Understanding Texts as Graphs

An inclusive approach to text modeling

Elli Bleeker
Bram Buitendijk
Ronald Haentjens Dekker
Astrid Kulsdom

R&D Humanities Cluster
Royal Science Academy of the Netherlands

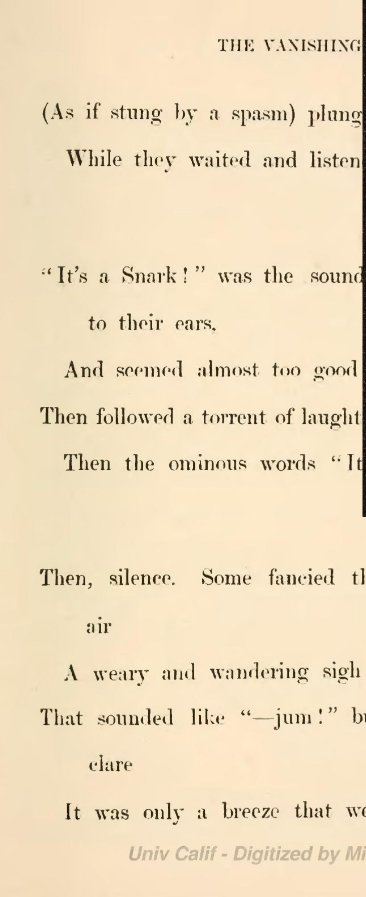
@ellibleeker
@ronald_dekker
@bram_buitendijk

COMHUM conference
Lausanne
June 4, 2018

Modeling Textual Objects

Components of text modeling:

1. A source text
2. A model of the source text.
3. A transcription of the source text



THE VANISHING

(As if stung by a spasm) plung

While they waited and listen

"It's a Snark!" was the sound

to their ears,

And seemed almost too good

Then followed a torrent of laught

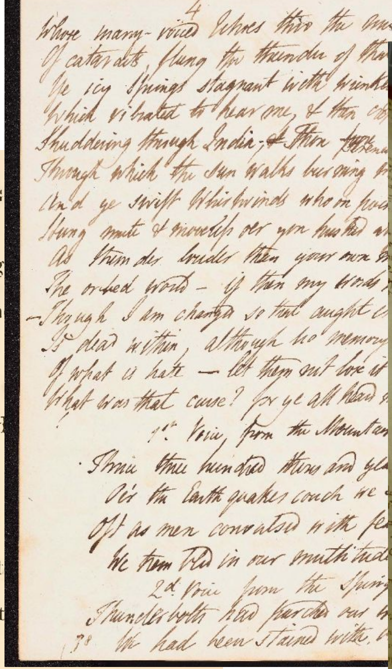
Then the ominous words "It

Then, silence. Some fancied they heard in the
air

A weary and wandering sigh

That sounded like "—jum!" but the others de-
clare

It was only a breeze that went by.



4

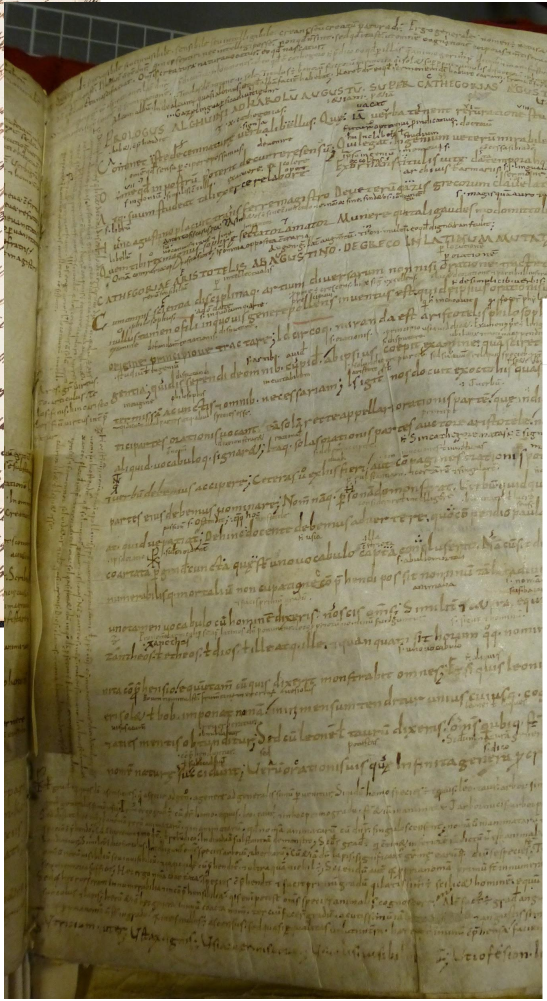
Whose many-voiced Echoes thro the mist
Of cataracts, flung the thunder of th
Ye icy Springs stagnant with wrinkl
Which vibrated to hear me, & then c
Shuddering through India;—& thou ~~[illegible]~~ serene
Through which the Sun walks burning
And ye swift Whirlwinds who on poi
Hung mute & moveless o'er yon hushed
As thunder louder than your own
The orbed world — if then my words
—Though I am changed so that aught
Is dead within, although no memory
Of what is hate — let them not lose it
What was that curse? for ye all heard

1st Voice from the Mountain

Thrice three hundred thousand yea
O'er the Earthquakes couch we
Oft as men convulsed with fe
We trembled in our multitude

2d Voice from the Spring

Thunderbolts had parched our
78 We had been stained with



PROLOGUS ALCHUINI AD KAROLU AUGUSTU. SUPER CATHEGORIAS AUGUSTINI.

Continet iste decem naturae verbalibellus: Quae vim verba tenent rerum ratione stupenda:

Omne qd in nostrum poterit decurrere sensum; Qui legat. ingenium veterum mirabile laudat.

[illegible] ce re labore [illegible] archivis armarii

A rq. suum studeat taliter [illegible] Graecorum clave latina.

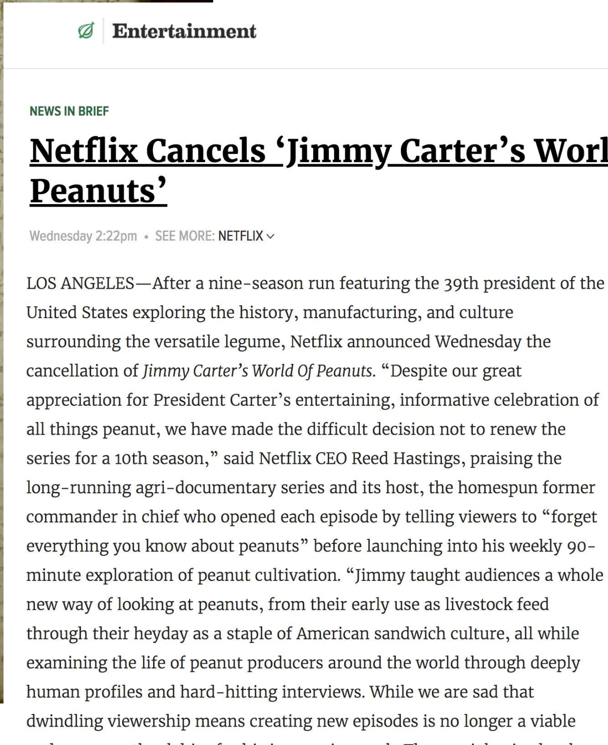
Hunc agustino placuit transferre magistro De veterum gazis

Quem tibi rex magnus sophiae sectator amator Munere qui tali gaudes modo domne legendi

CATHEGORIAE ARISTOTELIS AB AGUSTINO. DE GRECO IN LATINUM MUTATE.

Cum omnis scientia disciplinaq. artium diversarum non nisi oratione tractetur!

[illegible]



Entertainment

NEWS IN BRIEF

Netflix Cancels 'Jimmy Carter's Worl Peanuts'

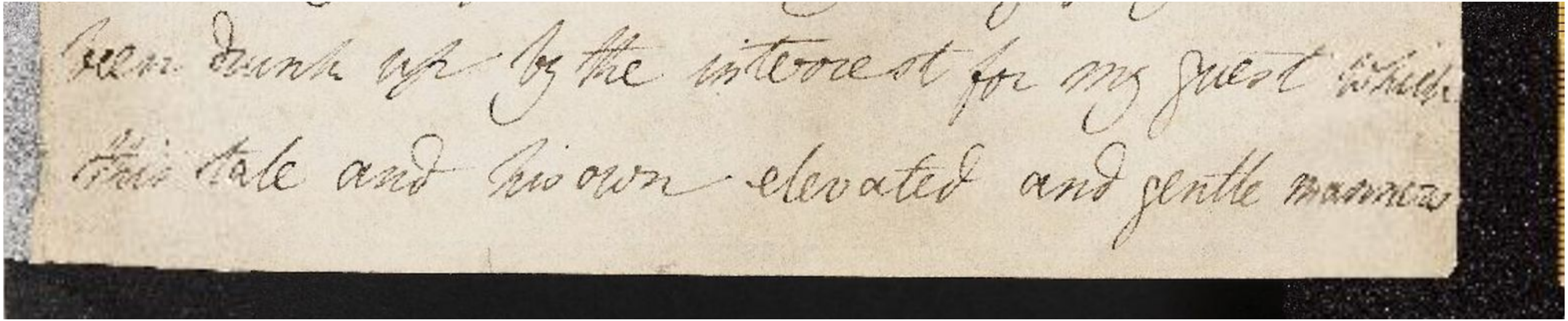
Wednesday 2:22pm • SEE MORE: NETFLIX

LOS ANGELES—After a nine-season run featuring the 39th president of the United States exploring the history, manufacturing, and culture surrounding the versatile legume, Netflix announced Wednesday the cancellation of *Jimmy Carter's World Of Peanuts.* "Despite our great appreciation for President Carter's entertaining, informative celebration of all things peanut, we have made the difficult decision not to renew the series for a 10th season," said Netflix CEO Reed Hastings, praising the long-running agri-documentary series and its host, the homespun former commander in chief who opened each episode by telling viewers to "forget everything you know about peanuts" before launching into his weekly 90-minute exploration of peanut cultivation. "Jimmy taught audiences a whole new way of looking at peanuts, from their early use as livestock feed through their heyday as a staple of American sandwich culture, all while examining the life of peanut producers around the world through deeply human profiles and hard-hitting interviews. While we are sad that dwindling viewership means creating new episodes is no longer a viable

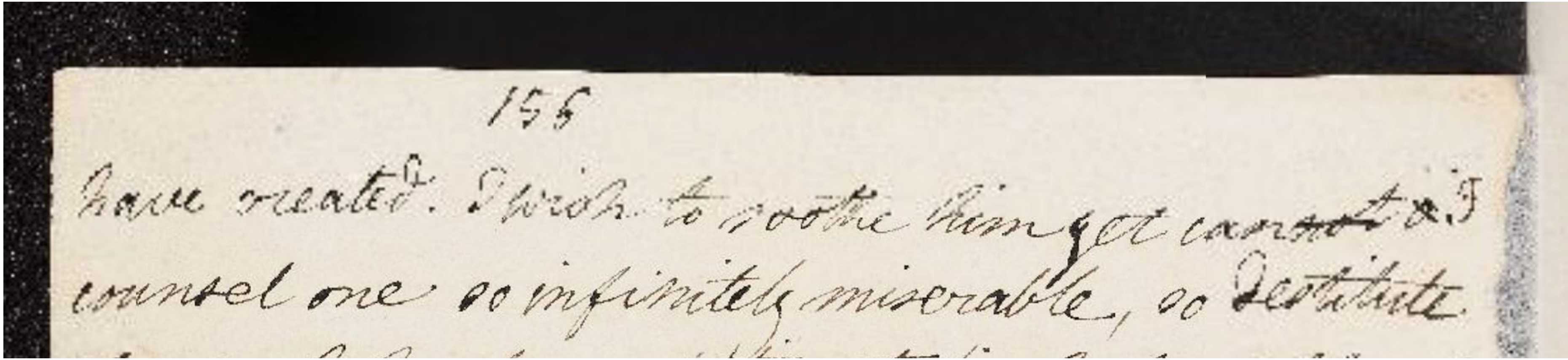
Modeling textual objects

(1) Source texts:

- Overlapping structures (including self-overlap)
- Non-linearity
- Discontinuity



been drunk up by the interest for my guest which this tale and his own elevated and gentle manners

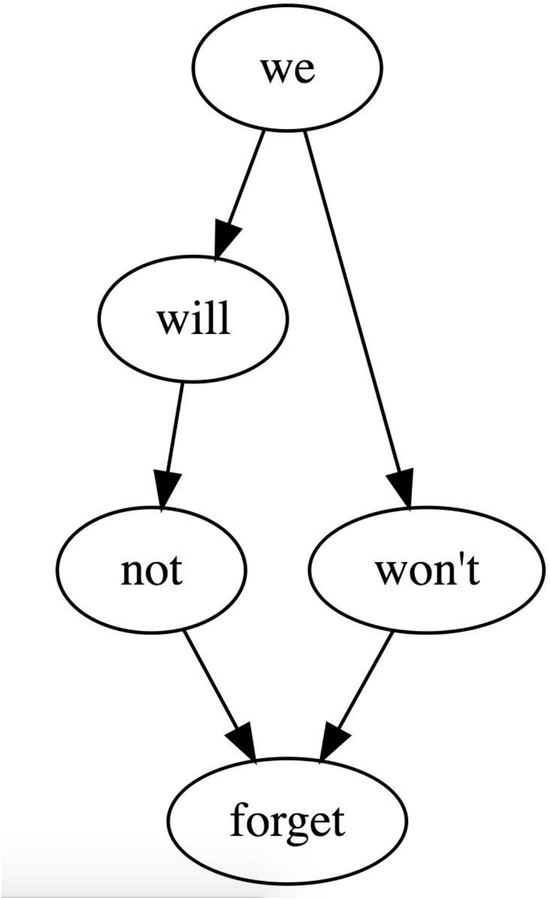


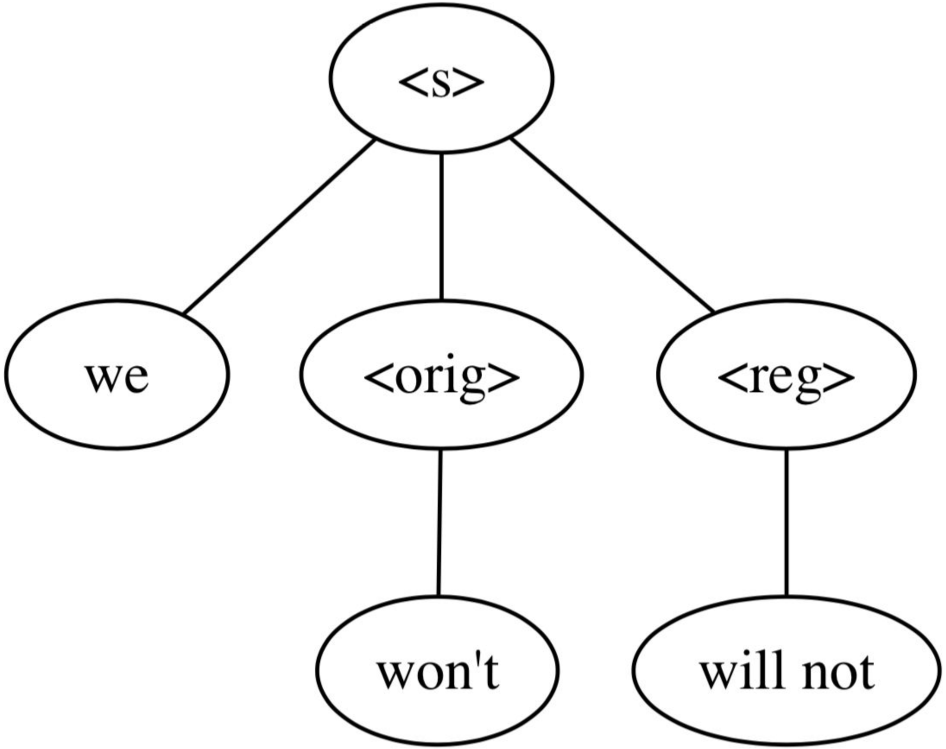
156

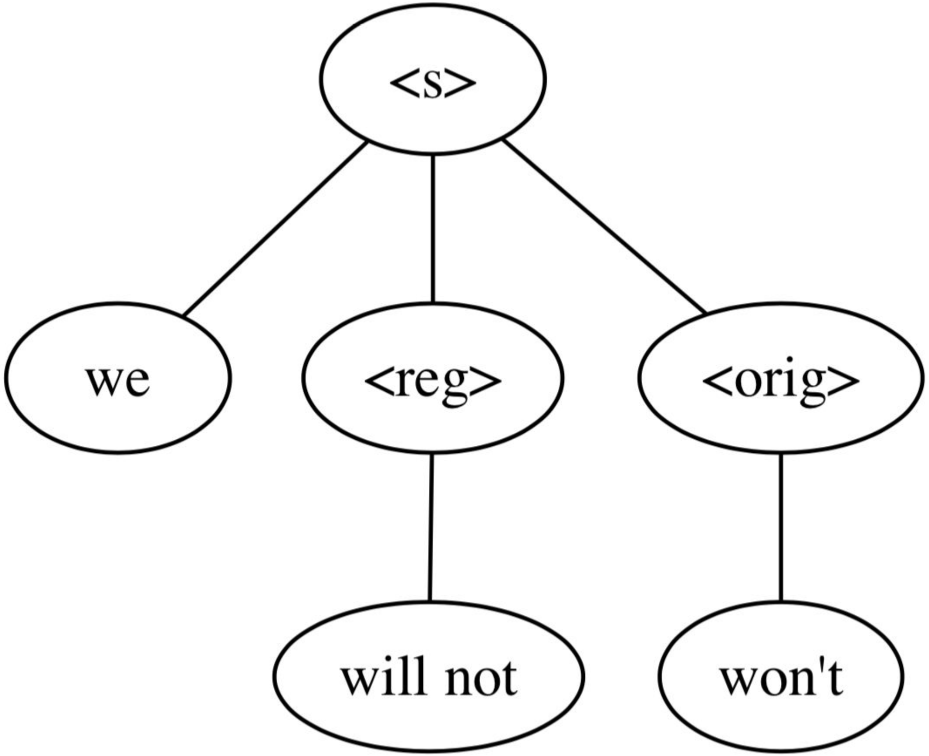
have created. I wish to soothe him yet can~~not~~ I counsel one so infinitely miserable, so destitute

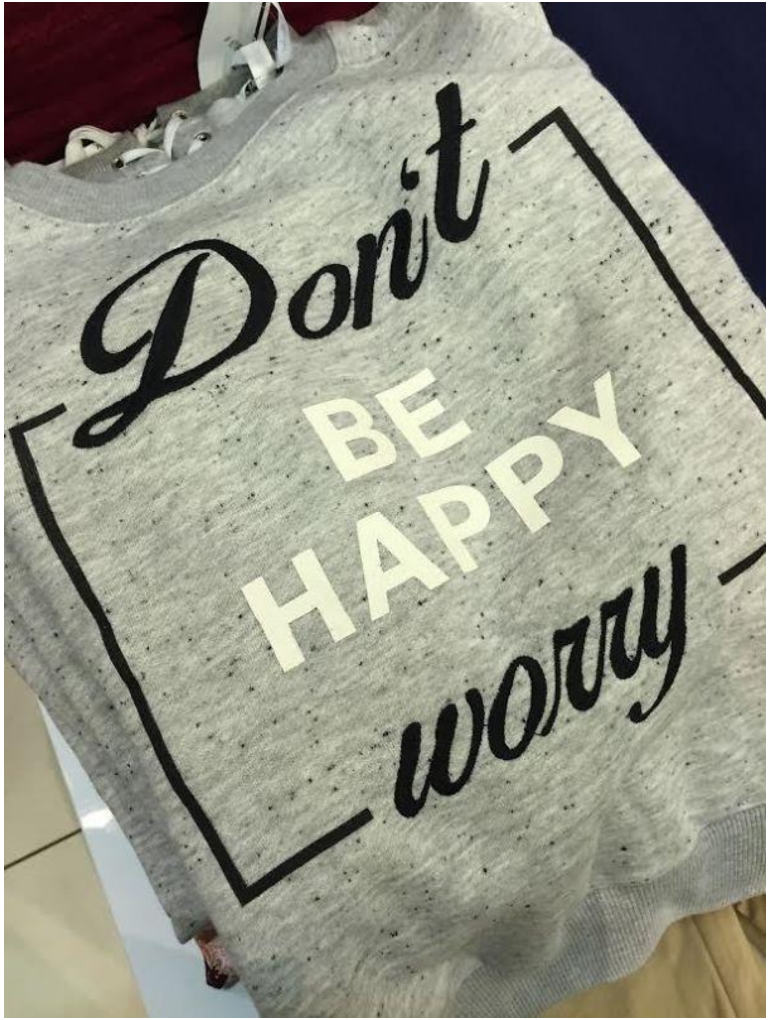
Source: Shelley, M. W. "Frankenstein, MS. Abinger C. 58", in The Shelley-Godwin Archive, MS. Abinger c. 58, 22r

```
we won't forget
```




<s>
we
<orig>
<reg>
won't
will not


<s>
we
<reg>
<orig>
will not
won't


Don't
BE
HAPPY
worry

Modeling textual objects

(2) Models and data structures

- String (e.g. plain text)
- JSON
- Tree (e.g. XML)
- Graph (e.g. RDF or GODDAG)
- Hypergraph (e.g. TAG)

Modeling textual objects

(3) Markup

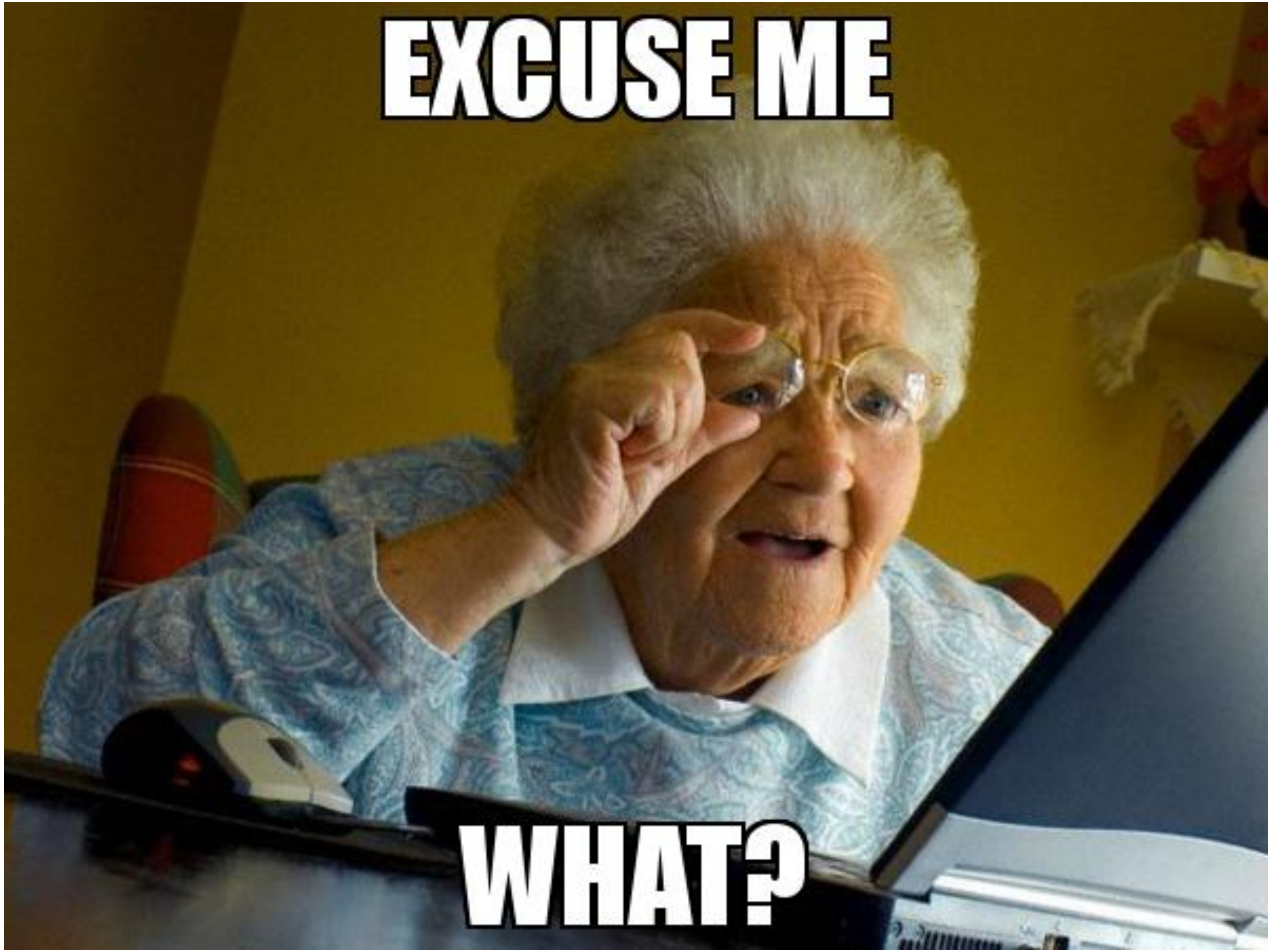
A serialization of the model

Structural capturing of information

Adding "layers" of additional information

What text really is

A multi-layered, non-linear object containing information which is at times ordered, partially ordered, and unordered.


EXCUSE ME
WHAT?

Multi-layered

Non-linear

Ordered

- Textual content (nodes)

Partially ordered

- Textual variation

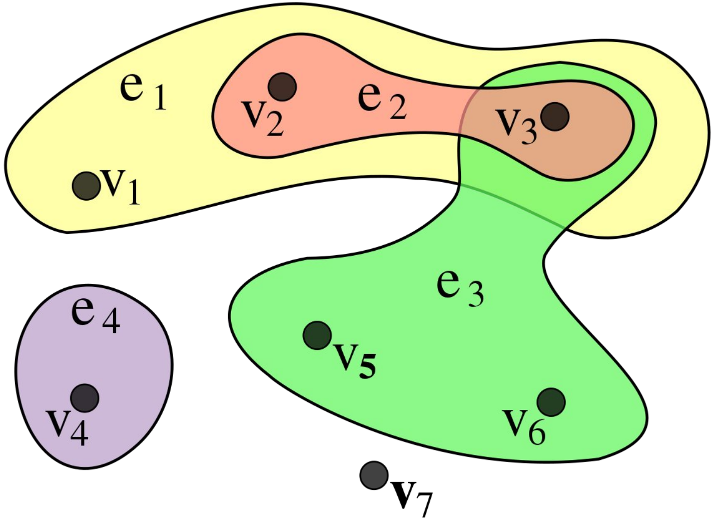
Unordered

- Metadata (e.g. name:value pairs)

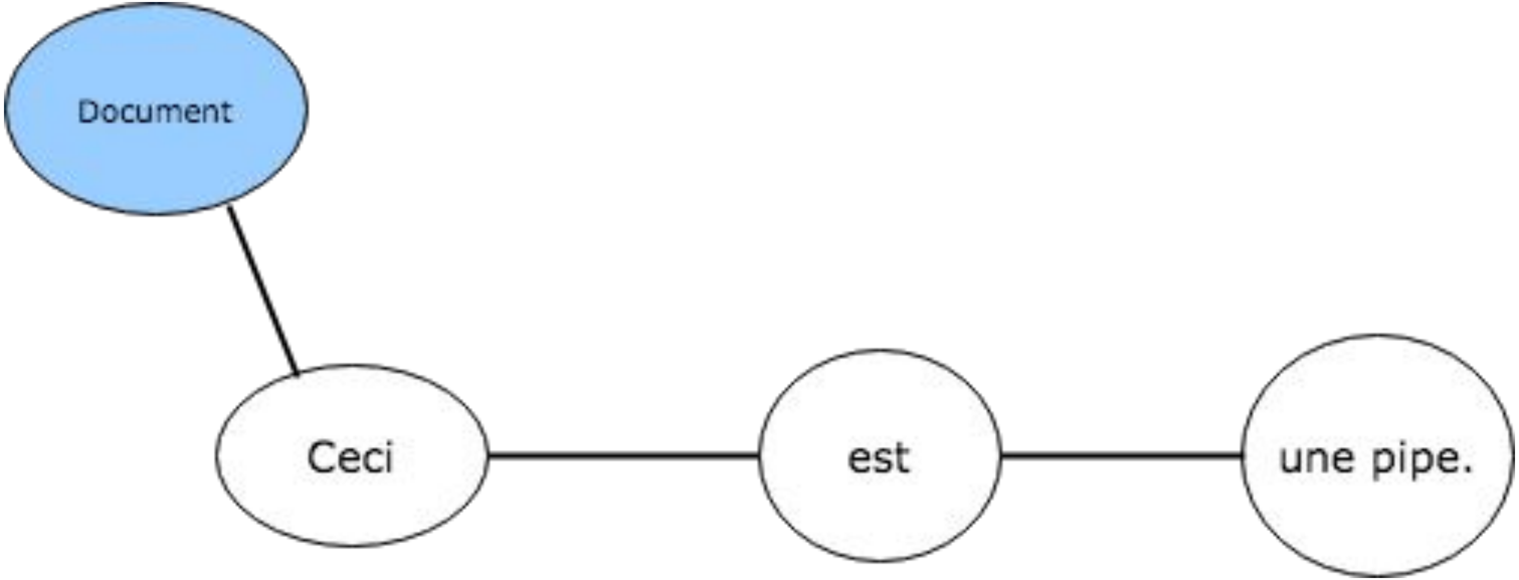
TAG

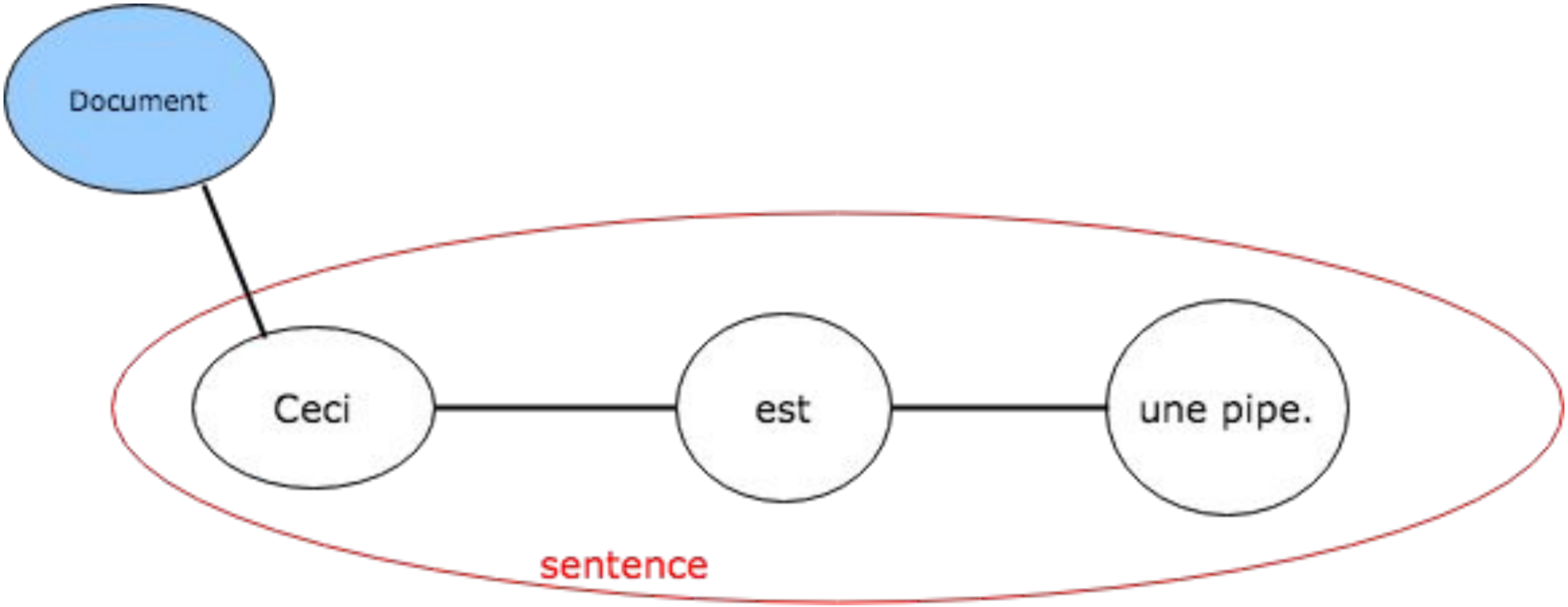
TAG data model: non-uniform cyclic property hypergraph of text

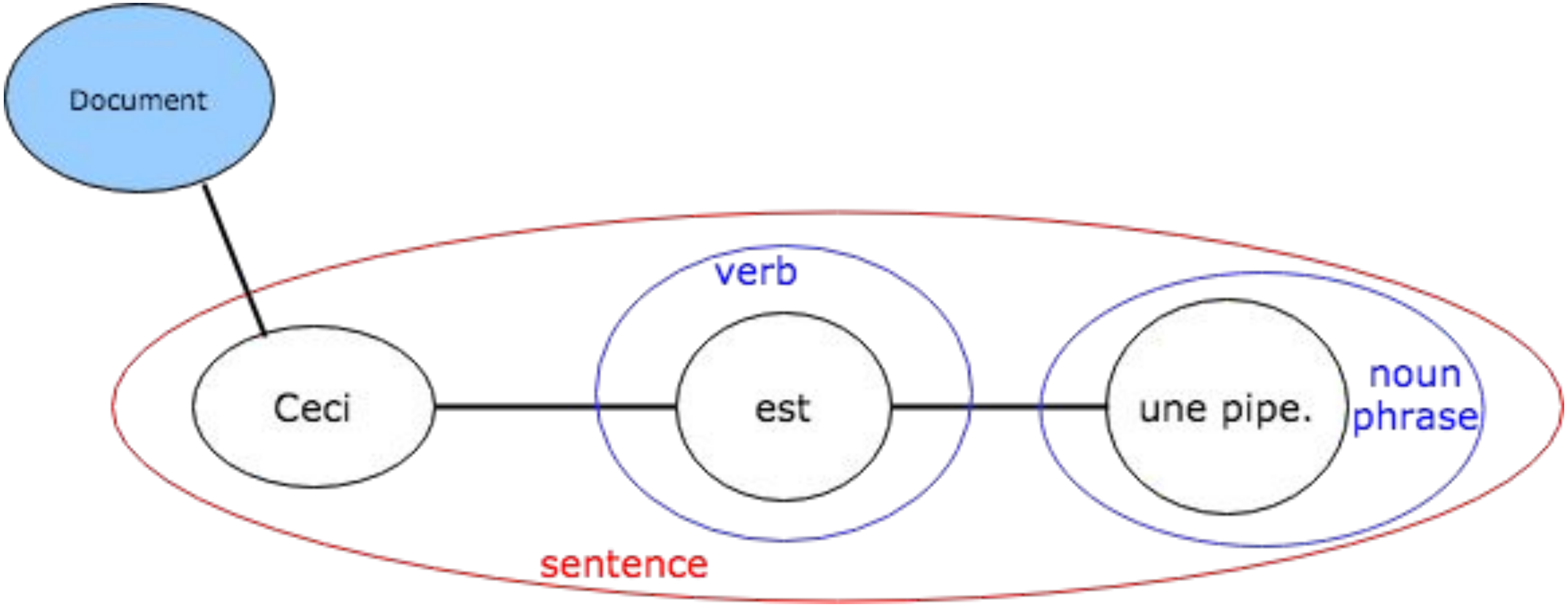
- Document Node
- Text Nodes
- Markup Nodes
- Annotation Nodes

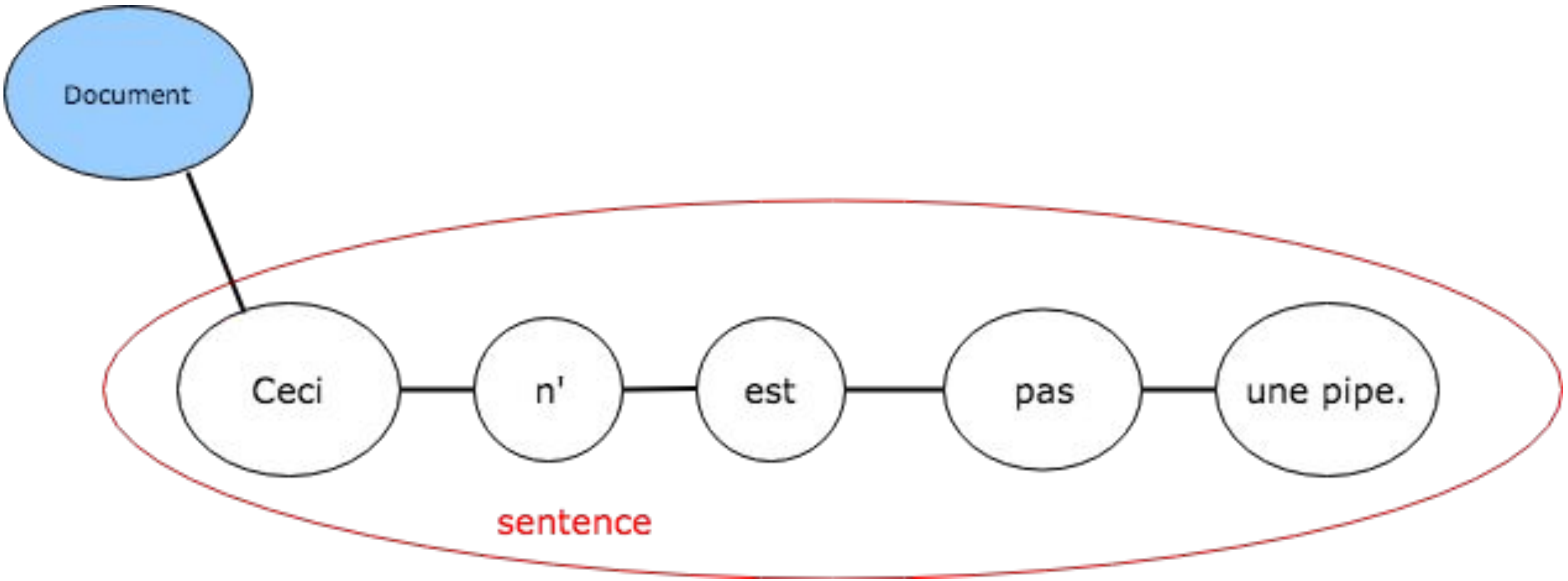


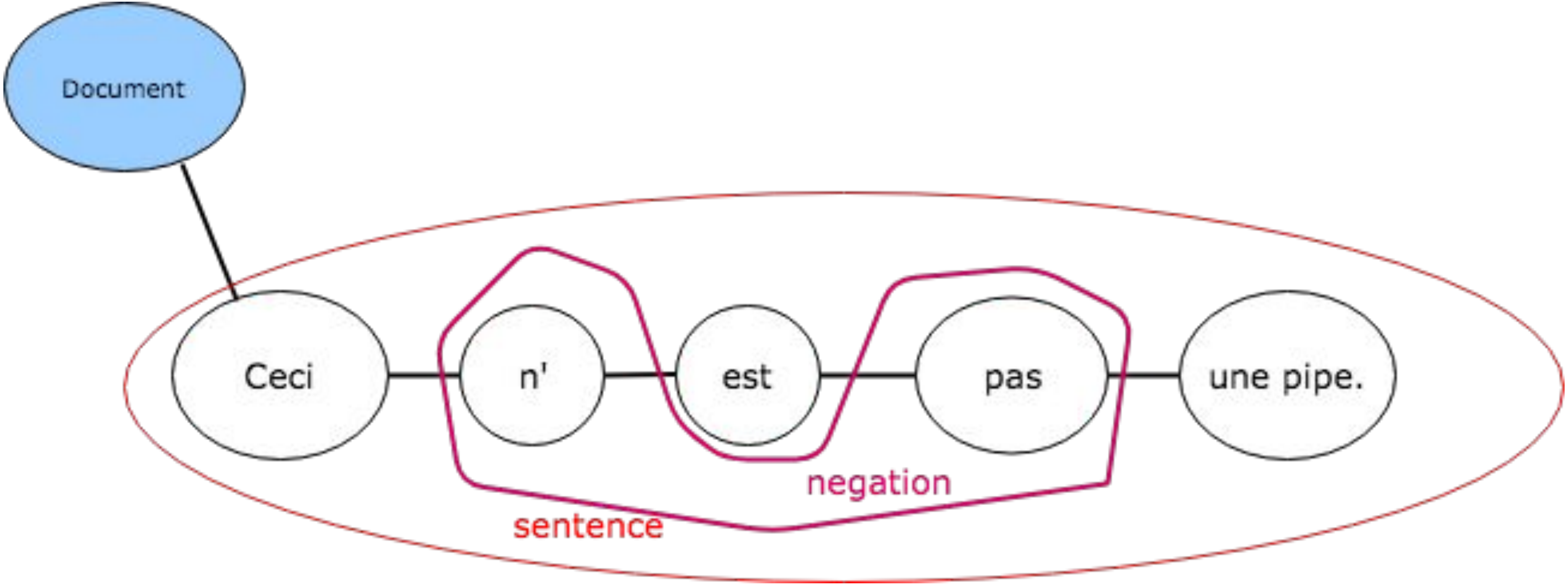
Markup as *sets* on text nodes

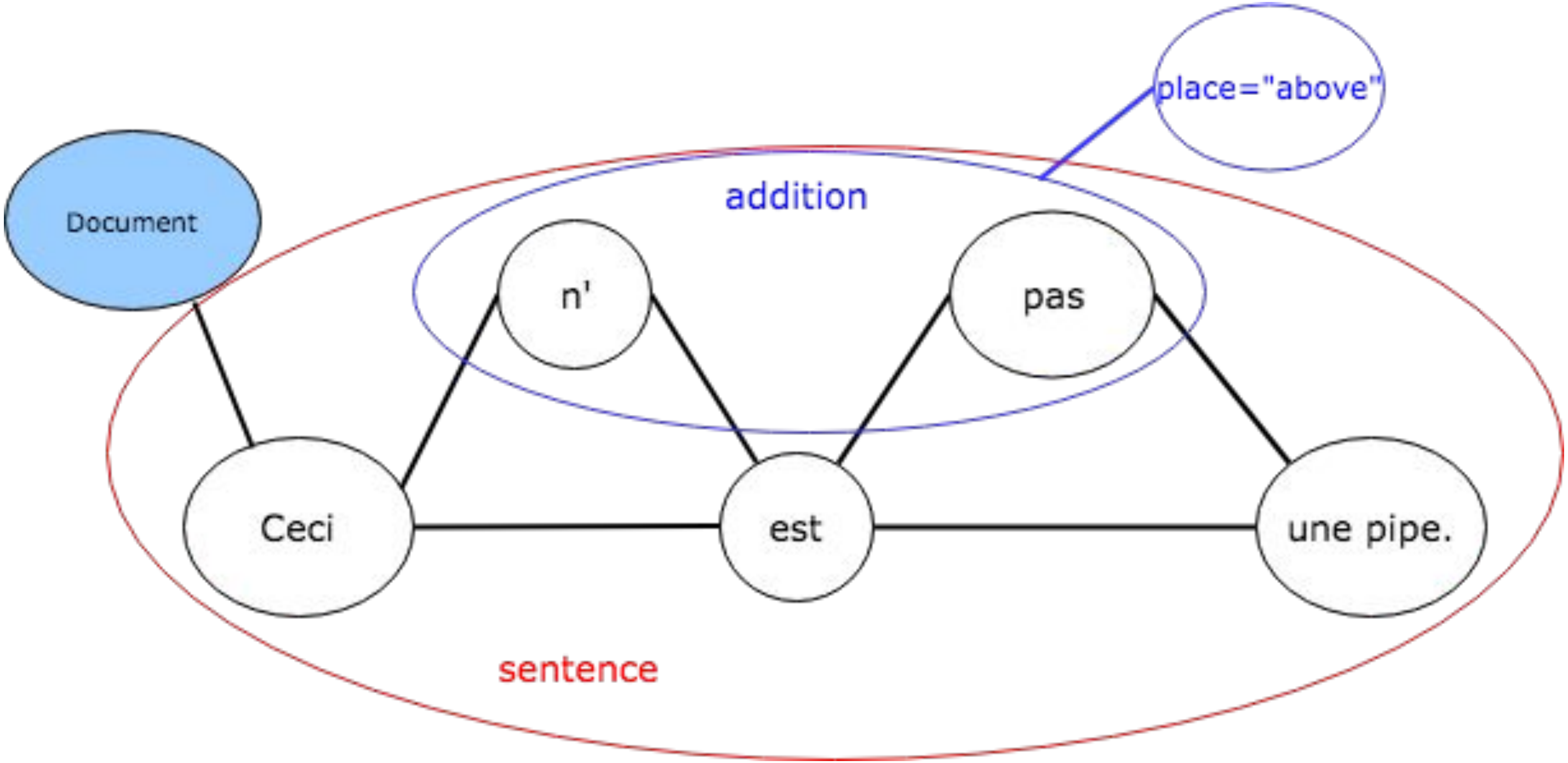

Document
Ceci
est
une pipe.


Document
Ceci
est
une pipe.
sentence


Document
verb
Ceci
est
une pipe.
noun phrase
sentence

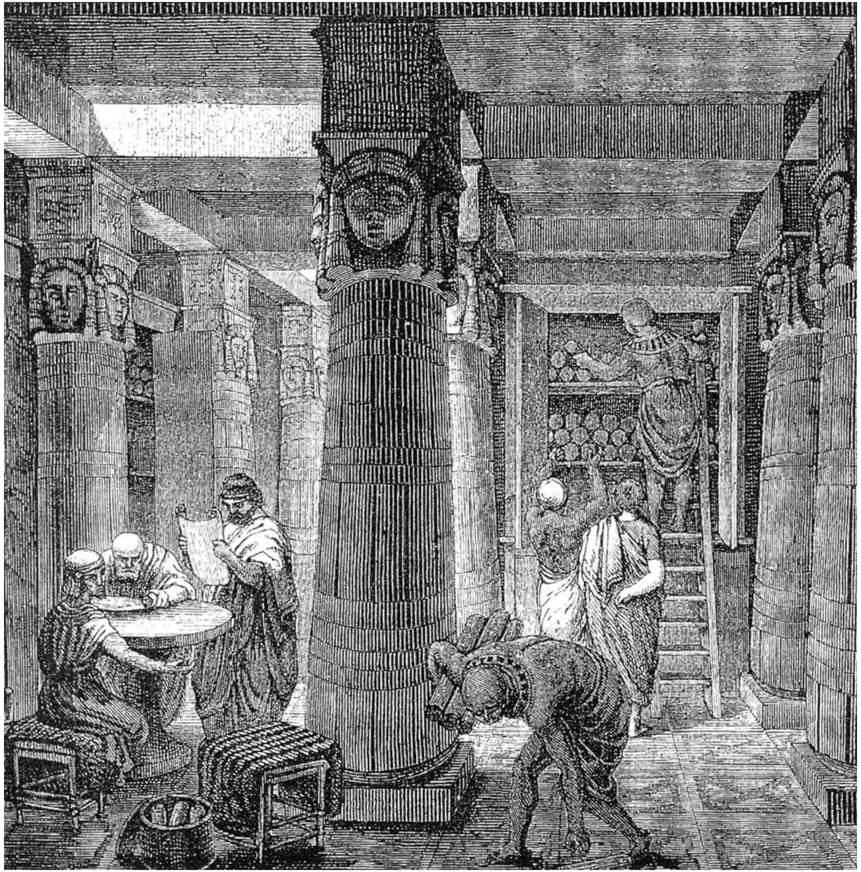

Document
Ceci
n'
est
pas
une pipe.
sentence


Document
Ceci
n'
est
pas
une pipe.
negation
sentence


place="above"
Document
addition
n'
pas
Ceci
est
une pipe.
sentence

Alexandria

Storing and querying TAG documents

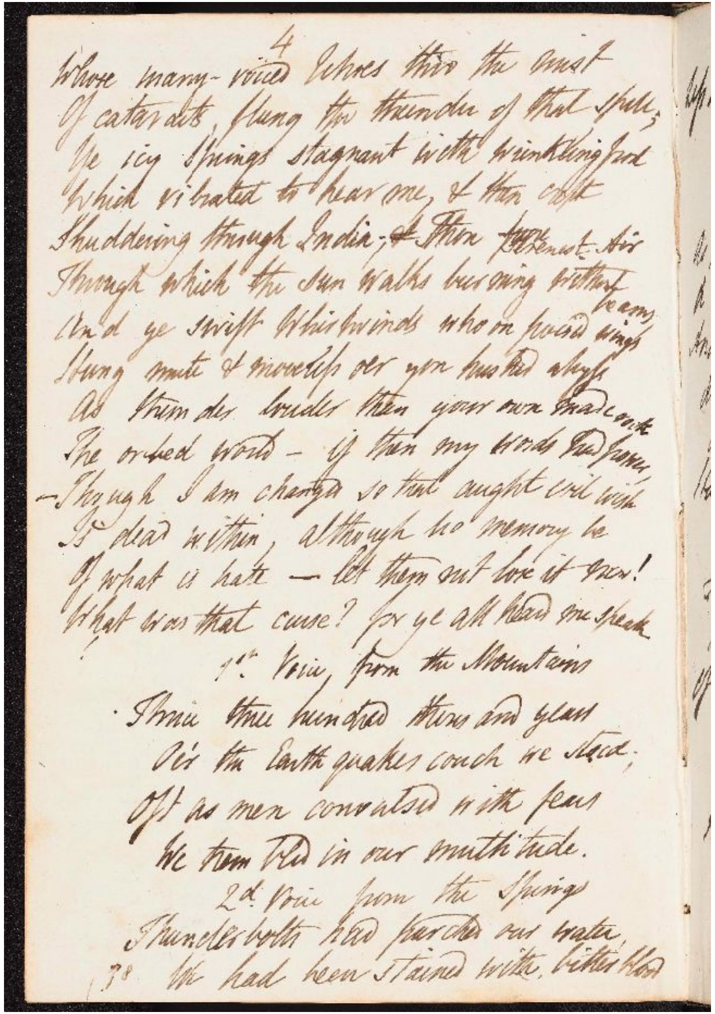


Analytical perspectives on textual objects

Perspective: a family of theories, methodologies, and analytical practices

- Dramatic
- Prosodic
- Discourse
- Syntactic
- ...

(Renear *et al*., 1993, "Refining Our Notion of What Text Really Is")



Whose many-voiced Echoes thro the mist
Of cataracts, flung the thunder of that spell;
Ye icy Springs stagnant with wrinkling frost
Which vibrated to hear me, & then crept
Shuddering through India; & Thou ~~pure~~ serenest Air
Through which the Sun walks burning without beams
And ye swift Whirlwinds who on poised wings
Hung mute & moveless oer yon hushed abyss
As thunder louder than your own made rock
The orbed world – if then my words had power
Though I am changed so that aught evil wish
Is dead within, although no memory be
Of what is hate – let them not lose it now!
What was that curse? for ye all heard me speak

1st Voice, from the Mountains

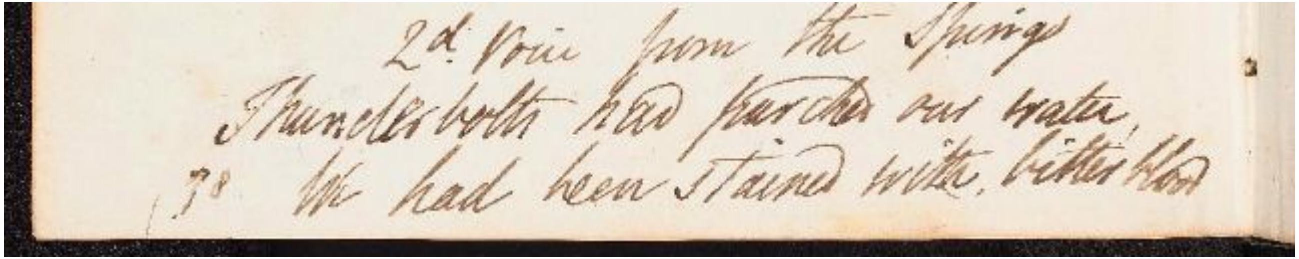
Thrice three hundred thousand years
O'er the Earthquakes couch we stood;
Oft as men convulsed with fears
We trembled in our multitude.

2d Voice from the Springs

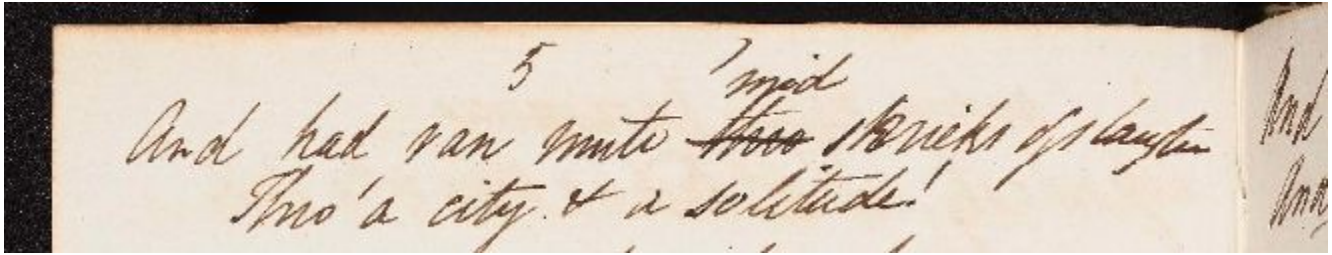
Thunderbolts had parched our water
We had been stained with bitter blood

- **Dramatic**: act, scene, speech, ...
- **Prosodic**: poem, verse, stanza, line, …
- **Material**: page, paragraph, line, ..
- **Discourse**: opening, topic, ending, ...
- **Syntactic**: sentence, clause, noun phrase, verbal phrase..

Alexandria - workflow



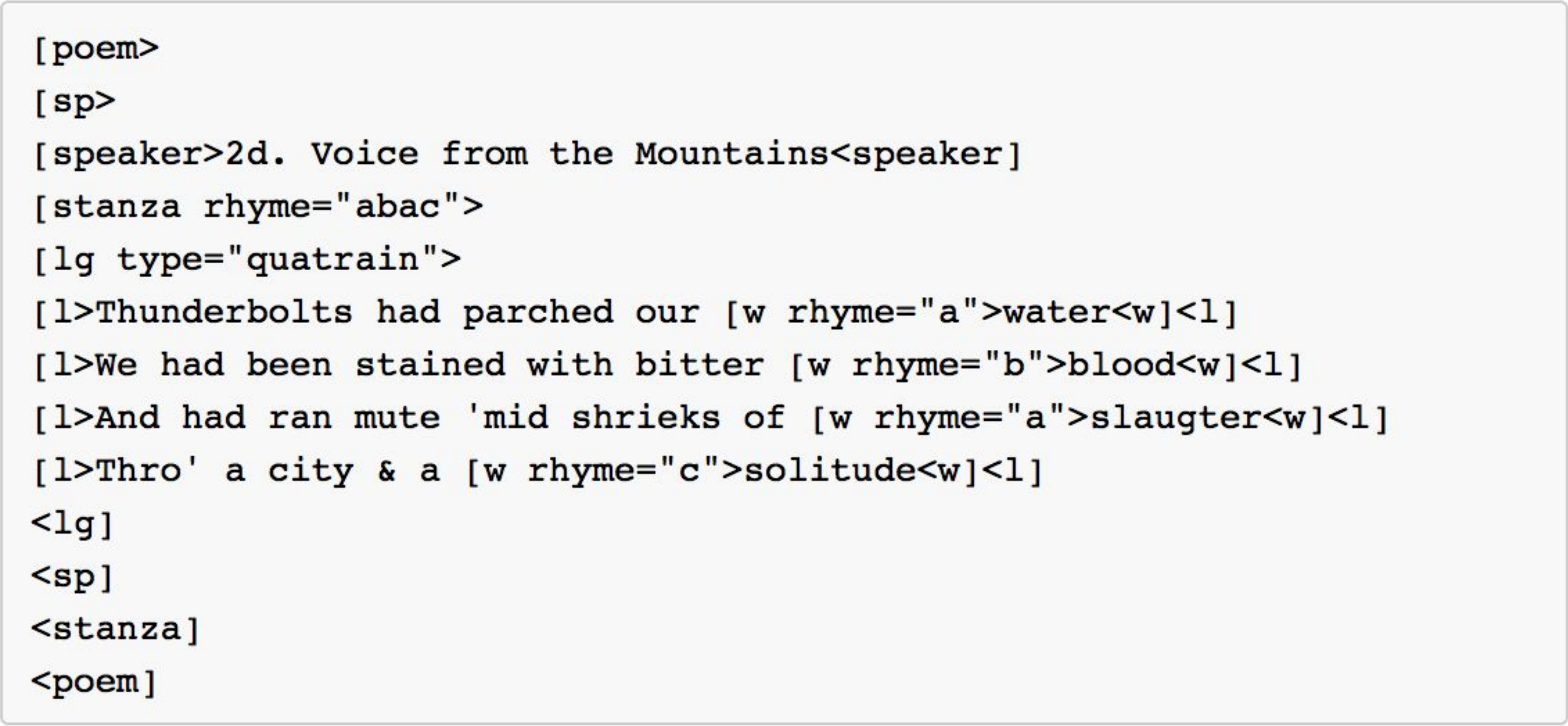
2d. Voice from the Springs
Thunderbolts had parched our water,
78 We had been stained with bitter blood



5
And had ran mute ~~thro~~ 'mid shrieks of slaughter
Thro' a city & a solitude!

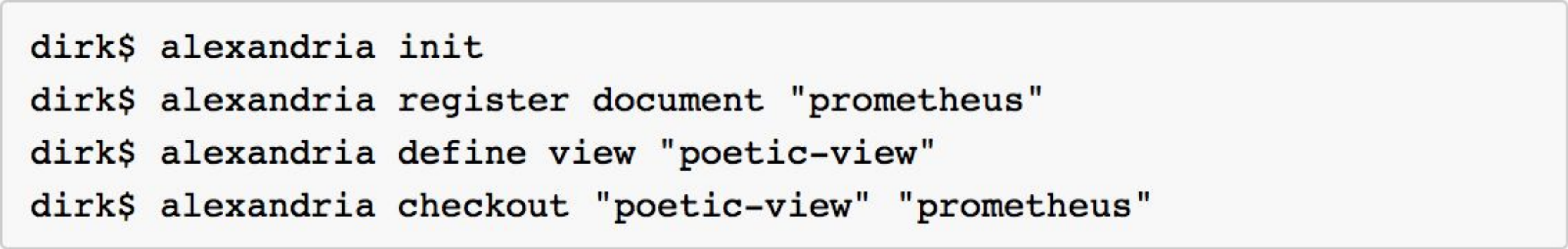
And
And

Alexandria - workflow



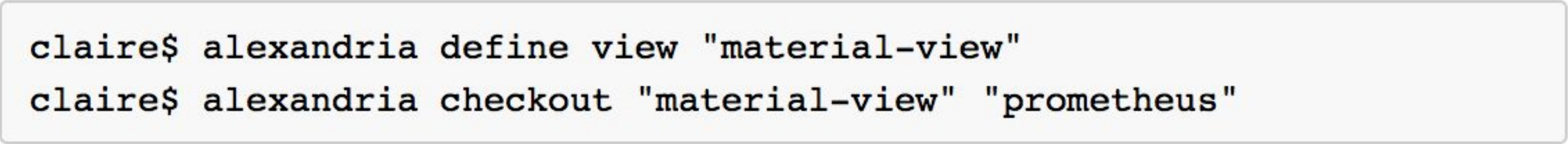
```
[poem>
[sp>
[speaker>2d. Voice from the Mountains<speaker]
[stanza rhyme="abac">
[lg type="quatrain">
[l>Thunderbolts had parched our [w rhyme="a">water<w]<l]
[l>We had been stained with bitter [w rhyme="b">blood<w]<l]
[l>And had ran mute 'mid shrieks of [w rhyme="a">slaugter<w]<l]
[l>Thro' a city & a [w rhyme="c">solitude<w]<l]
<lg]
<sp]
<stanza]
<poem]
```

Alexandria - workflow



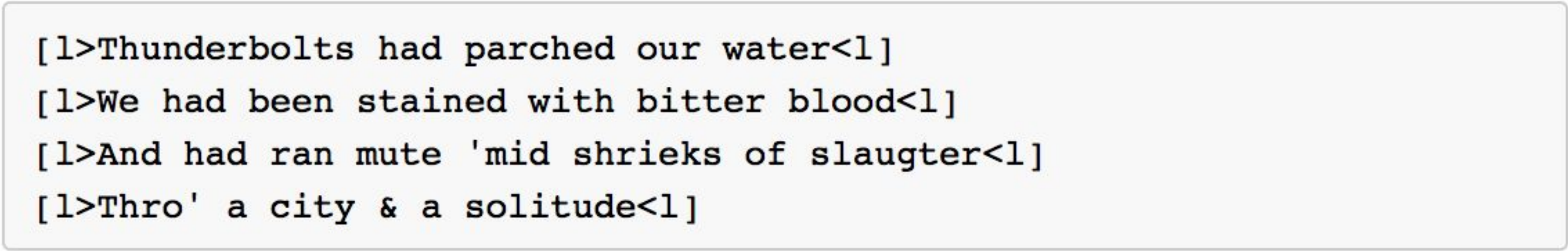
```
dirk$ alexandria init
dirk$ alexandria register document "prometheus"
dirk$ alexandria define view "poetic-view"
dirk$ alexandria checkout "poetic-view" "prometheus"
```

Alexandria - workflow



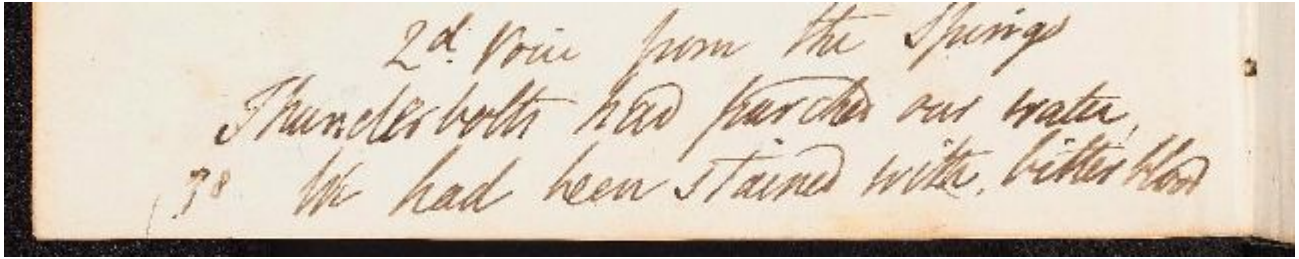
```
claire$ alexandria define view "material-view"
claire$ alexandria checkout "material-view" "prometheus"
```

Alexandria - workflow



```
[l>Thunderbolts had parched our water<l]
[l>We had been stained with bitter blood<l]
[l>And had ran mute 'mid shrieks of slaugter<l]
[l>Thro' a city & a solitude<l]
```

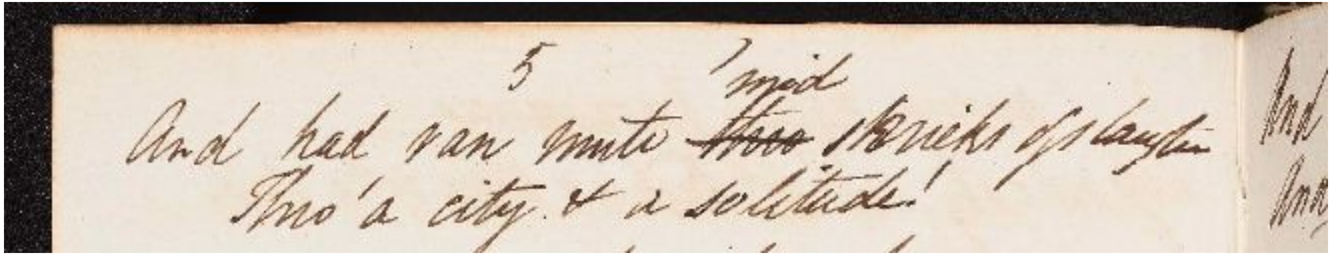
Alexandria - workflow



2d. Voice from the Springs

Thunderbolts had parched our water,

78 We had been stained with bitter blood

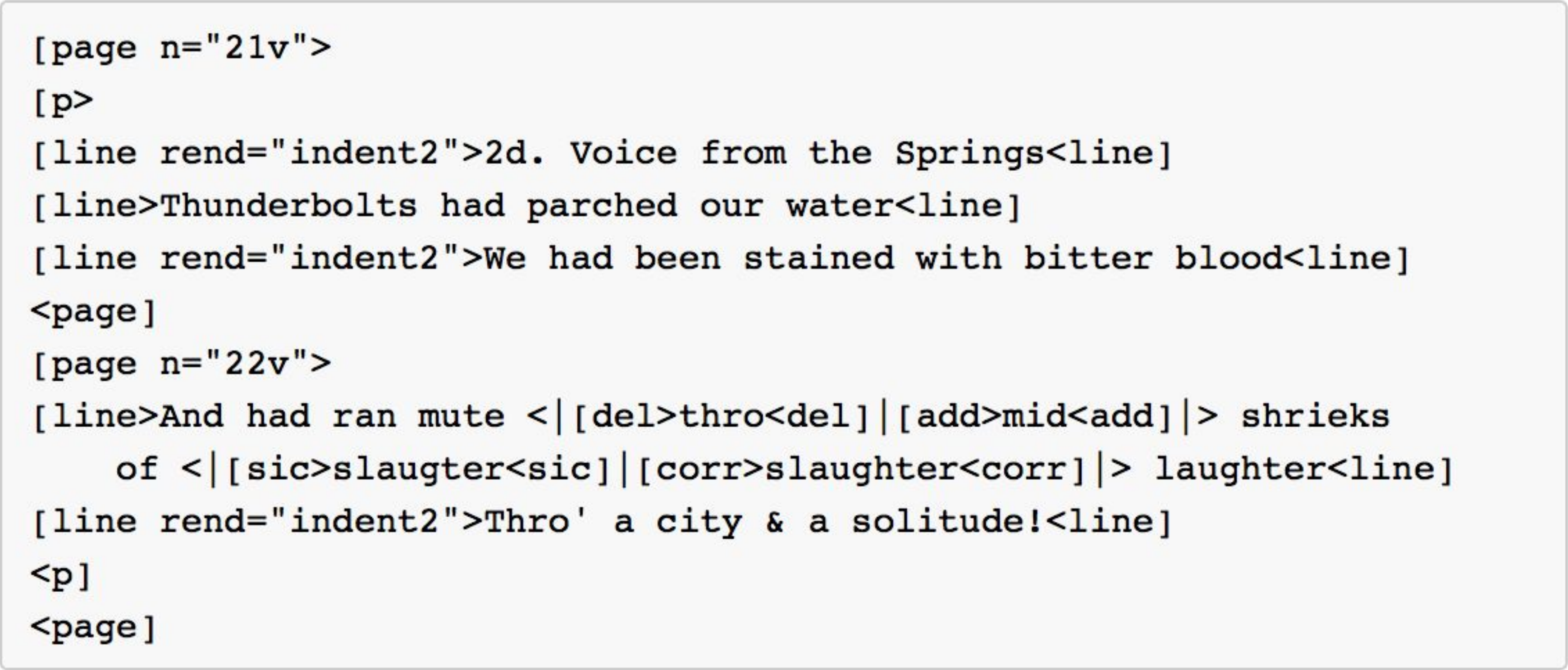


5

And had ran mute ~~thro~~ 'mid shrieks of slaughter

Thro' a city & a solitude!

Alexandria - workflow



```
[page n="21v">
[p>
[line rend="indent2">2d. Voice from the Springs<line]
[line>Thunderbolts had parched our water<line]
[line rend="indent2">We had been stained with bitter blood<line]
<page]
[page n="22v">
[line>And had ran mute <|[del>thro<del]|[add>mid<add]|> shrieks
    of <|[sic>slaugter<sic]|[corr>slaughter<corr]|> laughter<line]
[line rend="indent2">Thro' a city & a solitude!<line]
<p]
<page]
```

Discussion

Technical implications

Diffing TAG documents

Merging TAG documents

Conceptual implications

Multiple views on text

Textual awareness

Redefinition of "layer" of information

Inclusive workflow

Discussion

Technical implications

Diffing TAG documents

Merging TAG documents

Conceptual implications

Multiple views on text

Textual awareness

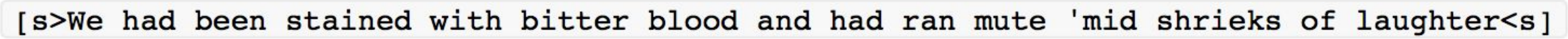
Redefinition of "layer" of information

Inclusive workflow

Alexandria - technical implications

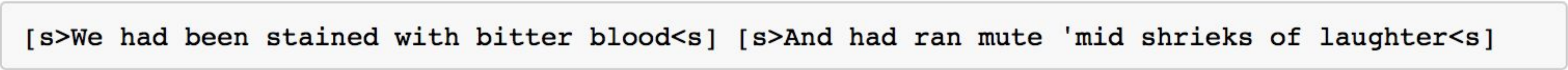
Diffing TAG documents

Dirk:



```
[s>We had been stained with bitter blood and had ran mute 'mid shrieks of laughter<s]
```

Claire:



```
[s>We had been stained with bitter blood<s] [s>And had ran mute 'mid shrieks of laughter<s]
```

Edit operations on text: deletion, addition, replacement

Edit operations on markup: deletion, addition, replacement, split, join

Alexandria - technical implications

Merging TAG documents (views and layers)

Views can contain multiple layers

In each layer, information is hierarchically structured

TAG supports the overlapping structures

Discussion

Technical implications

Diffing TAG documents

Merging TAG documents

Conceptual implications

Multiple views on text

Textual awareness

Redefinition of "layer" of information

Inclusive workflow

TAG

understanding texts as hypergraphs

Elli Bleeker
Bram Buitendijk
Ronald Haentjens Dekker
Astrid Kulsdom

R&D Humanities Cluster
Royal Science Academy of the Netherlands