Including XML Markup in the Automated Collation of Literary Text

Elli Bleeker, Bram Buitendijk, Ronald Haentjens Dekker, Astrid Kulsdom

R&D - Huygens Institute for the History of the Netherlands
KNAW Humanities Cluster

@ellibleeker
@ronald_dekker
@bram_buitendijk

Context: XML in the Humanities

Text modelling outside OHCO model: TEI (Text Encoding Initiative)

Textual phenomena (variation, overlap, etc.)

Focus: computational philology and textual variation

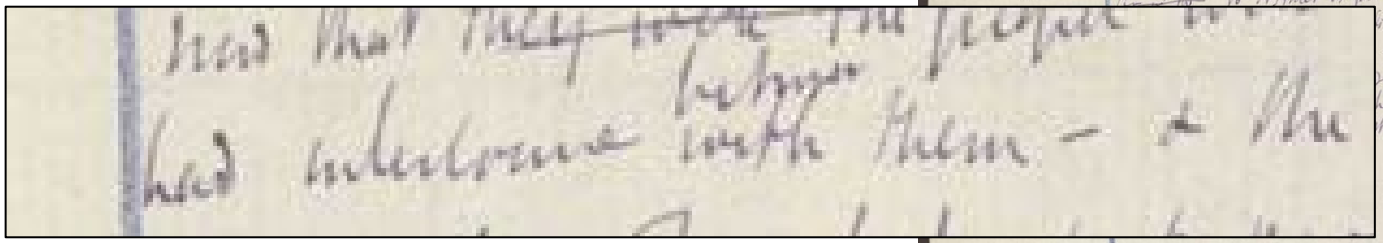
"Collation": the comparison of two or more text versions

Problem Statement

How can we make optimal use of the potential of XML for computational philology?

1. Modelling textual variation in TEI
2. Hypergraphs and hypergraph merging
3. Visualise/export results

Open Variants



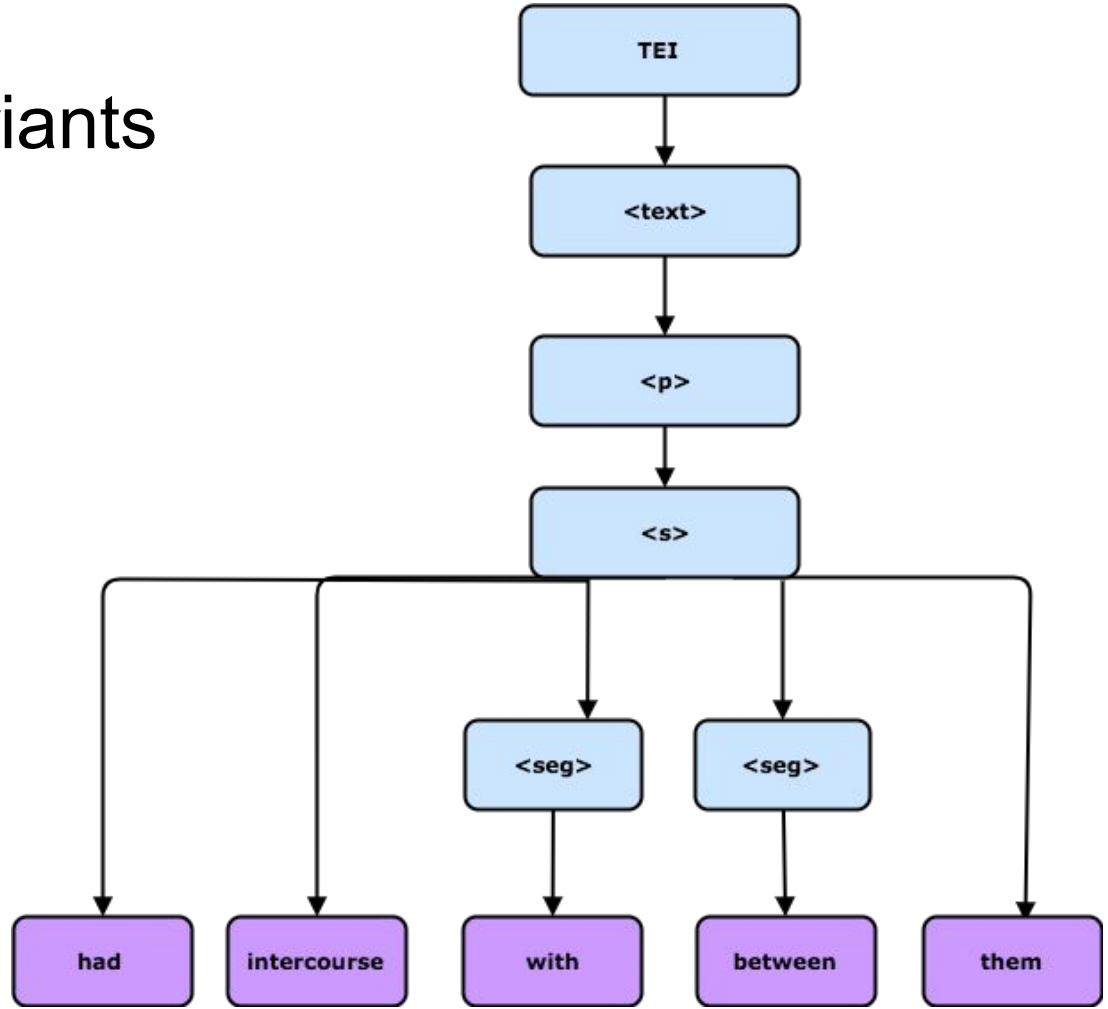
now that ~~they were the~~ people wer
between
had intercourse with them -- & th

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

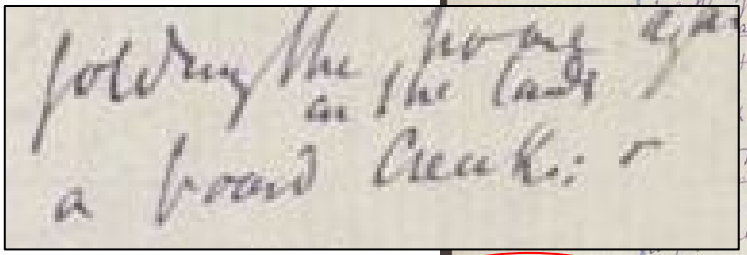
Open Variants

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI>
    <text>
        <p>
            <s>were open to them, had intercourse <seg type="alternative" xml:id="alt1">with</seg>
                <seg type="alternative" xml:id="alt2">between</seg> them</s>
        </p>
    </text>
</TEI>
```

Open Variants



Simultaneity

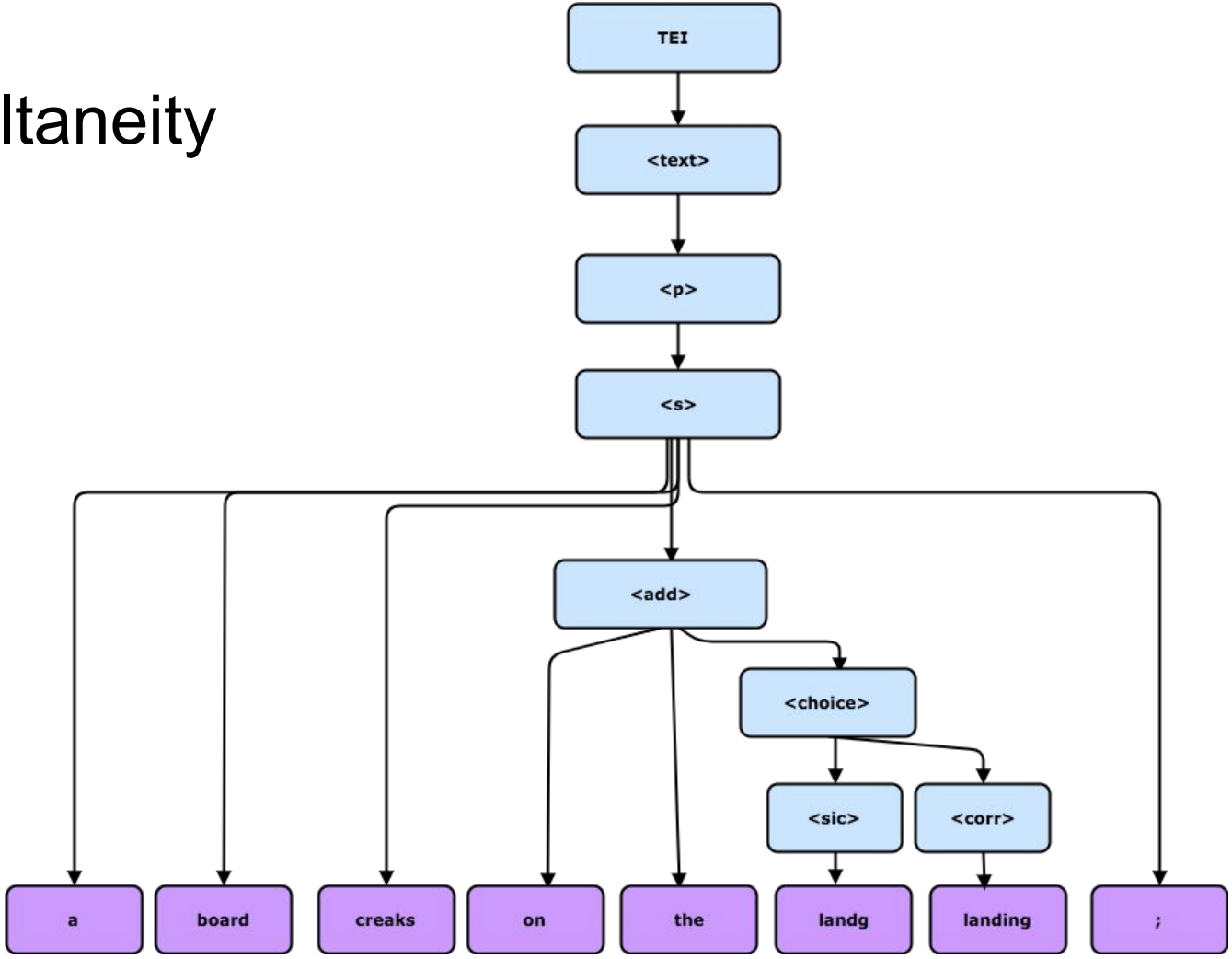


~~[?] pulse.~~ & then gently ~~mak~~ muffling & folding the house again.
on the landg
Only suddenly & for no perceptible reason, a board creaks; &

Simultaneity

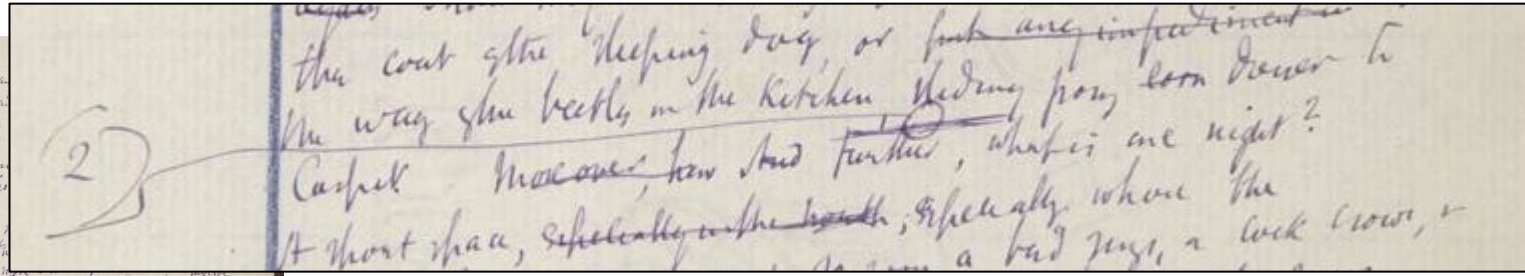
```
<?xml version="1.0" encoding="UTF-8"?>
<TEI>
    <text>
        <p>
            <s>a board creaks <add>on the <choice><sic>landg</sic><corr>landing</corr></choice></add>;</s>
        </p>
    </text>
</TEI>
```

Simultaneity



Structure

Witness 1



2

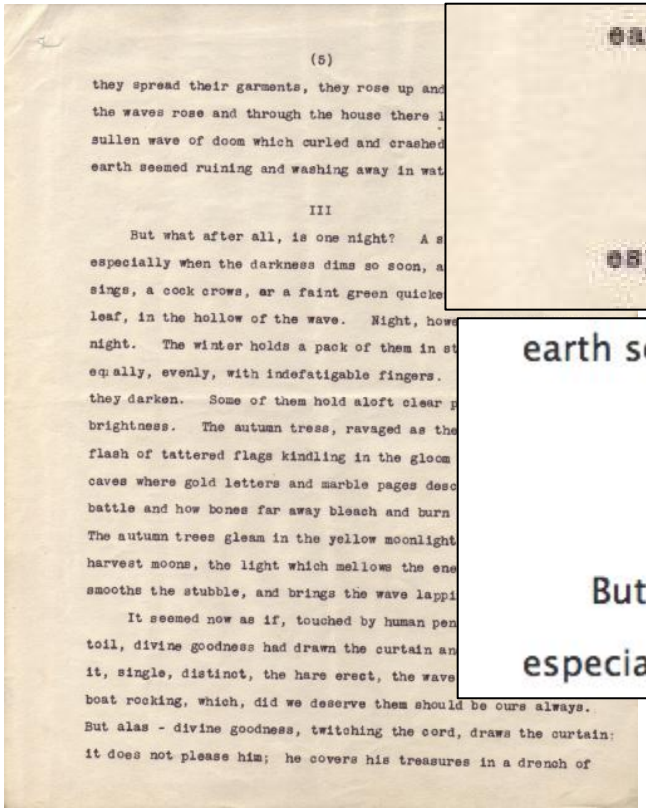
the way of the beetles in the kitchen, sliding from ~~corn~~ dresser to carpet. ~~Moreover, how~~ ~~And~~ Further, what is one night?

A short space, ~~especially in the north~~; especially where the

Source: Woolf, V. "Time Passes". Initial Holograph Draft Manuscript p.157, in "Woolf Online"

Witness 2

Structure



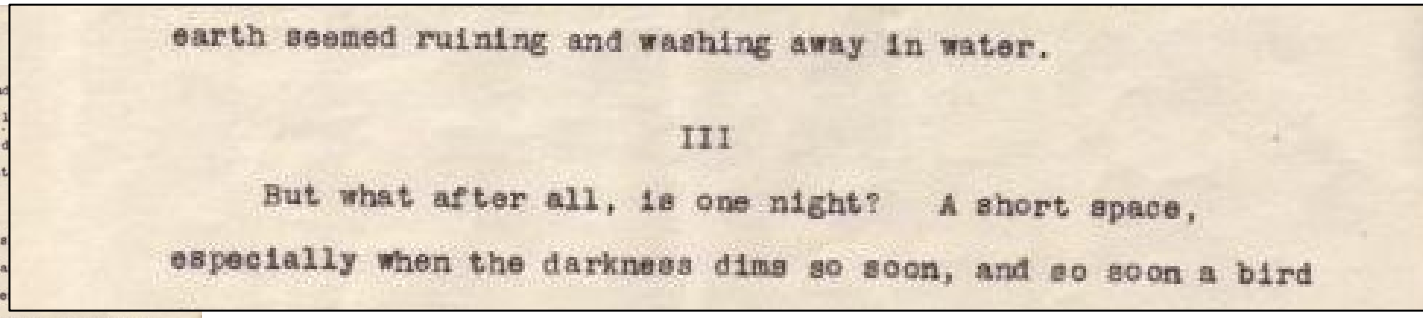
(5)

they spread their garments, they rose up and
the waves rose and through the house there
sullen wave of doom which curled and crashed
earth seemed ruining and washing away in wat

III

But what after all, is one night? A s
especially when the darkness dims so soon, a
sings, a cock crows, ar a faint green quicke
leaf, in the hollow of the wave. Night, how
night. The winter holds a pack of them in st
equally, evenly, with indefatigable fingers.
they darken. Some of them hold aloft clear
brightness. The autumn tress, ravaged as the
flash of tattered flags kindling in the gloom
caves where gold letters and marble pages desc
battle and how bones far away bleach and burn
The autumn trees gleam in the yellow moonlight
harvest moons, the light which mellows the ene
smooths the stubble, and brings the wave lappi

It seemed now as if, touched by human pen
toil, divine goodness had drawn the curtain an
it, single, distinct, the hare erect, the wave
boat rocking, which, did we deserve them should be ours always.
But alas - divine goodness, twitching the cord, draws the curtain:
it does not please him; he covers his treasures in a drench of

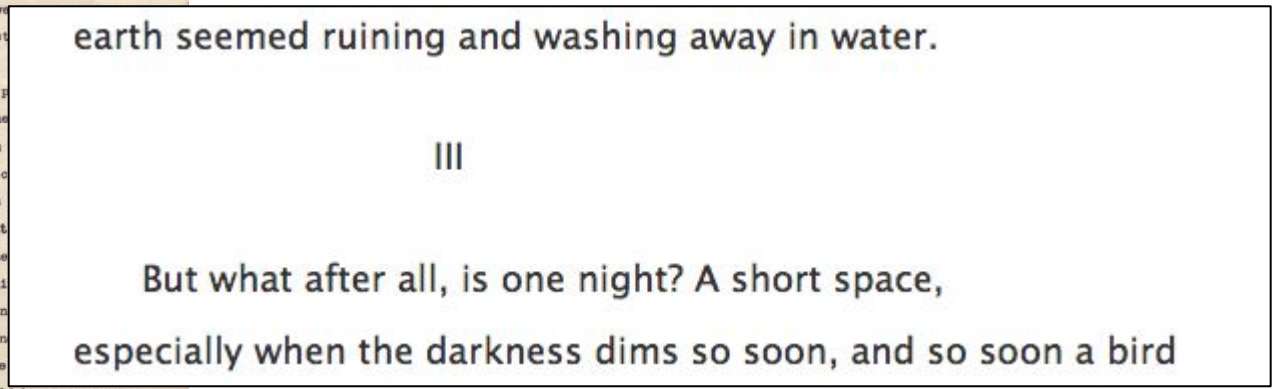


earth seemed ruining and washing away in water.

III

But what after all, is one night? A short space,

especially when the darkness dims so soon, and so soon a bird



earth seemed ruining and washing away in water.

III

But what after all, is one night? A short space,

especially when the darkness dims so soon, and so soon a bird

Source: Woolf, V. "Time Passes". Typescript p.5, in "Woolf Online"

Structure

Witness 1

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI>
    <div type="chapter" n="1">
        <p>
            <s>dresser to carpet.</s>
            <add place="margin"><head>2</head></add>
            <s>Further, what is one night?</s>
        </p>
    </div>
</TEI>
```

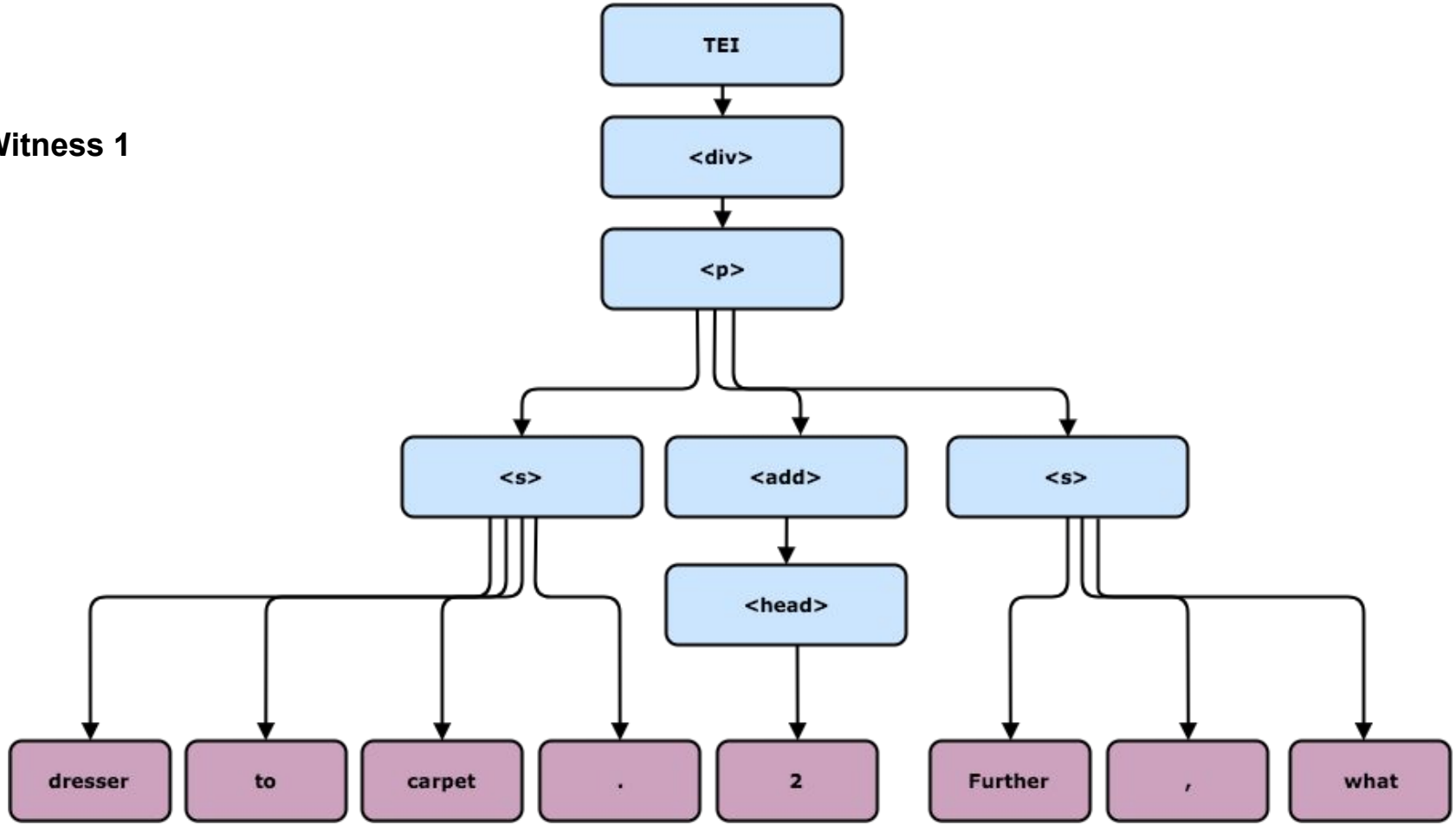
(simplified) TEI-XML transcription of manuscript p.157

Witness 2

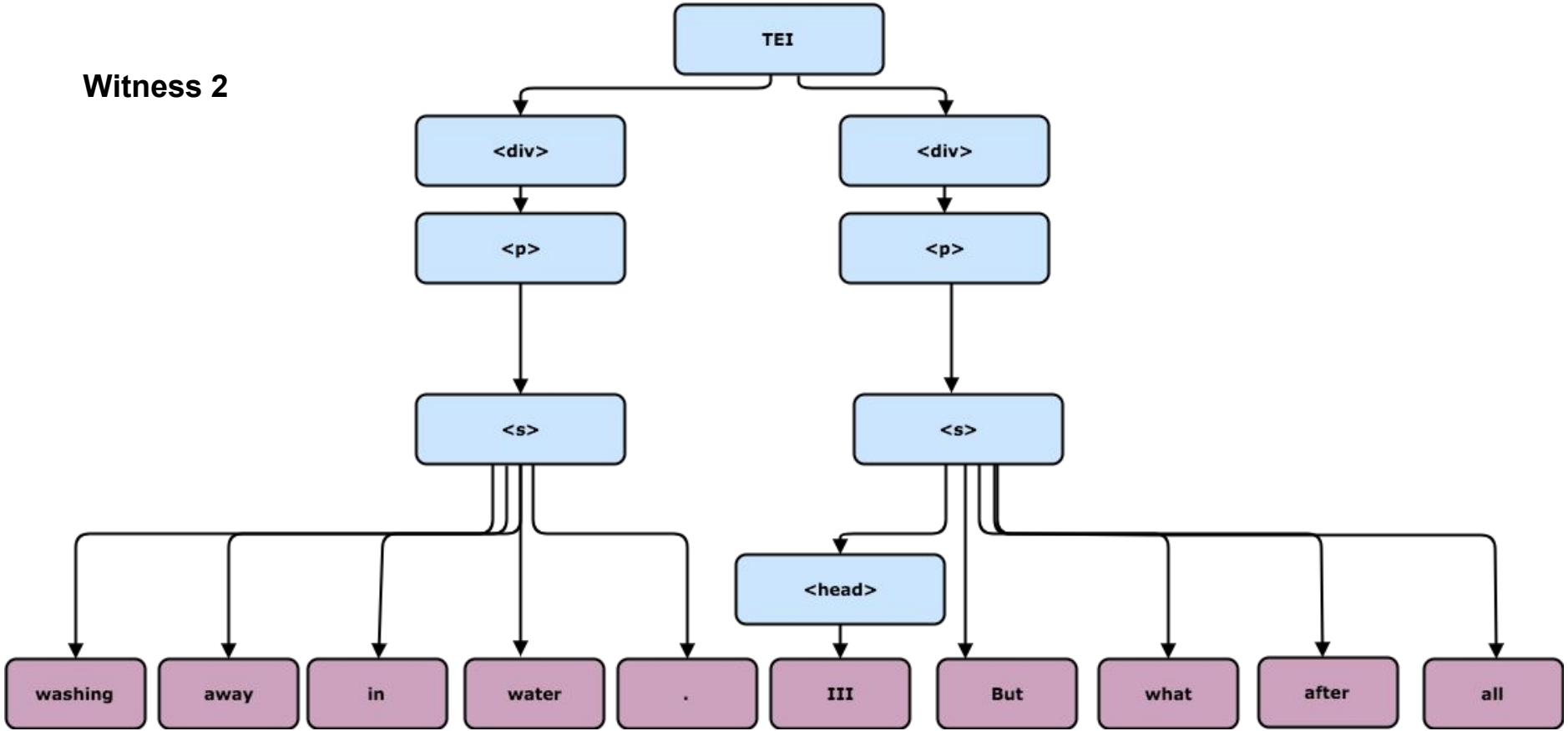
```
<?xml version="1.0" encoding="UTF-8"?>
<TEI>
    <div type="chapter" n="2">
        <p><s>washing away in water.</s></p>
    </div>
    <div type="chapter" n="3">
        <head>III</head>
        <p><s>But what after all, is one night?</s></p>
    </div>
</TEI>
```

(simplified) TEI-XML transcription of typescript p.5

Witness 1

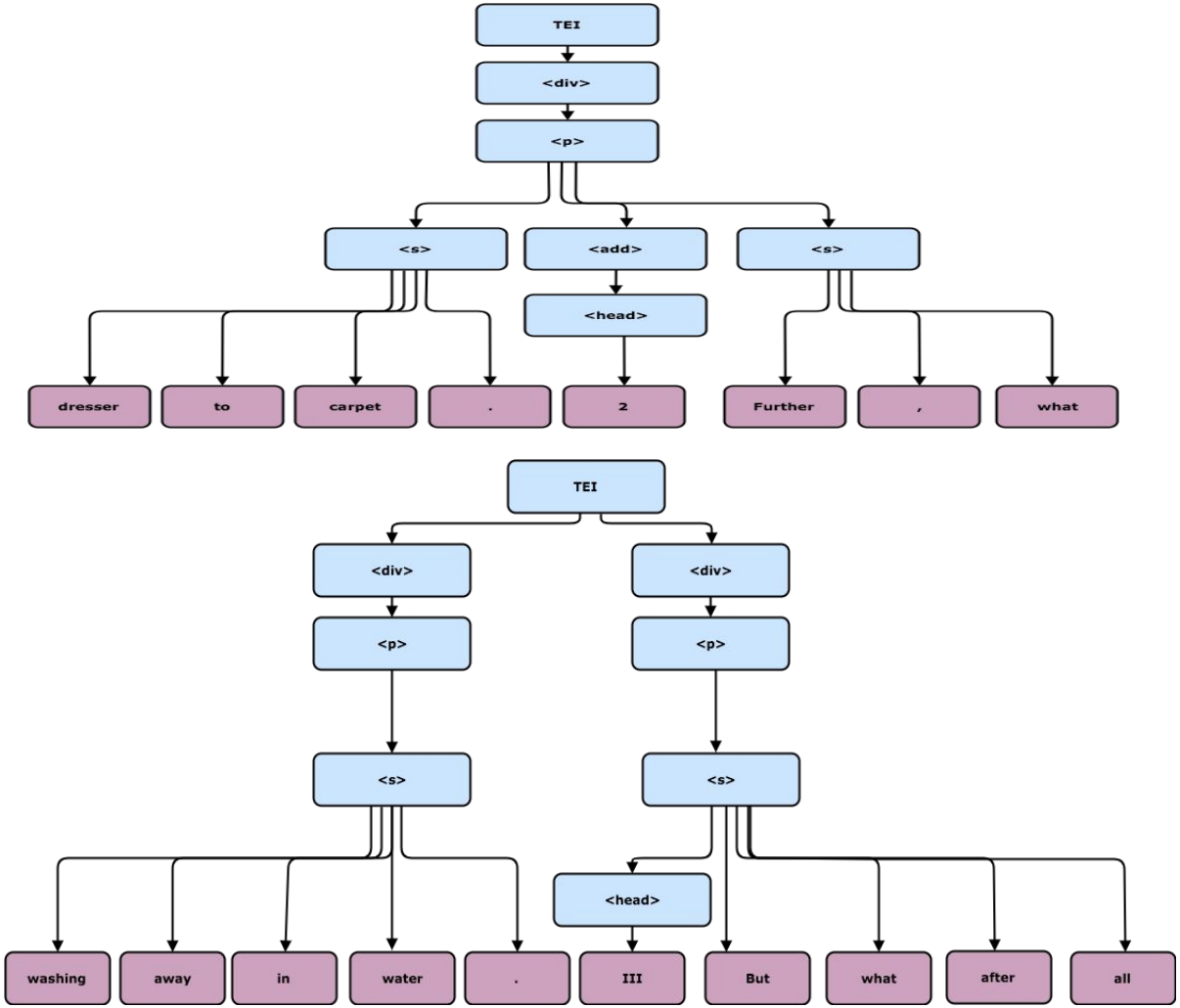


Witness 2



Witness 1

Witness 2



Modelling Texts in the Humanities

Beyond OHCO:

- Intradocumentary variation ⟶ multiple paths through one text, encoded with TEI tags
- Structure ⟶ comparing different documents with different structures results in conflicting hierarchies

Conditions

Different types of XML data:

- fully ordered
- unordered

Text-centric XML of literary texts (in short: "TEI text") = partially ordered data

Functional Requirements

Processing and analysing TEI text in a native way

Automated collation: finding the minimum set of changes needed to turn one document into the other.

Outset:

- documents are partially ordered
- treat structure and text equally
- compare more than 2 documents

Technical Requirements

Data model:

- schema-aware
- store multiple hierarchies

Processing:

- find the smallest amount of differences between two TEI texts
- distinguish between words, punctuation, markup
- respect non-linearity of TEI text

HyperCollate

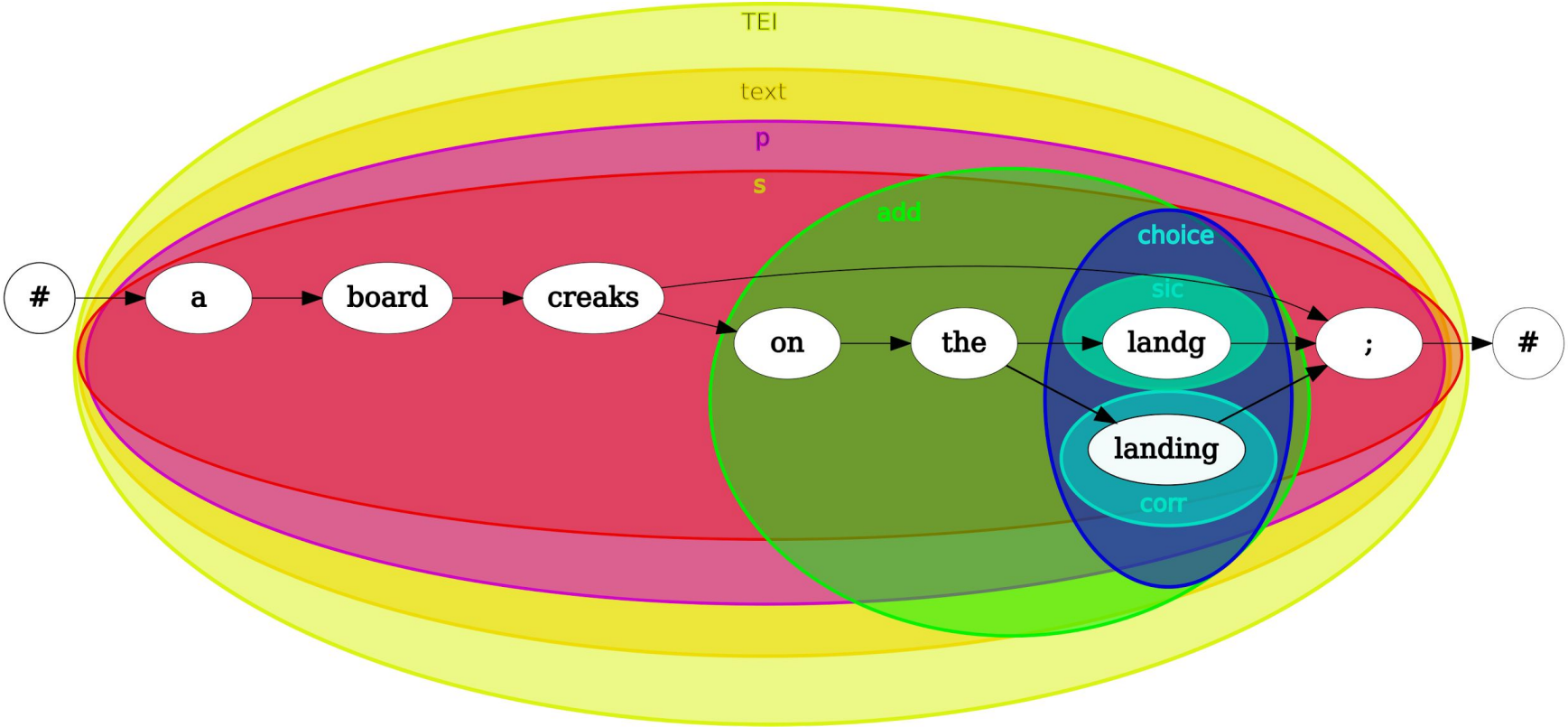
HyperCollate = automated collation tool that uses a hypergraph model for variation

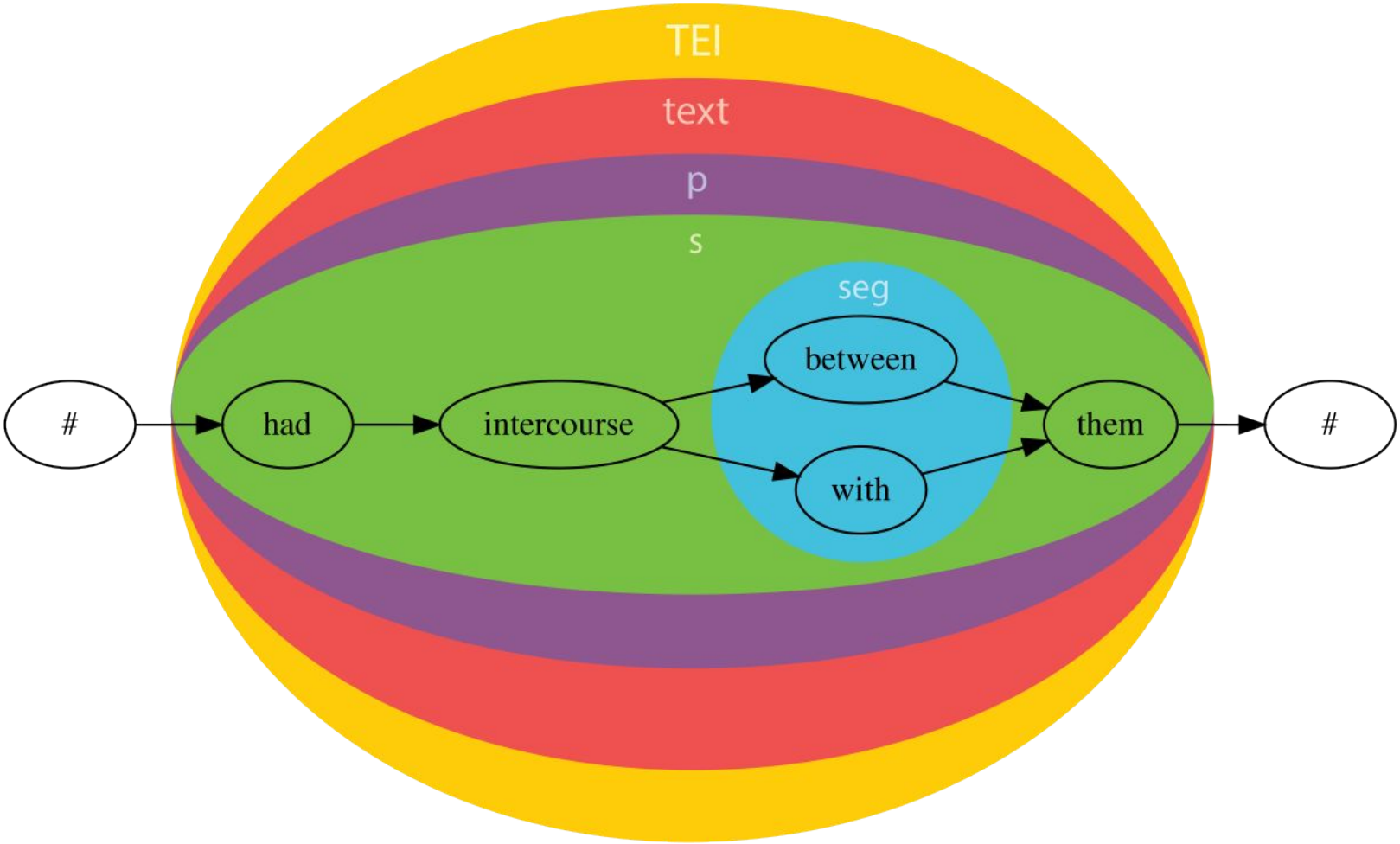
Hypergraph: nodes, hyperedges (one-to-one; one-to-many; many-to-many)

HyperCollate processes complex features of text, including:

- overlapping hierarchies
- intradocumentary variation

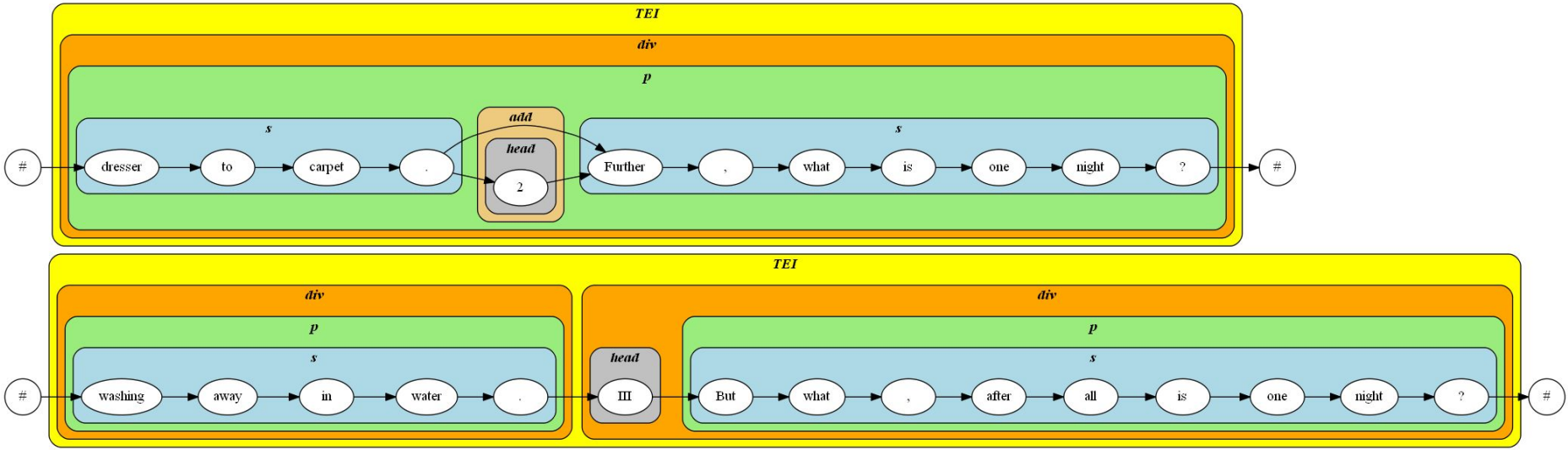
Hypergraph Model for Textual Variation



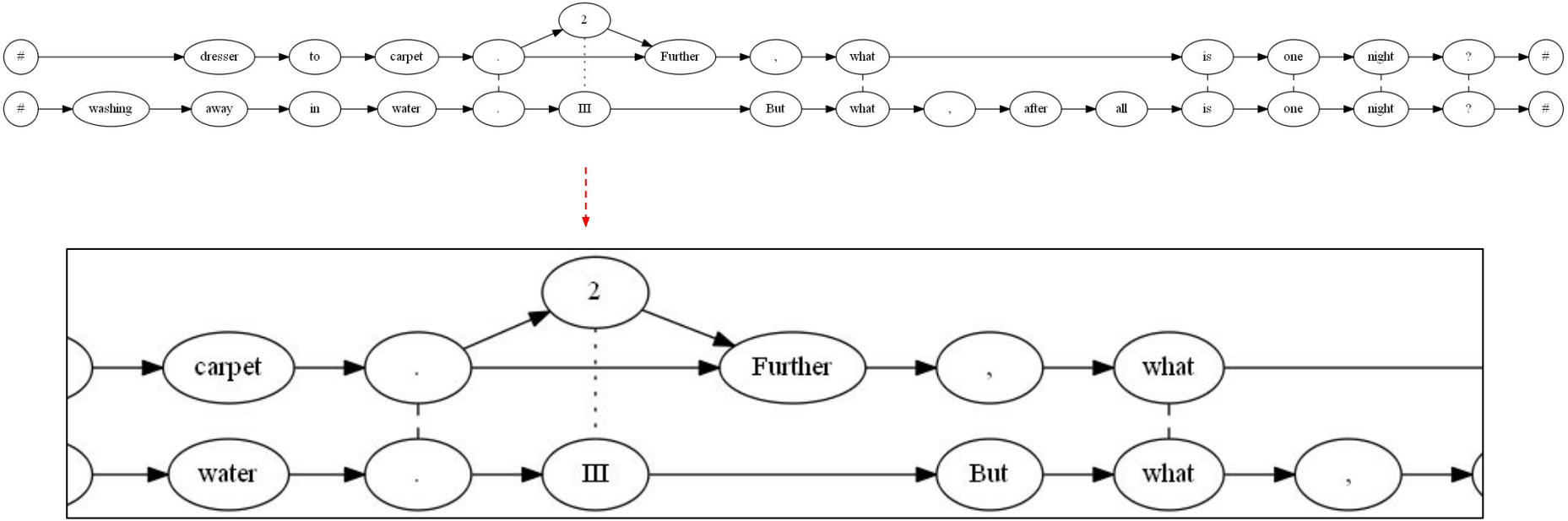

TEI
text
p
s
seg
#
had
intercourse
between
with
them
#

Steps of HyperCollate

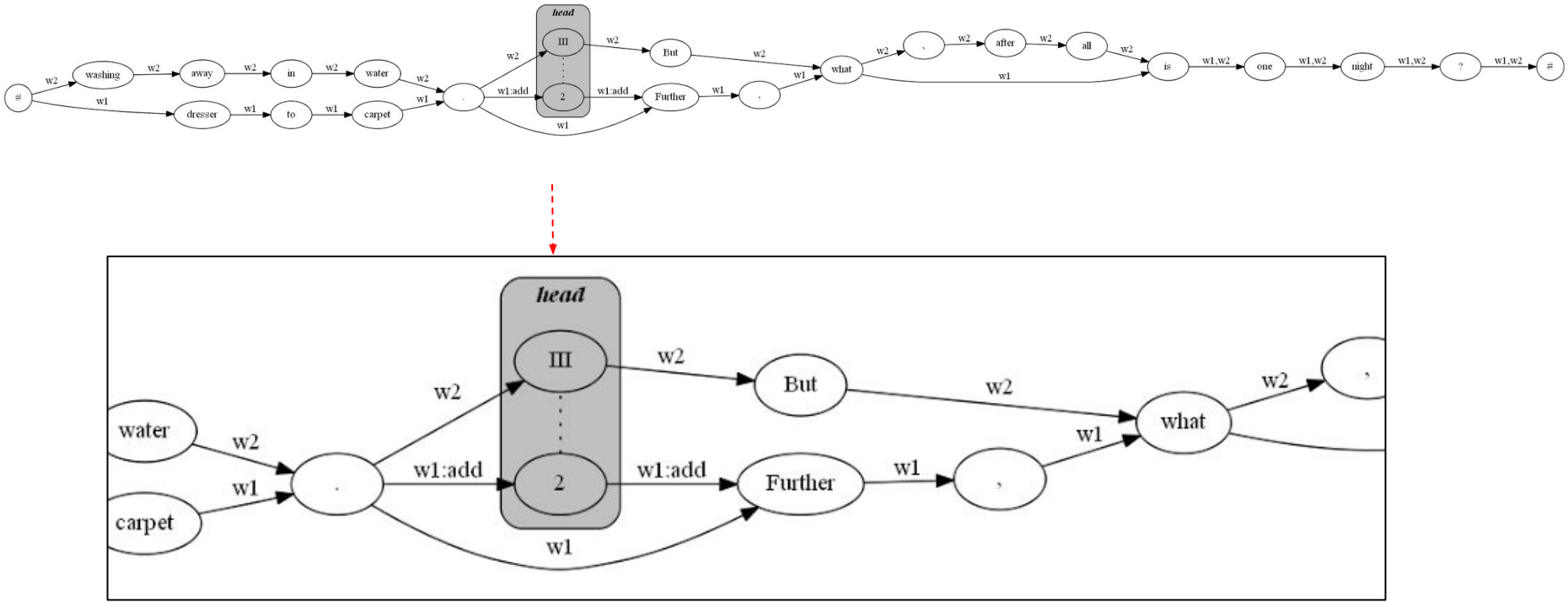
1. Convert XML tree of each witness into a variant hypergraph
2. Align two variant hypergraphs
3. Merge two hypergraphs in one collation hypergraph
4. Repeat in case of >2 witnesses
5. Visualise/export collation hypergraph



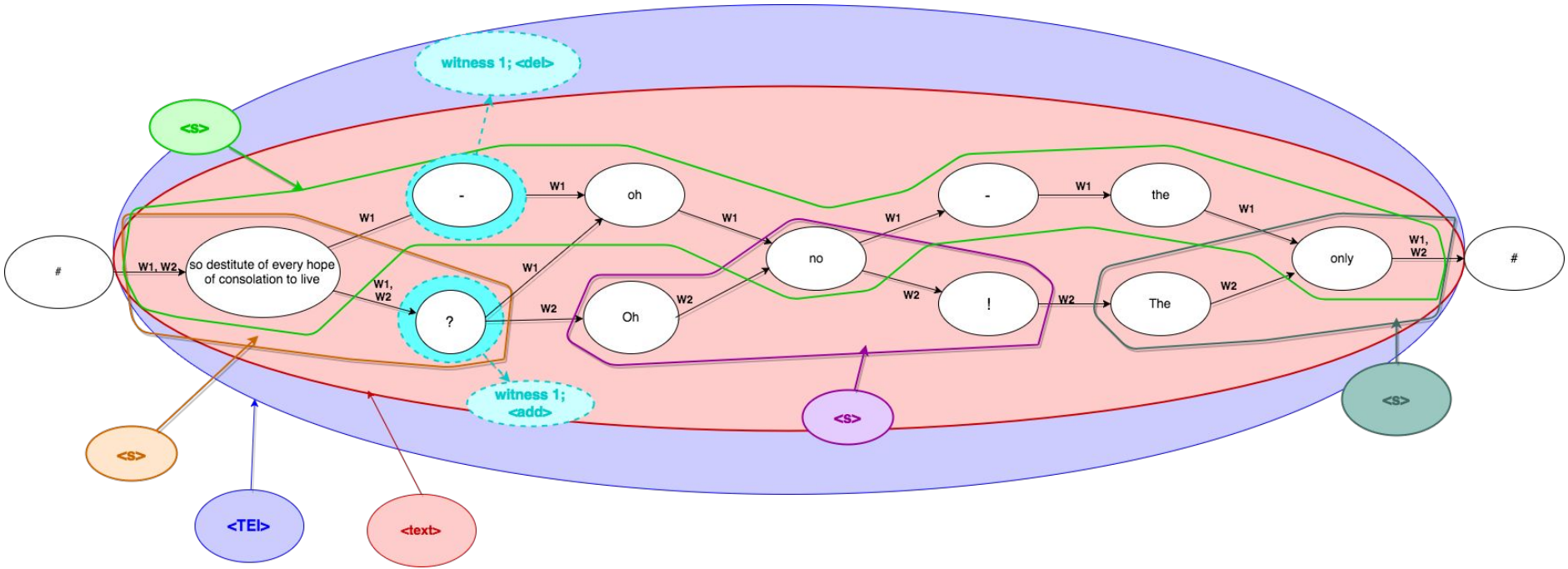
1. Transform TEI text witnesses into variant hypergraphs



2. Align the variant hypergraphs



3. Merge variant hypergraphs into one collation hypergraph


witness 1; <del>
<s>
-
W1
oh
W1
no
W1
-
W1
the
W1
only
W1, W2
#
W1
W1, W2
so destitute of every hope of consolation to live
W1, W2
?
W1
W2
Oh
W2
W1
W2
!
W2
The
W2
#
witness 1; <add>
<s>
<s>
<s>
<TEI>
<text>

Conclusion

TEI text, being partially ordered data, is especially challenging for processing.

Requirements for analysis of textual variation:

- process multiple paths (i.e. be schema-aware)
- store multiple hierarchies

A hypergraph data model allows us to make optimal use of XML's potential for humanities research.

Elli Bleeker, Bram Buitendijk, Ronald Haentjens Dekker, Astrid Kulsdom

R&D - Huygens Institute for the History of the Netherlands

KNAW Humanities Cluster

@ellibleeker
@ronald_dekker
@bram_buitendijk