

Sunoikisis Digital Classics - Fall 2020 Session 7

Using Treebanks

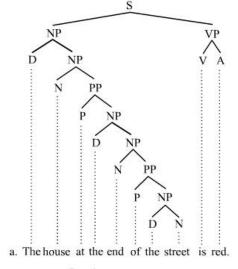
Querying Treebanked Corpora for Linguistic Phenomena

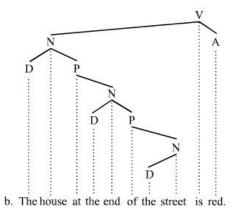




Introduction to Treebanking

- 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
 - o DG focus is individual word





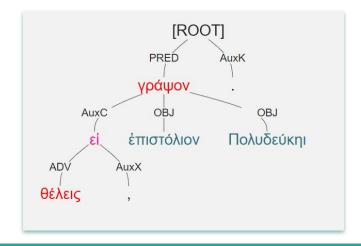
Constituency structure

Dependency structure

Source: Tjo3ya (https://commons.wikimedia.org/wiki/File:The house at the end of the street.ipg)

Introduction to Treebanking

- Main finite verb (or a coordinator) is the head of the tree (ROOT)
- Head <-> dependent relationships between all other words
 - Each word can have only one head
 - Words can have one, multiple or zero dependents
 - Head = parent, governor;
 - Dependent = child, descendent;
 - Siblings = words sharing the same parent word
- Relation labels indicate the syntactic function



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
 - o Duke-nlp
 - PapyGreek



Annotation

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

```
[ROOT]
                PRED
  SBJ
                 OBJ
                             COORD
            Άσκληπιάδηι
Ίσιδώρα
                              Kai.
                 ATR
                           OBJ CO
                                       OBJ CO
              άδελφῶι
                           χαίρειν
                                      ὑνιαίνειν
                 ATR
                                         AuxP
                                         διὰ
                                         ADV
                                       παντός
```

```
<sentence id="1">
    <word id="1" form="Τσιδώσα" lemma="Τσιδώσα" postag="n-s---fn-" relation="SBJ" wordid="242186" head="11"/>
    <word id="2" form="Ασκληπιάδηι" lemma="Ασκληπιάδης" postag="n-s---md-" relation="OBJ" wordid="242187" head="11"/>
    <word id="3" form="πῶι" lemma="ὁ" postag="l-s---md-" relation="ATR" wordid="242188" head="4"/>
    <word id="4" form="ἀδελφῶι" lemma="ἀδελφῶς" postag="n-s---md-" relation="ATR" wordid="242189" head="2"/>
    <word id="5" form="χαίσειν" lemma="χαίσω" postag="v--pna---" relation="OBJ_CO" wordid="242190" head="6"/>
    <word id="6" form="καὶ" postag="c------" relation="COORD" wordid="242191" head="11"/>
    <word id="7" form="ψιαίνειν" lemma="ὑγιαίνειν" postag="v--pna---" relation="OBJ_CO" wordid="242192" head="6"/>
    <word id="8" form="διὰ lemma="ὑγιαίνειν" postag="r-------" relation="AuxP" wordid="242193" head="7"/>
    <word id="9" 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>
```



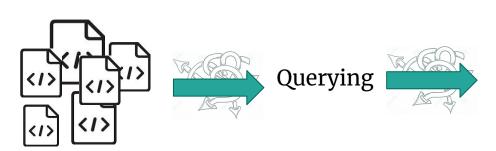


```
<sentence id="1">
 <word id="1" form="Ισιδώρα" lemma="Ισιδώρα" postag="n-s---fn-" relation="SBJ" head="11"/>
 <word id="2" form="Ασκληπιάδηι" lemma="Ασκληπιάδης" postag="n-s---md-" relation="OBJ" head="11"/>
 <word id="3" form="τωι" lemma="o" postag="l-s---md-" relation="ATR" head="4"/>
 <word id="4" 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="καί" 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="\deltaı\dot{\alpha}" lemma="\deltaı\dot{\alpha}" 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>
```



Finding one's way in the digital forest

From angle brackets to statistics and interpretation



Instances

Word Distance	Subtree Distance \$	Lemma ¢	Governor ¢	Hyperbaton ¢	Instance
7	3	άργαλέος	Noun	true	Έχων τις γυναίκα πρός πάντας τοὺς οἰκείους λίαν τὸ ἤθος ἀργαλέαν ήβουλήθη γνῶναι εί καὶ πρός τοὺς πατρώους οἰκέτας ὁμοίως διάκειται - [334]
6	3	άλόγιστος	Noun	true	Πρός ἄνδρα μεγολοπρεπή μέν σώμαπ , κατά ψυχήν δέ άλόγιστον ό λόγος εύκοιρος . έστί [303]
5	1	άψυχος	Pronoun	true	λιής εκιστρόμων συνορισήμετος πρός τον υμπλές έπρθέρει το δεν Δελφοίς μοντίδον, ώς ένέτει η προθεσμία, λιαβών στρομθών οις ήν γέρο και Τούτο το βωταίν αυτεπάστα, έγενει έχει δερό και ότα ός αντιβρός επρώρια πότερο ή εξυπνών έχει μετά χέρος ή θιμέγον, βουλόμενος, έδιν μέν διμέγο έται, ξωόν τό στρουθίον έπιθείξαι, έδιν δέ έμπνουν , όποιπίξος προσυγμέν. Δίπ (Ε.41)
4	2	πᾶς	Noun	true	Καὶ παραλαβών οἴκαδε , πάντα μέν αὐτῷ προστίγε τὰ <mark>ρύμματα</mark> , πᾶσι δὲ λούτροις ἐπειρᾶτο καθαίρειν . [86]
2	1	δωρητικός	Noun	true	μηδενός δέ ώνητοῦ προσιόντος , ἀκιαλέσασθαί τινας βουλόμενος , έβδα ώς ἀγαθοποιόν <mark>δαίμονα</mark> καί κέρδους δωρητικόν πιπράσκει . [10]
2	2	πολύς	Noun	false	Τοῦ δέ και τοῦτο τὶς ιὐφέλειαν τῶν δεσποτῶν πράπειν φήσαντος · πολλά γὰρ αὐτοῖς ιὐὰ τίκτεσθοι παρασκευάζει , διαπορηθείς ἐκεῖνος ἔφη · [94]
2	1	μέγας	Noun	false	Ο λόγος δηλοΐ ότι οι τούς κρείπτονας χλευάζοντες λανθάνουσι μείζονας ἐαυτοῖς τὰς <mark>ἀνίας</mark> ἐξ αὐτῶν ἐπισπώμενοι . [146]
2	2	περισσός	Noun	false	Καὶ δή ἀπάσας ἀθροίσασα παρήγει αύταῖς τὰς οὐρὰς ἀποκόπτειν , λέγουσα ώς οὐκ ἀπρεπὲς μόνον τοῦτο , ἀλλὰ καὶ περισσόν τ α ύταῖς βάρος προσήρηται . ἐστί [291]
1	1	ξύλινος	Noun	false	Ξύλινόν τις Έρμη ν κατασκευάσας καὶ προσενεγκών εἰς άγορὰν ἐπώλει · [9]
1	1	σφοδρός	Noun	true	οὖ κομισθέντος ἐπὶ τὴν καλιάν , σφοδρὸς ἐμπεσών ἄνεμος ἐκ λεπτοῦ καὶ παλαιοῦ κάρφους λαμπρὰν φλόγα ἀνῆψε . [25]
1	1	ἔτερος	Pronoun	false	Έτερος δέ τις τοῦτον έξαγοράσας ἀνεπτέρωσε τὸν ἀετὸν αὐτίκα . [50]
1	1	παλαιός	Noun	false	* Ού θέλω τὴν λύπην τῶν παλαιῶν μου συμφορῶν μεμνῆσθαι , καὶ διὰ τοῦτο τὰς ἐρήμους οἰκῶ . * [73]
1	1	μάνος	Pronoun	true	Ως δὲ οὐκ ἔπειθε , προσαγαγών ὖν ῆν εἶχε μόνην , παρόντος αύτοῦ ἐπώλει . [77]
1	1	πᾶς	Noun	false	Καὶ παραλαβών οἴκαδε , πάντα μὲν αὐτῷ προσήγε τὰ ῥύμματα , πᾶσι δὲ λούτροις ἐπειρᾶτο καθαίραν . [86]
1	1	πρόσφορος	Noun	true	Αίλουρος άκούσας δτι έν τινι έπαύλα δρνεις νοσοθοι , σχηματίσας έαυτόν είς Ιατρόν καὶ τὰ τῆς έπιστήμης πρόσφορα άναλαβώ ν έργαλεία , παρεγένετο , καὶ στὰς πρό τῆς έπαύλεως έπυνθάνετο αύτῶν πῶς ἔχοιεν . [106]
1	1	περισσός	Noun	false	Τοῦ δέ ποιμένος άχοριστίαν αύτιῶν κατηγοροῦντος , είγε περιποτέρας αύταὶ τημελείας έπιτυχοῦσαι καταλείπουσιν αύτόν , Εφασαν έπιστραφείσαι - [131]
1	1	φιλέρημος	Noun	true	Άλκύων <mark>δρνεόν έστι φιλέρημον δ</mark> ιά ποντός έν θαλάτη διαπώμενον . [198]
1	1	έτερος	Noun	false	Ετέρα δὲ ἀλώπηξ τῆδε παριούσα , ώς ῆκουσεν αὐτῆς τὸν στεναγμόν , προσελθούσα έπυνθάνετο τὴν αίτίαν . [216]
1	1	μέλας	Noun	false	Καὶ ή μὲν προβεβηκινῖα αίδουμένη νεωτέρω αὐτῆς πλησιάζειν , διετέλει , εί ποτε πρὸς αὐτὴν παρεγένετο , τὸς μελαίνας αὐτοῦ τρίχας περιαιρουμένη . [356]





Querying options - (some) tools

- INESS infrastructure for querying PROIEL
 - Tailored for PROIEL's treebanks <u>https://clarino.uib.no/iness/page</u>
- 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 <u>http://ufal.mff.cuni.cz/tred/</u>
- 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



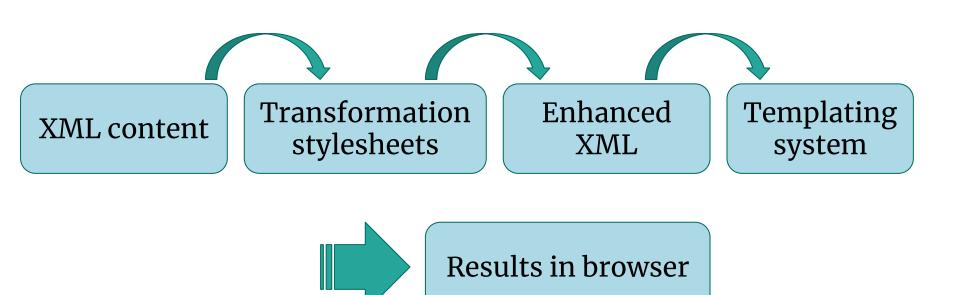
Kiln platform



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



How it works:





First things first...

```
ADV
▼<sentence document id="https://sematia.hum.helsinki.fi/edit/644" ic
                                                                            ἐπιμελόμενος
   <word form="yaριεῖ" head="2" id="1" lemma="yaρίζω" postag="v2sfin Auxz"</pre>
                                                                                OBJ
                                                                                      AuxC
   <word form="δὲ" head="0| id="2" lemma="δέ" postag="c----" rel καὶ</pre>
                                                                              σώματος
   <word form="καὶ" head="6" id="3" lemma="καί" postag="c-----" r</pre>
                                                                                ATR
                                                                                      ADV
   <word form="τοῦ" head="5" id="4" lemma="o' postag="l-s---ng-" rel</pre>
                                                                                     ὑνιαίνηις
   <word form="σώματος" head="6" id="5" lemma="σῶμα" postag="n-s---r</pre>
   <word form="έπιμελόμενος" head="1" id="6" lemma="έπιμελέομαι" postag="v-sppemn-" relation="ADV"/>
   <word form="ĭv'" head="6" id="7" lemma="ĭvα" postag="c----- relation="AuxC"/>
   <word form="ὑγιαίνηις" head="7" id="8" lemma="ὑγιαίνω" postag="v2spsa---" relation="ADV"/>
   <word form="." head="0" id="9" lemma="punc1" postag="u----- relation="AuxK"/>
 </sentence>
```



[ROOT]

COORD

OÈ

PRED_CO

XAPIEĨ

First things first...

```
▼<sentence document_id="https://sematia.hum.helsinki.fi/edit/644" id="6" span="" subdoc="">
    <word form="χαριεῖ" 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="καῖ" head="6" id="3" lemma="καῖ" postag="c-----" relation="AuxZ"/>
    <word form="τοῦ" head="5" id="4" lemma="ö" postag="l-s---ng-" relation="ATR"/>
    <word form="σώματος" head="6" id="5" lemma="σώμα" postag="n-s---ng-" relation="OBJ"/>
    <word form="έπιμελόμενος" head="1" id="6" lemma="έπιμελέομαι" postag="v-sppemn-" relation="ADV"/>
    <word form="ἴν'" head="6" id="7" lemma="ἴνα" postag="c------" relation="AuxC"/>
    <word form="ὑγιαίνηις" head="7" id="8" lemma="ὑγιαίνω" postag="v2spsa---" relation="ADV"/>
    <word form="." head="0" id="9" lemma="punc1" postag="u-----" relation="AuxK"/>
    </sentence>
```

```
[ROOT]

COORD AuxK

δὲ .

PRED_CO

Χαριεῖ

ΑDV

ἐπιμελόμενος

ΑυχΖ ΟΒJ ΑυχC

Καὶ σώματος ἵν΄

ΑΤR ΑDV

τοῦ ὑγιαίνηις
```



```
v<sentence 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="χαριεῖ" head="2" id="1" lemma="χαρίζω" postag="v2sfim---" relation="PRED CO">
     ▼<word form="ἐπιμελόμενος" head="1" id="6" lemma="ἐπιμελέομαι" postag="v-sppemn-" relation="ADV">
        <word form="καί" head="6" id="3" lemma="καί" postag="c-----" relation="AuxZ"/>
       ▼<word form="σώματος" head="6" id="5" lemma="σῶμα" postag="n-s---ng-" relation="0BJ">
          <word form="to0" head="5" id="4" lemma="0" postag="1-s---ng-" relation="ATR"/>
        </word>
       ▼<word form="ĭv'" head="6" id="7" lemma="ĭvα" postag="c------ relation="AuxC">
          <word form="ὑνιαίνηις" head="7" id="8" lemma="ὑνιαίνω" postag="v2spsa---" relation="ADV"/>
        </word>
       </word>
     </word>
   </word>
   <word form="." head="0" id="9" lemma="punc1" postag="u----- relation="AuxK"/>
 </sentence>
```





Querying linguistic phenomena - GA

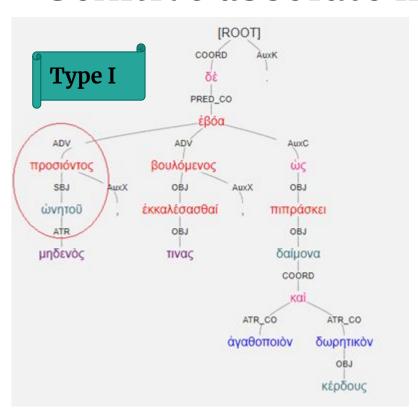
What's a "genitive absolute"?

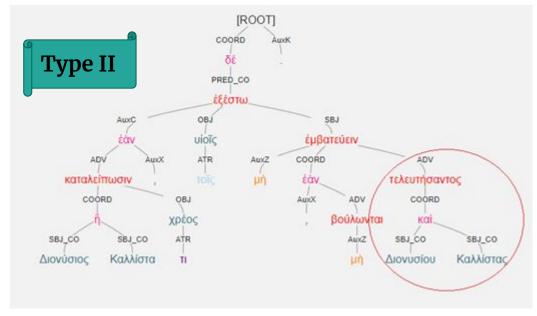


- Similar to Latin's ablative absolute
- Genitive participle serves as the subordinate conjugated verb
- A substantive in the genitive is the agent
- "Loose" not referring to elements from the main clause
 - ... or does it?
- Let's focus on the standard cases



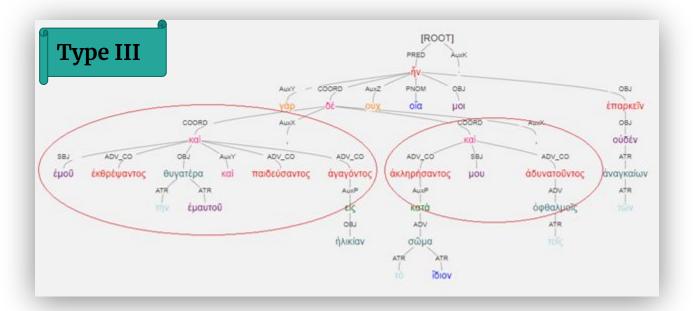
Genitive absolute in trees





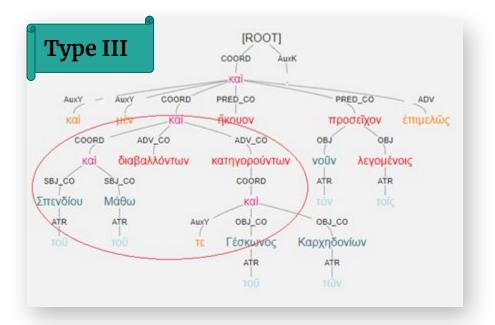
















Querying for linguistic phenomena - GA

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

```
| Volume |
```



Set the rules:

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



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:

Overview of preprocessing:

For the whole corpus:

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

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

Agent before participle	35
Participle before agent	28
Mixed	6

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

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

Туре	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

						-							tanc				-	-							-			Constituents order		
Text document •	Category ¢	Date -	Sentences &	GA	Φ -	Туре		7.000											Plural				Wo		rder	_	+			
	1000 5 050 (5	1000000000	0.030.000.000.000			1 4	♦ II	+	III		IV	\$	m	\$	F	\$ 1	1 0	m		f	n		CV	•	VC	•	URGH \$	ap ¢	pa ¢	mixed
							Т											Г			Т									
p-cair-zen-1- 59019	letter	-3	16	1		1	0		0		0		0		0	()	1		0	0		0		0		1	1	0	0
p-cair-zen-1- 59027	letter_business	-3	10	1		1	0		0		0		1		0	()	0		0	0		1		0		0	0	1	0
p-cair-zen-1- 59028	letter_business	-3	10	1		1	0		0		0		1		9	()	0		0	0		1		0		0	1	0	0
p-cair-zen-1- 59031	letter_business	-3	9	1		1	0		0		0		1	0	0	()	0		0	0		1		0		0	1	0	0
p-cair-zen-1- 59036	letter_business	-3	20	1		1	0		0		0		0		0	-)	0		0	0		1		0		0	1	0	0
p-cair-zen-1- 59059	None	3	10	2		2	0		0		0		1		0	-)	0		0	0		1		0		1	2	0	0
p-cair-zen-2- 59245	letter	-3	11	1		1	0		0		0		1	-	0)	0		0	0		1		0		0	0	1	0
p-col-4-66	letter	-3	16	4		4	0		0		0		2	- 1	0	-)	0		0	0		4		0		0	3	1	0
bgu-3-994-m1	contract_sale	-2	8	1		0	1		0		0		0	-	0	()	1		1	0		1		0		0	0	1	0
p-dryton-1-11	contract_loan	-2	4	1		1	0		0		0		1	0	0	()	0		0	0		1		0		0	0	1	0
p-dryton-1-32	petition	-2	4	1		1	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	1	0	-)	1		1	0		1		0		0	0	1	0
p-hels-1-2	petition	-2	6	4		3	0		1		0		1		0	13	1	2		0	0		3		1		0	2	1	1
upz-1-10	petition	-2	4	3		3	0		0		0		0	- 6	2	-)	0		0	1		3		0		0	2	1	0
upz-1-111	None	-2	4	2		2	0		0		0		0		0	-)	2		0	0		1		1		0	0	2	0
upz-1-12	petition	-2 -	4	4		3	0		1		0		4	9	0	-)	0		0	0		4		0		0	1	2	1
upz-1-13	petition	-2	4	1		1	0		0		0		1		0	-)	0		0	0		1		0		0	1	0	0
upz-1-18r	petition	-2	12	2		2	0		0		0		1	- 6	0	-	1	0		0	0		2		0		0	1	1	0
upz-1-2	petition	-2	6	3		3	0		0		0		0		1	-	2	0		0	0		1		0		2	3	0	0
upz-1-22	petition	2	11	3		3	0		0		0		0	-	1	-)	1		0	0		1		0		2	1	2	0
upz-1-29-m1-	account_receipt		2	1		0	1		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)	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		1		0		0	0	1	0
upz-1-35	petition	2	12	1		1	0		0		0		0	-	0	-)	0		0	0		1		0		0	0	1	0
upz-1-36-m1	petition	2	6	1		1	0		0		0		0	-	0)	0		0	0		1		0		0	0	1	0
upz-1-59-m1-	letter_private	-2	6	6		5	1		0		0		2	-	0	()	1		0	0		3		3		0	5	0	1
upz-1-6	petition	-2 -	9	7		2	0		4		1		2	- 0	0		1	4		0	0		4		0		3	4	2	1
upz-1-60	letter_private	-2	8	1		1	0		0		0		0	-	0	-	1	0		0	0		1		0		0	1	0	0
upz-1-69	letter_private	-2	11	1		1	0		0		0		1	-	9	()	0		0	0		1		0		0	0	1	0
upz-1-7-m1-2	petition	-2	6	2		1	1		0		0		0	- 2	1)	1		0	0		0		2		0	0	2	0
upz-1-8	petition	-2 -	7	4		3	0		1		0		1	-	0	()	2		0	0		1		3		0	2	1	1
p-flor-3-367	letter_private	3	7	1		0	0		1		0		1	-	0)	0		0	0		1		0		0	0	0	1
p-oxy-14- 1666	letter_private	3	13	1		1	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		1		0	0	1	0
p-lond-6-1915	letter_private	4	9	1		1	0		0		0		0	-	0	()	0		0	1		1		0		0	1	0	0
p-mich-13- 668-m1	account_receipt	6	3	1		1	0		0		0		1		0	-)	0		0	0		1		0		0	1	0	0

... all displayed on a single web page.