CollateX

DiXiT Camp 2 – Graz

Ronald Haentjens Dekker

ronald.dekker@huygens.knaw.nl

19-09-2014

Collation 1

Compare multiple witnesses against each other

- Finding differences and similarities between witnesses
- Additions, omissions, modifications, transpositions

Collation 2

Finding relations between witnesses

- Parallel segments
- Transposed segments

CollateX Features

- Compare multiple witnesses against each other
  - Baseless
  - Not just pairwise comparison
  - Order independent
- Supports multiple output formats
  - Can output a variant graph
  - Can output TEI Parallel Segmentation alike format
- CollateX is free software
  - source code is available
  - source code is hosted on github
- Available in Python (desktop) and Java (server and desktop)

Darwin Origin of Species 1859

WHEN we look to the individuals of the same variety or sub-variety of our older cultivated plants and animals, one of the first points which strikes us, is, that they generally differ much more from each other, than do the individuals of any one species or variety in a state of nature. When we reflect on the vast diversity of the plants and animals which have been cultivated, and which have varied during

Darwin Origin of Species 1869

Causes of Variability. WHEN we compare the individuals of the same variety or sub-variety of our older cultivated plants and animals, one of the first points which strikes us is, that they generally differ from each other more than do the individuals of any one species or variety in a state of nature. And if we reflect on the vast diversity of the plants and animals which have been cultivated,

1859	1860	1861	1866	1869	1872
-	-	-	Causes of Variability.	Causes of Variability.	Causes of Variability.
WHEN we	WHEN we	WHEN we	WHEN we	WHEN we	WHEN we
look to	look to	look to	look to	compare	compare
the individuals of the same variety or sub- variety of our older cultivated plants and animals, one of the first points which strikes us	the individuals of the same variety or sub- variety of our older cultivated plants and animals, one of the first points which strikes us	the individuals of the same variety or sub- variety of our older cultivated plants and animals, one of the first points which strikes us	the individuals of the same variety or sub- variety of our older cultivated plants and animals, one of the first points which strikes us	the individuals of the same variety or sub- variety of our older cultivated plants and animals, one of the first points which strikes us	the individuals of the same variety or sub- variety of our older cultivated plants and animals, one of the first points which strikes us
,	,	,	,	-	-
is, that they generally differ	is, that they generally differ	is, that they generally differ	is, that they generally differ	is, that they generally differ	is, that they generally differ
much	-	-	-	-	-
more	more	more	more	-	more
from each other	from each other	from each other	from each other	from each other	from each other
,	-	-	-	more	-
than do the individuals of any one species or variety in a state of nature.	than do the individuals of any one species or variety in a state of nature.	than do the individuals of any one species or variety in a state of nature.	than do the individuals of any one species or variety in a state of nature.	than do the individuals of any one species or variety in a state of nature.	than do the individuals of any one species or variety in a state of nature.
When	When	When	When	And if	And if
we reflect on the vast diversity of	we reflect on the vast diversity of	we reflect on the vast diversity of	we reflect on the vast diversity of	we reflect on the vast diversity of	we reflect on the vast diversity of

How to install

- Install Python 2.7
  - Included out of the box on Mac OS X and Linux
  - Windows: Active Python
- Install PIP (python package manager)
  - $ sudo easy_install pip
- Install CollateX
  - $ sudo pip install --pre collatex
- or upgrade:
  - $ sudo pip upgrade --pre collatex

How to use it

```
from collatex import *
collation = Collation()
collation.add_witness("A", "The quick brown fox jumps over the dog.")
collation.add_witness("B", "The brown fox jumps over the lazy dog.")
collate(collation)
```

How to use it (continued)

```
from collatex import *
collation = Collation()
collation.add_witness("A", "The quick brown fox jumps over the dog.")
collation.add_witness("B", "The brown fox jumps over the lazy dog.")
collate(collation)
```

A	The	quick	brown fox jumps over the	-	dog.
B	The	-	brown fox jumps over the	lazy	dog.

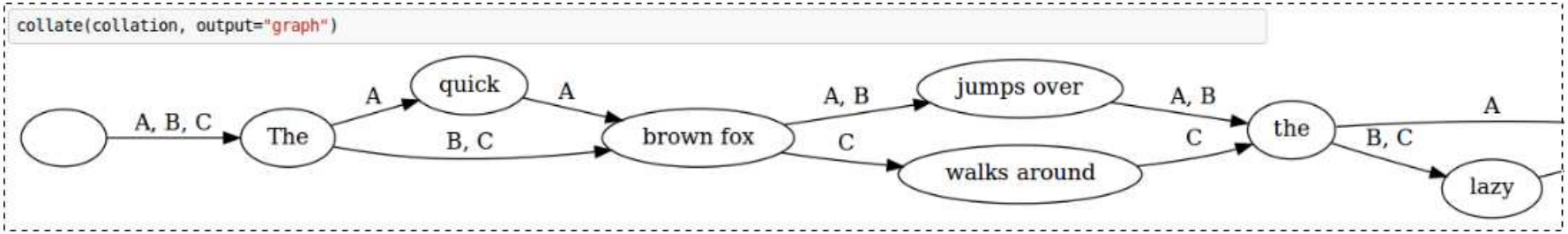
How to use it (continued)

```
collation.add_witness("C", "The brown fox walks around the lazy dog.")
collate(collation)
```

A	The	quick	brown fox	jumps over	the	-	dog.
B	The	-	brown fox	jumps over	the	lazy	dog.
C	The	-	brown fox	walks around	the	lazy	dog.

How to use it (continued)

```
collate(collation, output="graph")
```



How to deal with markup

<l 249.01> [init]D[/init]eR valscheite widr sazz.
<l 249.02> cherte v[sup]^[/sup]f der hv[sup]o[/sup]fslege chrazz.
<l 249.03> sin sceiden dan daz riwet mich.
<l 249.04> alrest nv Aventiwertez sich.
<l 249.05> do begvnde chrenchen sich ir spor.
<l 249.06> sich scieden di da riten vor.
<l 249.07> ir sla wart smal div ê was breit.
<l 249.08> er verlosse gar daz was im leit.
<l 249.09> mære vriesch do der ivnge man.
<l 249.10> da von er herce not gewan.
<l 249.11> [minit]D[/minit]o erhorte der degen ellens rich.
<l 249.12> einer fro[sup]v[/sup]wen stimme iœmerlich.
<l 249.13> ez was dennoch von to[sup]v[/sup]we naz.
<l 249.14> vor im v[sup]^[/sup]f einer linden. saz.
<l 249.15> ein magt der fv[sup]o[/sup]gte ir triwe not.
<l 249.16> ein gebalsemt rittr tot.
<l 249.17> lent ir zwiscen den armn.
<l 249.18> swenz niht wolt erbarmn.
<l 249.19> der si so sizzen sœhe.
<l 249.20> vntriwen ich im iœhe.
<l 249.21> [minit]S[/minit]in ors do gein ir wante.
<l 249.22> der wenich si bechante.
<l 249.23> si was doch siner mv[sup]o[/sup]men kint.

How to deal with markup (continued)

```
BERSTE .HOOFDSTUK <b>Telehaka</b>
&WR+
<i>In law an &APO+infant, and in years a boy</i>, <i>In mind a slave to every vicious joy</i>;<p/>
<i>From every sense of shame and virtue wean&APO+d</i>;<p/>
<i>In lies an adept, in deceit a fiend</i>;<p/>
<i>Versed in hypocrisy, while yet a child</i>;<p/>
<i>Fickle as wind, of inclinations wild</i>;<p/>
<i>Woman his dupe, his heedless friend a tool</i>;<p/>
<i>Old in the world, though scarcely broke from school&APO+, Damaetas ran through all the maze of s
<i>Even still conflicting passions shake his soul, And bid him drain the dregs of pleasure&APO+s bo
<i>But, pali&APO+d with vice, he breaks his former chain, And what was once his bliss, appears his
&WR+
<i>BYRON</i>
&WR+
```

How to deal with markup (continued)

```
    than grey to enfold me.
  </seg>
▼<seg n="MS-HRC-SB-5-10,[0480]" zone="zonets_104" xml:id="inntsd0e5122" corresp="#20r">
    What
    <del type="crossOut" hand="#SB" resp="#SW" rend="blueblack ink">nonsense</del>
    <add rend="blueblack ink" resp="#SW" hand="#SB" place="supralinear">rubbish</add>
    <lb rend="it"/>
    all this
    <del type="crossOut" hand="#SB" resp="#SW" rend="blueblack ink">business</del>
    <add rend="blueblack ink" resp="#SW" hand="#SB" place="supralinear">stuff</add>
    about light and dark.
  </seg>
▼<seg n="MS-HRC-SB-5-10,[0481]" zone="zonets_104" xml:id="inntsd0e5137" corresp="#20r">
    And how I have
    <lb rend="it"/>
    wallowed in it.
```

Beckett

ET1, typed text:	'What nonsense all this business about light and dark.'
ET1, 1st revision:	'What nonsense all this ~~business~~ stuff about light and dark.'
EN1 (manuscript):	'What nonsense all this stuff about light and dark.'
ET1, 2nd revision:	'What ~~nonsense~~ rubbish all this ~~business~~ stuff about light and dark.'
ET2:	'What rubbish all this stuff about light and dark.'

Beckett (continued)

variant, invariant.
Doubleclick or select a word to highlight it in the Synoptic Sentence View.

☑ Version 1 ☑ Version 2 ☑ Version 3 ☑ Version 4 Recollate

Version 1: MS-HRC-SB-5-9-1	What	nonsense		all this		stuff about light and dark	!
Version 2: MS-HRC-SB-5-10	What	~~nonsense~~	rubbish	all this	~~business~~	stuff about light and dark	.
Version 3: MS-WU-MSS008-3-71	What		rubbish	all this		stuff about light and dark	.
Version 4: MS-1958	What		rubbish	all this		stuff about light and dark	.

Tokenization

Preparation

- tokenization
- regularization

"The same clock as when for example Magee once died."

Tokenization (continued)

"The same clock as when for example Magee once died."

Preparation

- tokenization
- regularization

|The|same|clock|as|when|for|example|Magee|once|
died|.|

Normalization

Preparation

- tokenization
- regularization

"The same clock as when for example Magee once died."

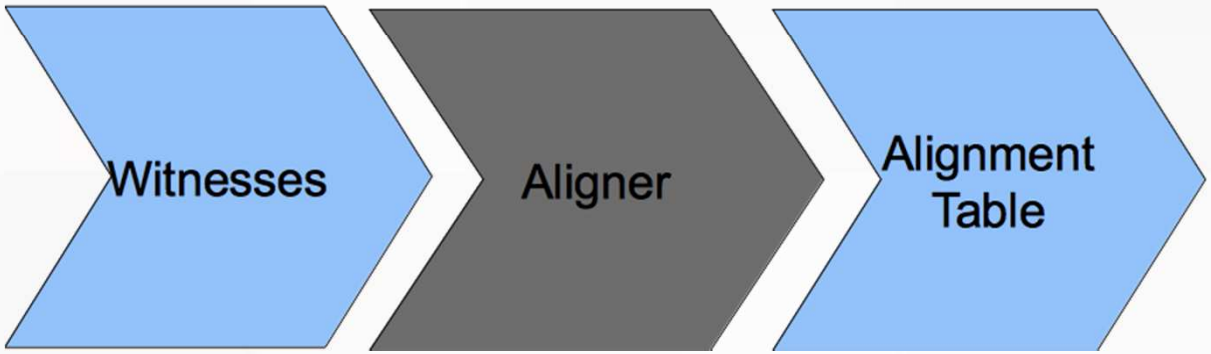
|The|same|clock|as|when|for|example|Magee|once|died.|

|the|same|clock|as|when|for|example|magee|once|died|.|

How it works

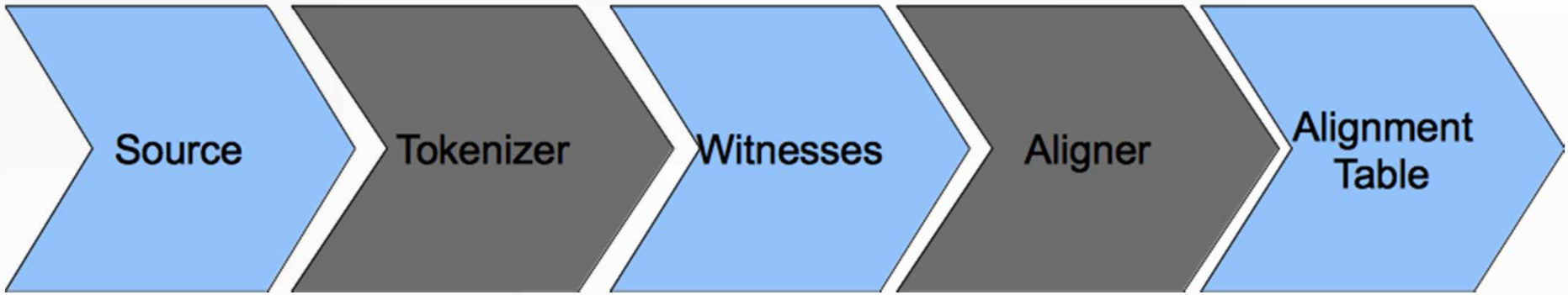
huygens ing

Basic pipeline for collation



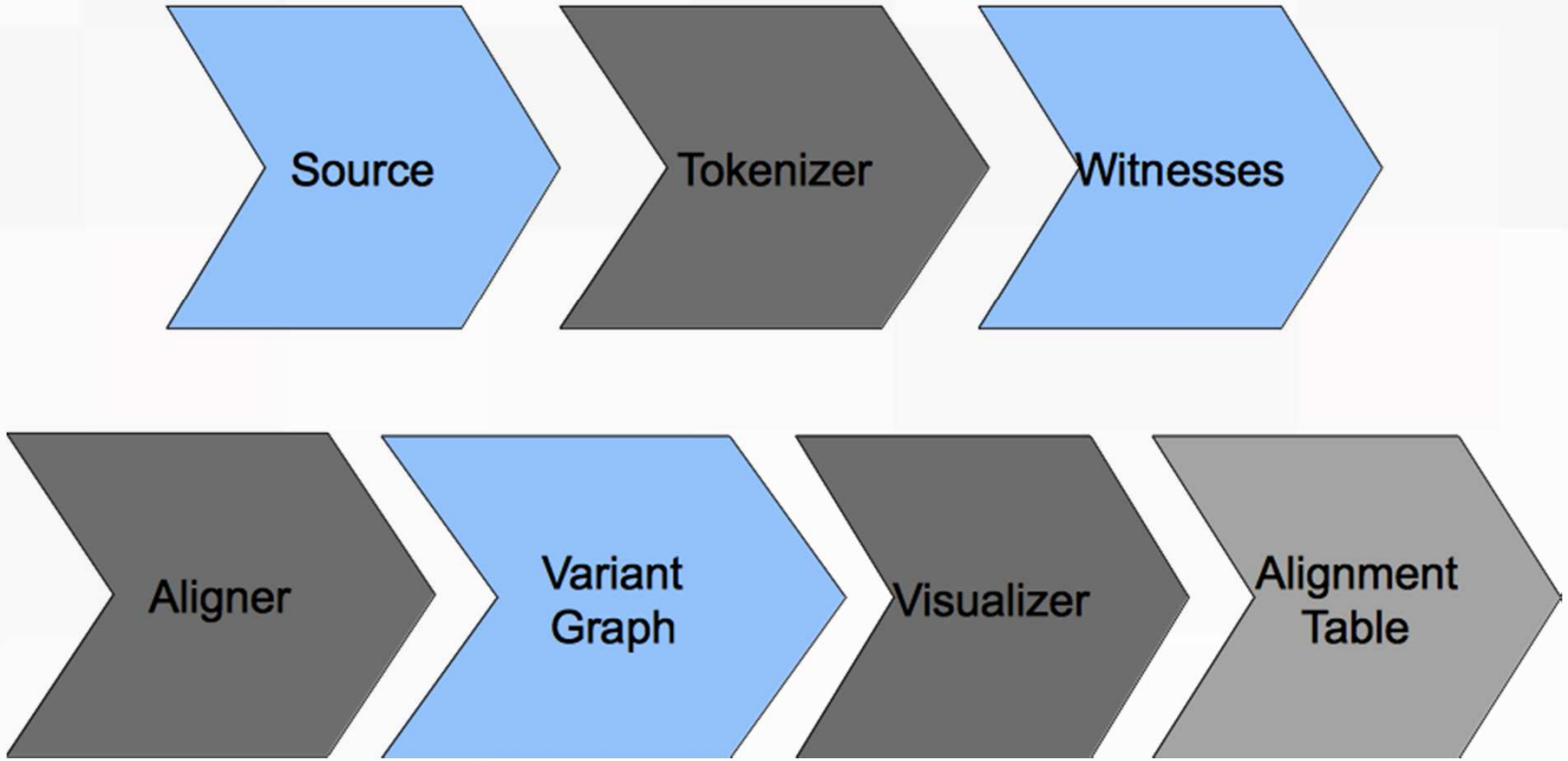
huygens ing

Extended pipeline with tokenizer

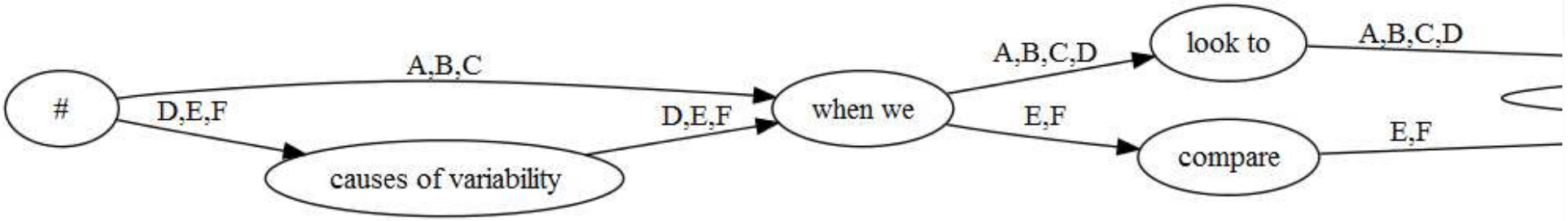


huygens ing

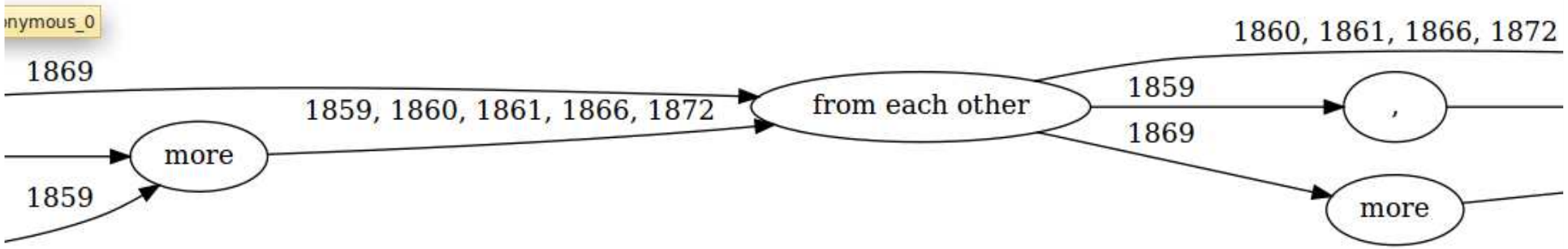
Extended pipeline with custom visualizer



How it works: variant graph



How it works: more -> from each other transposition



Visualization

```
A:  |the|same|clock|as|when|for|example|(...)
B:  |the|same|as|when|for|example|(...)
C:  |the|same|as|when|among|others|(...)
```

Visualization

- alignment table
- apparatus

```
|the same|clock|as when|for   example|  (...)
|the same|      |as when|for   example|  (...)
|the same|      |as when|among others |  (...)
```

Visualization (continued)

```
A:  |the|same|clock|as|when|for|example|(...)
B:  |the|same|as|when|for|example|(...)
C:  |the|same|as|when|among|others|(...)
```

Visualization

- alignment table
- apparatus

```
- <collatex:apparatus>
    the same
  - <app>
      <rdg wit="#A">clock</rdg>
      <rdg wit="#B #C"/>
    </app>
    as when
  - <app>
      <rdg wit="#A #B">for example</rdg>
      <rdg wit="#C">among others</rdg>
    </app>
  </collatex:apparatus>
```

Future

- In October a new version of CollateX will be released:

  CollateX 2.0

  Will be shown at the Collation workshop in Muenster, 3, 4 October 2014.

  Features non progressive multiple witness alignment

    Analyses all the witness before making alignment decisions.

  Features a scoring function

    Not only works with perfect matches, but also allows near matches (spelling variation) and synonyms.

Questions?