TEITOK

MAARTEN JANSSEN

JOZEF STEFAN INSTITUTE

23 SEPTEMBER 2015

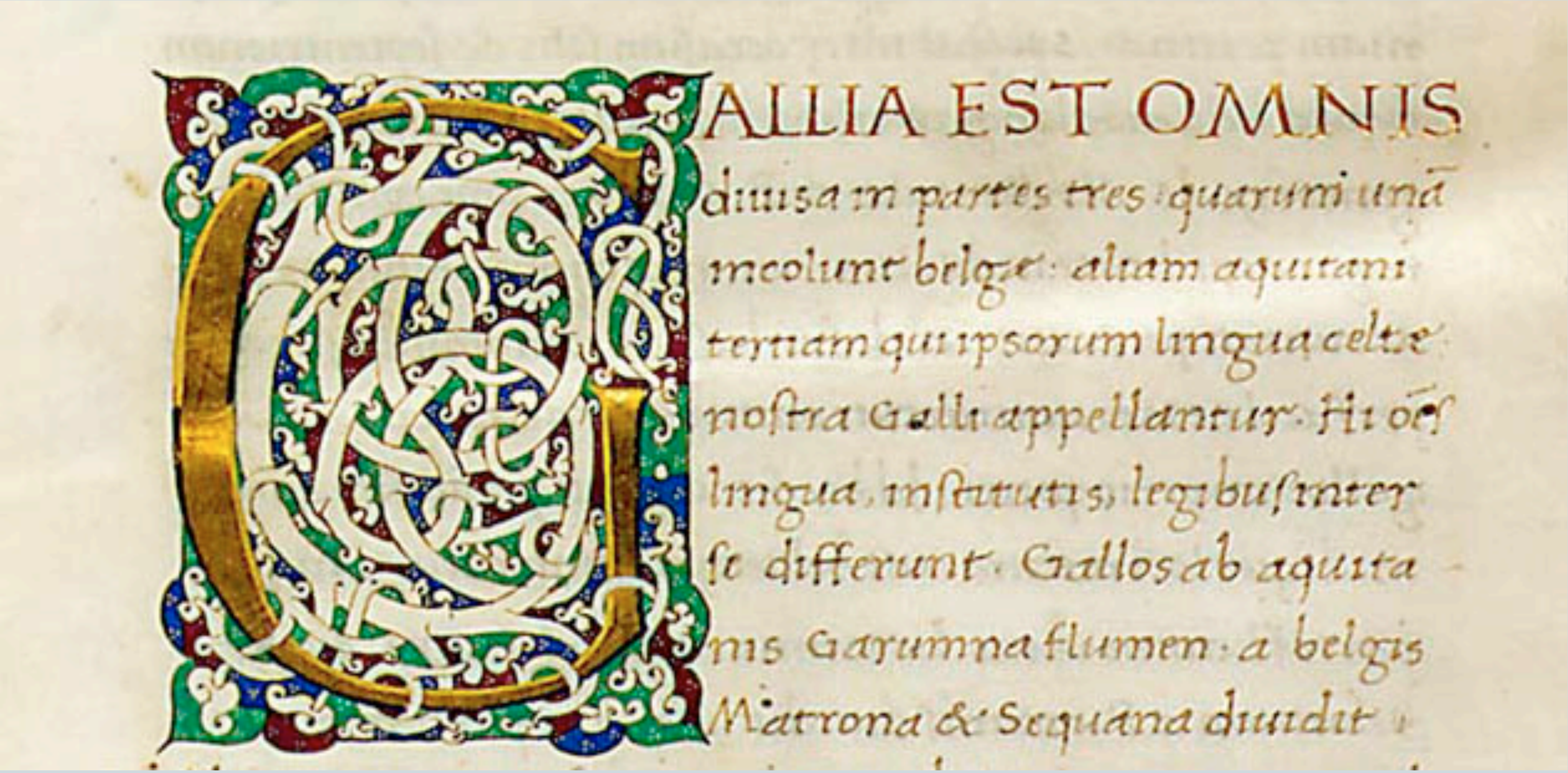
Introduction

- Traditionally 2 types of historic corpora

1. Paleographic corpora / Diplomatic corpora
    - Focus on representing textual representation
    - Deletions, rendering, hand, location, etc.
2. Linguistic corpora
    - Focus on linguistic analysis
    - Part-of-speech, lemma, syntax, semantics, etc.

- Hardly ever combined
    - Different interest groups
    - No tools to create combined corpora

Manuscript



GALLIA EST OMNIS
diuisa in partes tres: quarum unā
incolunt belgæ: aliam aquitani
tertiam qui ipsorum lingua celtæ
noſtra Galli appellantur. Hi oēs
lingua, inſtitutis, legibuſ inter
ſe differunt. Gallos ab aquita
nis Garumna flumen: a belgis
Matrona & Sequana diuidit.

Paleographic Corpus (TEI)

```
<hi rend="dropcap" n="9">G</hi>ALLIA EST OMNIS
<lb/>diuisa in partes tres quarum unã
<lb/>incolunt belgae: altam aquitani
<lb/>tertiam qui ipsorum lingua celtae
<lb/>noſtra Galli appellantur. Hi o<ex>mn</ex>e˜ſ
<lb/>lingua, inſtitutis, legibuſ inter
<lb/>ſe differunt. Gallos ab aquita
<lb/>nis Garumnea flumen, a belgis
<lb/>Matrona & Sequana diuidit.
```

Linguistic corpus

Gallia	NP	Gallia
est	VA	sum
omnis	P	omnis
divisa	VM	divido
in	SP	in
partes	NC	parte
tres	Z	tres
,	F	,
quarum	P	qui
unam	Z	unus
incolunt	VM	incolo
Belgae	NC	Belga
,	F	,

Combined Representation

```
<w lemma=“Gallia” ana=“NP”>
        <hi rend="dropcap" n="9">G</hi>ALLIA
</w>
<w lemma=“sum” ana=“VA”>
        EST
</w>
<w lemma=“omnis” ana=“P”>
        OMNIS
<lb/>
<w lemma=“divido” ana=“VM”>
        diuisa
</w>
```

Common Solution

- Create a paleographic corpus first
  - Raw XML editors
- Export text to txt file
  - Selecting a specific textual representation
  - Normalized orthography
- Create linguistic corpus based on this text
  - Independent platform
- Postscriptum
  - Oxygen XML editor to create TEI documents
  - Tycho-Brahe tool eDictor for the normalization and tagging

Drawback

- The two corpora are independent
  - Changes in the one do not affect the other
- Transcription is not (never) final
  - There are always transcription error or new decisions
- Normalization leads to changes
  - Number of words not the same in the two corpora
- Linking not possible
  - Changes on both sides make the two corpora incompatible

Combining Annotations (1)

<hi rend="dropcap" n="9">G</hi>ALLIA EST OMNIS <lb/>diuisa in partes tres quarum unã <lb/>incolunt belgae: altam aquitani <lb/>tertiam qui ipsorum lingua celtae <lb/>noſtra Galli appellantur. Hi o<ex>mn</ex>e˜ſ <lb/>lingua, inſtitutis, legibuſ inter <lb/>ſe differunt. Gallos ab aquita<lb/>nis Garumnea flumen, a belgis <lb/>Matrona & Sequana diuidit.

Combining Annotations (2)

```
<tok><hi rend="dropcap" n="9">G</hi>ALLIA</tok>
<tok>EST</tok> <tok>OMNIS</tok> <lb/><tok>diuisa</
tok> <tok>in</tok> <tok>partes</tok> <tok>tres</tok>
<tok>quarum</tok> <tok>unã</tok> <lb/><tok>incolunt</
tok> <tok>belgae</tok><tok>:</tok> <tok>altam</tok>
<tok>aquitani</tok> <lb/><tok>tertiam</tok> <tok>qui</
tok> <tok>ipsorum</tok> <tok>lingua</tok> <tok>celtae</
tok> <lb/><tok>noſtra</tok> <tok>Galli</tok>
<tok>appellantur</tok><tok>.<tok> <tok>Hi</tok>
<tok>o<ex>mn</ex>e˜ſ</tok> <lb/><tok>lingua</tok><tok>,
</tok> <tok>inſtitutis</tok><tok>, <tok> <tok>legibuſ</tok>
<tok>inter</tok> <lb/><tok>ſe</tok> <tok>differunt</
tok><tok>. </tok> <tok>Gallos</tok> <tok>ab</tok>
<tok>aquita<lb/>nis</tok> <tok>Garumnea</tok>
<tok>flumen</tok><tok>, </tok> <tok>a</tok> <tok>belgis</
tok> <lb/><tok>Matrona</tok> <tok>&</tok>
<tok>Sequana</tok> <tok>diuidit</tok><tok>.</tok>
```

Combining Annotations (3)

<tok **form=“GALLIA”**><hi rend="dropcap" n="9">G</hi>ALLIA</tok> <tok>EST</tok> <tok>OMNIS</tok> <lb/><tok **form=“divisa”**>diuisa</tok> <tok>in</tok> <tok>partes</tok> <tok>tres</tok> <tok>quarum</tok> <tok **form=“unam”**>unã</tok> <lb/><tok>incolunt</tok> <tok>belgae</tok><tok>:</tok> <tok>altam</tok> <tok>aquitani</tok> <lb/><tok>tertiam</tok> <tok>qui</tok> <tok>ipsorum</tok> <tok>lingua</tok> <tok>celtae</tok> <lb/><tok **form=“nostra”**>noſtra</tok> <tok>Galli</tok> <tok>appellantur</tok><tok>.<tok> <tok>Hi</tok> <tok **fform=“omnes”**>oe˜ſ</tok> <lb/><tok>lingua</tok><tok>, </tok> <tok **nform=“institutis”**>inſtitutis</tok><tok>, <tok> <tok **nform=“legibus”**>legibuſ</tok> <tok>inter</tok> <lb/><tok **nform=“se”**>ſe</tok> <tok>differunt</tok><tok>. </tok> <tok>Gallos</tok> <tok>ab</tok> <tok **form=“aquitanis”**>aquita<lb/>nis</tok>

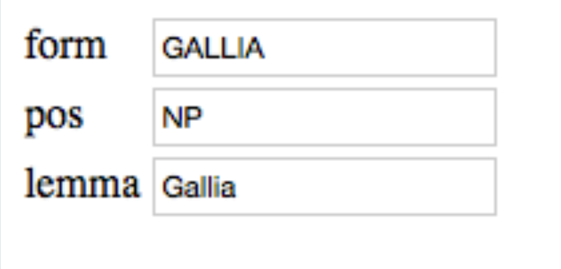
Combining Annotations (4)

```
<tok form=“GALLIA” pos=“NP” lemma=“Gallia”><hi
rend="dropcap" n="9">G</hi>ALLIA</tok> <tok
pos=“VA” lemma=“sum”>EST</tok> <tok pos=“P”
lemma=“omnis”>OMNIS</tok> <lb/><tok
form=“divisa” pos=“VM” pos=“divido”>diuisa</tok>
<tok pos=“SP” lemma=“in”>in</tok> <tok pos=“NC”
lemma=“parte”>partes</tok> <tok pos=“Z”
lemma=“tres”>tres</tok> <tok pos=“PR”
lemma=“qui”>quarum</tok> <tok pos=“Z”
lemma=“unus” form=“unam”>unã</tok> <lb/><tok
pos=“VM” lemma=“incolo”>incolunt</tok> <tok
pos=“NC” lemma=“belga”>belgae</tok><tok pos=“F”
lemma=“:”>:</tok> <tok pos=“P”
lemma=“altus”>altam</tok>
```

Graphical User Interface

- <tok form="GALLIA" **id="w-1"** pos="NP" lemma="Gallia"><hi rend="dropcap" n="9">G</hi>ALLIA</tok>

HTML Form



form	GALLIA
pos	NP
lemma	Gallia

Automated Processes

- As much computionally computed as possible
  - Scripts running behind the screens
  - Started from the Web-Based interface
- Tokenization
  - Calculation of predictable forms
  - Token (re)numbering
- POS-tagging
  - When POS tagger data available
  - Dedicated tagger (NeoTag) – other taggers need script
- Others under development/testing
  - Example-driven normalization module

Indexed Corpus

- XML corpus not searchable
  - Need for an indexed corpus
- Corpus Query Language (CQL)
  - Corpus Workbench (OpenCWB)
- Export all <tok>
  - With POS and lemma
- Import into CQP
  - Run queries from interface
  - CQPWeb, CUWI, [Sketchengine]

Link to XML

- Results link to original XML document
  - See full context
  - Can be restricted in case of copyright issues
  - Including all typesetting
- Direct lookup in XML document
  - Testing phase – too slow for larger XML documents
- XML and CQP remain linked
  - Frequent re-generation of CQP corpus
  - TEITOK mostly meant for “small” corpora (<300M)
  - Keeps corpus linked even after retokenization

Multiple forms

- Corpus always choice of form
  - Original orthography
  - Corrected errors
  - Expanded abbreviations
  - Critical form (normalized to author’s spelling)
  - Normalized form
- TEITOK has multiple forms
  - As many attributes on a <tok> as needed
  - Automatic switch between views (different text versions)
  - Inheritance tree

Form variation in CQP

- Various forms can be exported to CQP
  - Orthographic form and normalized form
- Searches by need
  - Original orthography or current orthography
- Comparative search
  - All word that used to be written with X but no longer are
- Learner corpus
  - spelling errors
- Historic corpus
  - orthographic changes

Customizable

- TEITOK used for various projects
  - Many things can be customized
- Interface design
  - Colours, logos, etc.
  - Interface language(s)
- Corpus settings
  - Which forms to use for each token
  - Which other attributes (pos, ana, etc.)
  - What to export to CQP
  - Which metadata to display
- Custom scripts, functions, etc.

Tokenization Differences

- Glued or separated words
  - “prav za prav” => “pravzaprav”
  - 3 original words, 1 normalized word
- Current day contraction
  - “aux” => “a” + “les”
  - 1 orthographic word, 2 grammatical words
- “Deep” or “Dependent” tokens
  - Orthographic unit = aux
  - Grammatical unit 1 = a
  - Grammatical unit 2 = les