

Between Freedom and Formalisation

A hypergraph model for representing the nature of text

Elli Bleeker, Bram Buitendijk, Ronald Haentjens Dekker
R&D group, Humanities Cluster
KNAW Amsterdam

TEI 2019 Conference
September 16-21, 2019
University of Graz, Austria



@ellibleeker
@bram_buitendijk
@ronald_dekker

Introduction

“We have paper minds dealing with electronic realities”

(Terry Cook 1994)

“Textual structures predate digitally representable information collections”

(Fabio Vitali 2016)

Introduction

Data models to express textual information

- **Plain text** (string)
- **CSV** (tabular data in plain text)
- **MS Word** or **Open Office**
- **JSON** (key:value pairs)
- **XML** (hierarchical tree structure)
- **RDF** (statements as triples *subject-predicate-object*)
- **TAG** (hypergraph)

<i>with handovers & workarounds</i>	Data	Text	Hierarchies	Presentation	Validation	References	Annotations	Overlapping
CSV								
JSON								
RDF								
Markdown								
HTML								
HTML+RDFa								
XML								
Overlapping fmts								

Source: Vitali 2016 (<https://bit.ly/2jWm96t>)

*with handovers
& workarounds
& some coding*

Data

Text

Hierarchies

Presentation

Validation

References

Annotations

Overlapping

CSV

JSON

RDF

Markdown

HTML

HTML+RDFa

XML

Overlapping fomats

Introduction

Modeling text in a data model that is in close agreement with the kind of text, the scholar's orientation, and the research objectives

Overview

1. The TAG model for text
 - a. Definition of text
 - b. TAGML
2. Textual genetic research and digital editing
 - a. Research objectives
 - b. Challenges
 - i. Complex textual features
 - ii. Text mode and document mode
3. Textual genetic editing in TAG
- 4-5. Recap and discussion
 - a. Modeling textual genetic information in TAG
 - b. Future work

1. TAG - definition of text

Text is

“a multi-layered, non-linear object containing information that is at times unordered, fully ordered, and partially ordered”

1. TAG - markup language

1. [**tag**> text <**tag**]

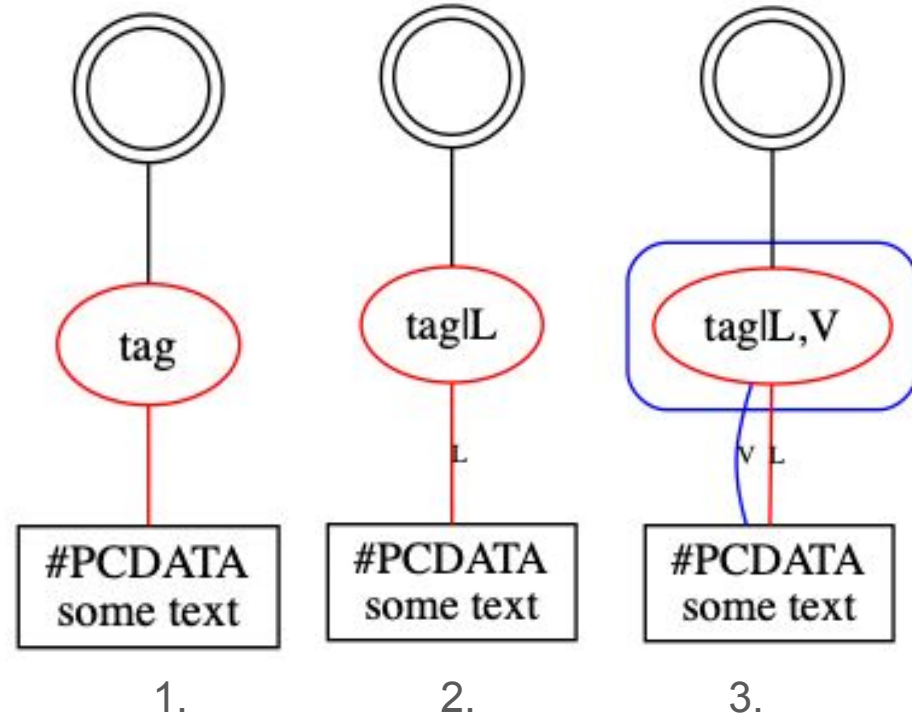
The TAGML node is in the default layer

2. [**tag**|+**L**> some text <**tag**]

The TAG node is in the L layer

3. [**tag**|+**L**,+**V**> some text <**tag**]

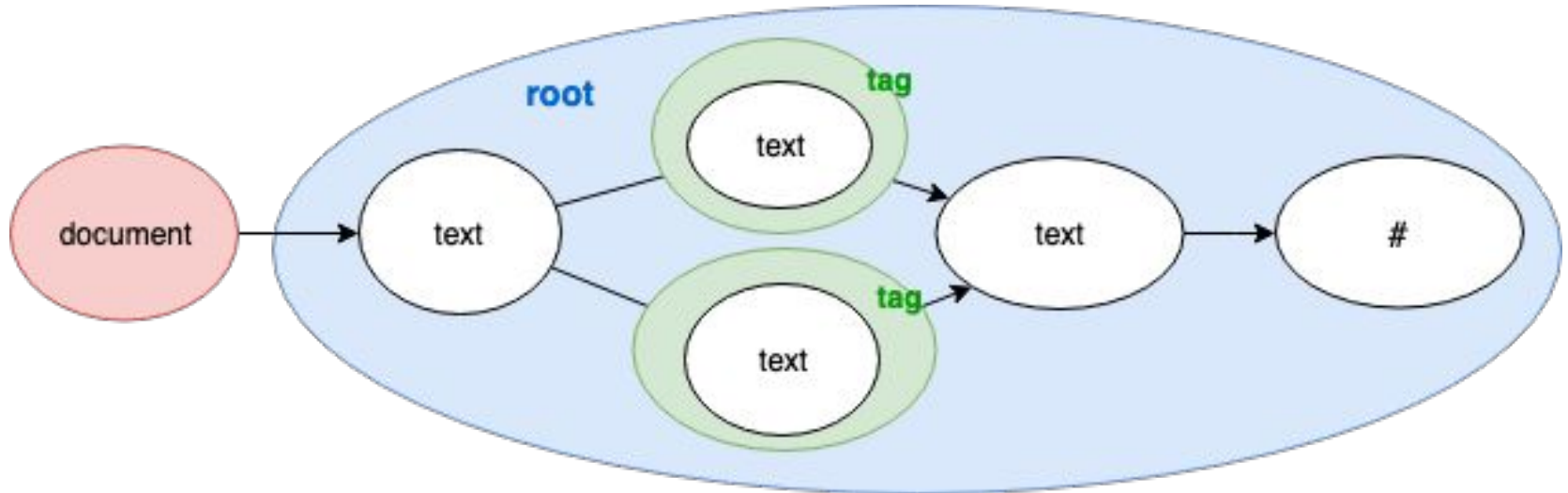
The TAG node is in the L and V layers



TAG - markup language

Branches

`[root>text <|[tag> text <tag]| [tag> text <tag]|> text <root]`



1. TAG - data model

- Text nodes, both ordered and partially ordered
- Markup as annotations on that text, grouped in layers
- Expressive and rich syntax (strings, boolean, integers/float, lists, nested annotations, ...)

An intuitive model for genetic text encoding?

2. Genetic criticism and digital editing

Research objectives:

- The production process of text (vs. its “final” state)
- Displaying multiple dimensions (textual, temporal, material, ...)
- Reconstruct the creative writing process (endo-, exo- and epigenesis)

Research requirements:

- A model to express nonlinear, non-hierarchical textual features
- A model that allows for multiple, co-existing structures

2. Genetic criticism and digital editing

Challenges

1. Complex textual features
 - 1.1. Non-linear text
 - 1.2. Discontinuous text
 - 1.3. Text fragments
2. Multiple overlapping structures
 - 2.1. Textual mode
 - 2.2. Documentary mode
 - 2.3. Temporal mode

3. Genetic editing in TAG

Challenges

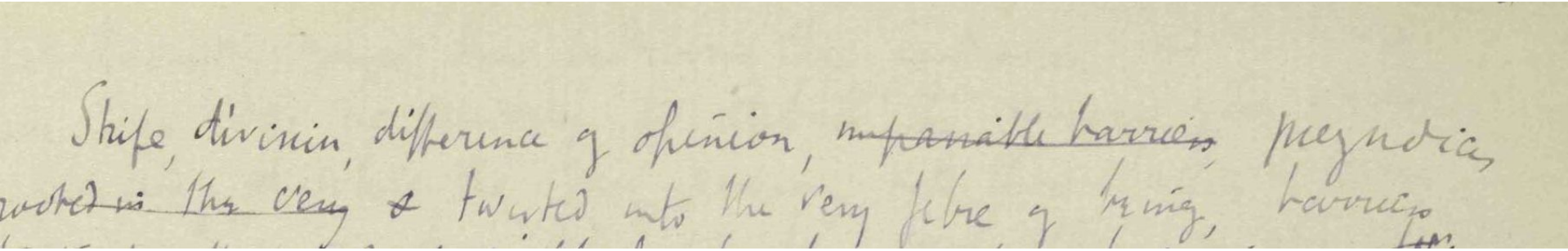
1. Complex textual features
 - 1.1. **Non-linear text**
 - 1.2. **Discontinuous text**
 - 1.3. Text fragments
2. Multiple overlapping structures
 - 2.1. **Textual mode**
 - 2.2. **Documentary mode**
 - 2.3. Temporal mode

3. Genetic editing in TAG

Non-linear text

- Ex. 1: single deletion/addition
- Ex. 2: revision grouped with <subst>
- Ex. 3: immediate revision (“*currente calamo*”)
- Ex. 4: open variants

3. Single deletion



Shife, divinin, difference of opinion, ~~unpardonable~~ barriers, meynodias,
rooted in the very & twined into the very fibre of being, barriers

(Source: Woolf, Virginia. *To the Lighthouse*. Holograph ms. Berg Collection. New York Public Library. Woolf Online. Ed. Pamela L. Caughie, Nick Hayward, Mark Hussey, Peter Shillingsburg, and George K. Thiruvathukal. Web. 16 September 2019. <<http://www.woolfonline.com>>)

3. Single deletion

XML TEI P5

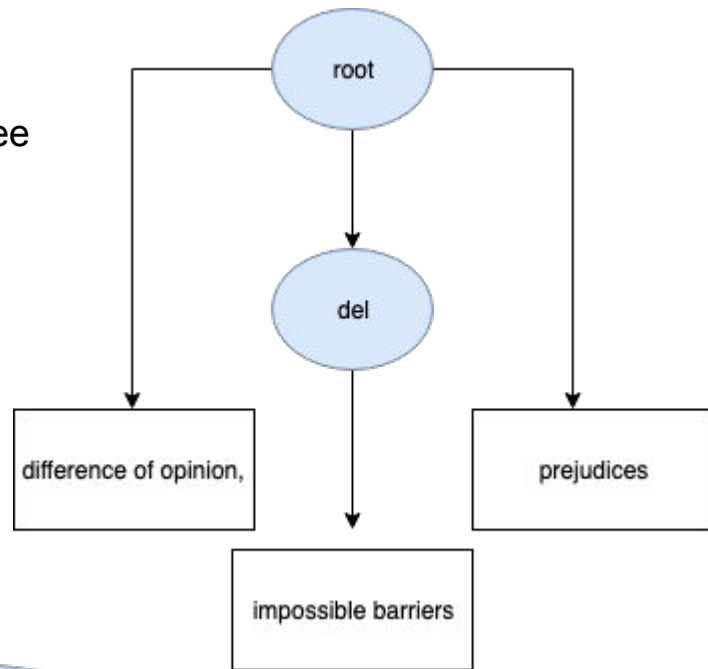
`<root>` difference of opinion,
``impossible barriers``
prejudices `</root>`

TAGML

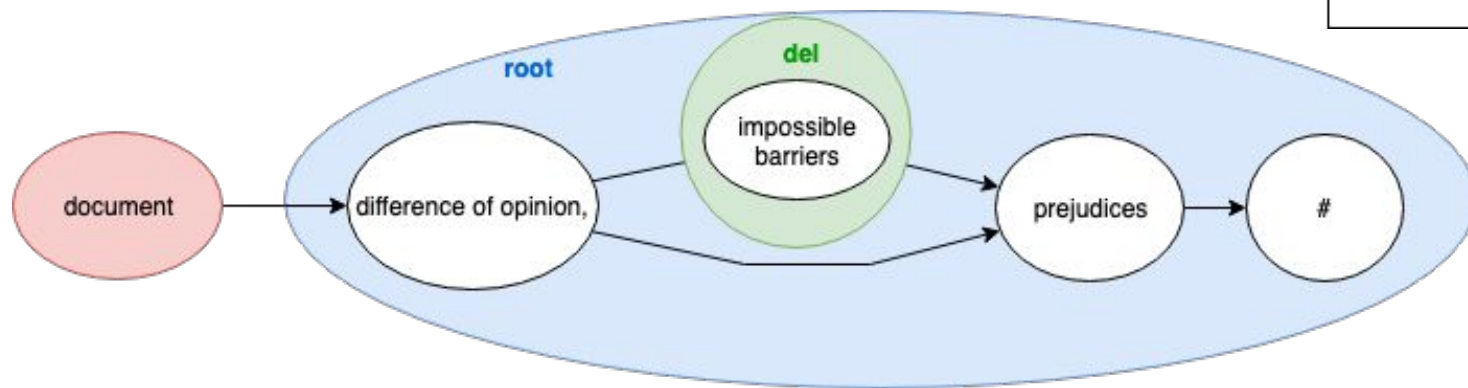
`[root>` difference of opinion,
`[?del>`impossible barriers`<?del]`
prejudices `<root]`

3. Single deletion

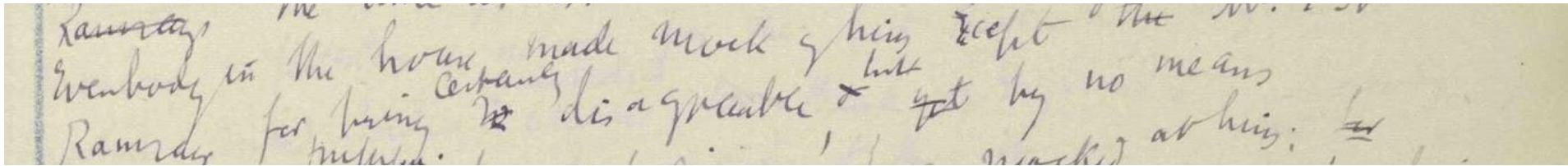
XML tree



TAGML hypergraph



3. Grouped revision



A snippet of a handwritten manuscript on aged paper. The text is written in cursive. The first line reads 'Ramsey, the ...'. The second line reads 'Everybody in the house made mock of him except the ...'. The third line reads 'Ramsey for being ~~so~~ ^{certainly} disagreeable & yet by no means ...'. The word 'so' is crossed out with a horizontal line, and 'certainly' is written above it. The word 'yet' is also crossed out with a horizontal line.

... for being ~~so~~ ^{certainly} disagreeable ...

3. Grouped revision

XML TEI P5: use <subst> or <mod>

```
<root>
```

```
... for being <subst> <del> so  
</del> <add> certainly </add>  
</subst> disagreeable ...
```

```
</root>
```

TAGML: branches

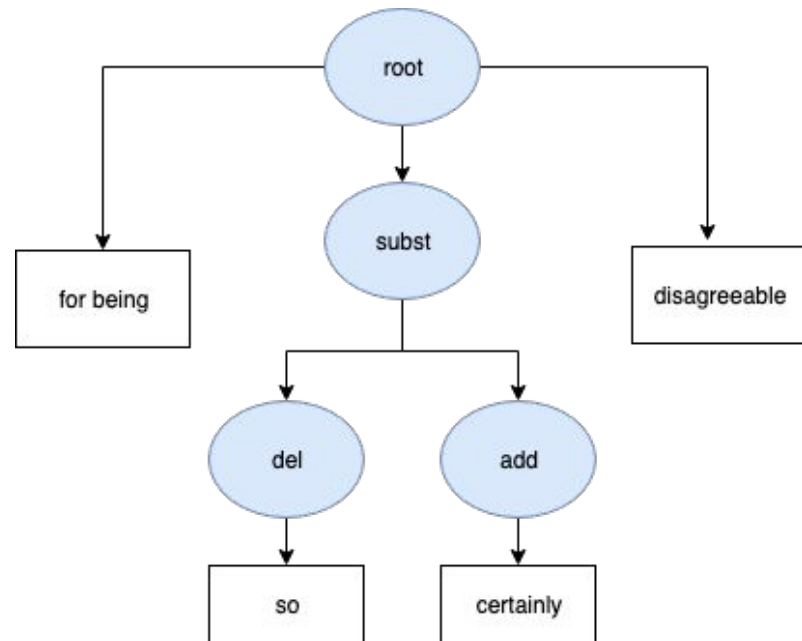
```
[root>
```

```
... for being  
<| [del>so<del] | [add>certainly<add] |>  
disagreeable
```

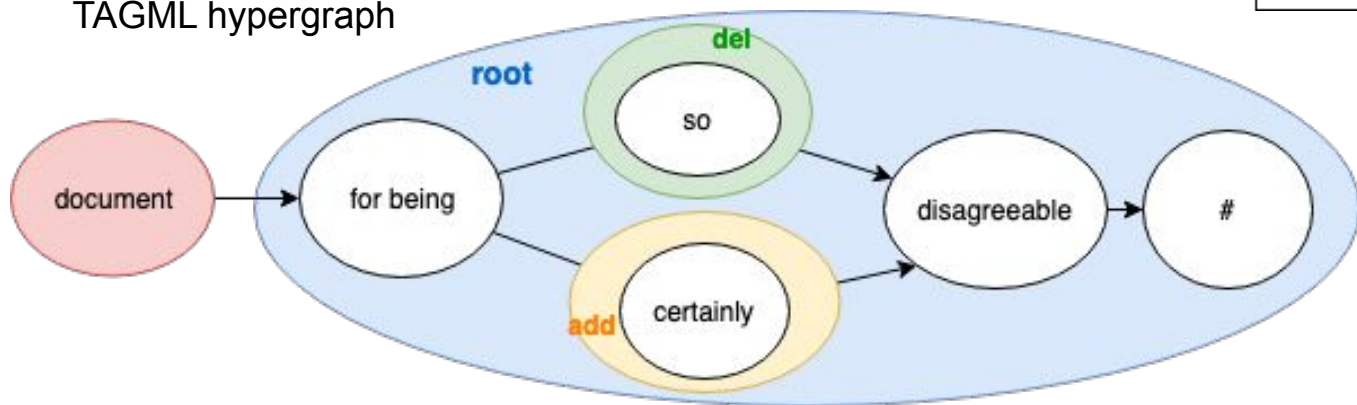
```
<root]
```

3. Grouped revision

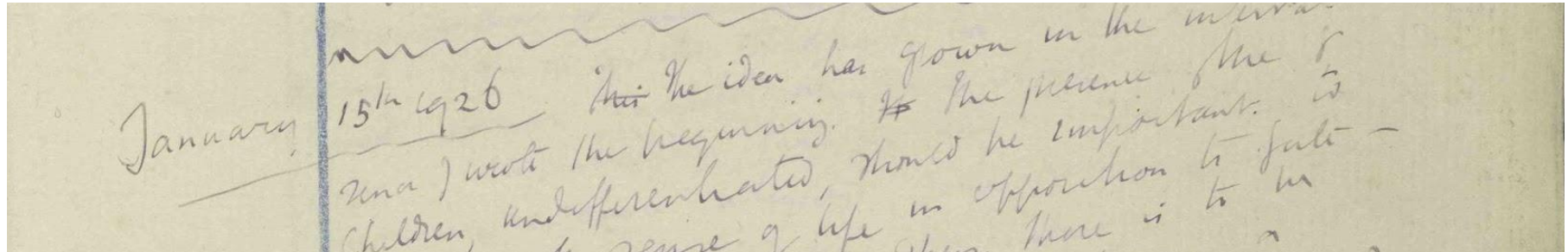
XML tree



TAGML hypergraph



3. Immediate revision



A photograph of a handwritten manuscript on aged, yellowed paper. The text is written in cursive ink. On the left, the date 'January 15th 1926' is written. To its right, the sentence 'This The idea has grown in the interval' is written. A horizontal line is drawn under the word 'This'. Below this line, the word 'And' is written, followed by 'I wrote the beginning. The presence of the 6 children, and indifferent, should be important. is a sense of life in opposition to fate - there is to be'.

January 15th 1926 This The idea has grown in the interval
And I wrote the beginning. The presence of the 6
children, and indifferent, should be important. is
a sense of life in opposition to fate - there is to be

January 15th 1926 ~~This~~ The idea has grown ...

3. Immediate revision

XML TEI P5

`<root>`

`<del seq="0">This` The idea
has grown ...

`</root>`

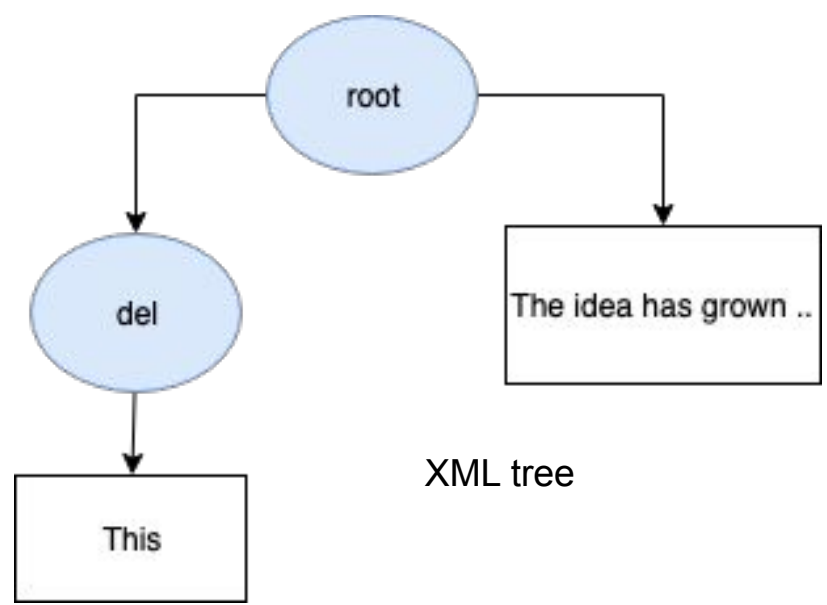
TAGML

`[root>`

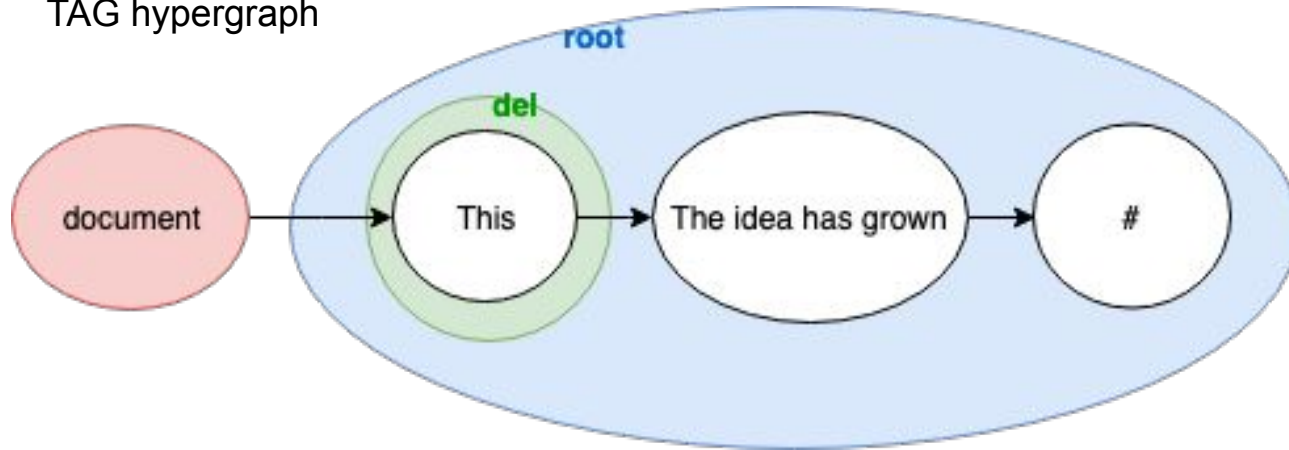
`[del>This<del]` The idea has grown

`<root]`

3. Immediate revision



TAG hypergraph



3. Open variants

That is ^{soon said} an easy thing to say.

Source: the BDMP encoding manual (<<http://uahost.uantwerpen.be/bdmp/>>, accessed 16 September 2019)

3. Open variants

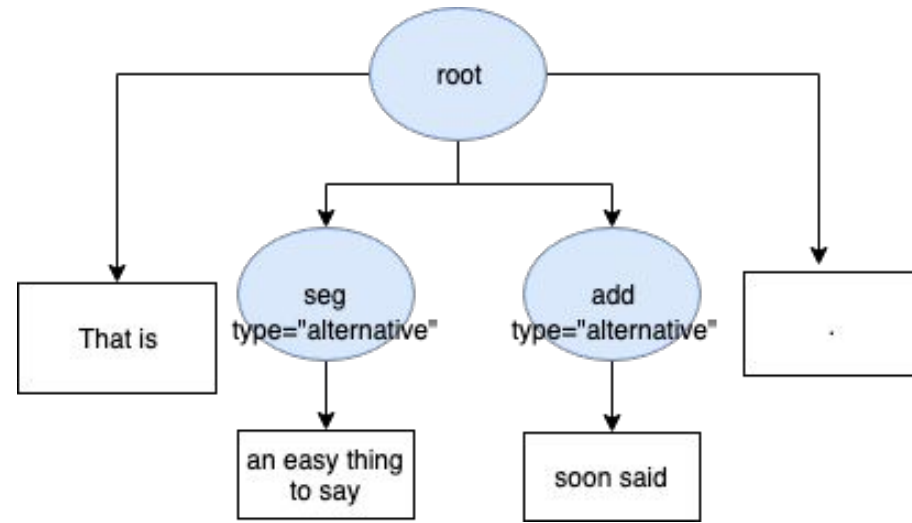
XML TEI P5

```
<root> That is  
  
<seg type="alternative"  
xml:id="alt1">an easy thing to  
say</seg> <add  
place="supralinear" xml:id="alt2"  
type="alternative">soon  
said</add>.  
  
</root>
```

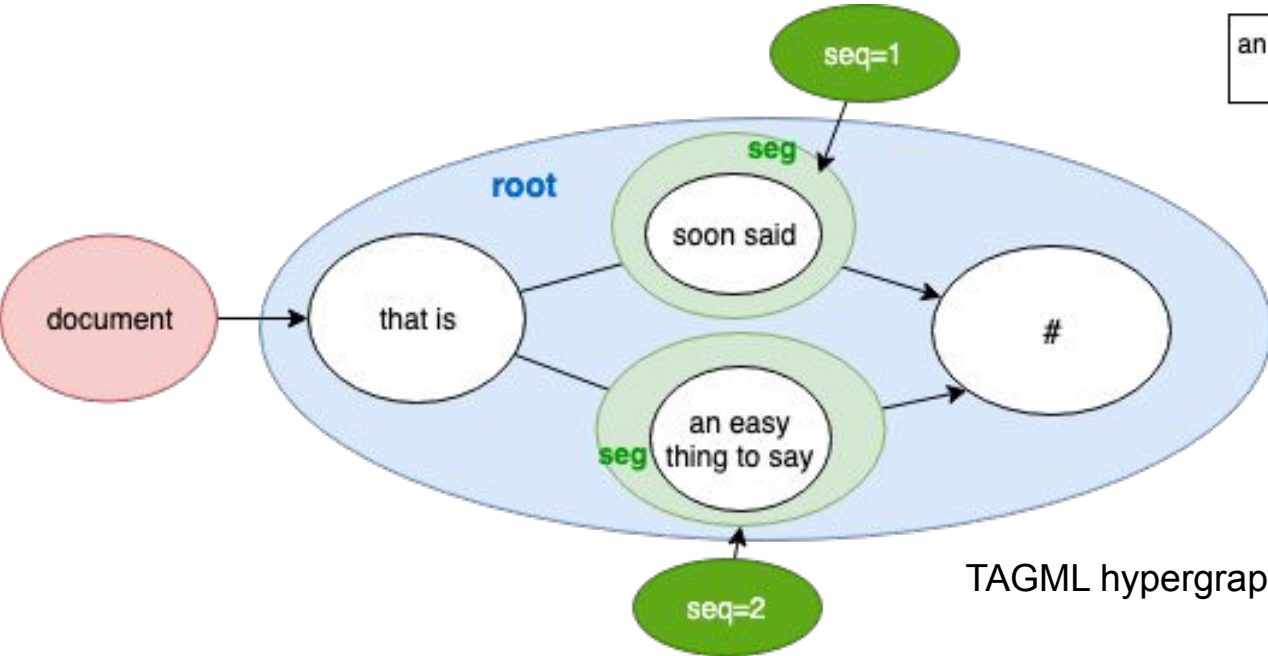
TAGML: branches

```
[root> That is  
  
<|[seg seq=1>an easy thing to  
say <seg]|[seg seq=2>soon  
said<seg]|>.  
  
<root]
```

3. Open variants

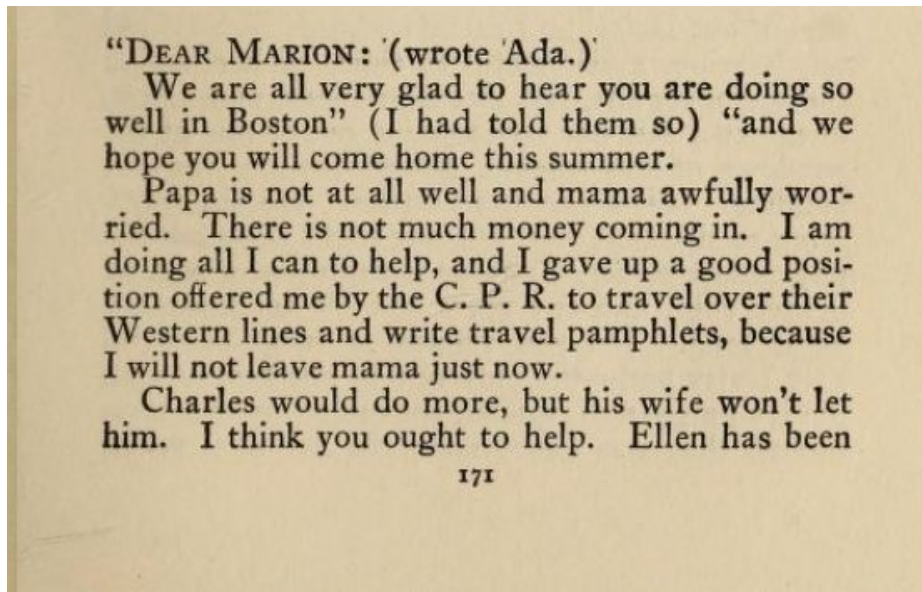


XML tree



TAGML hypergraph

3. Discontinuous text



“Dear Marion: (wrote Ada.) We are all very glad to hear ...”

3. Discontinuous text

XML TEI P5

```
<root>
```

```
<q xml:id="q1a"  
next="#q1b">"Dear  
Marion:</q> (wrote Ada.)
```

```
<q xml:id="q1b"  
prev="#q1a"> We are all  
very glad... </q>
```

```
</root>
```

TAGML

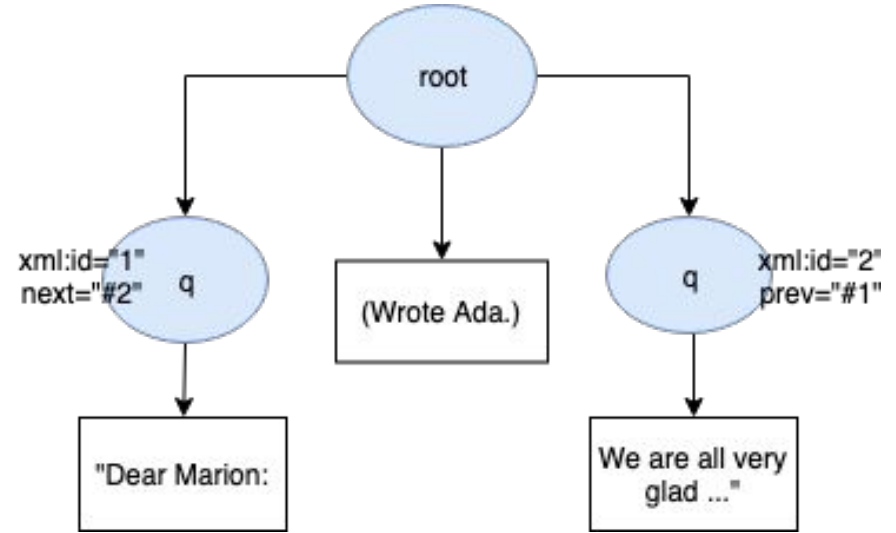
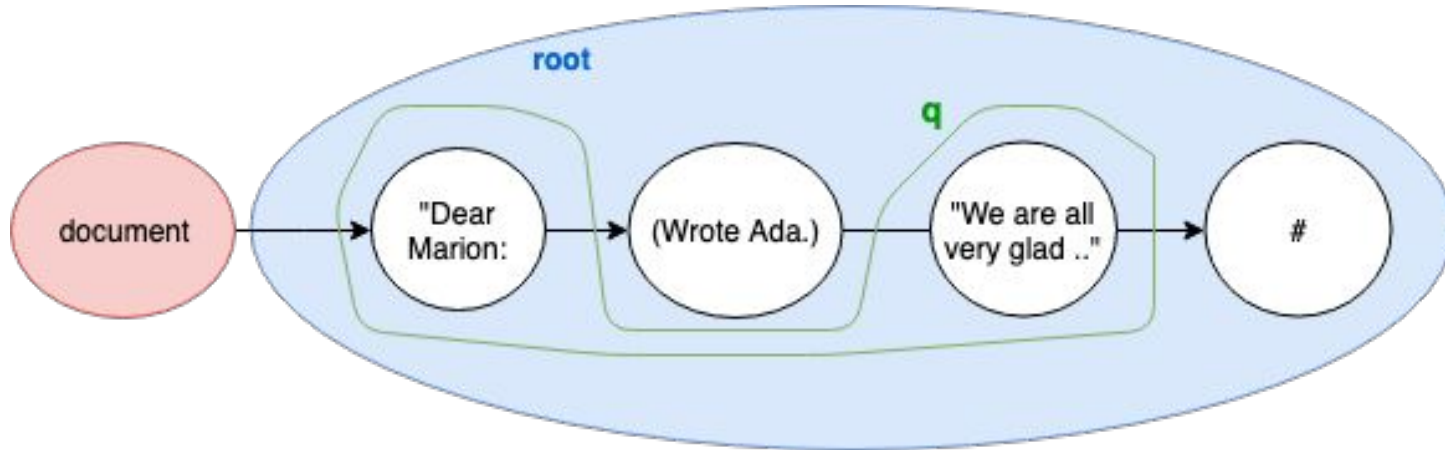
```
[root>
```

```
[q>"Dear Marion <-q] (wrote  
Ada.) [+q> We are all very  
glad..."<q]
```

```
<root]
```

3. Discontinuous text

TAG hypergraph



XML tree

3. Discontinuous text

het keizerlijk huis geleverd werden.

In den namiddag, wanneer de felste hitte voorbij was, kwam door ~~de~~ de luistille gangen van het paleis de uitverkorene, klopte bescheiden op de deur van de kamer waar de sultan ^{zijn} verveling geuwde. Zoals de mannequin in een Parijsch modewis den naam van het door haar gedragen toilet aan den klant opgeeft: "Afternoon Tea", "Blue Girl", "Un Caprice", of iets dergelijks, zoo zegde ze met in de stem de lichte beving der actrice die in de rats zit: "Koffe Verrassing", "Bloesembot" of "voor uw genoegen" - natuurlijk in het Arabisch.

Menigvuldig en soms wel van een frissche fantasie getuigend waren de benamingen die het "Keizerlijk College der Wellustcommissarissen" ^{afkond} Doch wat ze aankondigden kwam in den grond toch steeds op hetzelfde neer. Aldus was Shiriar mettertijd ook dit officieel geregeld minnespel beu geworden.

Inderdaad, wat is liefde die niet voorafgegaan is door verleiding? Helas hij, Shiriar, afstammeling der Profeten, krijgt den wellust voorgeschoteld als een dagelijkschen kost, waarvoor hij dan nog geen atjeid verricht heeft. Hoe zoet moet nochtans zijn dit ^{gelede van} ~~verwen~~ een vrouw! Die dagen van nerveuse verwachting vóór de liefelijke toetsterming.

Shiriar denkt aan de proletenvrijagies in de steegjes van Bagdad en op de hellingen der wallen. En dan die herinnering aan verleden zomer, toen hij in

- 3 -

geselschap van den vizier van Landbouw een inspectietocht deed in de streek van den Euphrates. Een ambtenaar-~~een~~ agronoom leerde hem juist ~~het verzen van de oude Arabische dichters~~ tarwe onderscheiden van rogge en haver als ze plots een jonge boer en boerin ~~in~~ ontdekten die zich in het koren liefkoosden. Als een opgeschrokken koppel patrijzen waren ze voor de statige heeren op de vlucht geslagen. De sultan had nog niet gezien hoe, al loopende, de blozende deerne de zware maakte ~~aan~~ borsten weer in haar jak wogstopte.

Waarachtig die hart ^{het leugen het minde} liefde in deze kamer van kussens en boeken kon den vorst min-en-min bekoren en, als een kind dat zijn zoetheid verwend is, begon hij de edaliken onverrichter zake terug te sturen. De aldus geaffronteerde meisjes slaakten bittere klachten en het keizerlijk College der ^{heeren} ~~wellust~~ commissarissen vroeg zich af hoe ze den sultan weer op het pad der traditioneele keizerlijke ~~borsten~~ zouden brengen.

Zij lieten exotische vrouwen komen: Spaansche ^{cigaretten} meisjes uit ~~cigarettenfabrieken~~ waarvan de huid okerblond is als de tabak die ze verwerken, Eoerache Walkirengestalten, schoon en koel als marmerbeelden, Geishas uit ^{de} Yoshiwara die een heele documentatie van technisch voortreffelijk erotische houtsneden meebrachten. Het kon al niet veel baten.

De ~~wellust~~ commissarissen nodigden de jongste "Miss Universe" uit. Doch daar ze vooreerst nog de

- 4 -

3. Discontinuous text

XML TEI P5

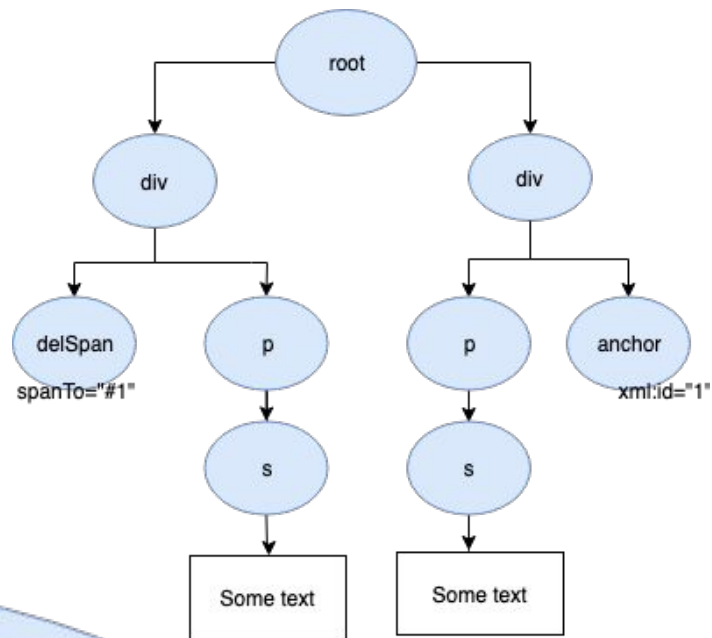
```
<root>
<div type="page">
<delSpan spanTo="#tsq_05r"/>
  <p>
    <s> some text </s>
  <p>
</div>
<div type="page">
  <p>
    <s> some text </s>
  </p>
<anchor xml:id="tsq_05r"/>
</div>
</root>
```

TAGML

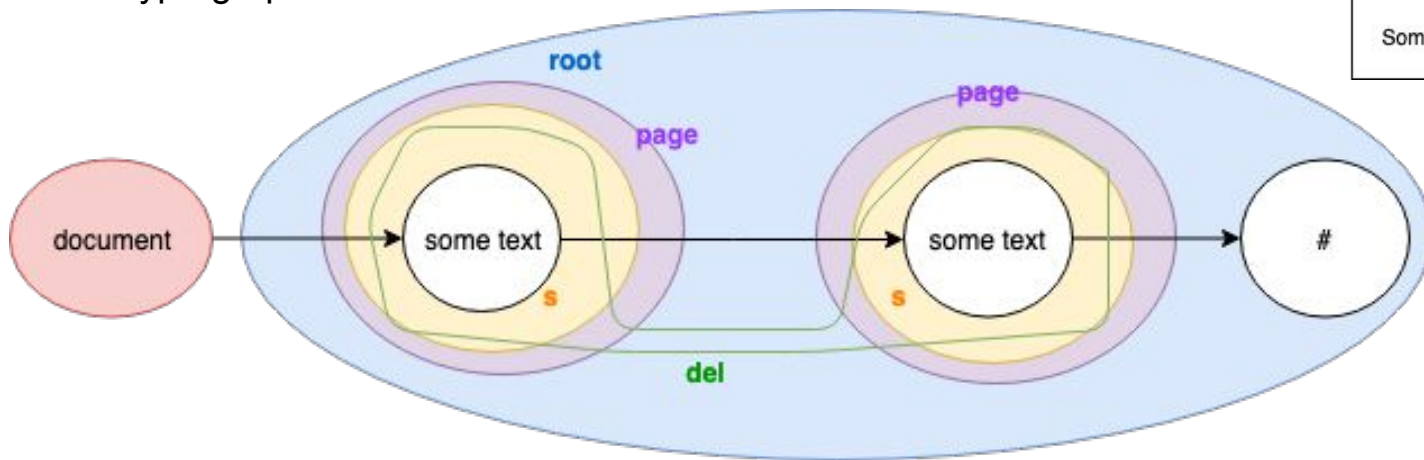
```
[root|+T,+D>
[page|D>
  [del|T>
    [s|T,D> some text <s]
  <-del]
<page]
[page|D>
  [+del|T>
    [s|T,D> some text <s]
  <del]
<page]
<root]
```


3. Discontinuous text

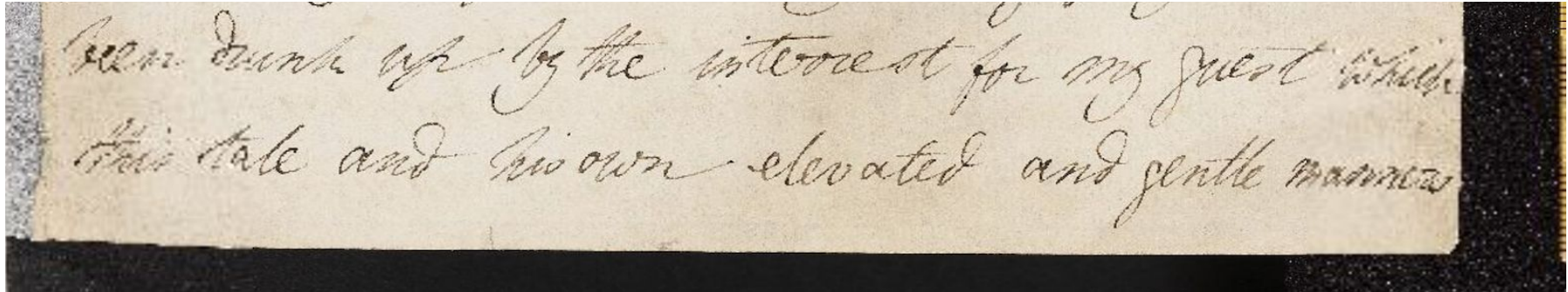
XML tree



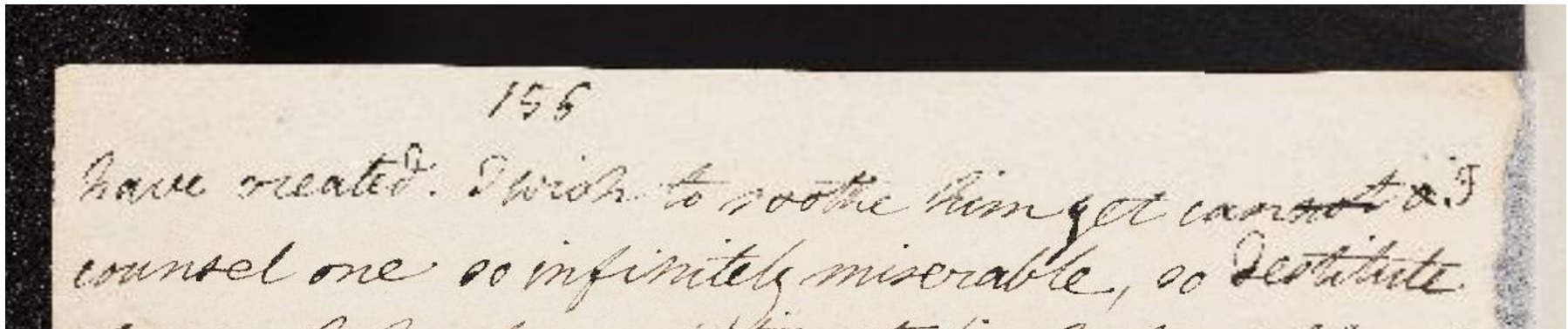
TAG hypergraph



3. Overlapping structures



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

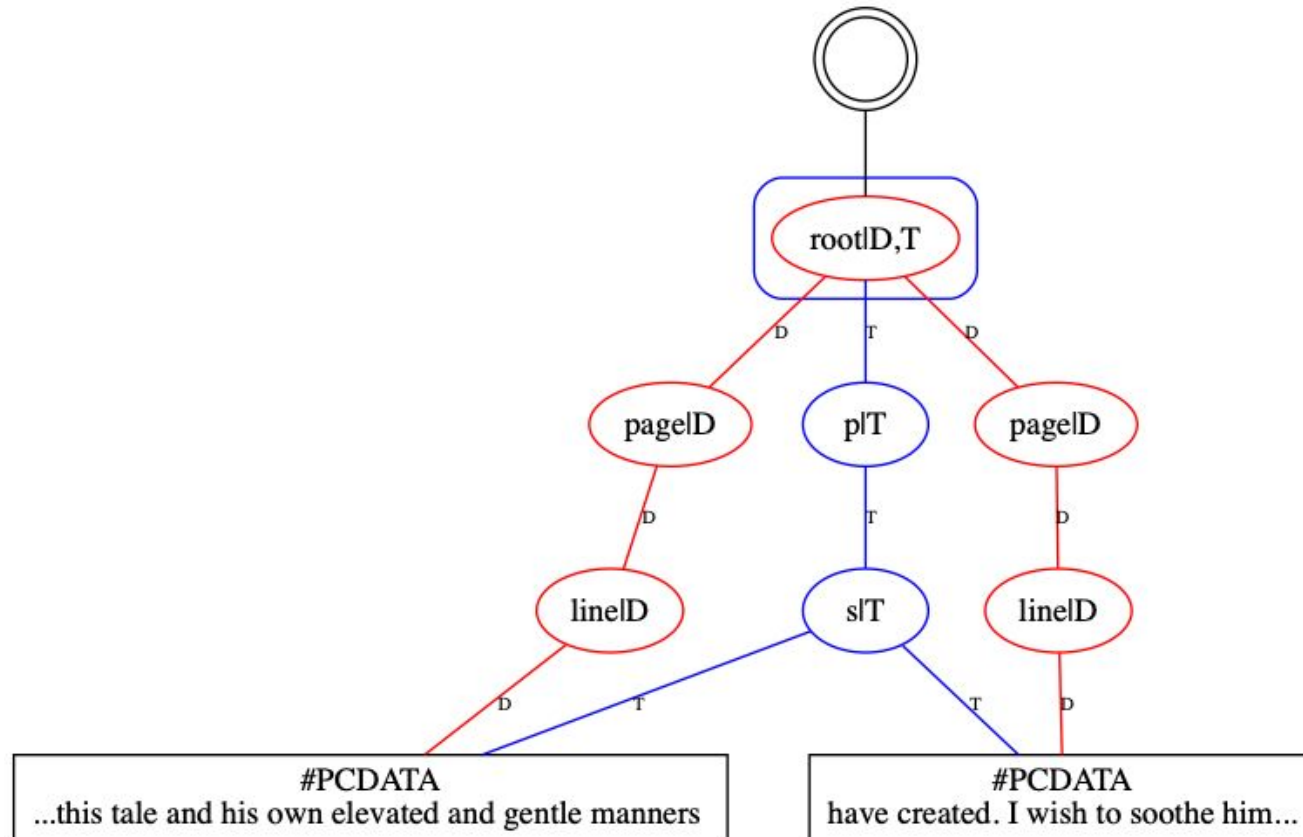


156
have created. I wish to soothe him yet cannot & I
counsel one so infinitely miserable, so destitute.

3. Overlapping structures

```
[root|+T,+D>
[page|D>
  [p|T>
    [line|D>
      [s|T>...this tale and his own elevated and gentle manners <line]
    <page]
  [page|D>
    [line|D>have created. I wish to soothe him... <line]
  <s]
  <p]
<page]
<root]
```

3. Overlapping structures



4. Recap

“the holy grail of computer science [is] to algorithmically express the natural world”

(Teresa Nakra, 2016)

“we encode texts and represent them digitally in order to present, examine, study, and reflect on the rich heritage of knowledge and expression presented to us in our cultural legacy”

(Wendell Piez, 2014)

4. Recap

The TAG model for text:

- Redefinition of text: multi-layered, non-linear, multiple orders
- Complex mix of information parsed and processed without workarounds
- More responsibility delegated to the syntax, less to the schema- or application level

The goals of textual genetic editing:

- Multiple coexisting dimensions of the text
- Making explicit what is on the manuscript page
 - Non-linear, non-hierarchical structures
 - Revisions *currente calamo*

More information

<https://huygensing.github.io/TAG/>

Get in touch!

elli.bleeker@di.huc.knaw.nl

Elli Bleeker
Ronald Haentjens Dekker
Bram Buitendijk

*R&D Humanities Cluster
Royal Science Academy of the Netherlands*



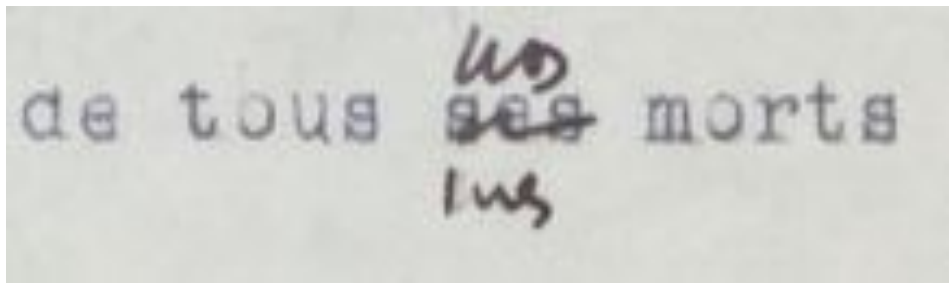
@ellibleeker
@ronald_dekker
@bram_buitendijk

**TEI conference
Graz
September 18, 2019**

5. Discussion

1. Open variants in TAGML and the need to address markup elements
2. Export functionality: implementing TAG in an existing (XML/TEI-based) workflow

5. Discussion: Open variants



Source: Gustave Roud's *Requiem*. GR MS 1H/16d, f. 1r

Encoding option 1

```
[root> de tous <|[seg n=1>ses<seg]| [seg n=2>nos<seg]| [seg  
n=3>les<seg]|> morts <root]
```

Encoding option 2

```
[root> de tous <|ses|nos|les|> morts <root]
```

5. Discussion: exporting TAGML

Why?

- tools for analysis
- visualisation
- publication frameworks

What?

- SVG
- XML
- DOT
- PNG

5. Discussion: Exporting TAGML to XML

Implications of information reduction

- TAGML can handle (self)overlap; XML cannot
- Overlap in TAGML is handled with layers
- A TAGML file with multiple layers will be exported to XML Trojan Horse

Extra: TAG design choices

- Data typing of annotations (string, integers, list, boolean, nested annotations)
- Recursive annotations (rich content)
- Git-like workflow
- Command-line interface; no editorial environment
- HyperGraph repository (*Alexandria*); distributed workflow

Open source: Apache License 2.0

Source code on Github

TAG portal: <https://github.com/HuygensING/TAG>

Extra: future work

Development agenda :

- Schema language validation
- Query language
- Diff-ing TAGML files
- ...

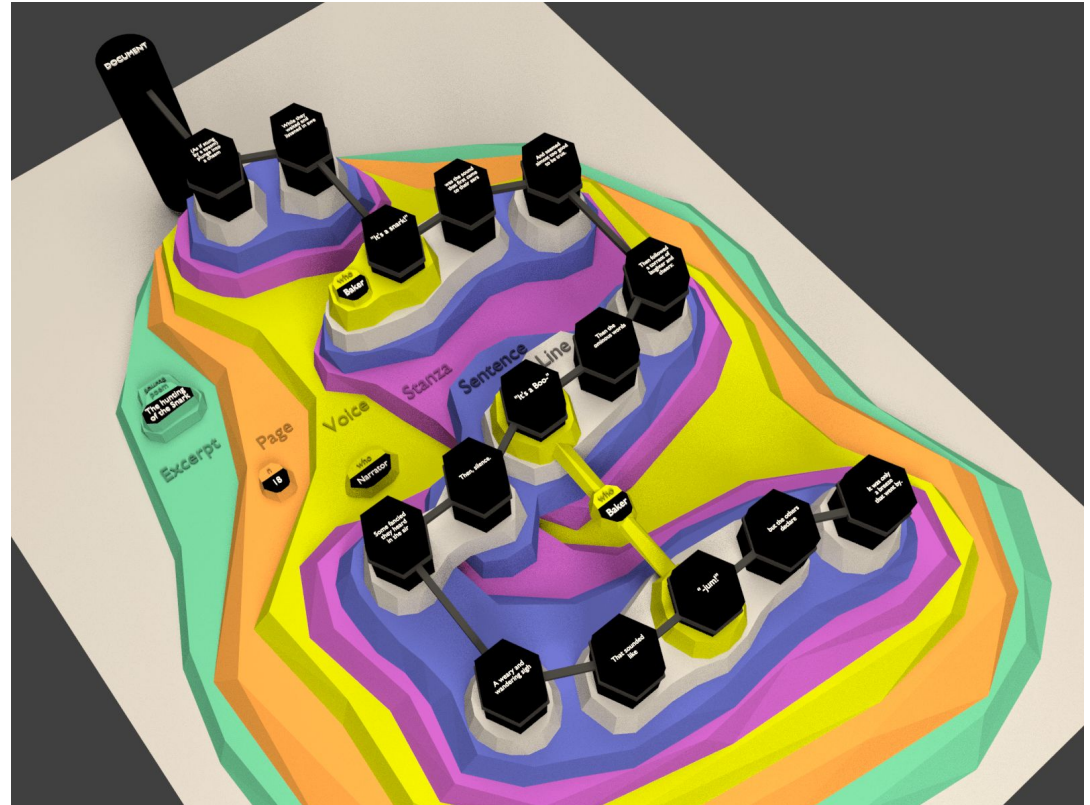
Extra: TAG - data model

Nodes

- Document node (root)
- Text nodes
- Markup nodes
- Annotation nodes

Edges (undirected)

- Document-Text
- Text-Text
- Markup-Text
- Annotation-Markup (multiple)
- Annotation-Annotation (multiple)
- Annotation-Text



Source: Dekker and Birnbaum 2017 (figure by Gijsjan Brouwer)