THE GOTHENBURG MODEL AND COLLATEX

OUTLINE

- The Gothenburg model
  - History
  - Goals
  - Components
- CollateX
  - What it is CollateX?
  - Collation pipeline in CollateX

THE GOTHENBURG MODEL: HISTORY

- Developers of CollateX and Juxta
- Joint workshop: Gothenburg 2009
- Sponsored by COST Action 32 and Interedition

GOALS

Identification of the core components of textual comparison at an abstract level

- common understanding
- facilitation of collaboration

COMPONENTS

1. Tokenization
2. Normalization/regularization
3. Alignmemt
4. Analysis
5. Visualization/output

Prerequisite: an electronic text version of each witness

1. TOKENIZATION

- Division of the continuous text into units to be aligned (tokens)

1. TOKENIZATION

- Division of the continuous text into units to be aligned (tokens)

```
Would you care for a sherbet lemon?
```

1. TOKENIZATION

- Division of the continuous text into units to be aligned (tokens)

```
Would you care for a sherbet lemon?
--> Would | you | care | for | a |
sherbet | lemon | ?
```

1. TOKENIZATION

- Division of the continuous text into units to be aligned (tokens)
- Any level of granularity
  - Typically: whitespace-delimited words
  - Other options: syllables, lines, phrases, verses, paragraphs, text nodes…

TOKENIZATION: CHALLENGES

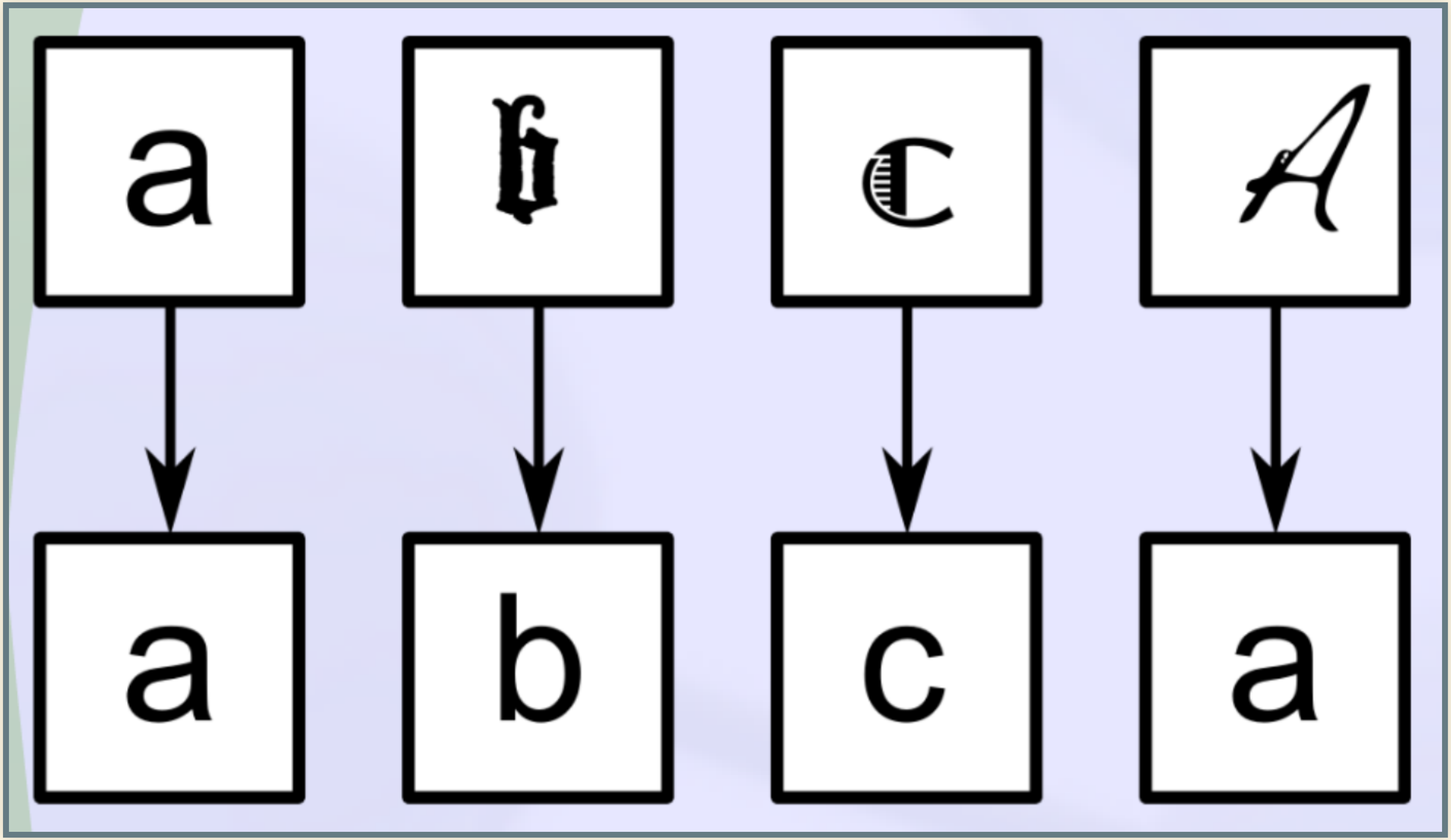
- Ambiguity
- Punctuation
- Language specific issues: contractions, superscription, etc.
- Markup

TOKENIZATION CHALLENGES: SOME EXAMPLES

- He remarked, “John said, ‘Bout starts at nine.’”
- He remarked, “John said, ‘It’s ’bout time.’”
- Tu es un %#@$!
- Oh d--n it!
- MASS.

2. NORMALIZATION/REGULARIZATION

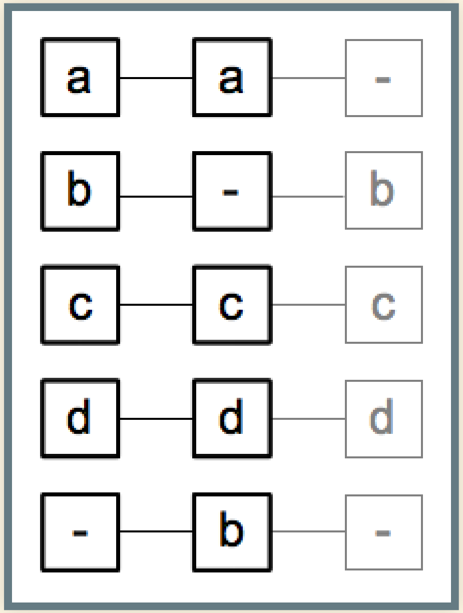
- Normalization during transcription *vs.* collation
- Ignore non-substantive variation for comparison
  - Punctuation
  - Upper/lower case
  - Orthographic variation
    - Allographs (letterforms)
    - Abbreviations



a
𝔟
c
𝒜
a
b
c
a

3. ALIGNMENT

- Find the tokens that match
- Introduce gap tokens when necessary (“omissions”)



a	a	-
b	-	b
c	c	c
d	d	d
-	b	-

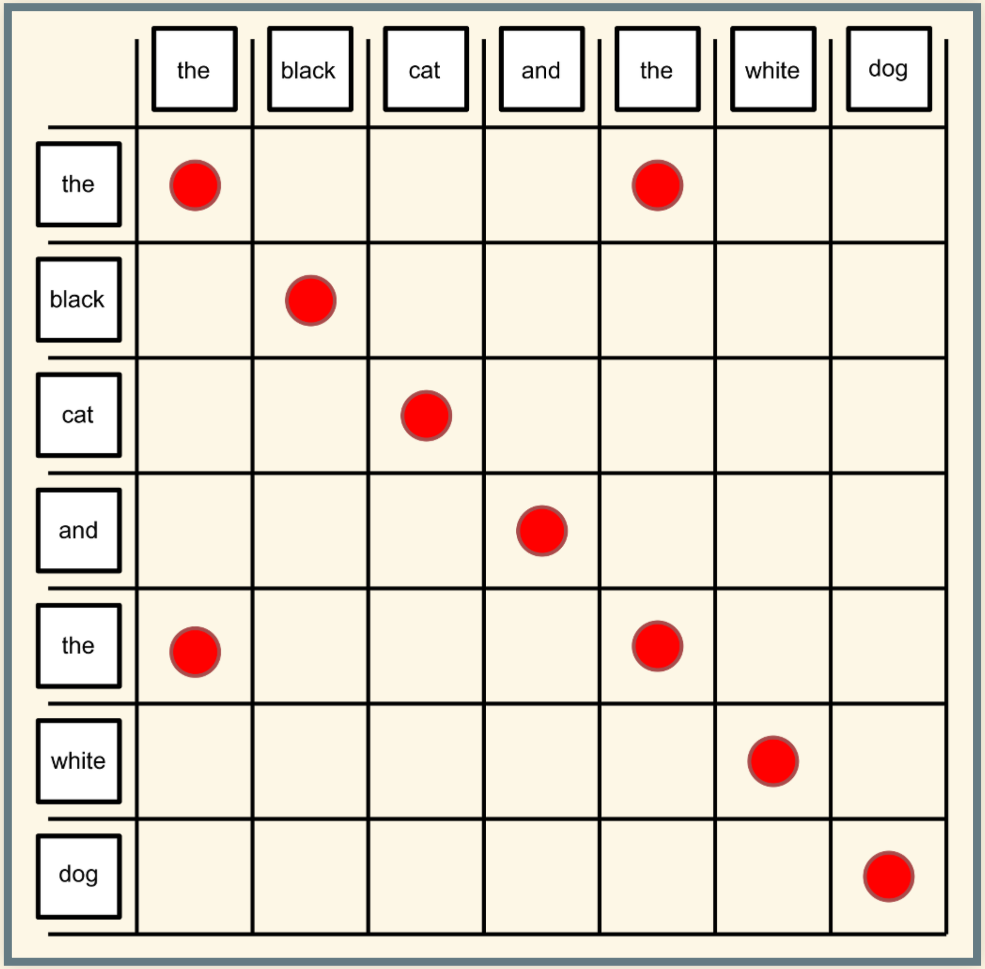
Alignment of tree witnesses

ALIGNMENT: CHALLENGES

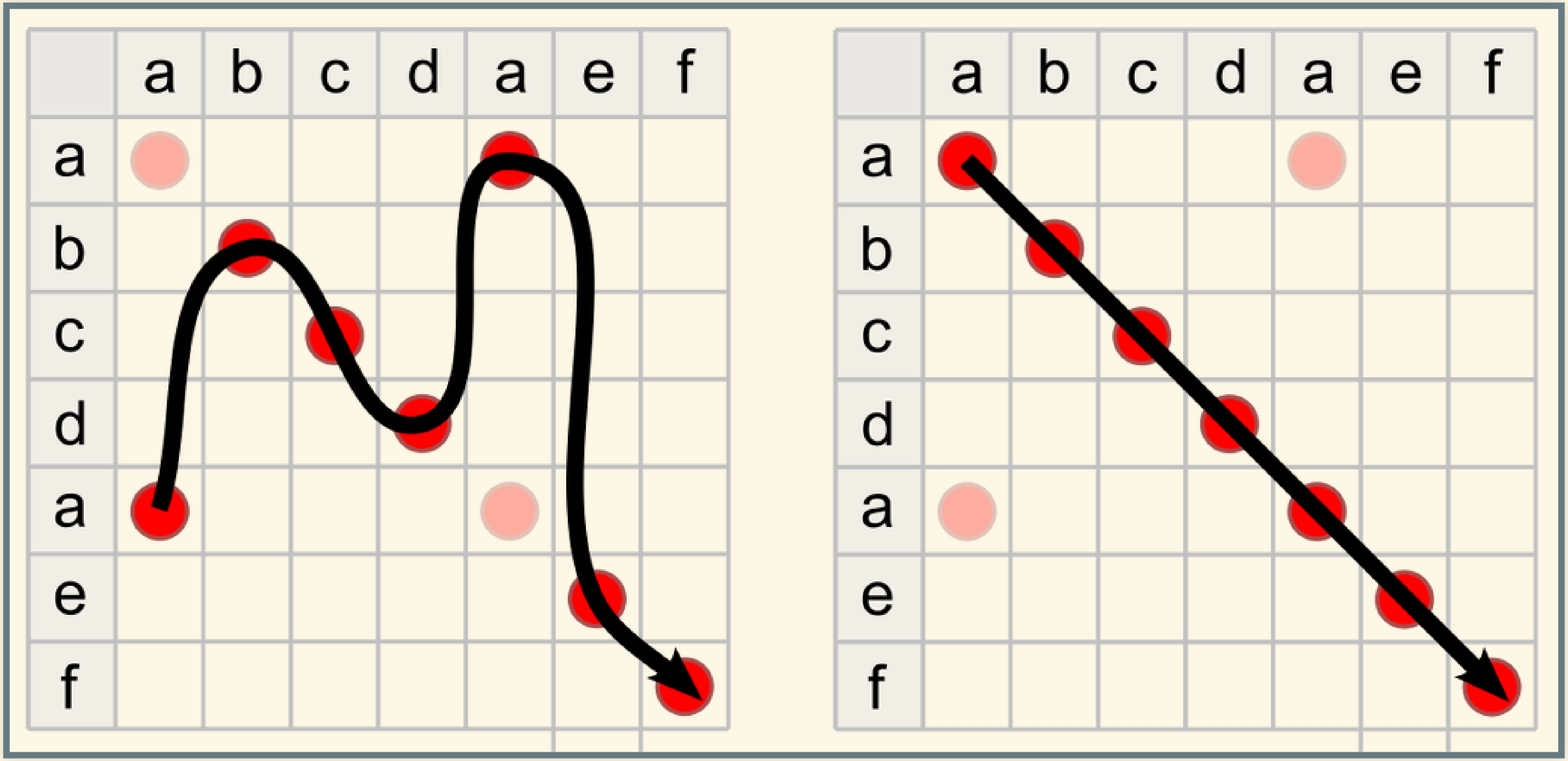
- Computational complexity

ALIGNMENT: CHALLENGES

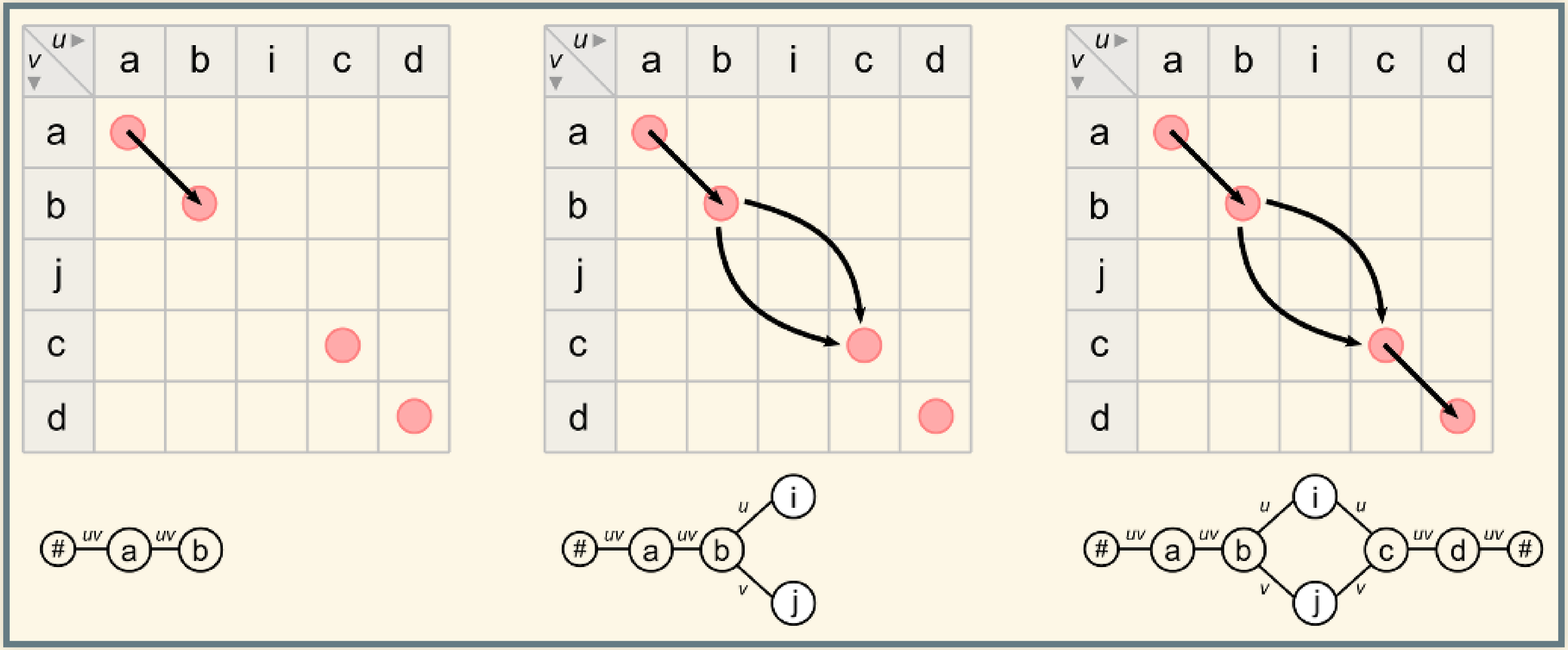
- Repetition
- Transposition



	the	black	cat	and	the	white	dog
the	●				●		
black		●					
cat			●				
and				●			
the	●				●		
white						●	
dog							●



a b c d a e f
a
b
c
d
a
e
f
a b c d a e f
a
b
c
d
a
e
f

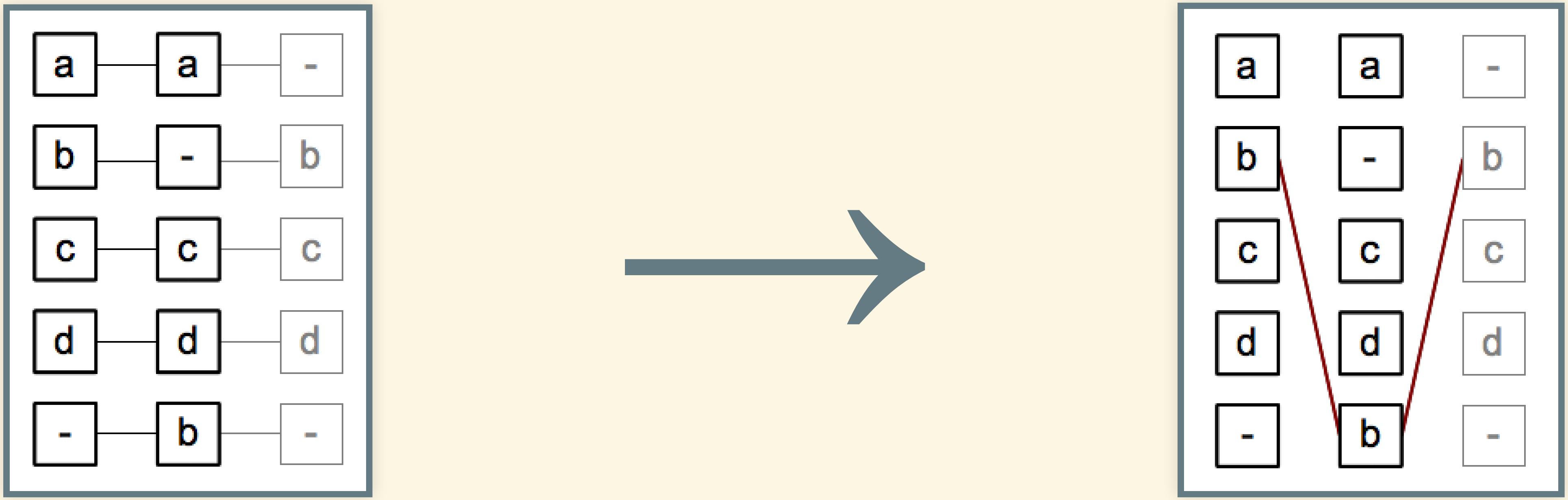

u
v
a
b
i
c
d
a
b
j
c
d
#
uv
a
uv
b
u
i
v
j
#
uv
a
uv
b
u
i
u
v
j
v
c
uv
d
uv
#

ALIGNMENT: CHALLENGES

- Order effects

4. ANALYSIS/FEEDBACK

- Intepretation beyond linear alignment
- Manual intervention?


a a -
b - b
c c c
d d d
- b -
a a -
b - b
c c c
d d d
- b -

5. VISUALIZATION/OUTPUT

- Markup for further processing
  - XML, TEI, JSON, GraphViz DOT, LaTeX, etc.
- Textual visualization, for examination and analysis
  - Textual alignment table
    - Plain text, HTML, PDF
  - Toolkits with additional functionalities: Juxta †
- Graphic visualization, for examination and analysis
  - Variant graph

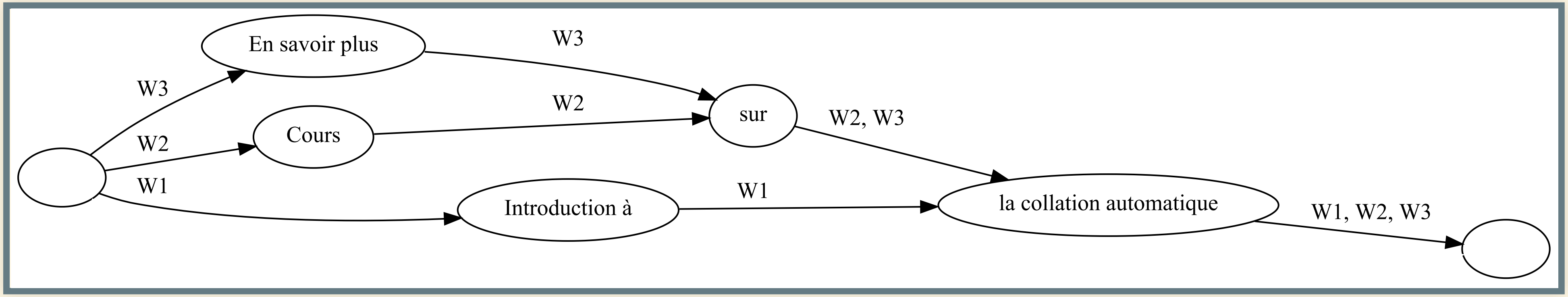
EXAMPLE

- *W1*: Introduction à la collation automatique
- *W2*: Cours sur la collation automatique
- *W3*: En savoir plus sur la collation automatique

ALIGNMENT TABLE

W1	Introduction à		la collation automatiq
W2	Cours	sur	la collation automatiq
W3	En savoir plus	sur	la collation automatiq

VARIANT GRAPH



COLLATION AND/OR VISUALIZATION TOOLS

- TRAViz
- CATview

COLLATEX

1. What it is CollateX?
2. Collation pipeline in CollateX

FLAVORS

- Java
- Web app
- Python

ADVANTAGES OF COLLATEX

- Data formats
  - **Input**: Anything and everything (JSON)
  - **Output**: Anything and everything (JSON)
- Control over each step of the pipeline

COLLATION PIPELINE IN COLLATEX

- Default behaviours
- Parameters

TOKENIZATION IN COLLATEX

- It divides the text into tokens using whitespaces as delimiter
- Punctuation is tokenized separately from alphanumeric characters

TOKENIZATION IN COLLATEX

- It divides the text into tokens using whitespaces as delimiter
- Punctuation is tokenized separately from alphanumeric characters

Example: *Peter's cat.*

Peter ' s cat .

NORMALIZATION IN COLLATEX

By default, it removes trailing white space at the end of tokens.

PRETOKENIZED AND NORMALIZED INPUT

JSON file as input: Each token may present a normalized version

COLLATEX ALIGNMENT PARAMETERS

- Different alignment algorithms
  - Dekker (Dekker & Middle 2011)
  - Needleman-Wunsch (Needleman & Wunsch 1970)
  - MEDITE (Bourdaillet & Ganascia 2007)

PROGRESSIVE ALIGNMENT

1. start by comparing two versions,
2. transform the result into a variant graph, then
3. compare another version against that graph, and
4. merge the result of that comparison into the graph;
5. repeate the procedure until all versions have been merged.

ANALYSIS IN COLLATEX

Exact *vs.* near (fuzzy) matching

- *A*: And Ron pulled out a fat grey rat
- *B*: And Ronald pulled out a gray rat

EXACT MATCHING

A	And	Ron	pulled	out	a	fat	grey	rat
B	And	Ronald	pulled	out	a	gray	-	rat

NEAR MATCHING

A	And	Ron	pulled	out	a	fat	grey	rat
B	And	Ronald	pulled	out	a	-	gray	rat

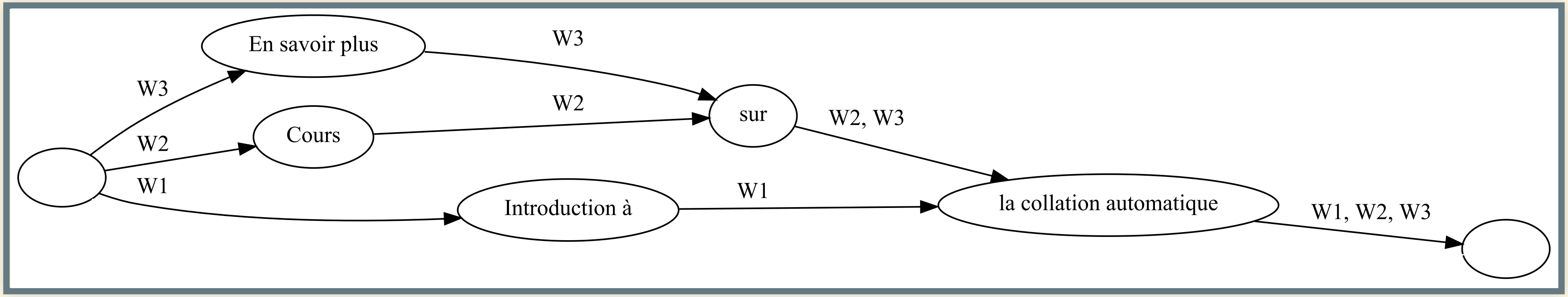
COLLATEX OUTPUTS

- Alignment table: ASCII, CSV, TSV, HTML,XML, XML-TEI, JSON
- Variant graph: SVG

ALIGNMENT TABLE

W1	Introduction à		la collation automatiq
W2	Cours	sur	la collation automatiq
W3	En savoir plus	sur	la collation automatiq

VARIANT GRAPH



TEI

```
<cx:apparatus xmlns:cx="http://interedition.eu/collatex/ns/1.0"
xmlns="http://www.tei-c.org/ns/1.0">
    <app>
        <rdg wit="W1">Introduction à</rdg>
        <rdg wit="W2">Cours</rdg>
        <rdg wit="W3">En savoir  plus</rdg></app>
    <app><rdg wit="W1"/><rdg wit="W2 W3">sur</rdg></app>
    la collation automatique</cx:apparatus>
```

BIBLIOGRAPHY

- Bourdaillet J. & Ganascia J.-G. (2007): “Practical block sequence alignment with moves.” *LATA 2007 - International Conference on Language and Automata Theory and Applications*, 3/2007.
- Dekker, R. H. & Middell, G. (2011): “Computer-Supported Collation with CollateX: Managing Textual Variance in an Environment with Varying Requirements.” *Supporting Digital Humanities 2011*. University of Copenhagen, Denmark. 17-18 November 2011.

- Interedition Development Group (2010-): *CollateX - Sofware for Collating Textual Sources*. https://collatex.net/
- Hoover, David L. (2015): “The Trials of Tokenization.” *DH2015*, University of Western Sydney, Australia, June 29–July 3, 2015.
- Needleman, Saul B. & Wunsch, Christian D. (1970): A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology 48* (3), 443–53.