Stemmatology

an R package for the computer-assisted analysis of textual traditions

Jean-Baptiste Camps and Florian Cafiero

1. Centre Jean-Mabillon, École nationale des chartes | PSL
2. I.I.A.C., École des hautes études en sciences sociales | PSL

jbcamps@hotmail.com

From copy errors to text genealogy

Before the printing press, the only way of spreading a written text was manual copying. In this process, accidents, errors and intentional modifications occurred, progressively modifying the text of each witness. For the philologist, it is imperative to study the variants of the witnesses, to assess their genealogical relations. Developped throughout the XIXth and XXth century, the Lachmannian method of common errors faces difficulties with phenomena such as horizontal transmission (contamination) or independent but identical variants (polygenesis) [6]. Since Dom Froger [4], computational procedures have been developed, based on textual criticism principles or inspired by other fields [1, 5]. In this package, we implement a method designed by Poole [7, 8] and extended by Camps & Cafiero [2].

Our philosophy

- Value interactions with the researcher: computer-assisted, not computer-produced results.
- Have as few assumptions as possible on the definition of basic units of variation (variant locations) or variant types.
- Stay as independent as possible from implementation choices.

Data model

- Each column stands for a witness, each line for a variant location;
- Each variant is given a numeric code: [*NA*] for not available, [0] for omission, [$1 \ldots n$] for variants.

Variants are rarely arranged in a simple linear succession. Instead, we have to deal with both localised (e.g. words) and macro-structural (e.g. verses, paragraphs) variations: omission/addition, different ordering, content modification (graphic, semantic, ...), as well as missing data (lacuna).

Users are free to choose the variation types and levels to be encoded. A cumulative encoding accounting for all levels and kinds of variation is possible by breaking them down into different variant locations.

Conversion from TEI

A sample from Chrétien de Troyes' Chevalier au lion (v. 3686)

H: Onques ne fu cil P: Onques chil ne fu M: Onques cil ne fut
V: Onques cil ne fu F: Cil ne fu onques S: Onques cil ne fu
G: Et cil ne fu pas A: Onques cil ne fu R: Onques cil ne fu

Encoded in tei

```
<l n="3686">
    <!-- First variant, Onques vs. Et -->
    <app xml:id="VL_3686.1" type="functionWord">
        <rdg wit="#H #P #V #A #S #R #M #F">
            <app xml:id="VL_3686.1.1">
                <!-- Subvariant: inversion of Onques -->
                <rdg wit="#H #P #V #A #S #R #M">Onques</rdg>
                <rdg wit="#F" corresp="#inv_F_01"/>
            </app>
        </rdg>
        <rdg wit="#G">Et</rdg>
    </app>
    <!-- Graphical variant chil / cil-->
    <app type="graphic" xml:id="VL_3686.2">
        <rdg wit="#P">chil</rdg>
        <rdg wit="#F #V #G #A #S #R #M #H">
            <!-- H has 'cil' at a different place,
            but with the same reading as FVGASRM -->
            <app xml:id="VL_3686.2.1">
                <rdg wit="#F #V #G #A #S #R #M">cil</rdg>
                <rdg wit="#H" corresp="#inv_H_01"/>
            </app>
        </rdg>
    </app>
    <!-- […] -->
    <!-- And here we account for the inversion -->
    <app type="functionWord" xml:id="VL_3686.4">
        <rdg wit="#H" xml:id="inv_H_01">cil</rdg>
        <rdg wit="#P #V #F #G #A #S #R #M"/>
    </app>
    <app type="functionWord" xml:id="VL_3686.5">
        <rdg wit="#G">pas</rdg>
        <rdg wit="#F" xml:id="inv_F_01">onques</rdg>
        <rdg wit="#H #P #V #A #S #R #M"/>
    </app>
</l>
```

Converted to the expected format with xslt:

	H	P	V	F	G	A	S	R	M
VL_3686.1	1	1	1	1	2	1	1	1	1
VL_3686.1.1	1	1	1	0	NA	1	1	1	1
VL_3686.2	2	1	2	2	2	2	2	2	2
VL_3686.2.1	0	NA	1	1	1	1	1	1	1
VL_3686.3	1	1	1	1	1	1	1	1	2
VL_3686.4	1	0	0	0	0	0	0	0	0
VL_3686.5	0	0	0	2	1	0	0	0	0

In this example, we assume that the user retains all variation types. By default, only variant locations labeled as `substantive` are kept in the transformation.

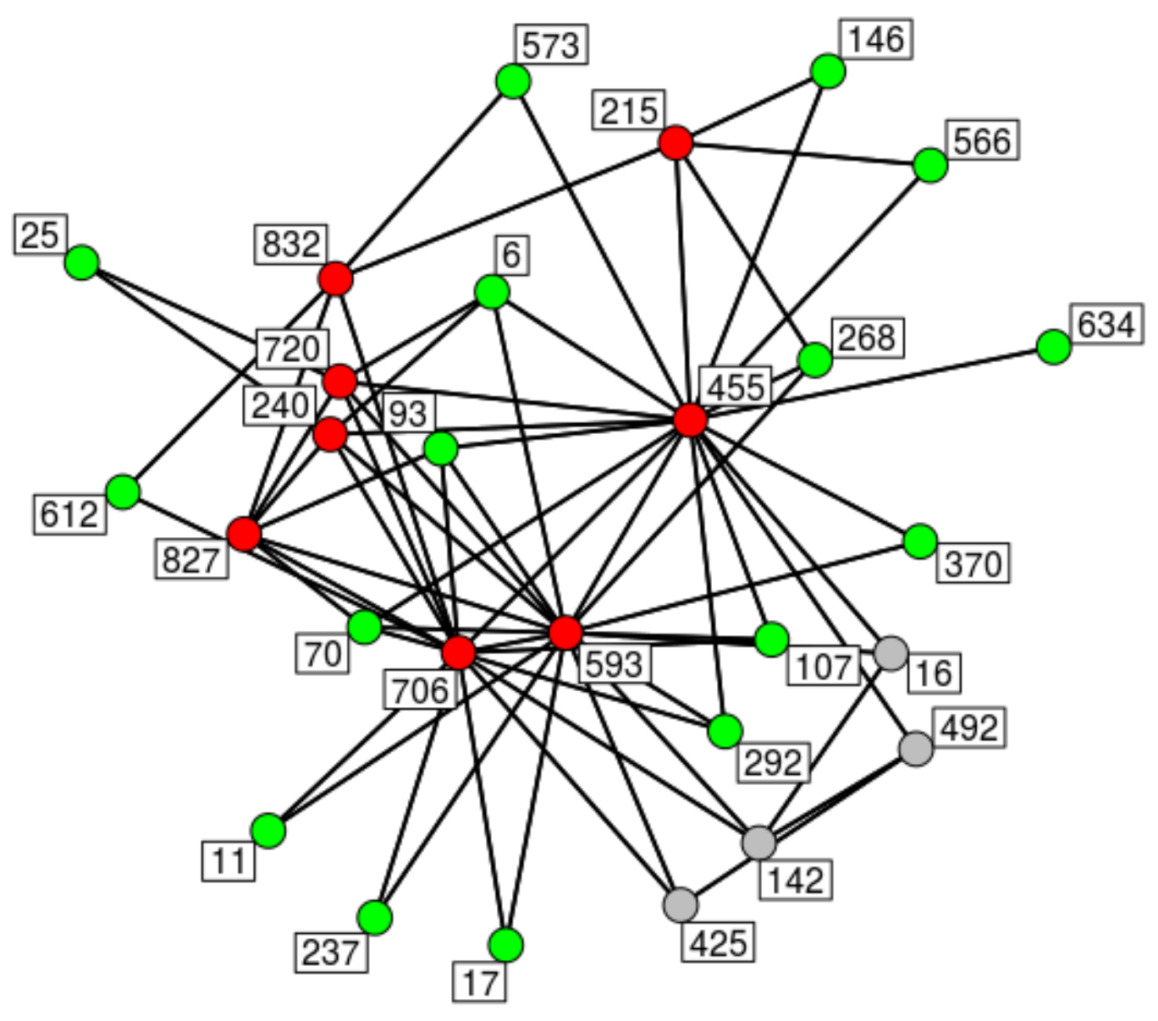
Exploratory analysis of a tradition

`PCC.conflicts`: identify contradictions in the variant locations' genealogical configurations, by comparing their readings two by two.

Intuition: a variant location in conflict with a large number of variant locations is unreliable. Symetrically, variant locations contradicted only by unreliable variant location are reliable.

We represent each conflicting variant location as a node on a graph, linked by an edge to nodes it is in conflict with.

The user is guided in determining the level of conflictuality that seems acceptable in his corpus. Through clustering, groups of nodes are defined, according to the value of their associated centrality.



Overconflicting variant locations isolated - in red - in the Parzival dataset

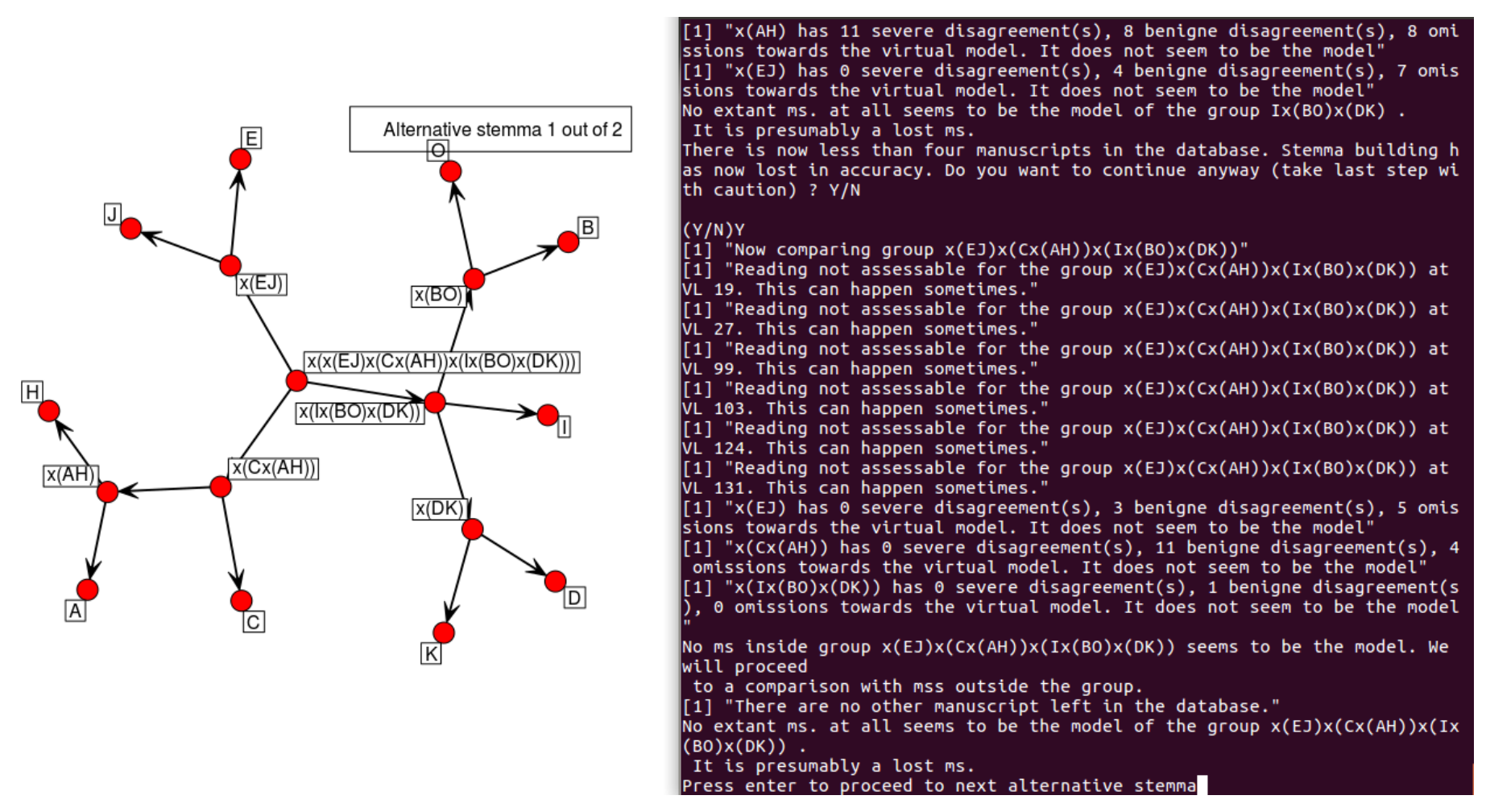
`PCC.elimination` eventually gets rid of unreliable variant locations. If contamination is suspected, the function `PCC.contam` tries to estimate the individual contribution of each witness to the number of contradictions between variant locations.

In the event of algorithmically undecidable situations, the function `PCC.equipollent` creates a separate database for each competing configuration.

Building a stemma

`PCC.Stemma` allows to build one or more stemmata, depending on the input. Our method relies on the transformation of the common error method into a disagreement-based algorithm.

The recursive algorithm first assesses coherent groups (`PCC.buildGroup`), then reconstructs or identifies their model.


Alternative stemma 1 out of 2
E
J
x(EJ)
O
B
x(BO)
x(x(EJ)x(Cx(AH))x(Ix(BO)x(DK)))
H
x(Ix(BO)x(DK))
I
x(AH)
x(Cx(AH))
x(DK)
A
C
D
K

```
[1] "x(AH) has 11 severe disagreement(s), 8 benigne disagreement(s), 8 omi
ssions towards the virtual model. It does not seem to be the model"
[1] "x(EJ) has 0 severe disagreement(s), 4 benigne disagreement(s), 7 omis
sions towards the virtual model. It does not seem to be the model"
No extant ms. at all seems to be the model of the group Ix(BO)x(DK) .
 It is presumably a lost ms.
There is now less than four manuscripts in the database. Stemma building h
as now lost in accuracy. Do you want to continue anyway (take last step wi
th caution) ? Y/N

(Y/N)Y
[1] "Now comparing group x(EJ)x(Cx(AH))x(Ix(BO)x(DK))"
[1] "Reading not assessable for the group x(EJ)x(Cx(AH))x(Ix(BO)x(DK)) at
VL 19. This can happen sometimes."
[1] "Reading not assessable for the group x(EJ)x(Cx(AH))x(Ix(BO)x(DK)) at
VL 27. This can happen sometimes."
[1] "Reading not assessable for the group x(EJ)x(Cx(AH))x(Ix(BO)x(DK)) at
VL 99. This can happen sometimes."
[1] "Reading not assessable for the group x(EJ)x(Cx(AH))x(Ix(BO)x(DK)) at
VL 103. This can happen sometimes."
[1] "Reading not assessable for the group x(EJ)x(Cx(AH))x(Ix(BO)x(DK)) at
VL 124. This can happen sometimes."
[1] "Reading not assessable for the group x(EJ)x(Cx(AH))x(Ix(BO)x(DK)) at
VL 131. This can happen sometimes."
[1] "x(EJ) has 0 severe disagreement(s), 3 benigne disagreement(s), 5 omis
sions towards the virtual model. It does not seem to be the model"
[1] "x(Cx(AH)) has 0 severe disagreement(s), 11 benigne disagreement(s), 4
 omissions towards the virtual model. It does not seem to be the model"
[1] "x(Ix(BO)x(DK)) has 0 severe disagreement(s), 1 benigne disagreement(s
), 0 omissions towards the virtual model. It does not seem to be the model
"
No ms inside group x(EJ)x(Cx(AH))x(Ix(BO)x(DK)) seems to be the model. We
will proceed
 to a comparison with mss outside the group.
[1] "There are no other manuscript left in the database."
No extant ms. at all seems to be the model of the group x(EJ)x(Cx(AH))x(Ix
(BO)x(DK)) .
 It is presumably a lost ms.
Press enter to proceed to next alternative stemma
```

User interface of the package: building the stemma (Fournival dataset)

The algorithm can compute and display a final configuration. Yet, the expert is incited to make his own decision regarding the very top of the stemma.

Further developments

- Improve the different visualisations;
- Implement new algorithms for exploratory procedures (cardiograms [3],...);
- Implement other methods for stemma construction.

Sources & Data

Sources: doi: 10.5281/zenodo.1117389
`http://github.com/Jean-Baptiste-Camps/stemmatology`.

Two datasets [2], representing the variant locations of an historical and an artificial tradition are included in the package, in `.rda` format.

Utils: doi, 10.5281/zenodo.1117181
`https://github.com/Jean-Baptiste-Camps/stemmatology-utils`

References

[1] Andrews T., Gershoni I., Imhof R., Kaufmann S., Schaerer J., Studer T., & Zumbrunn S., s.d., "Efficient Stemmatology: a Graph Database Application in the Digital Humanities"

[2] Camps J.-B. & Cafiero F., 2015, "Genealogical variant locations and simplified stemma: a test case", in Analysis of Ancient and Medieval Texts and Manuscripts: Digital Approaches, ed. Tara Andrews & Caroline Macé, Turnhout, p. 69-93 (Lectio, 1).

[3] Den Hollander, A.A., 2004, "How shock waves revealed successive contamination: A cardiogram of early sixteenth-century Dutch Bibles", in Studies in Stemmatology 2, ed. P. Van Reenen, A.A. Den Hollander & M.J.P. Van Mulken, Amsterdam, p. 99-112.

[4] Froger J., 1968, La critique des textes et son automatisation, Paris.

[5] Heikkilä T., & Roos, T., 2016, "Thematic Section on Studia Stemmatologica", Digital Scholarship in the Humanities 31-3, p. 520-22, doi : 10.1093/llc/fqw038.

[6] Parvum lexicon stemmatologicum - PLS - HIIT Wiki, ed. C. Macé & P. Roelli, 2015,

[7] Poole E., 1974, "The Computer in Determining Stemmatic Relationships", Computers and the Humanities, 8-4, p. 207-16.

[8] Poole E., 1979, "L'analyse stemmatique des textes documentaires", in La pratique des ordinateurs dans la critique des textes, p. 151-161.

[9] Roos T., & Heikkilä T., 2009, "Evaluating methods for computer-assisted stemmatology using artificial benchmark data sets", Literary and Linguistic Computing, 24-4, p. 417-433.

[10] Schmidt D., & Colomb R., 2009, "A data structure for representing multiversion texts online", International Journal of Human-Computer Studies, 67-6, p. 497-514, doi : 10.1016/j.ijhcs.2009.02.001.