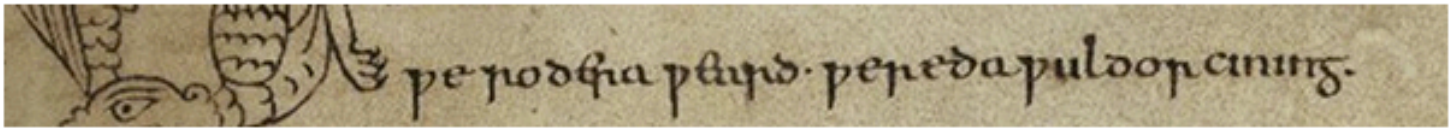
http://alfclul.clul.ul.pt/teitok/zographensis/index.php?action=file&cid=pag1

Junius 11 – Manuscript lines

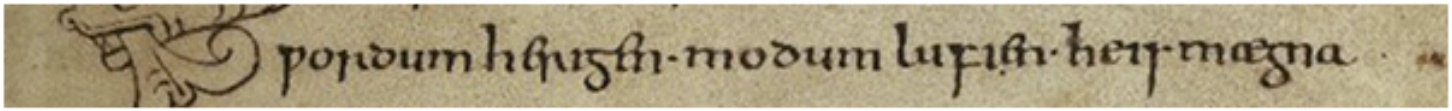
Folio 1

[1] 

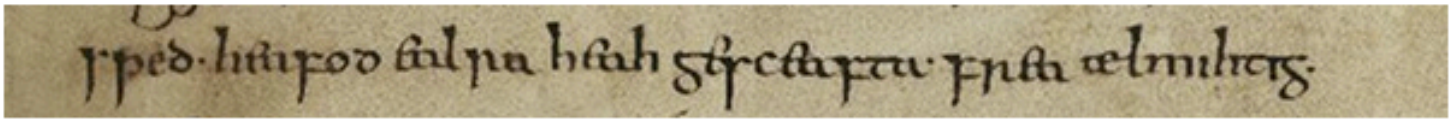
US IS RIHT MICEL ÐÆT

[2] 

pe rodera peard · pereda puldorcining ·

[3] 

þrym herigen · modum lufien · he ir mægna

[4] 

rped · heafod ealra heah gerceafta · frea ælmihtig ·

- Transcription can be aligned to manuscript line
- Good for verifying transcriptions against the original

Inline Tokenization

- TEI can be complex
  - 3 options to add information
- Export only words
  - Creates unrelated documents
- Stand-off annotation –
  - Typically position-based in the text: chars. 12-16 “Buch”
  - Difficult with XML, refutes robustness, and not editable
- Inline annotation
  - Create XML nodes for tokens
  - “Das Buch hier” => <token>Buch</token>
  - Unmodified TEI

XML Attributes

- Annotations marked over nodes
  - <tok pos="Noun">Buch</tok>
  - TEI: <w ana="N">Buch</w>
- White-space sensitive
  - Spaces in the text are meaningful
  - <c> </c>
- Inline verticalization
  - When looking only at <tok/> we get a list of tokens
  - Easy to convert to/from .vrt format

Multiple Orthographies

- Normalization typically required in historic corpus
  - Original text also needed
- Multiple orthographic layers
  - RIDGES has those as well
- No limit to the amount of forms
  - Original and normalized
  - Diplomatic, modernized, etc.
  - Romanized, transliterated
- Hierachical
  - Dependent layers, only filled when needed

Graphical User Interface

- XML of TEITOK not revolutionary
  - Very complex
  - <pb/> Von dem B.
  - <pb facs="p17.jpg" n="17" id="e-1"/> <lb id="e-2"/> <tok form="Von" nform="Von" pos="APPR" lemma="von" id="w-1">Von</tok> <tok form="dem" nform="dem" pos="ART" lemma="die" id="w-2">dem</tok> <tok form="B." nform="B." pos="NN" lemma="B." id="w-3">B.</tok> <lb id="e-3"/>
- Built to allow using complex XML
  - Simple HTML forms wherever possible

CQP Corpus

- A TEITOK corpus consists of XML files
  - Not easy to search
- Automatic exporting to CQP
  - Dedicated export module tt-cwb-export
  - Fast enough even for “large” copora (10 minutes)
- Corpus search as XML index
  - Added a level to CQP for byte-offset in XML file
  - CQL => byte-offset => XML
  - Full XML nodes shown in KWIC

Variable Tokenization

- Tokens are not identical at every level
  - hertz geſper => Hertzgesperr
  - Non-aligned tiers in ANNIS
- Spaces in the appropriate level
  - <tok nform="Hertzgesperr">hertz geſper</tok>
  - <tok nform="Auf dem">Auffm</tok>
- Grammatical words as subtokens
  - <tok>Auffm<dtok form="auf"/><dtok form="dem"/></tok>
  - Reversely usable for MWE <mtok><tok/><tok/></mtok>
- Requires a choice when exporting to CWB

Additional Modules

- Dependency Trees
  - http://alfclul.clul.ul.pt/teitok/test/index.php?action=deptree&cid=catconll2009.xml&sid=s-3
- Syntactic Trees
  - http://ps.clul.ul.pt/index.php?action=psdx&cid=CARDS0001&treeid=tree-1&node=node-8
- Stand-off Annotations (error annotation)
  - http://alfclul.clul.ul.pt/teitok/learnercorpus/index.php?action=annotation&annotation=errors&cid=nl001CVETD.xml&query=

Additional Modules

- Interlinear Glossed Text
  - http://90.171.34.41/teitok/typecraft/index.php?action=igt&id=ada/2175.xml