



Digital Grammar of Greek
Documentary Papyri (PapyGreek),
ERC Starting Grant, No. 758481



Sunoikisis Digital Classics – Fall 2020
Session 7

Using Treebanks

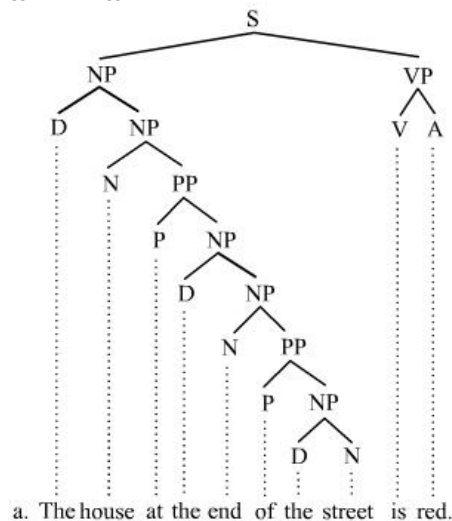
Querying Treebanked Corpora for Linguistic Phenomena

Polina Yordanova (Doctoral candidate, University of Helsinki)
polina.yordanova@helsinki.fi

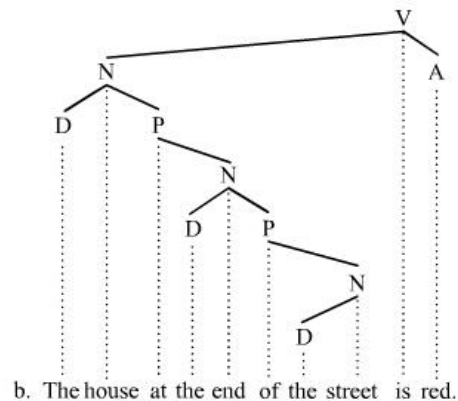


Introduction to Treebanking

- Linguistic analysis of a sentence, showing the syntactic relations between its words
- May include annotation of morphology and semantics
- Constituency vs. Dependency grammar
 - CG - focus is phrase structure
 - DG - focus is individual word



Constituency structure



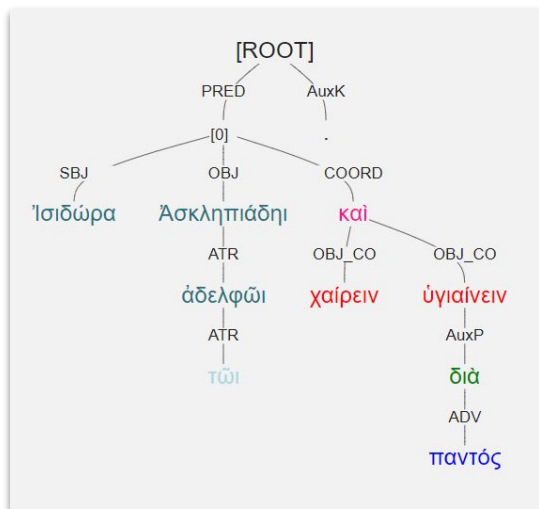
Dependency structure

Introduction to Treebanking

- Most Ancient Greek and Latin treebanks are dependency-based:
 - Index Thomisticus Treebank
 - Ancient Greek and Latin Dependency Treebank (AGLDT)
 - Gorman trees
 - PROIEL
 - Pedalion
 - Duke-nlp
 - PapyGreek

Annotation

- Manual annotation usually done through a user interface
- Information recorded as XML:



```
<sentence id="1">
  <word id="1" form="Ἰσιδώρα" lemma="Ἰσιδώρα" postag="n-s---fn-" relation="SBJ" wordid="242186" head="11"/>
  <word id="2" form="Ἀσκληπιάδῃ" lemma="Ἀσκληπιάδης" postag="n-s---md-" relation="OBJ" wordid="242187" head="11"/>
  <word id="3" form="τῶι" lemma="ὁ" postag="l-s---md-" relation="ATR" wordid="242188" head="4"/>
  <word id="4" form="ἀδελφῶι" lemma="ἀδελφός" postag="n-s---md-" relation="ATR" wordid="242189" head="2"/>
  <word id="5" form="χαίρειν" lemma="χαίρω" postag="v-pna---" relation="OBJ_CO" wordid="242190" head="6"/>
  <word id="6" form="καί" lemma="καί" postag="c-----" relation="COORD" wordid="242191" head="11"/>
  <word id="7" form="ὑγιαίνειν" lemma="ὑγιαίνω" postag="v-pna---" relation="OBJ_CO" wordid="242192" head="6"/>
  <word id="8" form="διὰ" lemma="διὰ" postag="r-----" relation="AuxP" wordid="242193" head="7"/>
  <word id="9" form="παντός" lemma="πᾶς" postag="a-s---mg-" relation="ADV" wordid="242194" head="8"/>
  <word id="10" form="." lemma="." postag="u-----" relation="AuxK" wordid="242195" head="0"/>
  <word id="11" insertion_id="0009e" artificial="elliptic" relation="PRED" form="[0]" head="0"/>
</sentence>
```

<sentence id="1">

<word id="1" form="Ἰσιδώρα" lemma="Ἰσιδώρα" postag="n-s---fn-" relation="SBJ" head="11"/>

<word id="2" form="Ἀσκληπιάδῃ" lemma="Ἀσκληπιάδης" postag="n-s---md-" relation="OBJ" head="11"/>

<word id="3" form="τῷ" lemma="ὁ" postag="l-s---md-" relation="ATR" head="4"/>

<word id="4" form="ἀδελφῷ" lemma="ἀδελφός" postag="n-s---md-" relation="ATR" head="2"/>

<word id="5" form="χαίρειν" lemma="χαίρω" postag="v--pna---" relation="OBJ_CO" head="6"/>

<word id="6" form="καὶ" lemma="καί" postag="c-----" relation="COORD" head="11"/>

<word id="7" form="ὑγιαίνειν" lemma="ὑγιαίνω" postag="v--pna---" relation="OBJ_CO" head="6"/>

<word id="8" form="διὰ" lemma="διὰ" postag="r-----" relation="AuxP" head="7"/>

<word id="9" form="παντός" lemma="πᾶς" postag="a-s---mg-" relation="ADV" head="8"/>

<word id="10" form="." lemma="." postag="u-----" relation="AuxK" head="0"/>

<word id="11" insertion_id="0009e" artificial="elliptic" relation="PRED" form="[0]" head="0"/>

</sentence>



Querying options – (some) tools

- INESS infrastructure for querying PROIEL
 - Tailored for PROIEL's treebanks
<https://clarino.uib.no/iness/page>
- DendroSearch (Alek Keersmaekers)
 - Works with a number of corpora using the AGDT annotation format; other texts can be added directly in the tool
<https://github.com/alekkeersmaekers/dendrosearch>
- TrEd
 - Graphical editor and tree visualizer
<http://ufal.mff.cuni.cz/tred/>
- Iliados, a.k.a. Structural Search (Nick Kallen)
 - <http://www.iliados.com/>

Querying options – DYI options

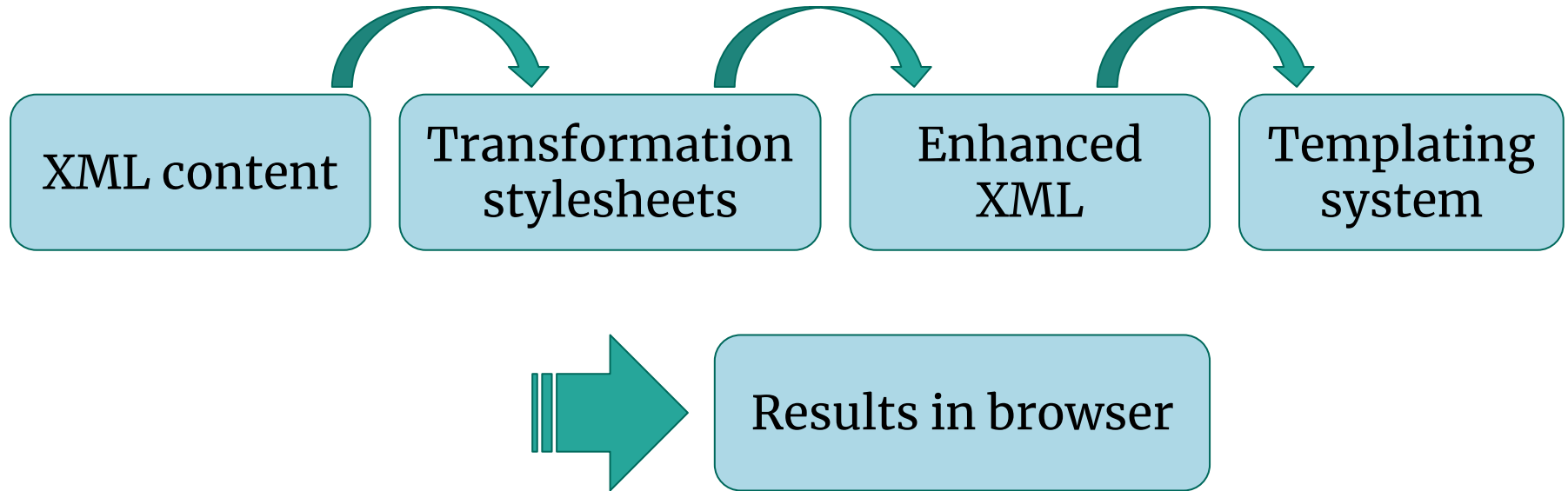
- BaseX – XML database engine with XQuery processor
- gAGDT – graph database based that includes the AGLDT data, developed by Francesco Mambrini. Based on SQL
- Kiln

Kiln platform



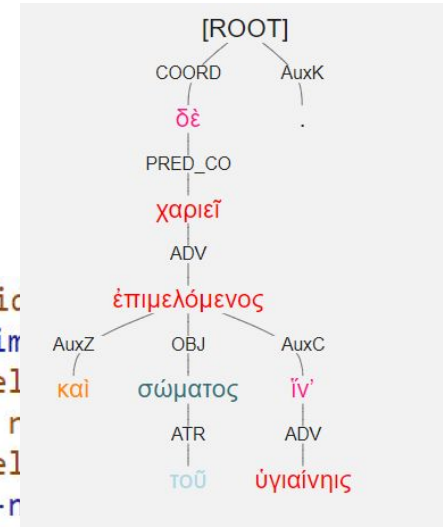
Other formats: HTML, plain text, PDF, PNG, etc.

How it works:

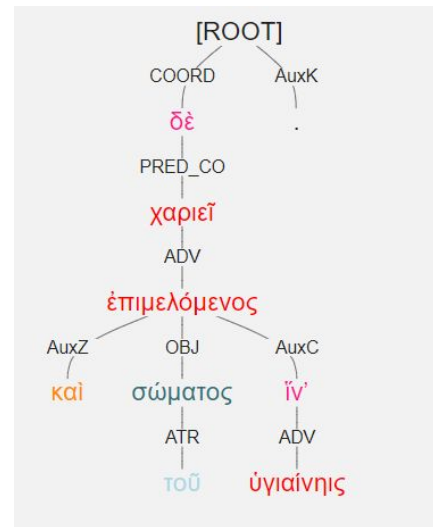


First things first...

▼ <sentence document_id="https://sematia.hum.helsinki.fi/edit/644" id
 <word form="χαριεῖ" head="2" id="1" lemma="χαρίζω" postag="v2sfin" relation="ADV" />
 <word form="δὲ" head="0" id="2" lemma="δέ" postag="c-----" relation="COORD" />
 <word form="καὶ" head="6" id="3" lemma="καί" postag="c-----" relation="AuxZ" />
 <word form="τοῦ" head="5" id="4" lemma="ὁ" postag="l-s---ng-" relation="ATR" />
 <word form="σώματος" head="6" id="5" lemma="σῶμα" postag="n-s---r" relation="OBJ" />
 <word form="ἐπιμελόμενος" head="1" id="6" lemma="ἐπιμελέομαι" postag="v-sppemn-" relation="ADV" />
 <word form="ἰν'" head="6" id="7" lemma="ἵνα" postag="c-----" relation="AuxC" />
 <word form="ὕγιαίνης" head="7" id="8" lemma="ὕγιαίνω" postag="v2spsa---" relation="ADV" />
 <word form="." head="0" id="9" lemma="punc1" postag="u-----" relation="AuxK" />
</sentence>



First things first...



```
<?xml version="1.0" encoding="UTF-8" ?>
<document document_id="https://sematia.hum.helsinki.fi/edit/644" id="6" span="" subdoc="">
  <word form="χάριτι" head="2" id="1" lemma="χαρίζω" postag="v2sfim---" relation="PRED_CO"/>
  <word form="δὲ" head="0" id="2" lemma="δέ" postag="c-----" relation="COORD"/>
  <word form="καὶ" head="6" id="3" lemma="καί" postag="c-----" relation="AuxZ"/>
  <word form="τοῦ" head="5" id="4" lemma="ὁ" postag="l-s---ng-" relation="ATR"/>
  <word form="σώματος" head="6" id="5" lemma="σῶμα" postag="n-s---ng-" relation="OBJ"/>
  <word form="ἐπιμελόμενος" head="1" id="6" lemma="ἐπιμελόμαι" postag="v-sppemn-" relation="ADV"/>
  <word form="τῆς" head="6" id="7" lemma="ἡ" postag="c-----" relation="AuxC"/>
  <word form="ὑγιεινῆς" head="7" id="8" lemma="ὑγιαίνω" postag="v2spsa---" relation="ADV"/>
  <word form="." head="0" id="9" lemma="punc1" postag="u-----" relation="AuxK"/>
</document>
```

```
<?xml version="1.0" encoding="UTF-8" ?>
<document document_id="https://sematia.hum.helsinki.fi/edit/644" id="6" span="" subdoc="">
  <word form="δὲ" head="0" id="2" lemma="δέ" postag="c-----" relation="COORD">
    <word form="χάριτι" head="2" id="1" lemma="χαρίζω" postag="v2sfim---" relation="PRED_CO">
      <word form="ἐπιμελόμενος" head="1" id="6" lemma="ἐπιμελόμαι" postag="v-sppemn-" relation="ADV">
        <word form="καὶ" head="6" id="3" lemma="καί" postag="c-----" relation="AuxZ"/>
        <word form="σώματος" head="6" id="5" lemma="σῶμα" postag="n-s---ng-" relation="OBJ">
          <word form="τοῦ" head="5" id="4" lemma="ὁ" postag="l-s---ng-" relation="ATR"/>
        </word>
        <word form="τῆς" head="6" id="7" lemma="ἡ" postag="c-----" relation="AuxC">
          <word form="ὑγιεινῆς" head="7" id="8" lemma="ὑγιαίνω" postag="v2spsa---" relation="ADV"/>
        </word>
      </word>
    </word>
  </word>
  <word form="." head="0" id="9" lemma="punc1" postag="u-----" relation="AuxK"/>
</document>
```

Querying linguistic phenomena – GA

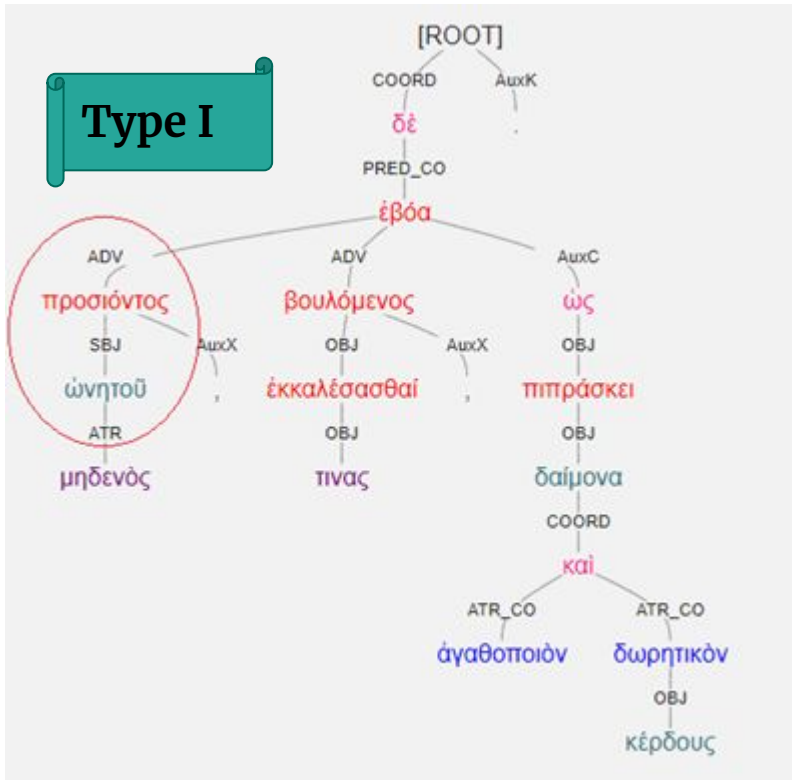
What’s a “genitive absolute”?



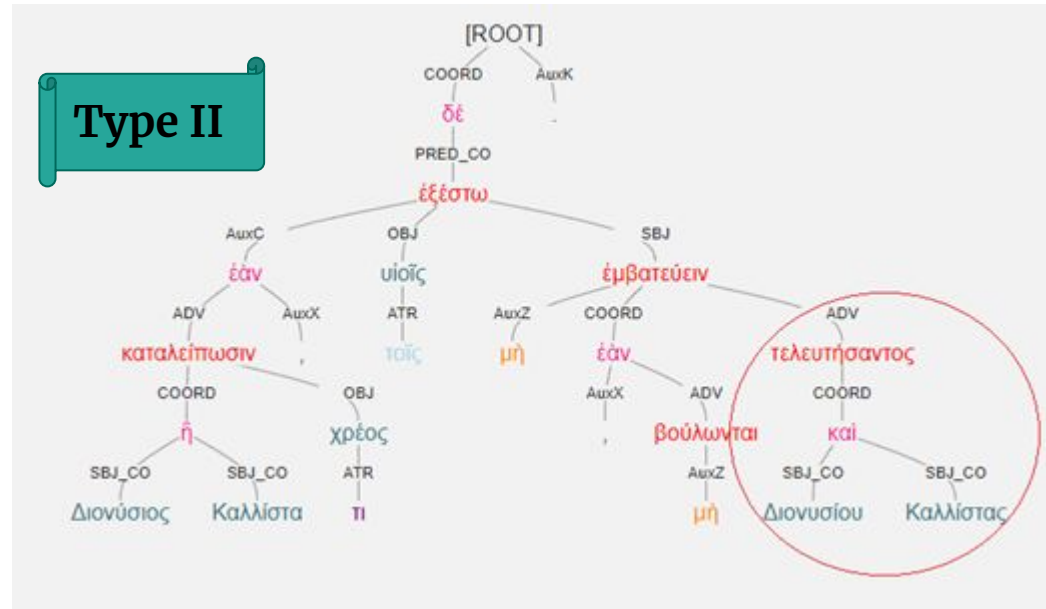
- Similar to Latin’s ablative absolute
- Genitive participle serves as the subordinate conjugated verb
- A substantive in the genitive is the agent
- “Loose” – not referring to elements from the main clause
 - ... or does it?
- Let’s focus on the standard cases

Genitive absolute in trees

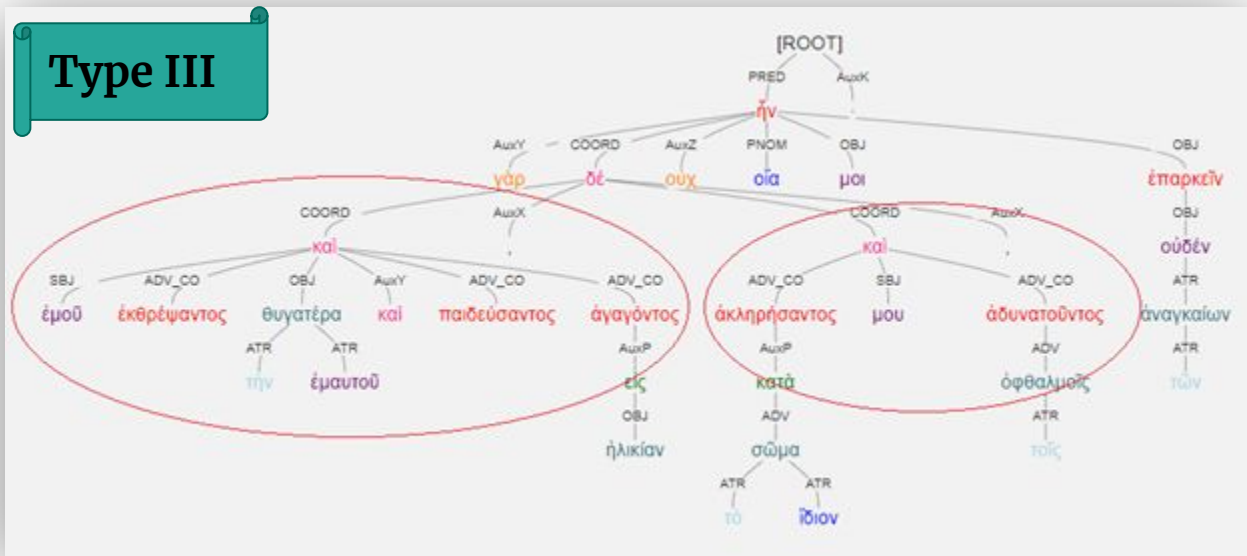
Type I



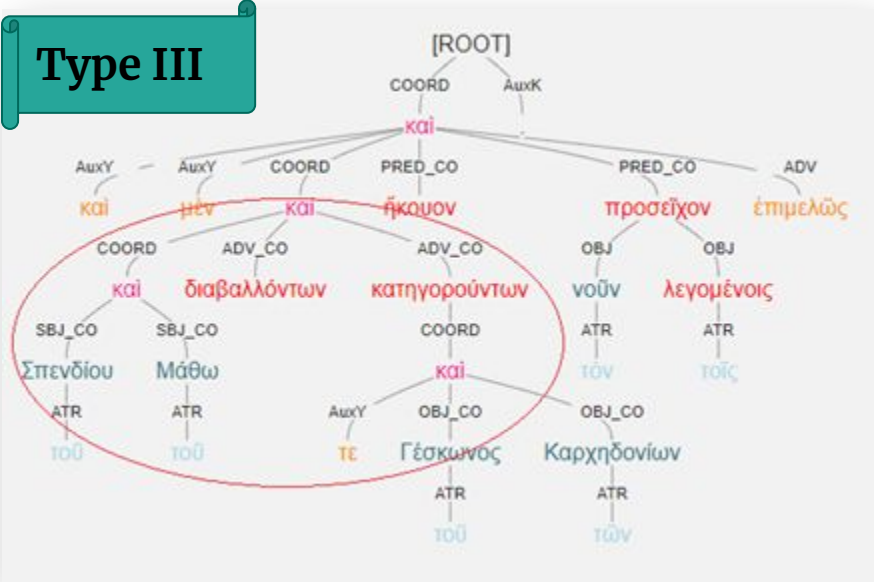
Type II



Type III



Type III



Querying for linguistic phenomena – GA

How can we make this more easily discoverable?



- Querying in XSLT is applying filters:

Can we give it a specific marker that will allow it to be caught by the appropriate sieve?

Set the rules:

- Possible relations between a participle and a genitive agent in GA
 - Type I – direct parent;
 - Type II – ancestor connected only via coordinating conjunction(s)
 - Type III – sibling, where the agent's parent is a conjunction
 - Type IV – i.e. anywhere in the tree, as long as the path between to it goes only through coordinators

```
▼<word id="4" form="ζῶντος" lemma="ζῶν" postag="v-sppamg-" relation="ADV" head="1" group="3">
  ▼<word id="5" form="καὶ" lemma="καὶ" postag="c-----" relation="COORD" head="4" group="3">
    ▼<word id="3" form="πατρὸς" lemma="πατήρ" postag="n-s---mg-" relation="SBJ_CO" head="5" group="3">
      <word id="2" form="τοῦ" lemma="ὁ" postag="1-s---mg-" relation="ATR" head="3"/>
    </word>
    ▼<word id="7" form="μητρὸς" lemma="μήτηρ" postag="n-s---fg-" relation="SBJ_CO" head="5" group="3">
      <word id="6" form="τῆς" lemma="ὁ" postag="1-s---fg-" relation="ATR" head="7"/>
    </word>
  </word>
</word>
```

Set the rules:

- The verb from the matrix clause is also annotated as governing the construction:

```
▼ <word id="1" form="λαμβάνωσι" lemma="λαμβάνω" postag="v3ppsa---" relation="PRED" head="0" governs-4="4">
  ▼ <word id="4" form="ζώντος" lemma="ζάω" postag="v-sppamg-" relation="ADV" head="1" group="3">
    ▼ <word id="5" form="καὶ" lemma="καί" postag="c-----" relation="COORD" head="4" group="3">
      ▼ <word id="3" form="πατρός" lemma="πατήρ" postag="n-s---mg-" relation="SBJ_CO" head="5" group="3">
        <word id="2" form="τοῦ" lemma="ὁ" postag="l-s---mg-" relation="ATR" head="3"/>
      </word>
      ▼ <word id="7" form="μητρός" lemma="μήτηρ" postag="n-s---fg-" relation="SBJ_CO" head="5" group="3">
        <word id="6" form="τῆς" lemma="ὁ" postag="l-s---fg-" relation="ATR" head="7"/>
      </word>
    </word>
  </word>
</word>
```

Summarize info per construction:

- Which sentence is the construction in?
- What is the value of @group defining all its participants?
- What is the agent's gender?
- What is the number of its governing participle?
- In word order, which one comes first, the agent or the participle?
- What “type” of construction is it, regarding the coordination of the dependency?

```
▼ <constructions>  
  <construction sentence="1" group="3" masculine="true" feminine="true" constituents-order="mixed" order-in-sentence="vc" number="s" type="2"/>  
</constructions>
```



Overview of preprocessing:

For an individual text:

- Get the preprocessed tree

```
<map:match id="preprocess-genitive-absolute-text" pattern="pp/construction/genitive-absolute/*/*.xml">
```

- Apply transformational stylesheets

```
<map:generate src="cocoon:/pp/tree/{1}/{2}.xml"/>
```

```
<map:transform src="../../stylesheets/treebank/pp-genitive-absolute-participants.xml"/>
```

```
<map:transform src="../../stylesheets/treebank/pp-genitive-absolute-participants-coords.xml"/>
```

```
<map:transform src="../../stylesheets/treebank/pp-genitive-absolute-participants-governing-verb.xml"/>
```

```
<map:transform src="../../stylesheets/treebank/pp-genitive-absolute-prune-annotations.xml"/>
```

```
<map:transform src="../../stylesheets/treebank/pp-genitive-absolute-constructions.xml"/>
```

- Create an XML document from the result

```
<map:serialize type="xml"/>
```

```
</map:match>
```

Overview of preprocessing:

For the whole corpus:

- Create a single giant XML from the result trees for each individual text file

```
<map:match id="preprocess-genitive-absolute-corpus" pattern="pp/construction/genitive-absolute/*.xml">
  <map:generate src="cocoon://_internal/dirlist/content/xml/treebank/{1}.xml"/>
  <map:transform src="../../stylesheets/admin/dir-to-xinclude.xml">
    <map:parameter name="prefix" value="cocoon:/pp/construction/genitive-absolute/{1}/"/>
  </map:transform>
  <map:transform type="xinclude"/>
  <map:serialize type="xml"/>
</map:match>
```

And finally...

Nice neat tables:

Number of sentences	1254
Sentences containing genitive absolute constructions	51
Number of files containing genitive absolute	36
Total number of genitive absolute constructions	69

Text type	Text count	Genitive absolute count
None	71	4
account_receipt	5	3
contract_loan	3	1
contract_sale	2	2
letter	20	6
letter_business	22	4
letter_private	68	12
petition	14	37

Agent before participle	35
Participle before agent	28
Mixed	6

Number	Gender	Count
Singular participle	Masculine	23
	Feminine	6
	Neuter	6
Dual participle	Masculine	0
	Feminine	0
	Neuter	0
Plural participle	Masculine	19
	Feminine	2
	Neuter	3

Type	Count
Type 1 - single participle governing single agent	54
Type 2 - single participle governing co-ordinated agents	6
Type 3 - co-ordinated participles sharing a single agent	8
Type 4 - co-ordinated participles governing co-ordinated agents	1

Text document	Category	Date	Sentences	GA	Type	Instances										Plural				Word order				Constituents order			
						I	II	III	IV	m	f	n	m	f	n	m	f	n	cv	vc	URGH	ap	pa	mixed			
p-calr-zen-1-59019	letter	-3	16	1	1	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0	0						
p-calr-zen-1-59027	letter_business	-3	10	1	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0						
p-calr-zen-1-59028	letter_business	-3	10	1	1	0	0	0	1	0	0	0	0	0	0	1	0	0	1	0	0						
p-calr-zen-1-59031	letter_business	-3	9	1	1	0	0	0	1	0	0	0	0	0	0	1	0	0	1	0	0						
p-calr-zen-1-59036	letter_business	-3	20	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0						
p-calr-zen-1-59059	None	--3	10	2	2	0	0	0	1	0	0	0	0	0	0	1	0	1	2	0	0						
p-calr-zen-2-59245	letter	-3	11	1	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0						
p-col-4-66	letter	-3	16	4	4	0	0	0	2	0	0	0	0	0	4	0	0	3	1	0	0						
bgu-3-994-m1	contract_sale	-2	8	1	0	1	0	0	0	0	0	0	1	1	0	1	0	0	0	1	0						
p-drylon-1-11	contract_loan	-2	4	1	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0						
p-drylon-1-32	petition	-2	4	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0						
p-grenf-2-15-m1	contract_sale	-2	6	1	0	1	0	0	0	0	0	0	1	1	0	1	0	0	0	1	0						
p-hels-1-2	petition	-2	6	4	3	0	1	0	1	0	1	2	0	0	3	1	0	2	1	1	0						
upz-1-10	petition	-2	4	3	3	0	0	0	0	2	0	0	0	1	3	0	0	2	1	0	0						
upz-1-111	None	-2	4	2	2	0	0	0	0	0	0	2	0	0	1	1	0	0	0	2	0						
upz-1-12	petition	-2	4	4	3	0	1	0	4	0	0	0	0	0	4	0	0	1	2	1	0						
upz-1-13	petition	-2	4	1	1	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0	0						
upz-1-18r	petition	-2	12	2	2	0	0	0	1	0	1	0	0	0	2	0	0	1	1	0	0						
upz-1-2	petition	-2	6	3	3	0	0	0	0	1	2	0	0	0	1	0	2	3	0	0	0						
upz-1-22	petition	--2	11	3	3	0	0	0	0	1	0	1	0	0	1	0	2	1	2	0	0						
upz-1-29-m1-1	account_receipt	-2	2	1	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0	1	0						
upz-1-30	account_receipt	-2	1	1	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0	1	0						
upz-1-34	petition	--2	6	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0						
upz-1-35	petition	--2	12	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0						
upz-1-36-m1	petition	--2	6	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0						
upz-1-59-m1-1	letter_private	-2	6	6	5	1	0	0	2	0	0	1	0	0	3	3	0	5	0	1	0						
upz-1-6	petition	-2	9	7	2	0	4	1	2	0	1	4	0	0	4	0	3	4	2	1	0						
upz-1-60	letter_private	-2	8	1	1	0	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0						
upz-1-69	letter_private	-2	11	1	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0						
upz-1-7-m1-2	petition	-2	6	2	1	1	0	0	0	1	0	1	0	0	0	2	0	0	0	2	0						
upz-1-8	petition	-2	7	4	3	0	1	0	1	0	0	2	0	0	1	3	0	2	1	1	0						
p-flor-3-367	letter_private	3	7	1	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	1						
p-oxv-14-1666	letter_private	3	13	1	1	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	0						
p-bour-25	letter_private	4-5	10	1	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	1	0						
p-lond-6-1915	letter_private	4	9	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0						
p-nich-13-668-m1	account_receipt	6	3	1	1	0	0	0	1	0	0	0	0	0	0	1	0	0	1	0	0						

... all displayed on a single web page.