

Text-as-Graph

A data model for complex texts

Elli Bleeker

Ronald Haentjens Dekker

R&D group - KNAW Humanities Cluster

Lorentz workshop "Processing Ancient Text Corpora"

February 17, 2020

Overview

1. Introduction: Modeling cultural heritage texts
 - 1.1. Why
 - 1.2. What
 - 1.3. How
2. Data models for complex texts
 - 2.1. Overview
 - 2.2. Text-as-Graph (TAG)
 - 2.3. Text modeling in TAG
3. Processing cultural heritage texts
 - 3.1. Automated collation
 - 3.2. Workflow
 - 3.3. Automated collation in TAG: HyperCollate

Part 1. Modeling cultural texts

1. What is text modelling?

A model is a formalised description of a real world object or concept.

Meant to study the object, test a theory.

1. Why is text modelling important?

The way you model the data determines how it can be queried and what kind of information can be extracted from it.

No model means no structure, no information.

Limited model means information is lost.

1. Text modelling

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

1. Text encoding = modelling

“The primary goal of text encoding in the humanities should not be to conform to standards ... Rather, 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

1. Text encoding

Truisms

Every transcription and every encoding is an *interpretation*

We use it to

- express our understanding of text(s)
- process, analyse, and/or publish text(s)

We hope it can be used to

- be reused by others

Whose many-voiced Whores thro' the mist
Of cataracts, flung the thunder of that fall,
The icy Springs stagnant with wrinkling fold
Which vibrated to hear one, & then with
Shuddering through India; & then ~~through~~ ^{the} air
Through which the sun walks busying without
And ye swift Whirlwinds who on ^{peers} ~~poor~~ wings
Hung mute & motionless o'er your hushed sleep
As thum der boules then your own ~~and~~ ^{and} ~~and~~ ^{and}
The orbed words - if then my words ~~had~~ ^{had} ~~had~~ ^{had}
- though I am changed so that ought ~~and~~ ^{and} ~~and~~ ^{and}
Is dead within, although no memory be
Of what is here - let them not look it over!
What was that cause? for ye all hear me speak
1st Voice from the Mountains
Ironic three hundred thousand years
O'er the earth quakes couch we stand;
Oft as men convulsed with fear
We tremble in our smothered trade.
2^d Voice from the Springs
Thunder with head ~~passed~~ ^{passed} ~~our~~ ^{our} ~~water~~ ^{water}
3^d We had been stained with ~~little~~ ^{little} ~~blood~~ ^{blood}

1. What do we want to encode?

Multiple perspectives

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

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

4
Whose many-voiced Whores thro' the mist
Of cataracts, flung the thunder of that fall,
The icy Springs stagnant with wrinkling fold
Which vibrated to hear one, & then with
Shuddering through India; & thro' ~~the~~ air
Through which the sun walks busying without
And ye swift Whirlwinds who on ^{peers} ~~peers~~ wings
Hung mute & motionless o'er your hushed sleep
As storm doth louder than your own ~~and~~ ^{and} ~~and~~
The orbed world - if then my words ~~had~~ ^{had} ~~had~~
- though I am changed so that ought ~~and~~ ^{and} ~~and~~ ^{and} ~~and~~
Is dead within, although no memory be
Of what is here - let them not look it over!
What was that cause? for ye all hear me speak
1st Voice from the Mountains
From three hundred throes and years
O'er the earth quakes couch we stand;
Oft as men convulsed with fear
We tremble in our smothered trade.
2^d Voice from the Springs
Thunders with head paraded our water,
3^d We had been stained with bitter blood

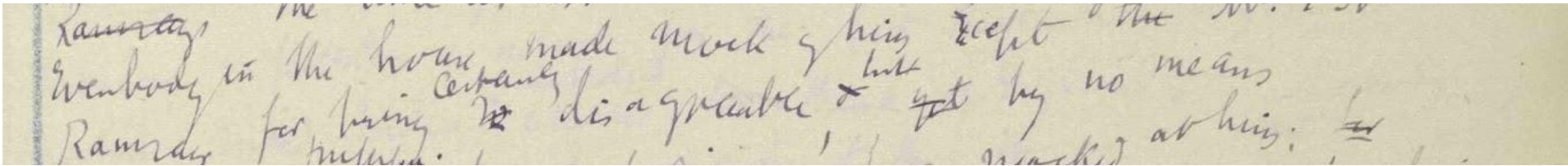
1. What do we want to encode?

Textual characteristics

- **Overlap**
- **Discontinuity**
- **Non-linearity**
- **Containment and dominance**
- **Self-overlap**
- ...

1. Textual characteristics

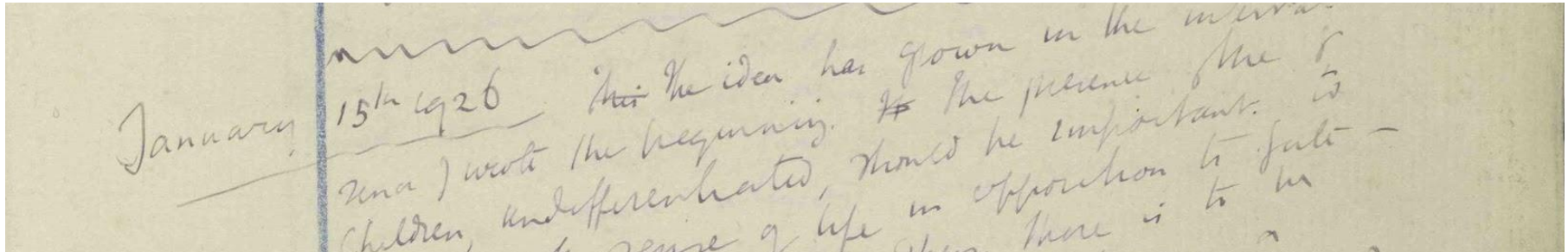
Grouped revision



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

1. Textual characteristics

Immediate revision



January 15th 1926

~~This~~ The idea has grown ...

1. Textual characteristics

Open variants

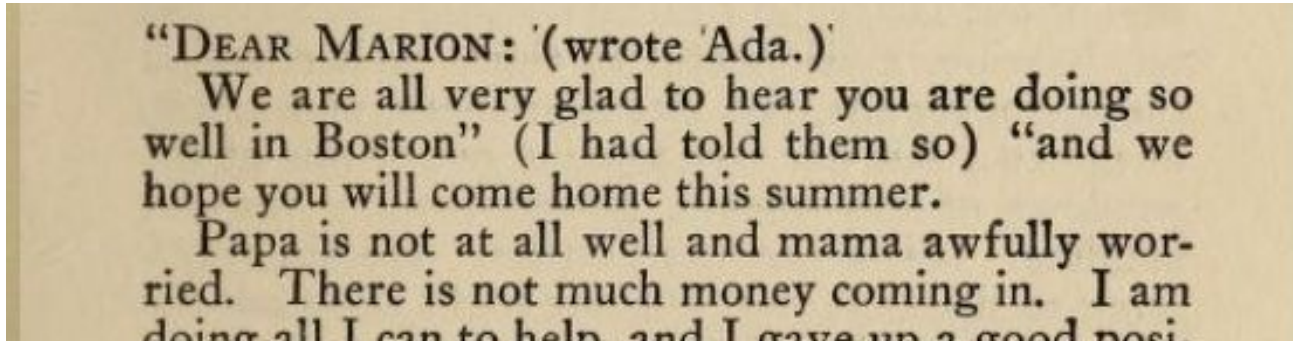


That is ^{soon said} an easy thing to say.

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

1. Textual characteristics

Discontinuous text

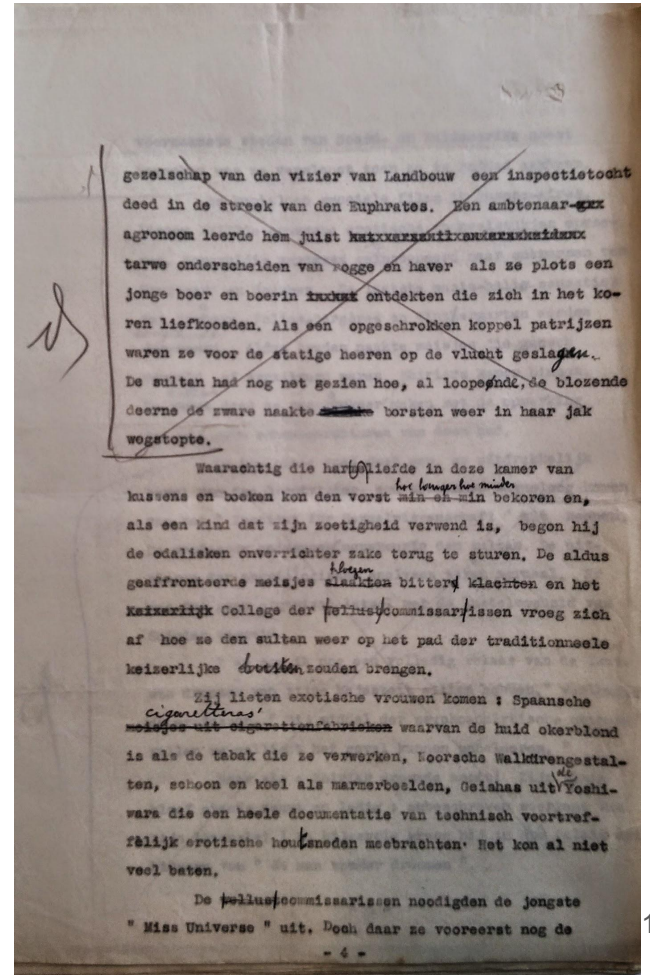
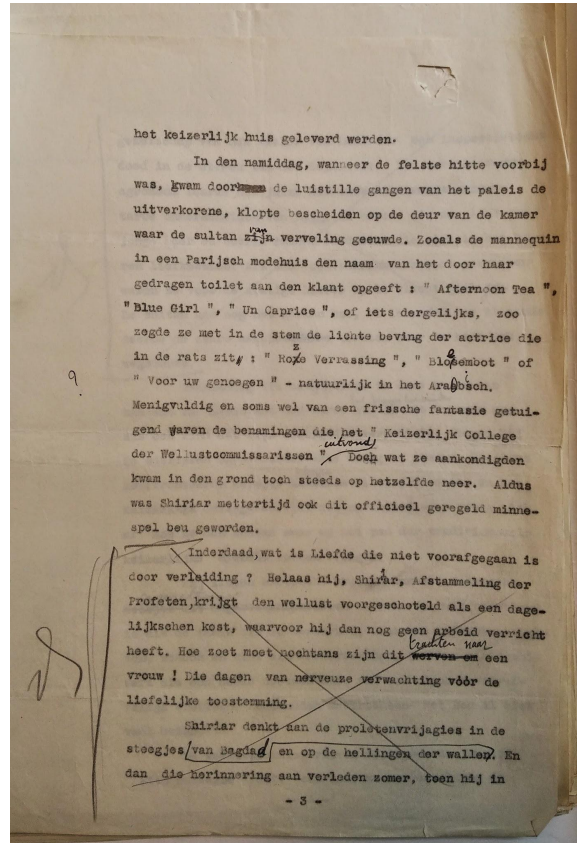


“DEAR MARION: (wrote Ada.)
We are all very glad to hear you are doing so well in Boston” (I had told them so) “and we hope you will come home this summer.
Papa is not at all well and mama awfully worried. There is not much money coming in. I am doing all I can to help, and I gave up a good posi...

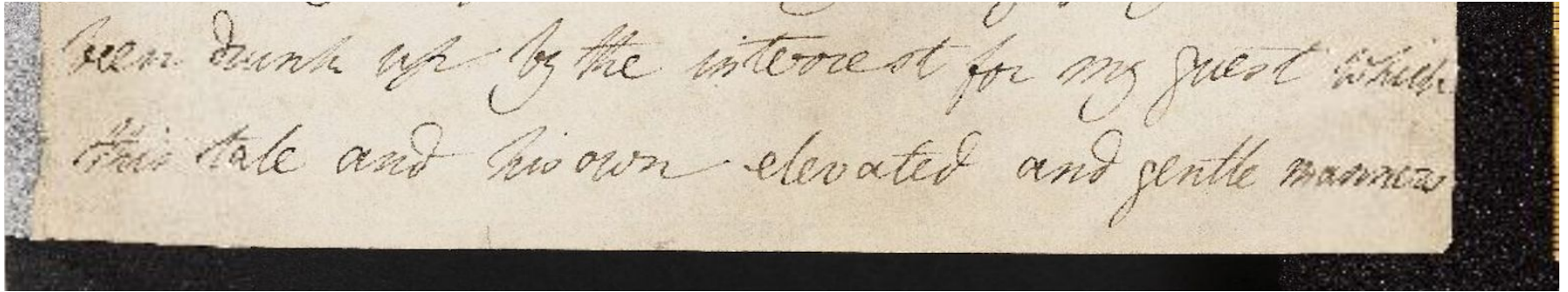
“Dear Marion: (wrote Ada.) We are all very glad to hear ...”

1. Textual characteristics

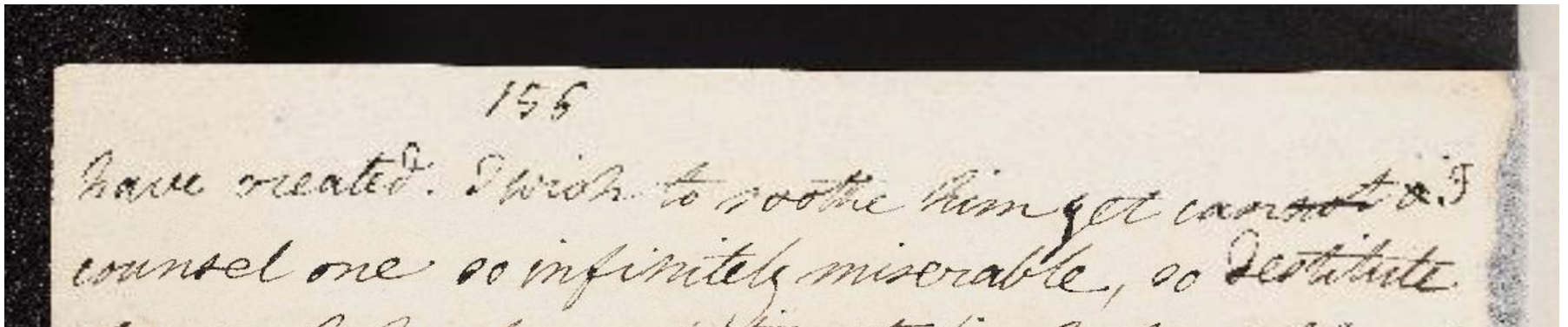
Discontinuous structures



1. Textual characteristics

A close-up photograph of a handwritten manuscript snippet on aged, yellowish paper. The text is written in a cursive hand and reads: "been drunk up by the interest for my quest which
his tale and his own elevated and gentle manners".

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

A close-up photograph of a handwritten manuscript snippet on aged, yellowish paper. The page number "156" is written at the top center. The text below is written in a cursive hand and reads: "have created. I wish to soothe him yet cannot & I
counsel one so infinitely miserable, so destitute".

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

Part 2. Data models for cultural texts

2. Data models for humanist texts

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

2. Text encoding workarounds

The use of workarounds is ingrained in text encoding practices. But *why wouldn't* we want to use workarounds?

- Workarounds are not part of a standard and “in-house solutions” hinder interoperability, reuse, and analysis
- The limits and potential of a data model influence how we look at text

2. TAG data model

In the Text-As-Graph (TAG) data model, we understand text to be

“a multilayered, nonlinear object containing information that is at times ordered, partially ordered and unordered.”

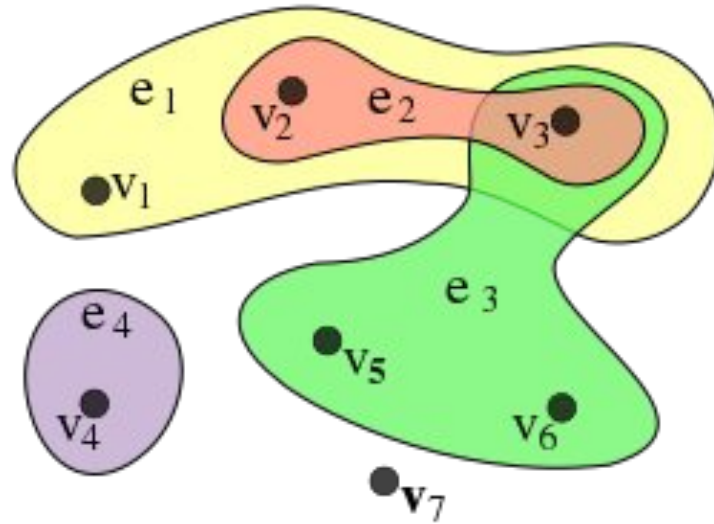
2. TAG data model

TAG uses a **hypergraph model for text**

An intuitive model for text encoding?

- Text
- Annotations on that text, grouped in layers
- Expressive and rich (strings, boolean, numbers, lists, nested annotations)

2. What is a hypergraph?



Source: hypergraph drawing from WikiCommons; <https://commons.wikimedia.org/wiki/File:Hypergraph-wikipedia.svg>

(As if stung by a spasm) plunge into a chasm,
While they waited and listened in awe.

“It’s a Snark!” was the sound that first came
to their ears,

And seemed almost too good to be true.
Then followed a torrent of laughter and cheers :
Then the ominous words “It’s a Boo—”

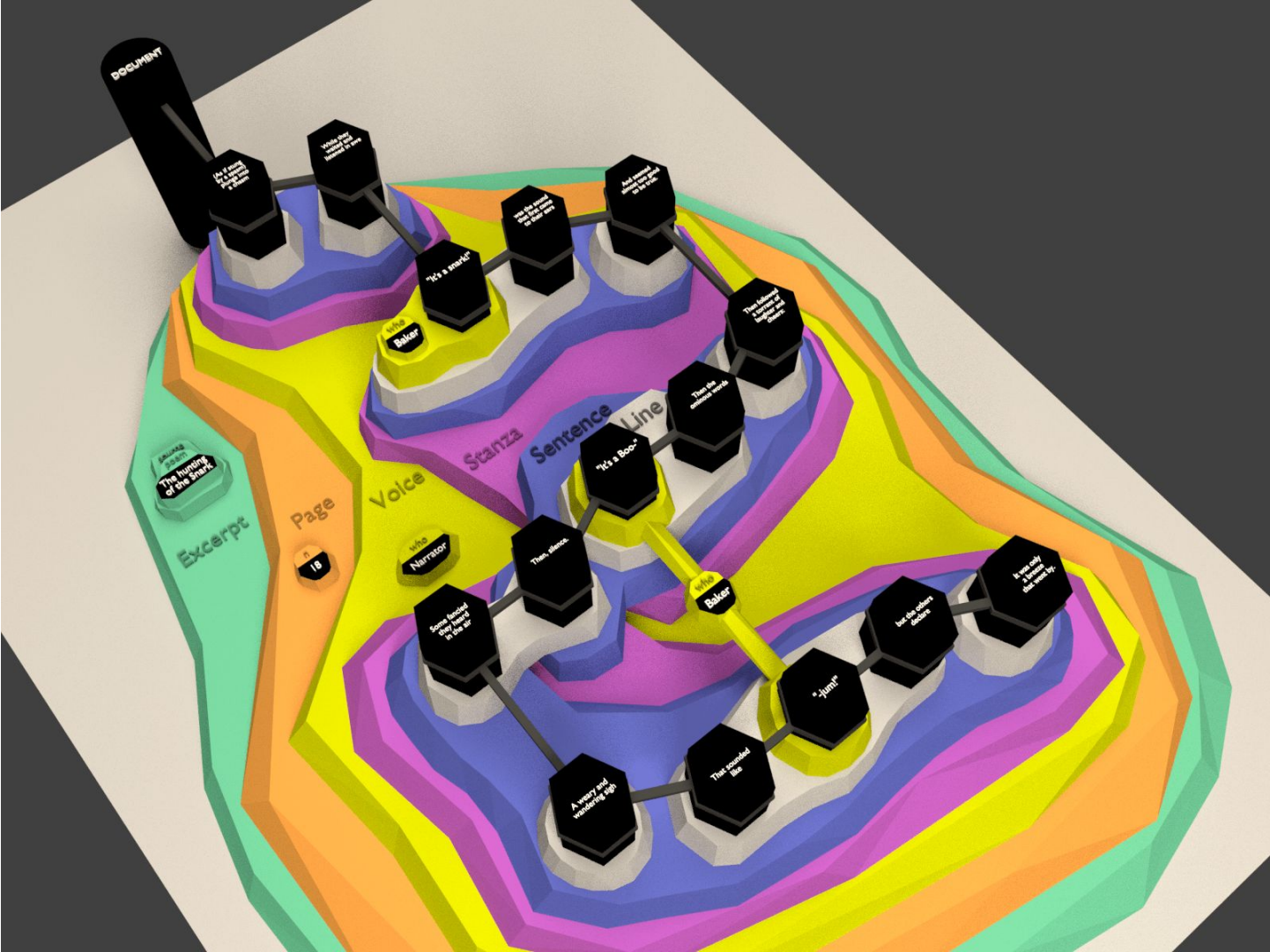
Then, silence. Some fancied they heard in the
air

A weary and wandering sigh
That sounded like “—jum!” but the others de-
clare

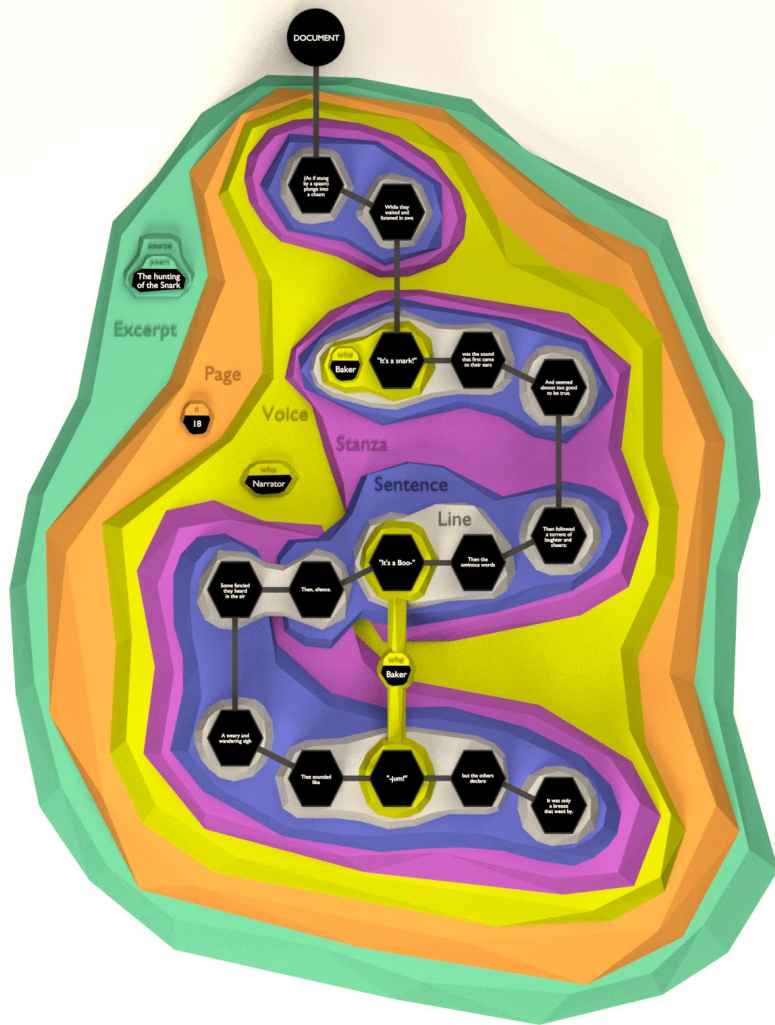
It was only a breeze that went by,

Source: Lewis Carroll 1876. *The Hunting of the Snark: an Agony in Eight Fits*, page 81. Available via

https://en.wikisource.org/wiki/The_Hunting_of_the_Snark







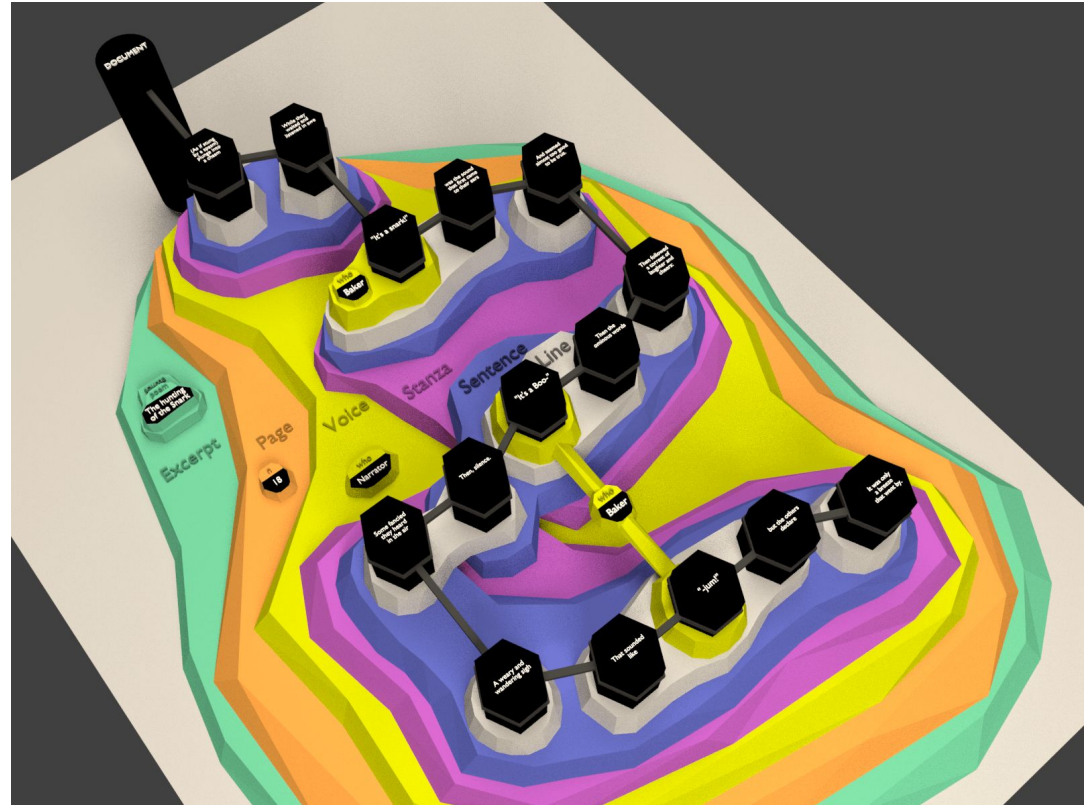
2. TAG - hypergraph data model for text

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

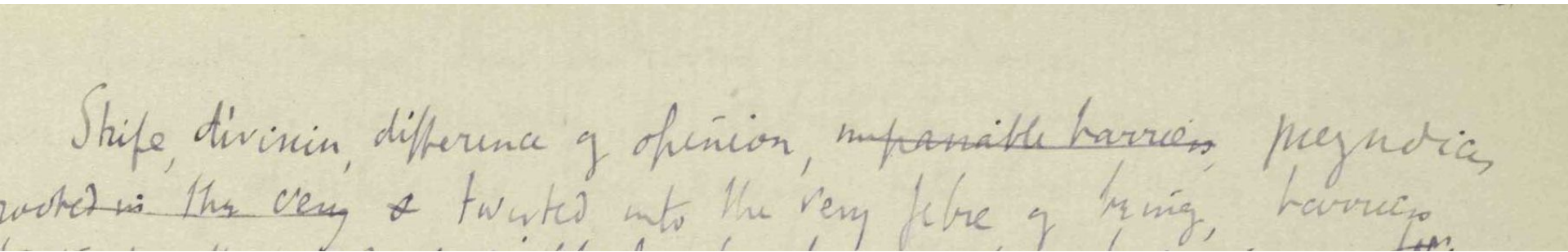


2. Text modelling in TAG

Modeling complex textual characteristics in TAG is easier,
conceptually (once you get used to it)

Modeling literary texts in TAG allows you to express your
understanding of text easily and in great detail

2. Single deletion



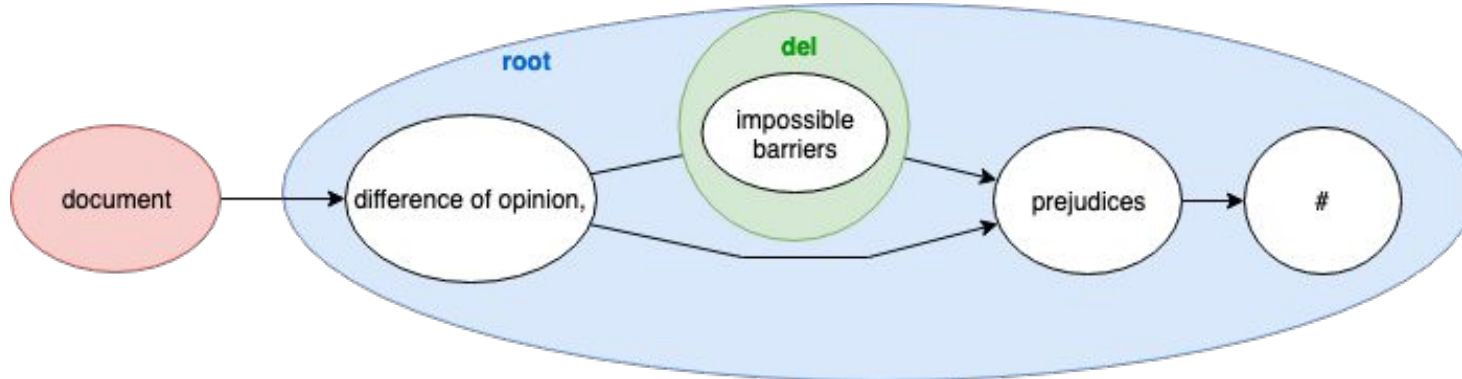
... difference of opinion, ~~impossible barriers~~, prejudices...

(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>>)

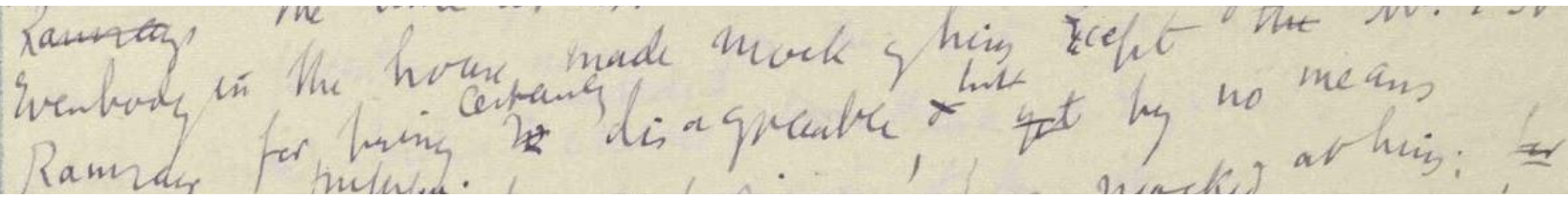
2. Single deletion

[root> difference of opinion, [~~impossible barriers~~], prejudices <root]

TAGML hypergraph



2. Grouped revision



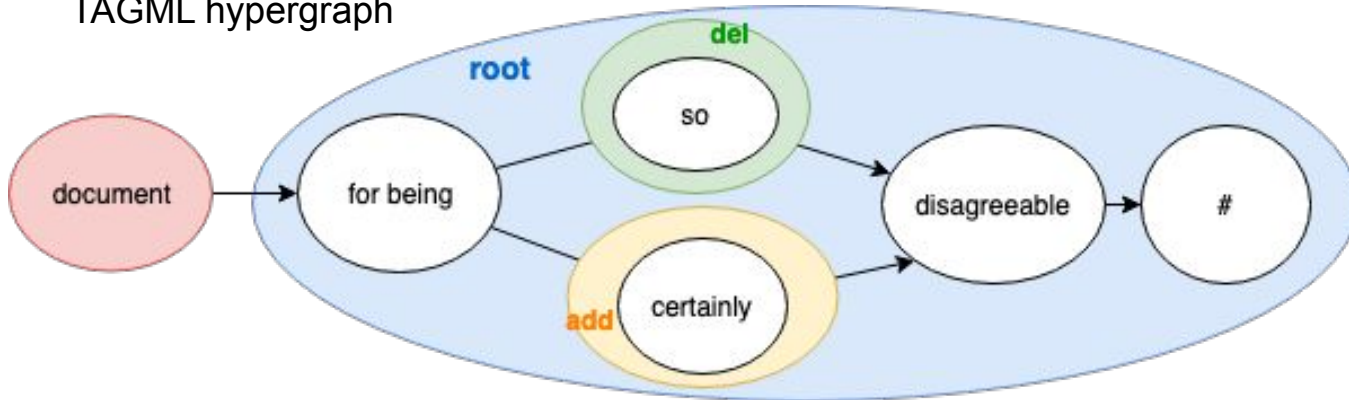
The image shows a horizontal strip of a handwritten manuscript on aged paper. The text is written in cursive and includes several corrections. The visible text is: "Everybody in the house made mock of him except the ...". Below this, there is a correction: "Ramrod for being ~~so~~ ^{certainly} disagreeable & 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. The text continues with "at him; ~~for~~".

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

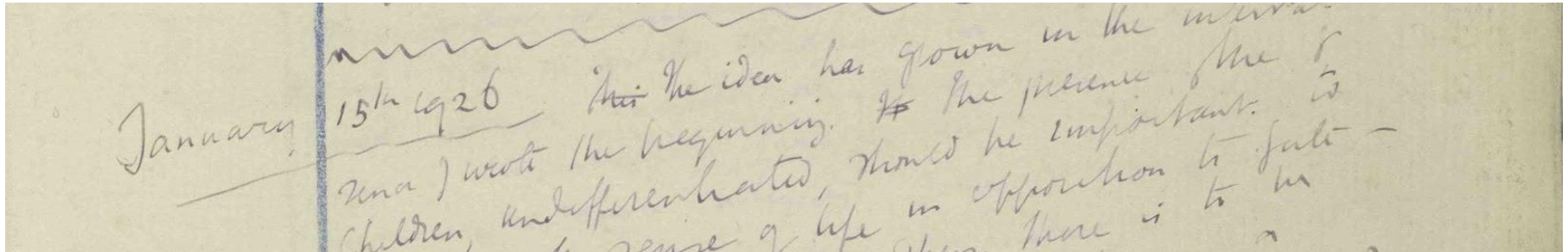
2. Grouped revision

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

TAGML hypergraph



2. Immediate revision



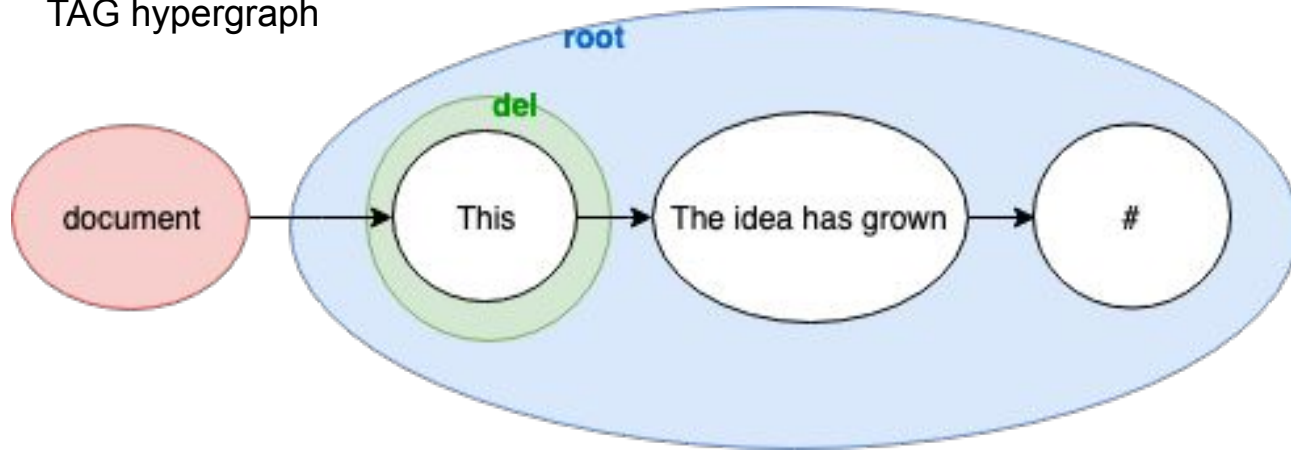
January 15th 1926

~~This~~ The idea has grown ...

2. Immediate revision

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

TAG hypergraph



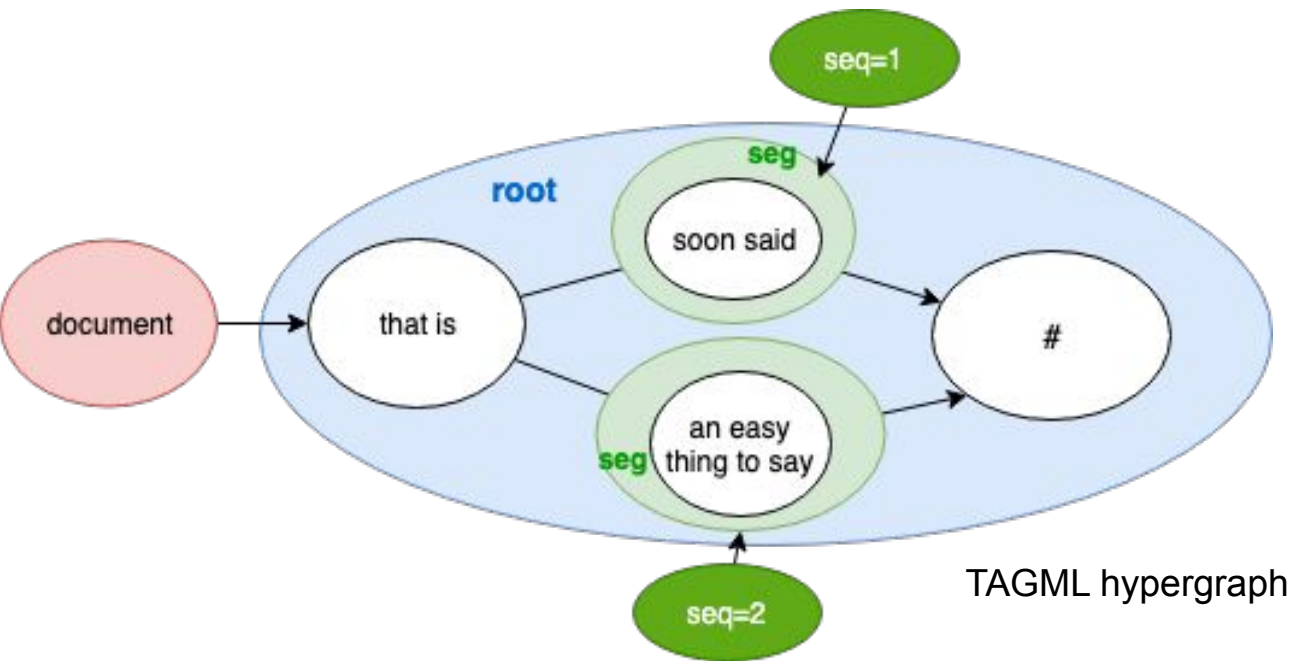
2. Open variants

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

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

2. Open variants

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



2. Discontinuous text

“DEAR MARION: (wrote 'Ada.)

We are all very glad to hear you are doing so well in Boston” (I had told them so) “and we hope you will come home this summer.

Papa is not at all well and mama awfully worried. There is not much money coming in. I am doing all I can to help, and I gave up a good position offered me by the C. P. R. to travel over their Western lines and write travel pamphlets, because I will not leave mama just now.

Charles would do more, but his wife won't let him. I think you ought to help. Ellen has been

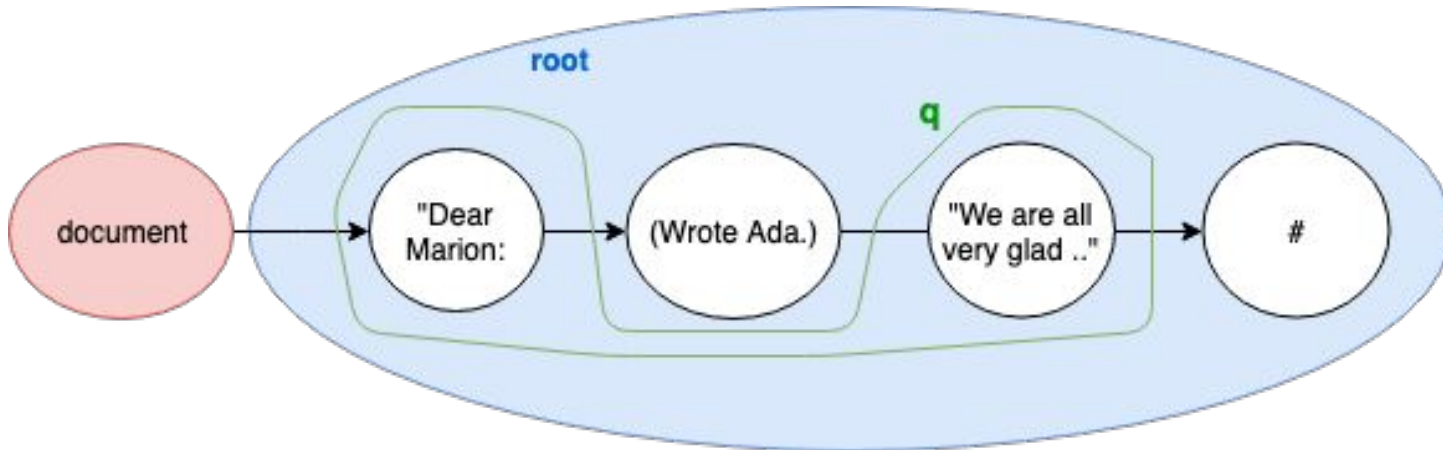
171

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

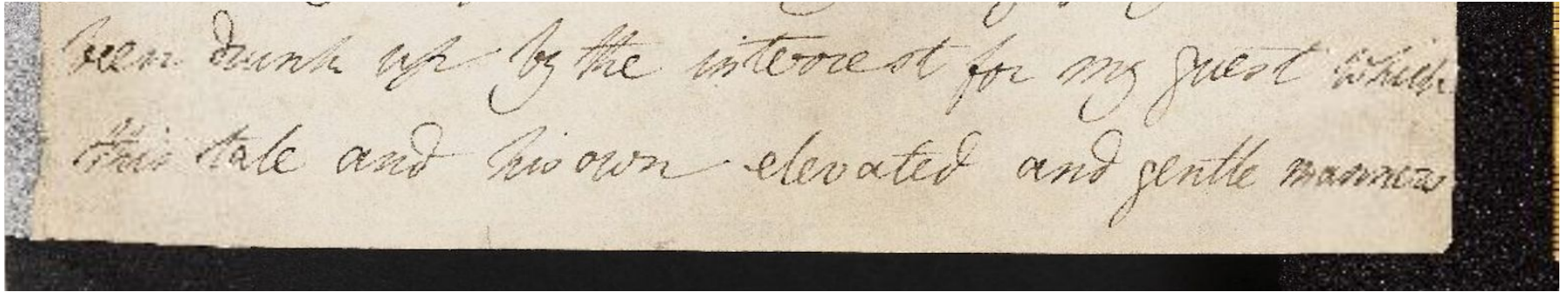
2. Discontinuous text

[root>[q>“Dear Marion:<-q] (wrote Ada.) [+q>We are all very glad...”<q]<root]

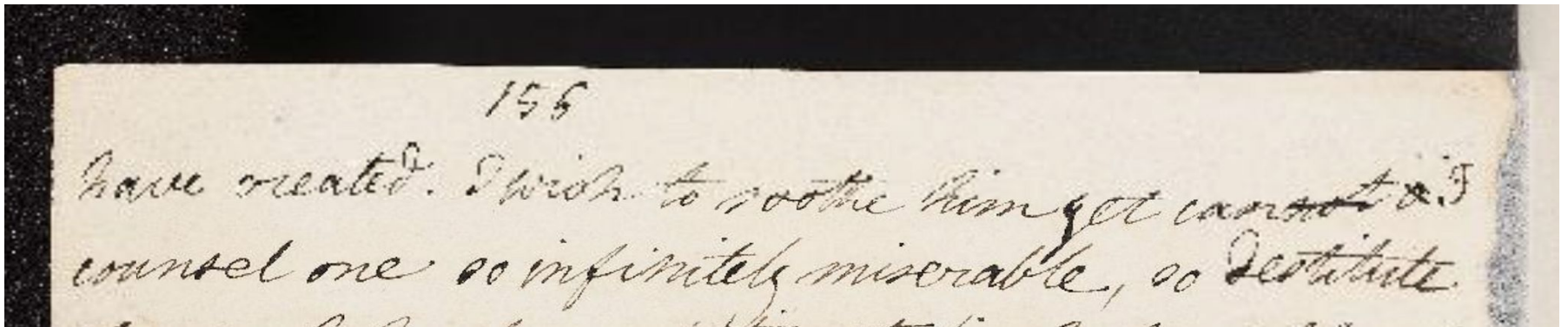
TAG hypergraph



2. Overlapping structures

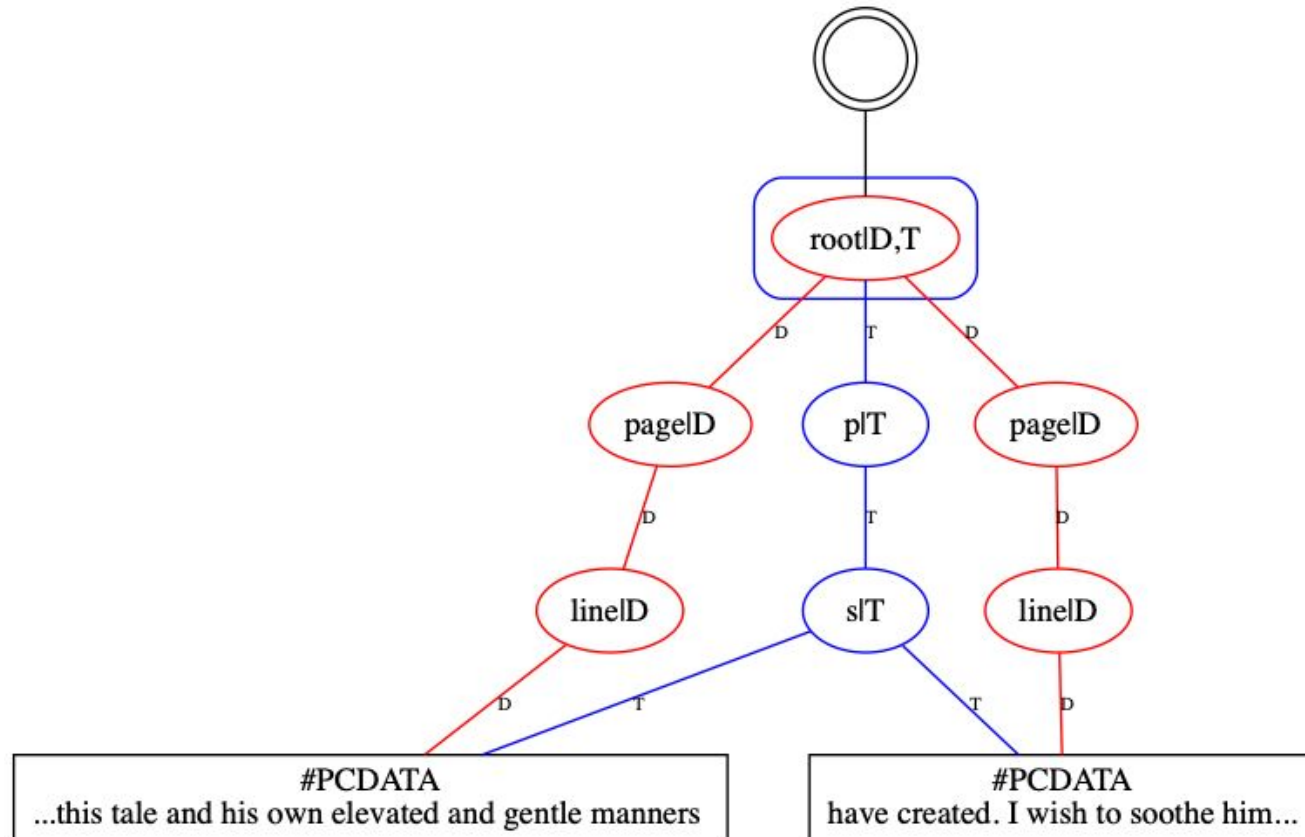
A close-up photograph of a handwritten manuscript snippet on aged, yellowish paper. The text is written in a cursive hand and reads: "been drunk up by the interest for my quest which
his tale and his own elevated and gentle manners".

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

A close-up photograph of a handwritten manuscript snippet on aged, yellowish paper. The page number "156" is written at the top center. The text is written in a cursive hand and reads: "have created. I wish to soothe him yet cannot & I
counsel one so infinitely miserable, so destitute".

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

2. Overlapping structures



Intermediate conclusion: TAG data model

The TAG definition of text facilitates the modelling of complex textual characteristics

- Text: multi-layered, non-linear, multiple orders
- Complex mix of information expressed in TAGML without workarounds

3. Processing complex texts

3. Processing complex texts

Collation: the comparison of two or more versions (witnesses) of text

"Error-prone, laborious, painstaking, time-consuming"...

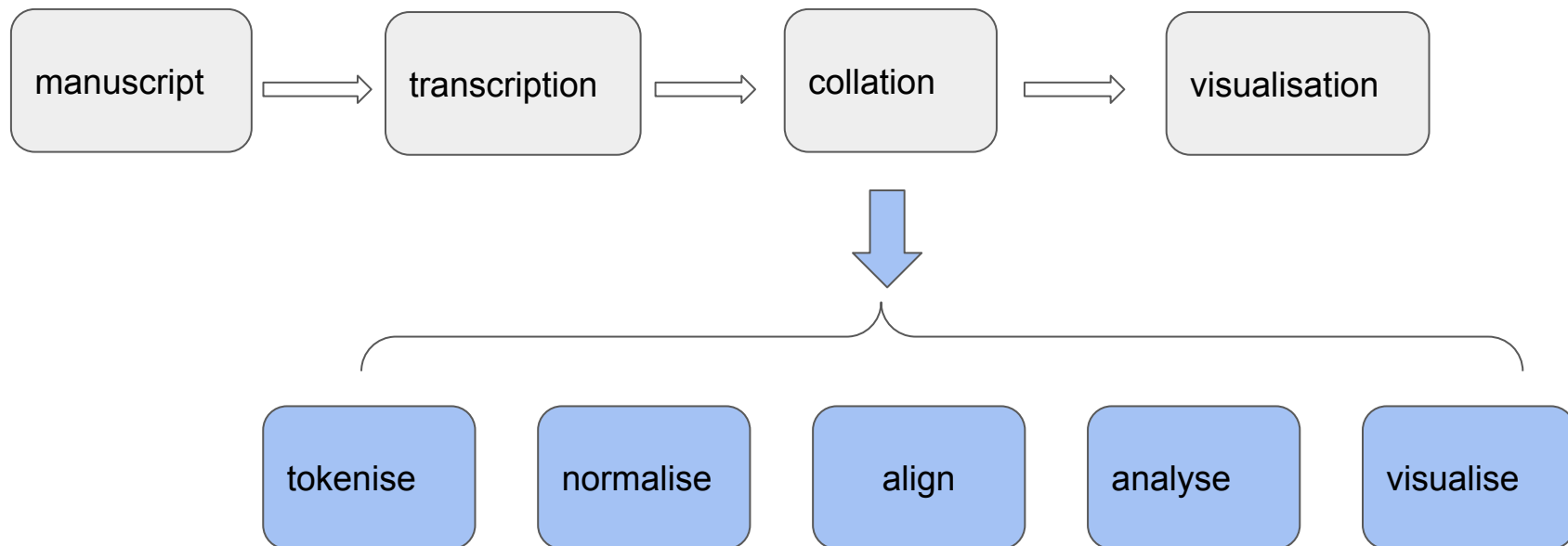
Semi-automated collation tools:

- CollateX
- Juxta
- TXStep
- nMerge

Mapping and classifying textual variation:

- StemmaWeb

3. Processing pipelines



3. Processing complex texts

Challenge / limitation:

Tools collate plain text, so the markup is either ignored or needs to be removed.

Implication:

Witnesses that contain textual variation are "flattened"; the richness of the manuscript is ignored.

3. Problem Statement

How can we use markup and obtain a more refined analysis of textual variation?

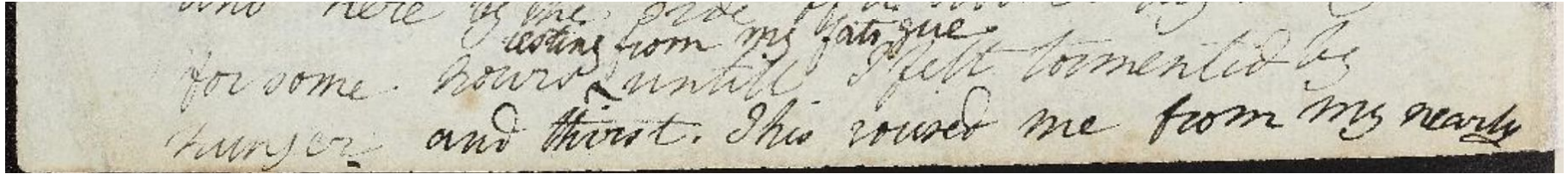
1. Modelling textual variation within one witness in TAGML
2. Advanced analysis of text through algorithms that use hypergraphs and hypergraph merging
3. Visualise/export results

3. Processing complex texts

What do we want to include?

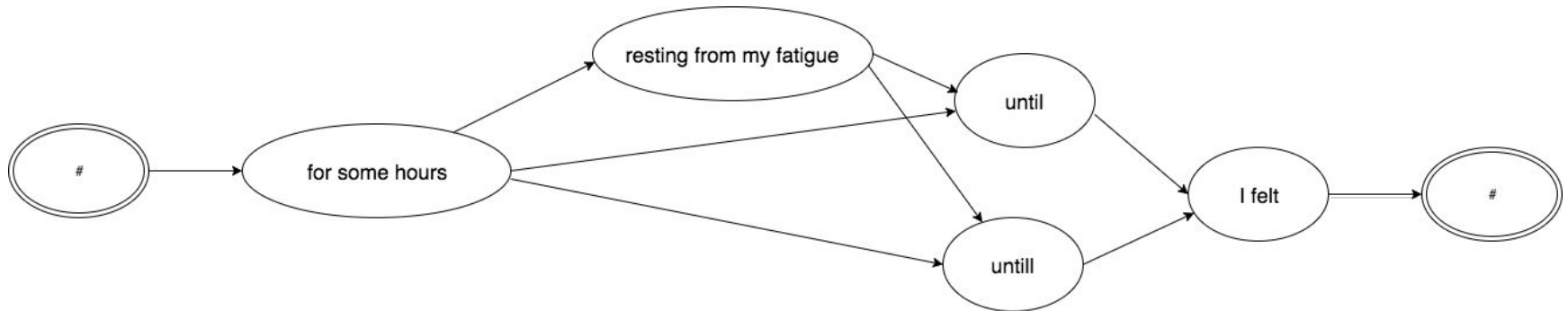
- **Textual variation** within one witness → *multiple paths* through one text, encoded tags with divergence and converge markup tags.
- **Structure** → comparing different documents with different structures results in conflicting hierarchies

3. Example of multiple paths



"... for some hours resting^{from my fatigue} untill[sic] I felt tormented by..."

[Witness A](#) ("Frankenstein", notebook C.57, p. 1r)



3. Processing complex texts

Functional Requirements

- Processing and analysing witnesses with multiple paths through the text
- Finding the minimum set of changes needed to turn one document into the other
- Compare more than 2 documents

3. Processing complex texts

Technical requirements

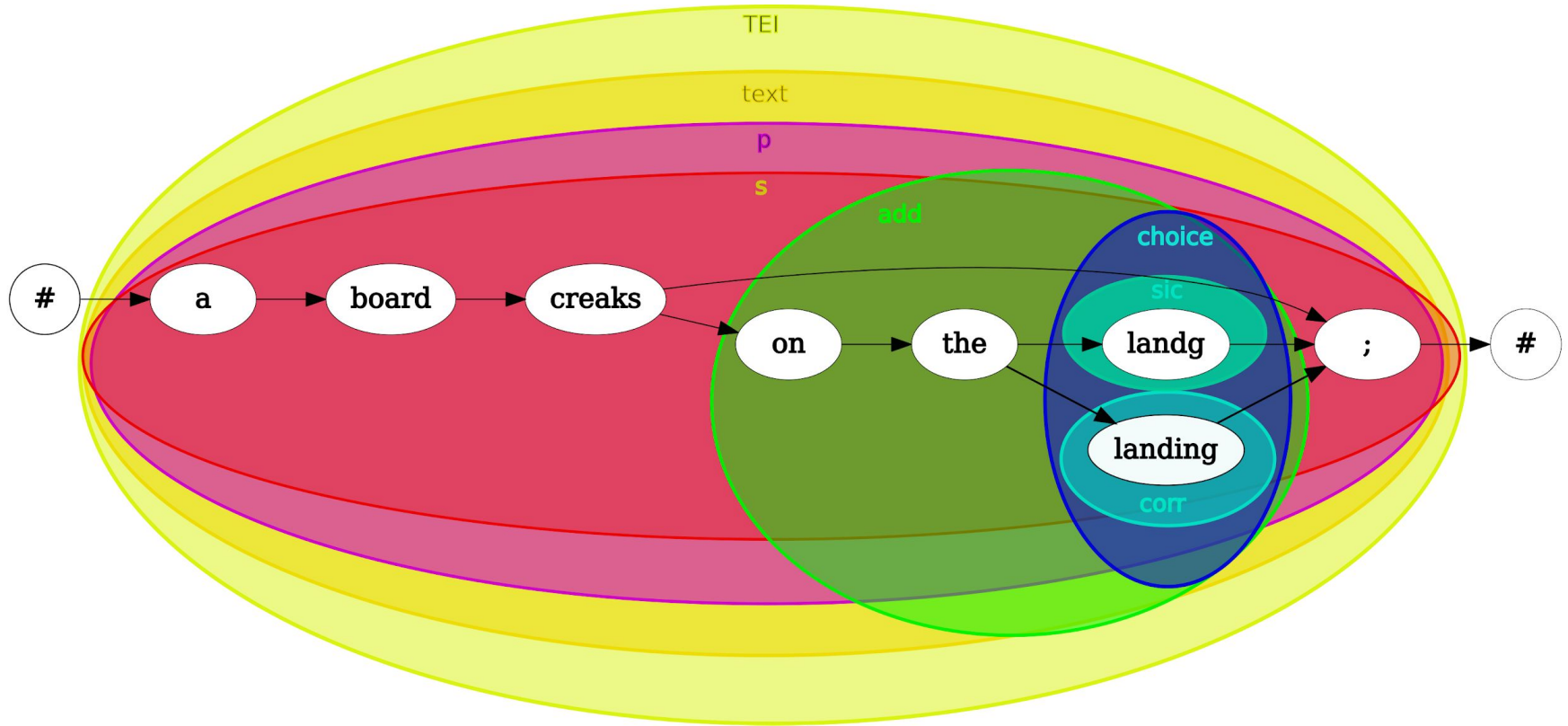
Collation tool

- Respect non-linearity of witness text
- Recognize specific markup tags
- Distinguish between words, punctuation, markup
- Find the smallest amount of differences between two witnesses
- Store results in a hypergraph data model

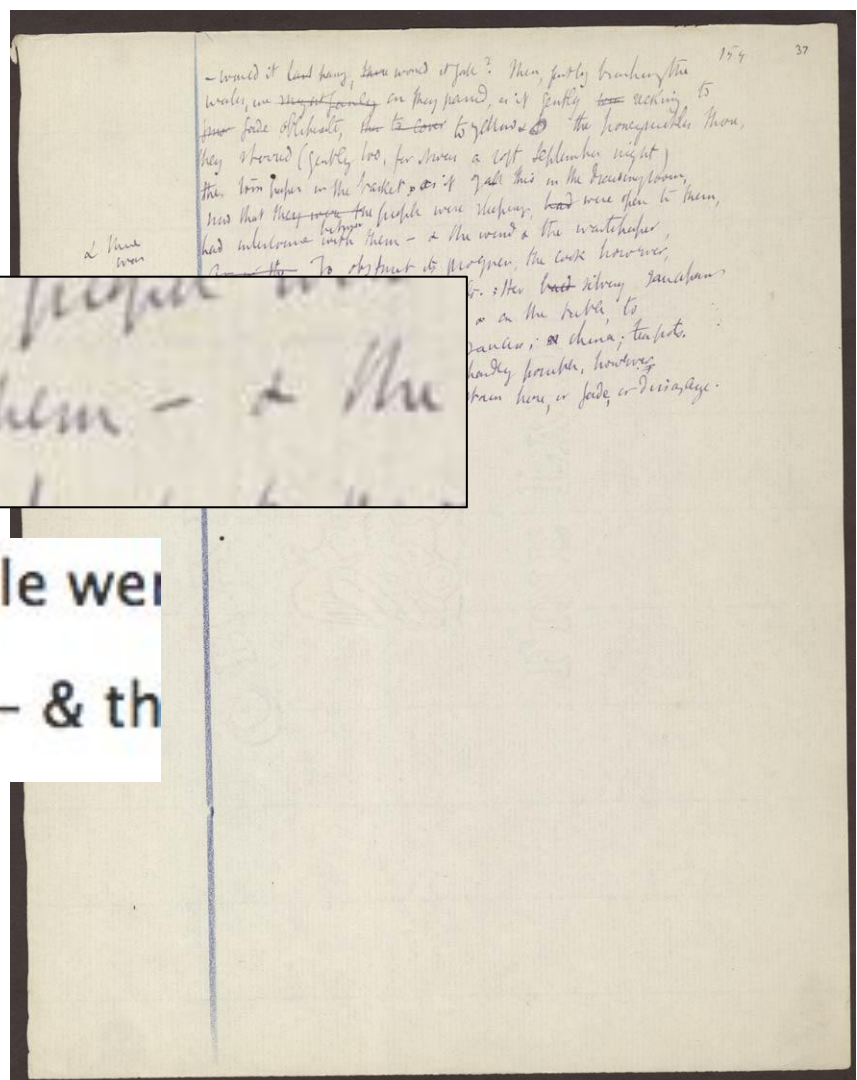


HyperCollate

3. Hypergraph Model for Textual Variation



3. Open Variants

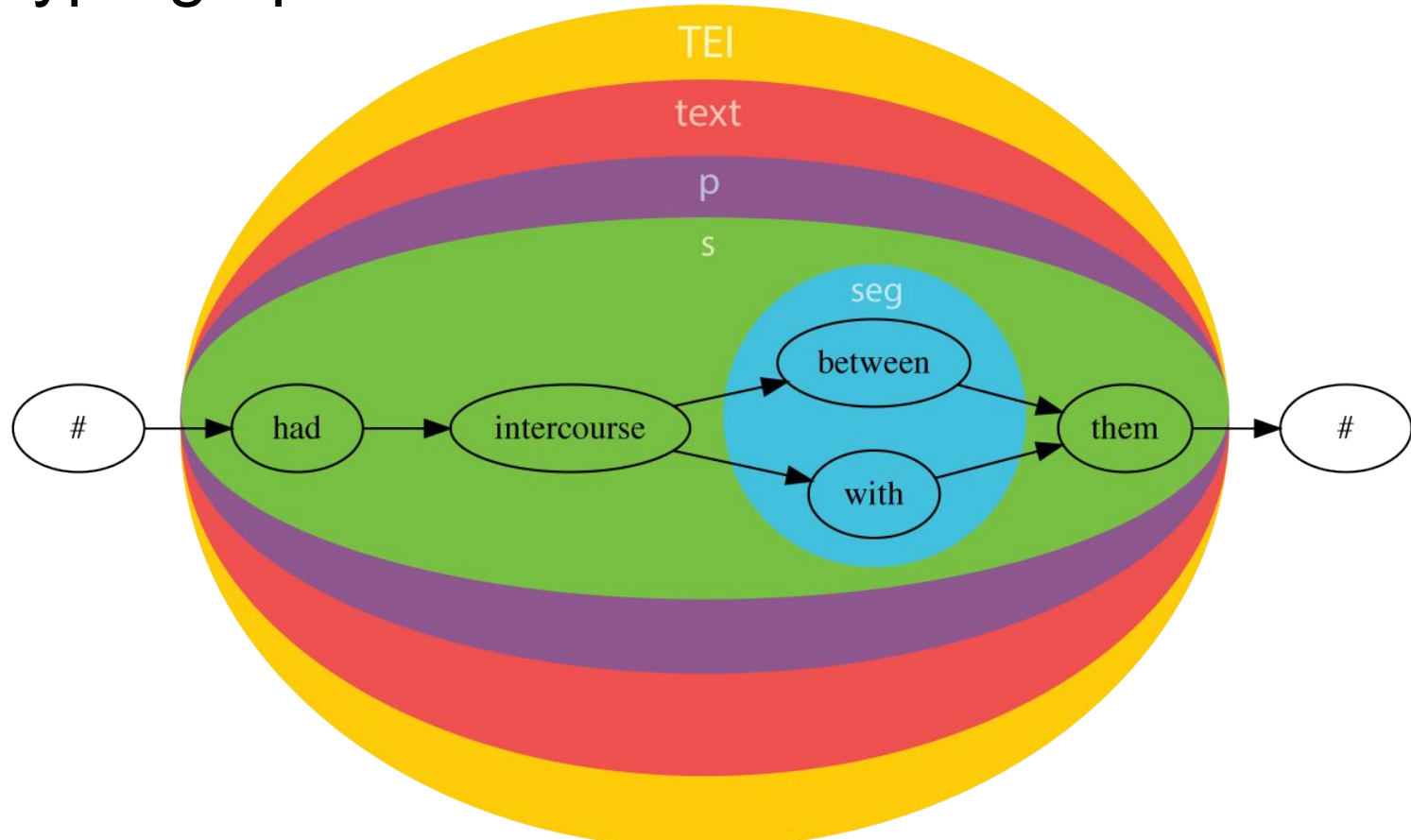


now that they were the people we
had intercourse with them -- & th

now that they were the people we
between
had intercourse with them -- & th

Source: Woolf, V. "Time Passes". Initial Holograph Draft Manuscript p.154, in "Woolf Online"

3. Hypergraph Model for Textual Variation



3. Steps of HyperCollate

1. Align two variant hypergraphs
2. Merge two hypergraphs in one collation hypergraph
3. Repeat in case of >2 witnesses
4. Visualise/export collation hypergraph

3. HyperCollate

Witness A

"... for some hours resting
from my fatigue **untill[sic]**
I felt tormented by..."

~~operations~~ operations of my various senses. By degrees
I remember a stronger light passed upon
my nerves and so that I was obliged to close
my eyes. Darkness then came over me
and troubled me. But hardly had I felt
this when / by opening my eyes and now ~~seeing~~
the light passed in upon me again. I walked
and I believe descended, but presently I found
a great ^{alteration} difference in my sensations; before
dark opaque bodies had surrounded me
impervious to my touch or sight. and
now found that I could wander on at liberty

and here by the side of a brook I lay ~~resting~~
resting from my fatigue
for some hours **untill** I felt tormented by
hunger and thirst. This roused me from my nearly

perceive shade. This was the foot near my foot,
and here by the side of a brook I lay ~~resting~~
resting from my fatigue
for some hours **untill** I felt tormented by
hunger and thirst. This roused me from my nearly

3. HyperCollate

Witness B

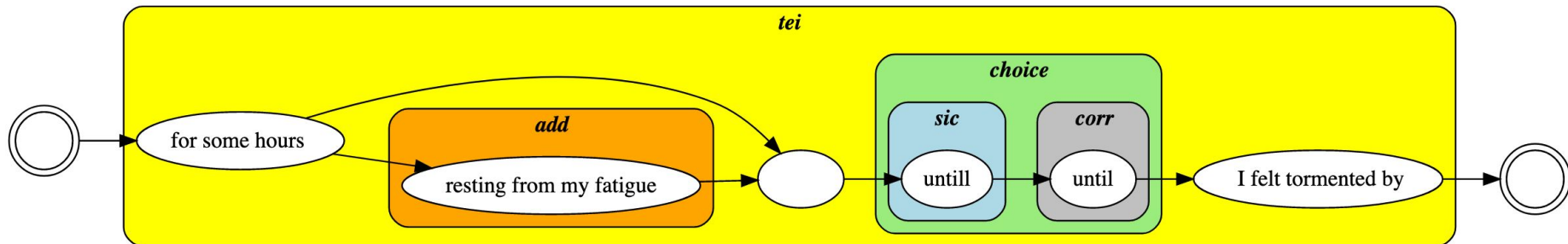
ation or denial of this opinion. For the first time, also, I felt what the duties of a creator towards his creature were, and that I ought to render him happy before I complained of his wickedness. These motives urged me to comply with his demand. We crossed the ice, therefore, and ascended the opposite rock. The air was cold, and the rain again began to descend : we entered the hut, the fiend with an air of exultation, I with a heavy heart, and depressed spirits. But I consented to listen ; and, seating myself by the fire which my odious companion had lighted, he thus began his tale.

could receive shade. This was the forest near Ingolstadt ; and here I lay by the side of a brook resting from my fatigue, until I felt tormented by hunger and thirst. This roused me from my nearly dormant state, and I ate some berries which I found hanging on the trees, or lying on the

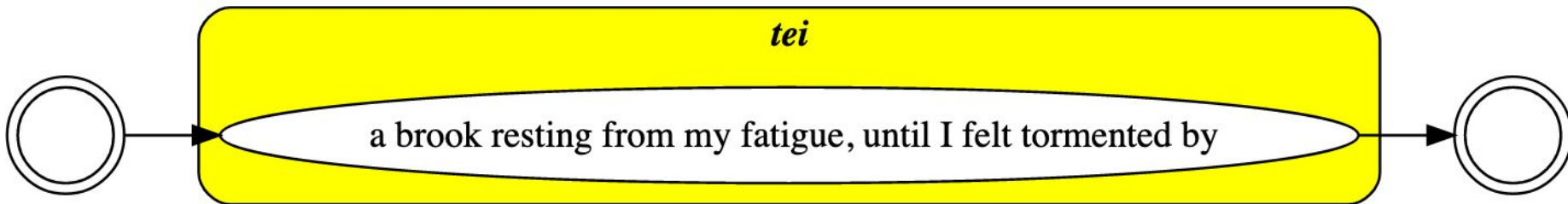
" ... a brook resting from my fatigue, until I felt tormented by ..."

I now found that I could wander on at liberty, with no obstacles which I could not either surmount or avoid. The light became more and more oppressive to me ; and, the heat wearying me as I walked, I sought a place where I could receive shade. This was the forest near Ingolstadt ; and here I lay by the side of a brook resting from my fatigue, until I felt tormented by hunger and thirst. This roused me from my nearly dormant state, and I ate some berries which I found hanging on the trees, or lying on the

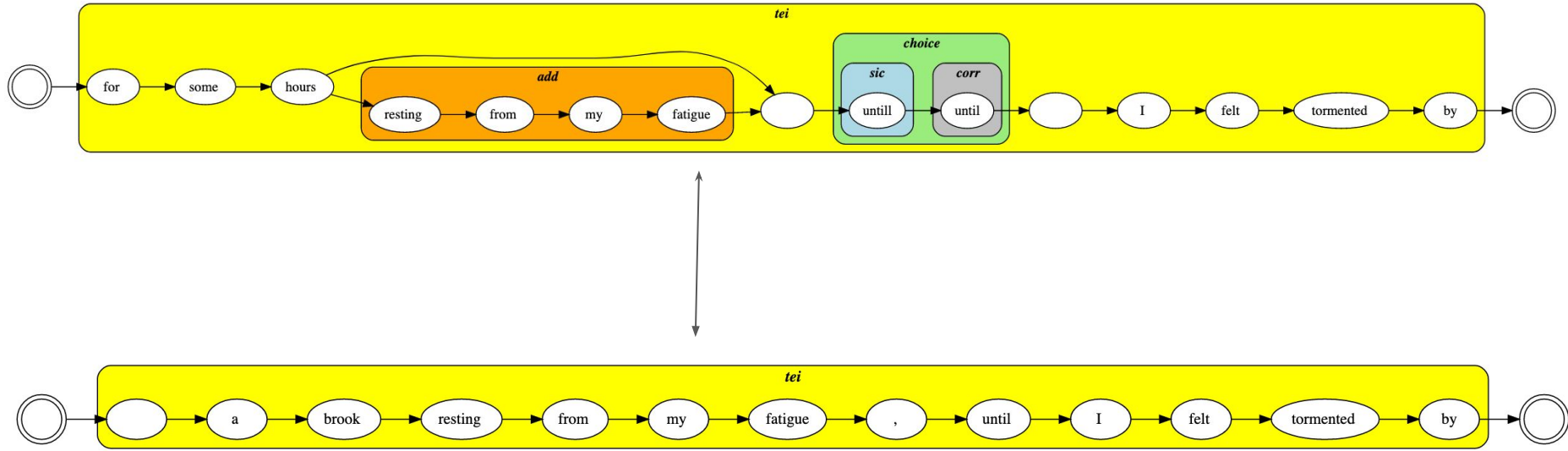
Witness A



Witness B

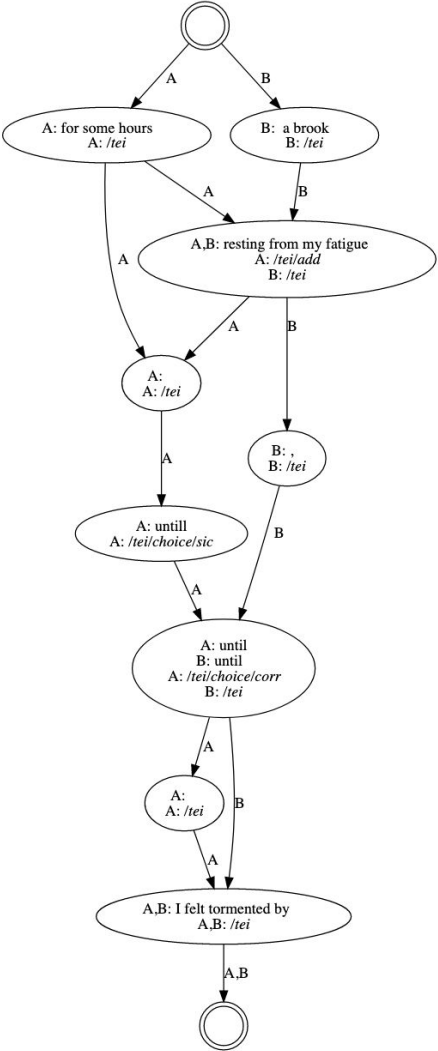


1. Transform TEI text witnesses into variant hypergraphs



2. Align the variant hypergraphs

3. Merge variant hypergraphs into one collation hypergraph



4. Visualise the collation hypergraph

Alignment table

[A]	for some hours	[+] resting from my fatigue		untill	until		I felt tormented by
[B]	a brook	resting from my fatigue	,		until		I felt tormented by

3. Intermediate conclusion: HyperCollate

Witnesses containing textual variation, being partially ordered data, are especially challenging for processing.

Requirements for analysis of this type of textual variation:

- process multiple paths (i.e. recognize markup tags indicating the start and end of a path)
- find the best alignment of all the paths

HyperCollate profits from editorial knowledge encoded in a transcription in order to come to a more refined alignment of witnesses.

4. Conclusion

“Tools always shape the hand that wields them; technology always shapes the minds that use it.”

Michael Sperberg-McQueen 1992

Tools and technologies

- Influence our editorial praxis
- Shape our thinking about text
- Affect how our texts are (re)used
- Reflect our orientations to text
- Digital models influence our research methods

4. Discussion

Visualisation

- How to visualise such an information-rich output?
 - Alignment table
 - Collation hypergraph with internal and external variation in text and markup
 - ... or?
- What do we expect from a collation tool?
- What is "text"?
- How can we visualise the collation output in a meaningful way?

More information

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

<https://huygensing.github.io/hyper-collate/>

Get in touch!

ronald.dekker@di.huc.knaw.nl

elli.bleeker@di.huc.knaw.nl

Elli Bleeker

Ronald Haentjens Dekker

Bram Buitendijk

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



@ellibleeker
@ronald_dekker
@bram_buitendijk

**Lorentz workshop
Leiden
Februari 17, 2020**