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

Where many- voices where this the must un my h Within . although he memory them aut that curse! for ye all Mean me shead 1. Veriu, form the Mountains . Shi the hundred there and years Ver the lasth quakes couch we stand; Oft as men convator with fear Wa in our muthitude. La Voie from the Spings " hundler both had barches

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

Whose many- voices whoes this the mist dit, flung the themales of that shell ics Unings stagnant with winkling find Shuddering through India; the Mun Was hom burn a VALIA umm Thin my f. the orded I Am chan Withen, although the memory be hate - let them gut love it frees! was that curse? for ye all hear on speed 1. Vein, form the Mountains . Imi the hindred there and year Vir the lasth quakes couch we stand; oft as men converted with fear We trem Red in our month tude. La Voie from the Spings I hunderboth had harder our water had been stained with

1. What do we want to encode?

Textual characteristics

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

Vol.TE Chap I "It is with Sifficulty that remember the cora of my to being. all the events of that period appear confused & indictinet. Anone only that I felt a strong a stronge demation see seized me and a it append I vaw, gett heard and milt at the same time and it was inseed a long time be fore I harnes to Distinguish between the minu poperations of my various serves. By Buges " remember a stronger light walked and mis nerves and rothant I was offiger to close my 13 s. Ewohnich then came over me and troubled one. But hardly had Ifelt This When I by opening my eyes as I note surprise The light horoced in whom me again I water and "believer legended; but presents gound a great difference in my serve ations; before back opaque bodies had surrounded me mpervious to my bruch a sight and now found that I could wander on at When with no obstacles which I could not either owmount or avoid the light to became more operfrive to me and the heat searging The as Swather Soonght a place shere Scould hereence shake. This was the freet near Saytstart and here to the perty that hourd I has the Hunder and thoust. This couver me from my nearly

Soundenstein the particulars of his formation but on this foint he was impenetrable oreg mar, my friend, with he, or whither bols gown Unsettpoursity lead son - would Soulalso oracte for sourcellos the world a damoniag eneny or to what Jose your questions tent Peace, neace learn my miseries Do not seek to encrease your own Frankenstin Dis lovered that ilitan Sormade note uncerning his and and and asher trace them & himse / worder them in many places but Juncifally in going giving the life & spirit of the the conver pations he held with his enemes the you have made an account found not that a mutilated one for down to Thus have ten bage haped was postente While Shave histories to the strangest tale that ever imagination formed. my thought & every feeling of my coul Whiten Jurk by by the interrest wind the fill the my quest worked to write him. own clevals those asing contormable swelt on hover must man so distitute of years hope of consolations we the no the only jog he sends was in composing his shattered menters created peace & death I her ma feeling he enjoy , the comfort, the second of the prostile V denin from the Allams he have helieved that boken invaluation miseries or inclusion factor himse to very

Bodleian Library, University of Oxford, 2011 MS, Abinger c, 57, fol, 1 © Bodleian Library, University of Oxford, 2011 MS. Abinger c. 57, fol. 85r

Grouped revision

Wenbody in the house made mock y him Ecchet the we

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

Immediate revision

January 15th 1926 their the idea has gown in the marker is January 15th 1926 their the idea has gown in the Infortant. is una) worth the beginning. It have the important full-thethen, undeflerentiated, there is the the

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

Open variants

That is an easy thing to say.

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

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 ..."

Discontinuous structures

het keizerlijk huis geleverd werden.

In den namiddag, wanneer de felste hitte voorbij was, gwam doorhown de luistille gangen van het paleis de uitverkorene, klopte bescheiden op de deur van de kamer waar de sultan zijn verveling geeuwde. Zooals de mannequin in een Parijsch modehuis 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 zity : " Roxe Verrassing ", " Blosembot " of " voor uw genoegen " - natuurlijk in het Arafbich. Menigvuldig en soms wel van een frissche fantasie getuigend waren de benamingen die het " Keizerlijk College der Wellustcommissarissen " Doch wat ze aankondigden kwem in den grond toch steeds op hetzelfde neer. Aldus was Shiriar mettertijd ook dit officieel geregeld minnespel beu geworden.

Inderdaad, wat is hiefde die niet voorafgegaan is door verlaiding ? Helaas hij, Shirkr, Afstammeling der Profeten, krijgt den wellust voorgeschoteld als een dagelijkschen kost, wearvoor hij dan nog geen apbeid verricht heeft. Hee soet moet oontans zijn dit wervicht vervient heeft. Hee soet moet oontans zijn dit wervien en een vrouw ! Die dagen van nerveuse verwachting voor de liefelijke toestomming.

Shiriar donkt aan de proldvonvrijagies in de stoegjes/van Beging on op de hellingen der wallen? En dan die Herinnering aan verleden somer, toen hij in

- 3 -

gezelsehap van den visier van Landbouw een inspectietoond deed in de streek van den Euphrates. Een ambtenaar-gez agronoom leerde hem juist kanvarantikanvaranteidanv tarwe onderscheiden van vogge en haver als ze plots een jonge boer en boerin invest ontdekten die zich in het koren liefkoosden. Als een opgeschrokken koppel patrijzen waren ze voor de statige heeren op de vlucht geslagte. De sultan has nog net gesien hoe, al loopendt, do blozende deerne de zware naakte state borsten weer in haar jak wogstopte.

Waarachtig die harfoliefde in doze kamer van hussens en boeken kon den vorst min oh min bekoren en, als een kind dat zijn zoetigheid verwend is, begon hij de odalisken onverrichter zake terug te sturen. De aldus geaffrenteerde meisjes slakiten bitterd klachten en het kainarligk College der fellusjoomsissarfissen vroeg zich af hee se den sultan weer op het pad der traditionneele keiserlijke obtiken zouden brengen.

Zij lieton exotische vrouwen komen : Spaansche Coportituis mission uit eigersteanfabrisken waarvan de huid okerblond is als de tabak die ze verwerken, Koersche walkdrengestalten, schoon en koel als marmerbeelden, Geiahas uit/Yoshiwara die een heele documentatie van technisch voortreffèlijk erotische houteneden meebrachten. Het kon al niet veel baten.

De pellusformissarissen noodigden de jongste " Miss Universe " uit. Doch daar ze vooreerst nog de

- 4 -

seen bunch up by the interve of for my quest while this tale and his own elevated and gentle mannes 156 have meated. Swich to roothe him get cannot as counsel one so infinitely miserable, of destitute.

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

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)

with handovers & workarounds	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								

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

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."

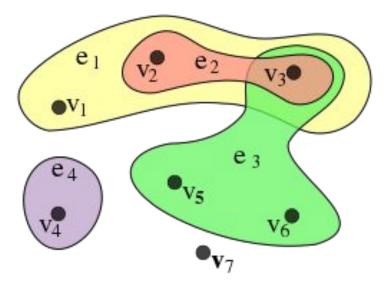
2. TAG data model

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

81

(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

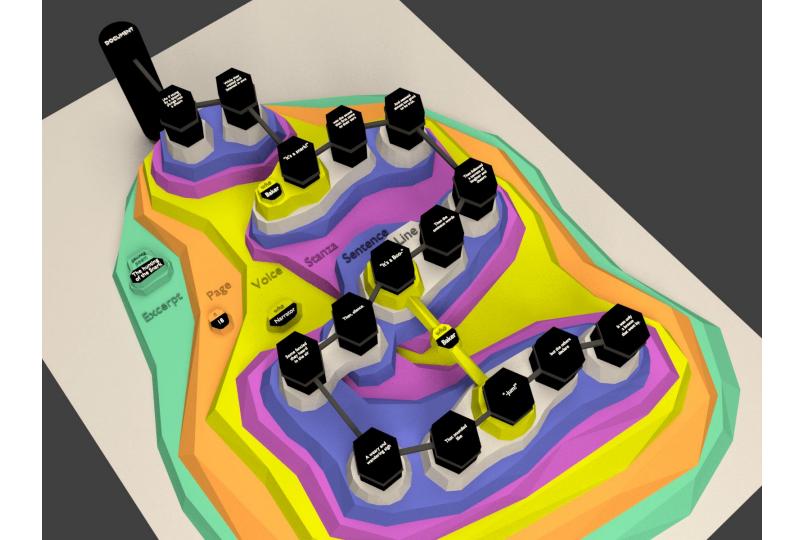
A weary and wandering sigh That sounded like "-jum!" but the others declare

It was only a breeze that went by,

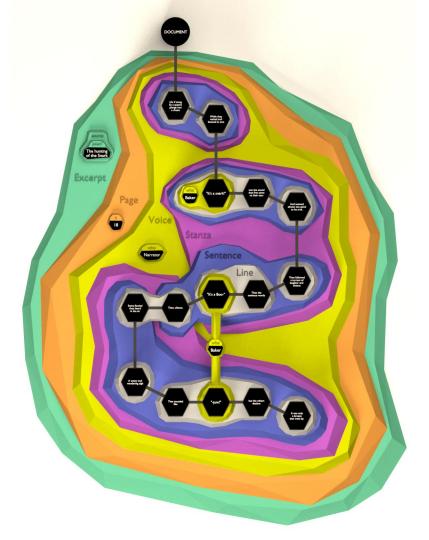
Univ Calif - Digitized by Microsoft ®

Source: Lewis Carroll 1876. *The Hunting of the Snark: an Agony in Eight Fits*, page 81. Available via <u>https://en.wikisource.org/wiki/The</u>

<u>https://en.wikisource.org/wiki/The</u> Hunting of the Snark







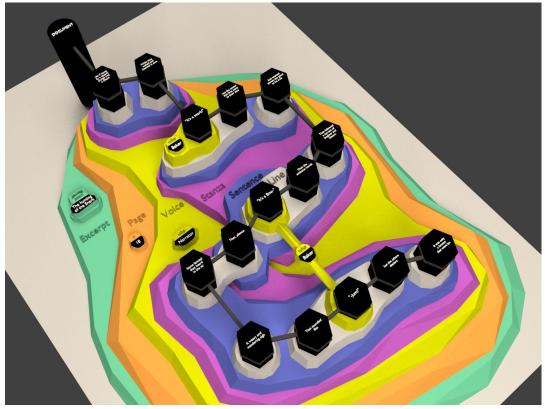
2. TAG - hypergraph data model for text

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



Source: Haentjens Dekker and Birnbaum 2017, 30 https://doi.org/10.4242/BalisageVol19.Dekker01 (figure by Gijsjan Brouwer)

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

Shipe divinin difference of opinion, impaniable havies pregnotices,

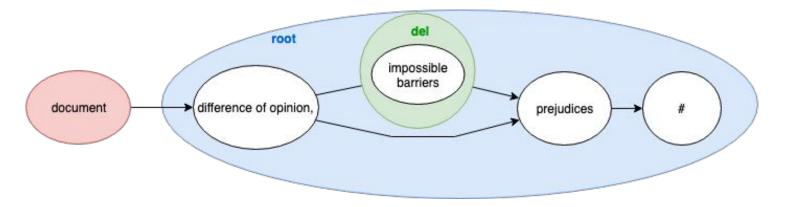
... difference of opinion, impossible barriers, prejudices...

(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, [?del>impossible barriers<?del], prejudices <root]</pre>

TAGML hypergraph



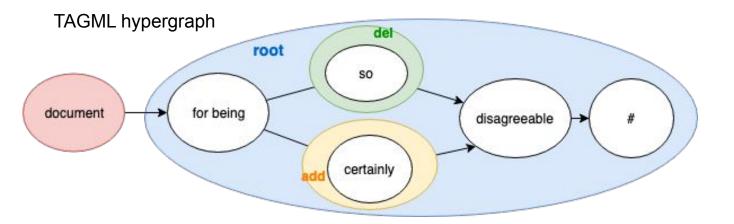
2. Grouped revision

Ramsdy in the hour made mock shing Ecchet the w.

... for being so certainly disagreeable ...

2. Grouped revision

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

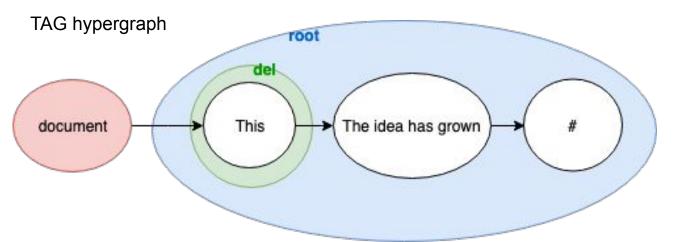


2. Immediate revision

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

2. Immediate revision

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



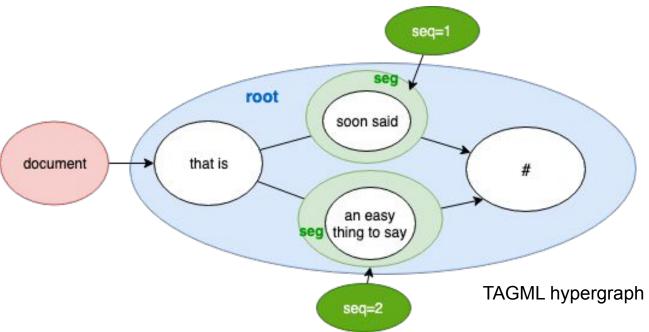
2. Open variants

That is an easy thing to say.

Source: the BDMP encoding manual (<<u>http://uahost.uantwerpen.be/bdmp/</u>>, 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]</pre>



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 wor-

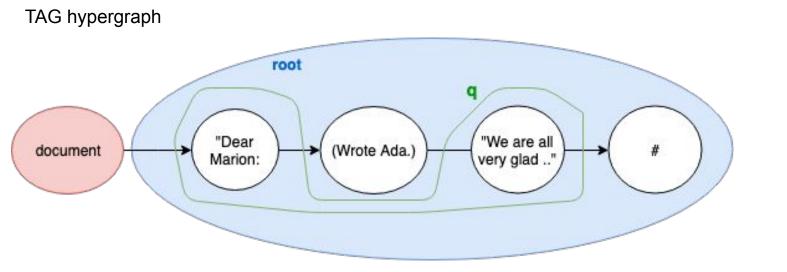
ried. There is not much money coming in. I am doing all I can to help, and I gave up a good position offered me by the C. P. R. to travel over their Western lines and write travel pamphlets, because I will not leave mama just now.

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

"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]</pre>

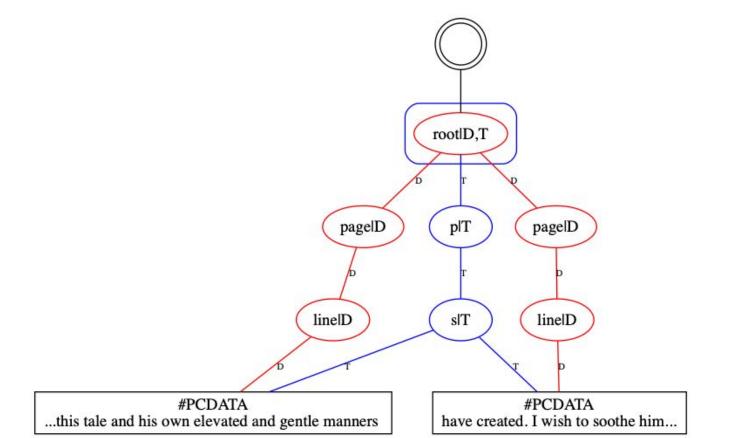


2. Overlapping structures

seen dunk up by the interve of for my quest while this tale and his own elevated and gentle mannes 156 have meated. Swich to roothe him get cannot as

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

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"...

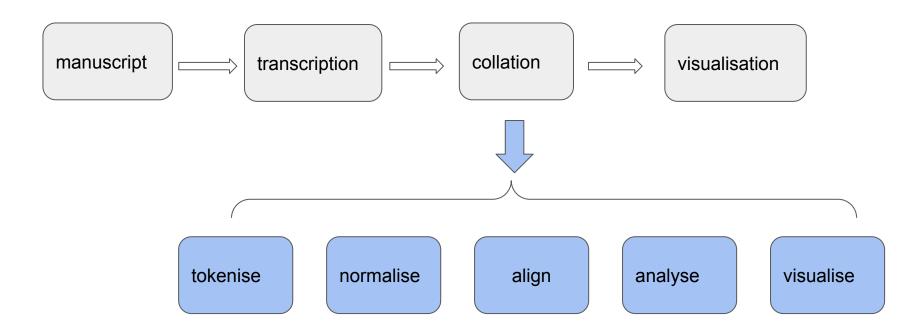
Semi-automated collation tools:

- 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?

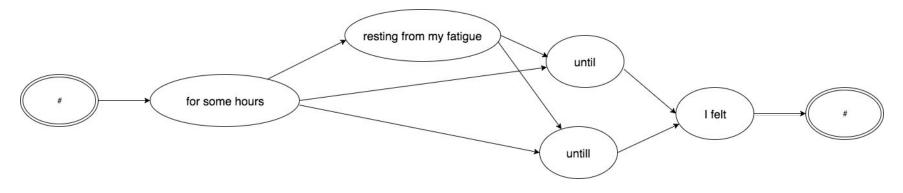
- Textual variation within one witness multiple paths through one text, encoded tags with divergence and converge markup tags.
- Structure comparing different documents with different structures results in conflicting hierarchies

3. Example of multiple paths

for some hours from my fatigue tomention

"... 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

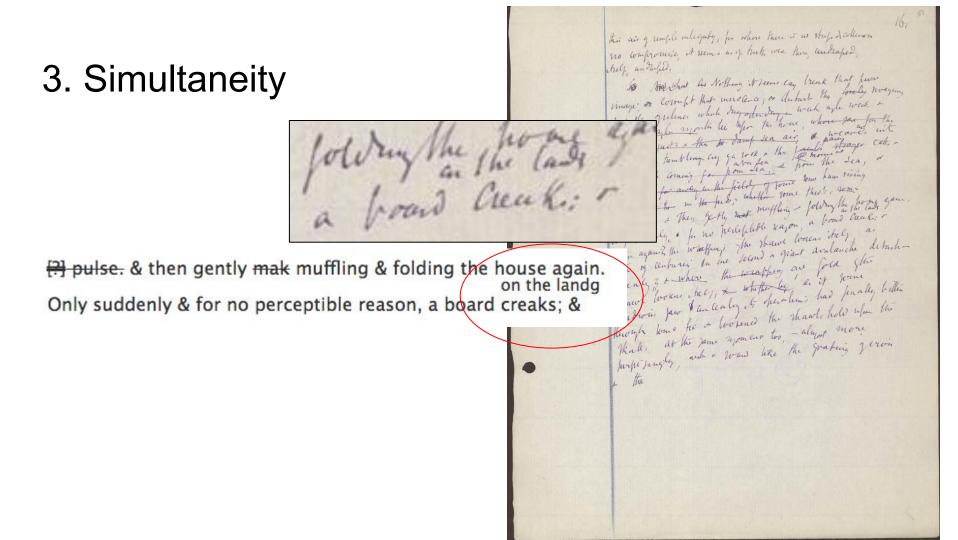
3. Processing complex texts

Technical requirements

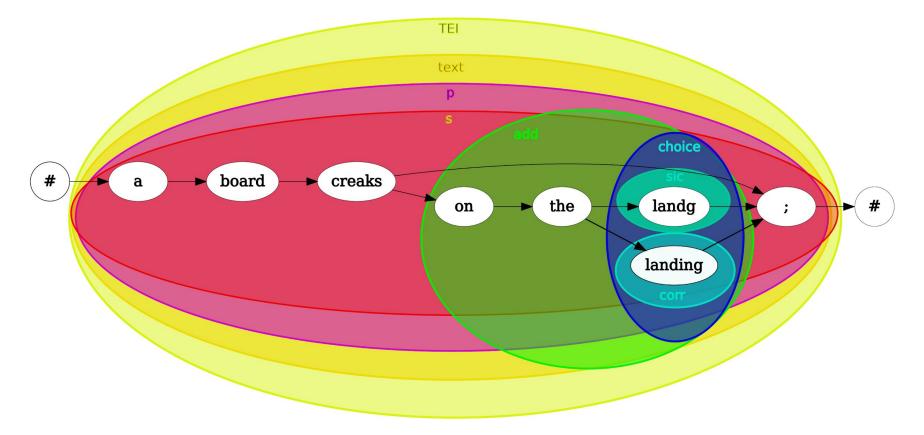
Collation tool

- 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



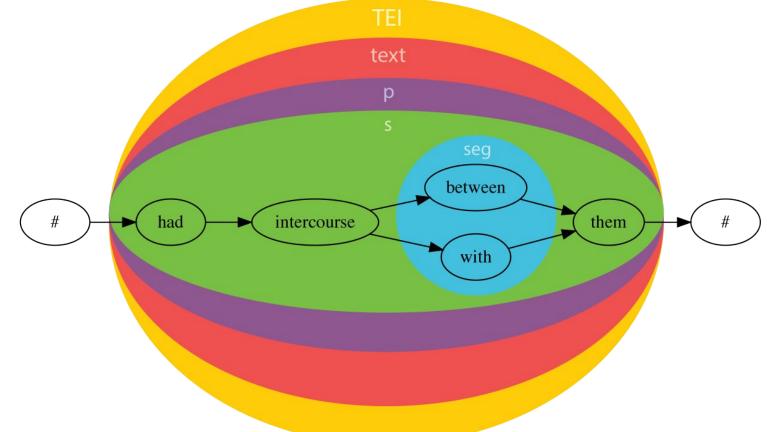


3. Hypergraph Model for Textual Variation



- would it land have there wond at Tak? Then, furthey branchery the wolls, we suggest from by on they period, an it finthy too recking, and force officialty that to cover to yellow a the proneyuchly thou 3. Open Variants iey showed (justly 100, for shows a left september night) Tom Judies in the racket , as if gal this in the Freudeny lower, nos that they were the fughte were shepeny had were then to them has intertourse with them - & the want & the wantchefer a three but hilvery gaucifian a on the public to santer , & china; tapet, hade louble, howing but here or file or Terragly. Inless now that they were the people were 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..."

some perations of my various senses. By degrees " remember a stronger light proper when my nerves and sothat I was alliget to close "Is eses. Farknip than came over me and troubled me. But hardly had Ifelt This When 1 by opening my eyes as I note scripting) the light horoced in whom me again wather and the here legended; but presenth found a great difference in my sensations; before book opaque bodies had sworounded me mpervious to my touch a sight and now found that could wander on at liberty my fato que

hunger and thout. This coused me from my nearly hereene shafe. This was the forest near Inglotant for some hours from my fatigues tormention for some . hours antil turyer and thoust. This courses me from my rearly

for some hours untill Shelt tormentides

3. HyperCollate

Witness B

86

FRANKENSTEIN; OR,

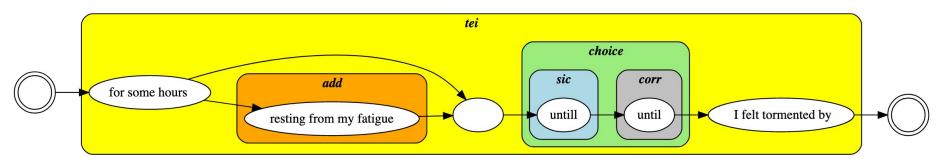
ation or denial of this opinion. For the first time, also, I felt what the duties of a creator towards his creature were, and that I ought to render him happy before I complained of his wickedness. These motives urged me to comply with his demand. We crossed the icc, therefore, and ascended the opposite 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.

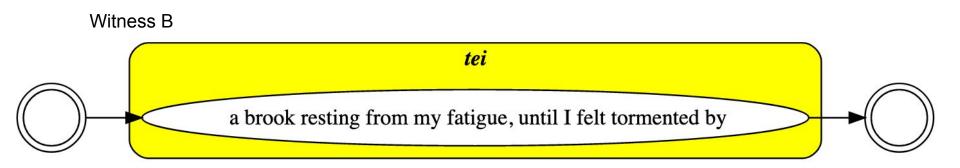
and here I lay by the side of a brook resting from my fatigue, until I felt tormented by hunger and thirst. This roused me from my nearly dormant state, and I ate some berries which I found hanging on the trees, or lying on the

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

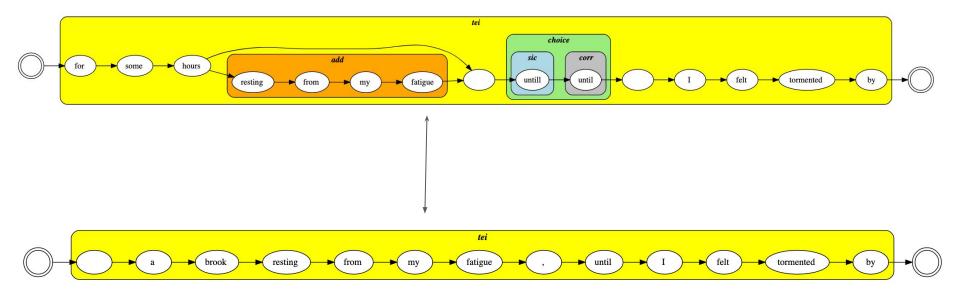
I now found that I could wander on at liberty, with no obstacles which I could not either surmount or avoid. The light became more and more oppressive to me; and, the heat wearying me as I walked, I sought a place where I could receive shade. This was the forest near Ingolstadt; and here I lay by the 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 herries which I found hanging on the trees, or lying on the

Witness A



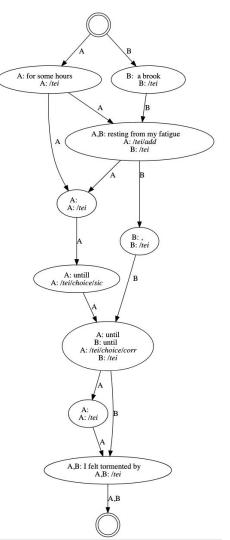


1. Transform TEI text witnesses into variant hypergraphs



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

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