Gefördert durch
DFG Deutsche Forschungsgemeinschaft

UNIVERSITÄT LEIPZIG

STANDOFF ANNOTATION FOR THE ANCIENT GREEK AND LATIN DEPENDENCY TREEBANK

04.6.2019, LiLa Conference, MILAN

Giuseppe G. A. Celano

DFG PROJECT

(https://git.informatik.uni-leipzig.de/celano/agldt1)

1. **Revise**: correct inconsistencies/errors

2. **Standardize**: make the AGLDT standoff as PAULA XML
    1. standoff for multiple annotations and multiple interpretations of the same token

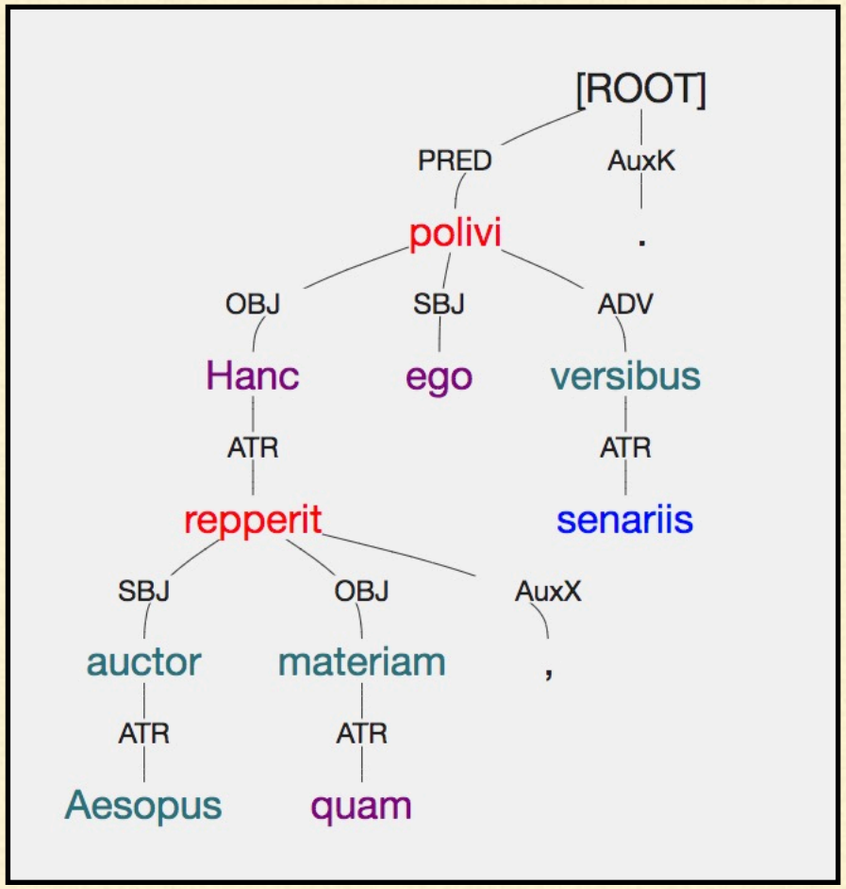
    2. standoff to overcome the problem of conflicting hierarchies

3. **Expand**: add new annotations

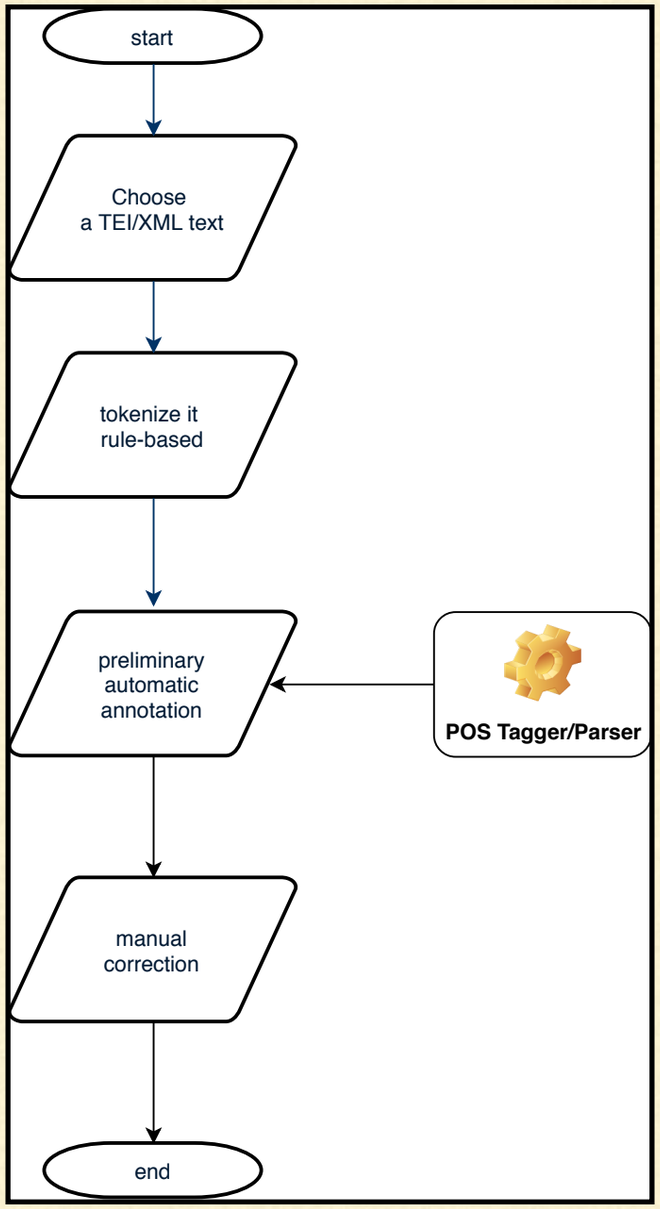
THE AGLT

- Ancient Greek texts: 557,922 tokens
- Latin texts: 79,697 tokens
- available in GitHub/GitLab:
  - https://perseusdl.github.io/treebank_data/
  - https://git.informatik.uni-leipzig.de/celano/agldt1

LABELED DIRECTED ACYCLIC GRAPHS



TREEBANK PIPELINE



THE PERSEUS TREEBANK: TEI XML TEXT

```
<text xml:lang="lat">
        <body>
<div type="edition" xml:lang="lat" n="urn:cts:latinLit:phi0448.phi002.perseus-lat2">
<div type="textpart" subtype="book" n="1"> <head>C. Iuli Caesaris Commentariorum De Bello Civili,
Liber Primus</head>
<div type="textpart" subtype="chapter" n="1">
<div type="textpart" subtype="section" n="1"><p><gap reason="lost"/> Litteris <del>a Fabio</del> C.
Caesaris consulibus redditis aegre ab his impetratum est summa tribunorum plebis contentione ut in
senatu recitarentur; ut vero ex litteris ad senatum referretur, impetrari non potuit.</p></div>
<div type="textpart" subtype="section" n="2"><p>referunt consules de re publica infinite. L.
Lentulus consul senatu rei publicae se non defuturum pollicetur, si audacter ac fortiter sententias
dicere velint; </p></div>
```

THE PERSEUS TREEBANK: INLINE ANNOTATION

```
<sentence id="2" paula_sent_id="clause2" senTxt="ut vero ex litteris ad senatum referretur, impetrari non potuit.">
    <word id="1" form="ut" lemma="ut" postag="c--------" head="11" relation="AuxC" paula_tok_id="tok20"/>
    <word id="2" form="vero" lemma="verus" postag="d--------" head="7" relation="AuxY" paula_tok_id="tok21"/>
    <word id="3" form="ex" lemma="ex" postag="r--------" head="7" relation="AuxP" paula_tok_id="tok22"/>
    <word id="4" form="litteris" lemma="littera" postag="n-p---mb-" head="3" relation="OBJ" paula_tok_id="tok23"/>
    <word id="5" form="ad" lemma="ad" postag="r--------" head="7" relation="AuxP" paula_tok_id="tok24"/>
    <word id="6" form="senatum" lemma="senatus" postag="n-s---ma-" head="5" relation="OBJ" paula_tok_id="tok25"/>
    <word id="7" form="referretur" lemma="refero" postag="v3sisp---" head="1" relation="ADV" paula_tok_id="tok26"/>
    <word id="8" form="," lemma="," postag="u--------" head="1" relation="AuxX" paula_tok_id="tok27"/>
    <word id="9" form="impetrari" lemma="impetro" postag="v--pnp---" head="11" relation="OBJ" paula_tok_id="tok28"/>
    <word id="10" form="non" lemma="non" postag="d--------" head="11" relation="AuxZ" paula_tok_id="tok29"/>
    <word id="11" form="potuit" lemma="possum" postag="v3sria---" head="0" relation="PRED" paula_tok_id="tok30"/>
    <word id="12" form="." lemma="punc" postag="u--------" head="11" relation="AuxK" paula_tok_id="tok31"/>
</sentence>
```

INLINE ANNOTATION: ADVANTAGES

1. easy to add
2. easy to query
3. well supported by annotation tools

INLINE ANNOTATION: DISADVANTAGES

1. the tokenized text becomes the new base text

2. after text extraction from a TEI text, links to the original text are virtually lost (e.g., *amabam-que* and content of some editorial markup)

3. it is unfeasible to connect such base texts to other annotation layers, e.g., with different tokenization schemes. For example:

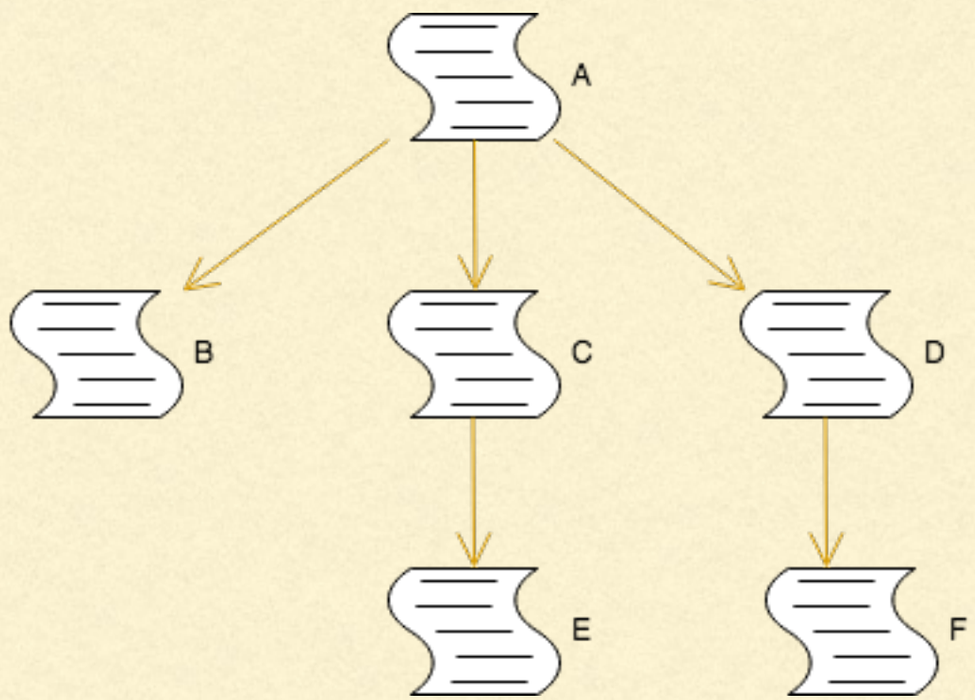
   - *amabamque*: one phonetic word

   - *amabam-que*: two syntactic words

STANDOFF ANNOTATION

1. each annotation layer is attached separately to the original text (i.e., the base text).

2. an annotation layer references the original text or another annotation layer which references the original text



STANDOFF ANNOTATION: PAULA XML

1. Open format based on the principles of LAF (ISO 24612:2012)
2. already employed in a number of historical language corpora
3. the base text is a bare xml text, which is virtually referenced only via offsets

THE CASE STUDY: DE BELLO CIVILI

(https://git.informatik.uni-leipzig.de/celano/latinnlp/tree/master/case-study))

1. the base text is a 'complex' TEI xml file'

    - reference is made via XPath coinciding with CTS references

```
<encodingDesc>
            <refsDecl n="CTS">
                <cRefPattern n="Section" matchPattern="(\w+).(\w+).(\w+)"
                    replacementPattern="#xpath(/tei:TEI/tei:text/tei:body/tei:div/
tei:div[@n='$1']/tei:div[@n='$2']/tei:div[@n='$3'])">
                    <p>This pointer pattern extracts Book and Chapter and Section</p>
                </cRefPattern>
                <cRefPattern n="Chapter" matchPattern="(\w+).(\w+)"
                    replacementPattern="#xpath(/tei:TEI/tei:text/tei:body/tei:div/
tei:div[@n='$1']/tei:div[@n='$2'])">
                    <p>This pointer pattern extracts Book and Chapter</p>
                </cRefPattern>
                <cRefPattern n="Book" matchPattern="(\w+)"
                    replacementPattern="#xpath(/tei:TEI/tei:text/tei:body/tei:div/
tei:div[@n='$1'])">
                    <p>This pointer pattern extracts Book</p>
                </cRefPattern>
```

TOKENIZATION/WORD SEGMENTATION

- Latin: rule-based
    - select the text to annotate from the TEI XML file
    - identify abbreviations (word list + regular expressions)
        - Cn. = Gnaeus
    - list of not-to-tokenize words (e.g., Antigone, aeque)
    - tokens ending with ne/que/ve
    - list of to-tokenize words (e.g., nequis, nobiscum)

PAULA:TEI BASE TEXT

```
<text xml:lang="lat">
        <body>
<div type="edition" xml:lang="lat" n="urn:cts:latinLit:phi0448.phi002.perseus-lat2">
<div type="textpart" subtype="book" n="1"> <head>C. Iuli Caesaris Commentariorum De Bello Civili,
Liber Primus</head>
<div type="textpart" subtype="chapter" n="1">
<div type="textpart" subtype="section" n="1"><p><gap reason="lost"/> Litteris <del>a Fabio</del> C.
Caesaris consulibus redditis aegre ab his impetratum est summa tribunorum plebis contentione ut in
senatu recitarentur; ut vero ex litteris ad senatum referretur, impetrari non potuit.</p></div>
<div type="textpart" subtype="section" n="2"><p>referunt consules de re publica infinite. L.
Lentulus consul senatu rei publicae se non defuturum pollicetur, si audacter ac fortiter sententias
dicere velint; </p></div>
```

PAULA:TOKENIZATION

```
<paula version="1.1">
  <header paula_id="phi0448.phi002.perseus-lat2.tok01"/>
  <markList xmlns:xlink="http://www.w3.org/1999/xlink" type="tok" xml:base="phi0448.phi002.perseus-lat2.xml">
    <mark id="tok1" xlink:href="#xpointer(string-range(/*:TEI[1]/*:text[1]/*:body[1]/*:div[1]/*:div[1]/*:div[1]/*:div[1],'',2,8))"/>
    <!-- Litteris -->
    <mark id="tok2" xlink:href="#xpointer(string-range(/*:TEI[1]/*:text[1]/*:body[1]/*:div[1]/*:div[1]/*:div[1]/*:div[1],'',19,2))"/>
    <!-- C. -->
    <mark id="tok3" xlink:href="#xpointer(string-range(/*:TEI[1]/*:text[1]/*:body[1]/*:div[1]/*:div[1]/*:div[1]/*:div[1],'',22,8))"/>
    <!-- Caesaris -->
    <mark id="tok4" xlink:href="#xpointer(string-range(/*:TEI[1]/*:text[1]/*:body[1]/*:div[1]/*:div[1]/*:div[1]/*:div[1],'',31,10))"/>
    <!-- consulibus -->
    <mark id="tok5" xlink:href="#xpointer(string-range(/*:TEI[1]/*:text[1]/*:body[1]/*:div[1]/*:div[1]/*:div[1]/*:div[1],'',42,8))"/>
    <!-- redditis -->
    <mark id="tok6" xlink:href="#xpointer(string-range(/*:TEI[1]/*:text[1]/*:body[1]/*:div[1]/*:div[1]/*:div[1]/*:div[1],'',51,5))"/>
    <!-- aegre -->
```

PAULA: SENTENCE SPLIT

```
<paula version="1.0">
  <header paula_id="phi0448.phi002.perseus-lat2.sentence_seg01"/>
  <markList xmlns:xlink="http://www.w3.org/1999/xlink" type="sentence_seg" xml:base="phi0448.phi002.perseus-lat2.tok01.xml">
    <mark id="clause1" xlink:href="#xpointer(id('tok1')/range-to(id('tok19')))"/>
    <!--Litteris C. Caesaris consulibus redditis aegre ab his impetratum est summa tribunorum plebis contentione ut in senatu recitarentur ;-->
    <!--Litteris C. Caesaris consulibus redditis aegre ab his impetratum est summa tribunorum plebis contentione ut in senatu recitarentur;-->
    <mark id="clause2" xlink:href="#xpointer(id('tok20')/range-to(id('tok31')))"/>
    <!--ut vero ex litteris ad senatum referretur , impetrari non potuit .-->
    <!--ut vero ex litteris ad senatum referretur, impetrari non potuit.-->
    <mark id="clause3" xlink:href="#xpointer(id('tok32')/range-to(id('tok38')))"/>
    <!--referunt consules de re publica infinite .-->
    <!--referunt consules de re publica infinite.-->
```

CURRENT CHALLENGES

- extraction of text from TEI texts may require different scripts
- what is the ideal tokenization/word segmentation?
- annotation tools do not support standoff annotation
- lack of support for XPointer

THANKS
FOR YOUR ATTENTION!