`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 | Paris Sciences & Lettres

2. Institut Inderdiscplinaire d'Anthropologie du Contemporain
École des hautes études en sciences sociales | Paris Sciences & Lettres

Abstract

Given a set of witnesses of a text, determining their relations and reconstructing a *stemma codicum* is one of the fundamental purposes of textual criticism and philology. For this task, various computer-assisted procedures and methods have been described since the 1950's, some elaborating on traditional principles (Lachmannian, Quentinian...), some borrowed from other fields such as phylogenetics. In this poster, we describe `Stemmatology`, a new open source package for the statistical software `R`, that implements procedures for the computer-assisted analysis of textual traditions. We have started implementation of stemmatological methods in the package by focusing, on one hand, on procedures derived from traditional textual criticism, the "Lachmannian" tradition in general, and particularly some of Eric Poole's methodological insights (Poole [13, 14]); and on the other hand, we made use of methods for the detection of contamination and polygenesis, two major issues for genealogical analysis.

1 Introduction

Before the appearance of the printing press, in the West, the only way of reproducing and spreading a text in written form was manual copying. During this process, accidents, errors and intentional modifications occurred, progressively modifying the text of each witness. For the philologist interested in the study of a textual tradition or the restoration of the original text, it has been imperative to study the

[*]The source code for the package is available on Github: Jean-Baptiste Camps & Florian Cafiero, `stemmatology` : an R stemmatology package, v. 0.2.2, 2014-..., `http://github.com/Jean-Baptiste-Camps/stemmatology`, DOI: 10.5281/zenodo.1117389.

[†]Corresponding author: `jbcamps@hotmail.com`

different variants of the witnesses, to assess their genealogical relations, at least since the beginning of the scientific age of philology in the XIX$^{\text{th}}$ century. As a result, the method of common errors (often deemed “Lachmannian”) took progressively form during the XIX$^{\text{th}}$ and the beginning of the XX$^{\text{th}}$ century. Yet, different phenomena such as horizontal transmission (contamination) or the independent appearance of identical variation in different witnesses (polygenesis) cause major difficulties to this method[1].

Dating back to the experiments of Dom Froger [8], various computational procedures have been used to help assess the genealogy of a textual tradition. These can be roughly divided between methods based on pre-existent textual criticism principles (Lachmannian, Quentinian or other) and methods inspired by fields other than textual criticism, such as phylogenetics or compression-based algorithms [17, 1]. Research in this specialised field is active; methodological contributions, in particular, are numerous, though it is out of our scope here to summarise them (see for instance the recent special issue on stemmatology in [11]).

Amongst these methods, we decided to start by implementing a method based on traditional philological principles, firstly because these are often less available or lack open-source and user-friendly implementations, secondly because, for now, classical methods (including non computerised ones), still seem to offer very satisfactory results when compared to others [17]. We chose to start with the approach designed by Poole [13, 14], extended by Camps & Cafiero [3]. It offered us an algorithmic and easy to compute transposition of the common error method. It also helped us addressing the major problems raised by contamination and polygenesis.

Beyond the methods implemented for now, this software package has been developed with a main objective: valuing the interactions with the researcher, allowing his or her insights to guide and enhance the results.

2 Data model

It is not our purpose here to decide which data model would be better to represent textual variations, nor to put constraints on the meaning given to the notions of *witness* or *variant location*. Moreover, the definition of the basic unit of variation, the inclusion or not of some types of variation, are all features that can deeply vary from one context to another, or between projects. A variety of models already exists, from the word-based collation table to the TEI encoded apparatus or the graph model [18], with the addition of local and project-specific models. The data themselves can be stored according to various implementations, including graph databases [1].

To stay as independent as possible from all these choices, we adopted a simple and abstract representation, with very few hard constraints for the user. In our data

[1]For a definition of the notions specific to textual criticism, see Duval [6], especially “Contamination”, “Erreur polygénétique” and “Polygenèse” (p. 88-89, 134-135 and 218). See also Macé, Roelli *et al.* [12], for English language definitions.

model, each column stands for a witness, and each line for a variant location. Each variant is given a numeric code (*NA* for not acquired, 0 for omission, $1 \ldots n$ for variants), as shown in table 1. The exact meaning to give to *variant location* and *variant* is defined by the user, according to his or her approach and the nature of the materials being analysed. The choice to consider omission as readings or not is also left to the user through an option (`omissionsAsReadings`, which can be set to `TRUE` or `FALSE`).

	W_1	W_2	W_3
VL_1	1	1	2
VL_2	0	1	2
VL_n	NA	1	2

Table 1: Tabular data model

To illustrate the flexibility of this approach, let us take as example the case of the potential combination of macro-structural and localised variants: the order of a few paragraphs or verses (or books, etc.) may be different in two groups of witnesses, while the paragraphs themselves also contain *varia lectio*. The user may then choose to:

1. consider only the variations in the order of paragraphs, and encode it as a single variant location, with each observed order taken as a variant and given a numeric code;

2. consider only the variations in content, and create a variant location for each of them, ignoring macro-structural variation;

3. do both, and encode successively a variant location for the change of order, and others for the variations in content.

This flattened approach can also be used for smaller inversions, and there is no limit in terms of depth or imbrication: if there is, in a given tradition, variation in the order of books, in the order of paragraphs inside those books, of sentences inside paragraphs and words inside sentences, along with localised variations in content, a user may very well decide to create separate variant locations to record separately variation of order at each of those levels, in addition to variant locations recording localised variants. Methodological choices in terms of alignment or segmentation are outside the scope of this package.

Since the principal purpose of our software is stemmatological analysis, and not data representation or manipulation, we do not include functions to collate texts or encode variants. Yet, to maximise interoperability, we offer a configurable and easily customised `xslt` stylesheet to transform a TEI-encoded parallel-segmentation apparatus to our data format. We welcome the contribution of other stylesheets to the repository[2]. To illustrate one of the possible transformations,

[2]Available on Github, Jean-Baptiste Camps, `stemmatology-utils`, `https://github.com/Jean-Baptiste-Camps/stemmatology-utils`, DOI: 10.5281/zenodo.1117181.

including cases of inversion, we can take as example this short sample from the beginning of v. 3686 of Chrétien de Troyes' *Chevalier au lion*:

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

In this sample we find:

- inversions, at various levels: permutation of *Onques* and *cil ne fu*, as well as permutations of *cil* and *ne fu*;
- variants on function-words: *Onques* against *Et* . . . *pas*, that can be decomposed, if need be, in a substitution (*Onques*/*Et*) and an addition/omission (*pas*);
- graphic (diatopic, diachronic) variation (*cil*/*chil*; *fu*/*fut*).

This can be expressed in TEI in the following way (among other possibilities). To account for displacement of one word, we decompose it into two simpler forms of changes, that is one deletion at a first location and one addition in a second location (but we link the two, and still indicate the content subvariants inside):

```
<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' but at a different place, we nonetheless
                     indicate that its reading is the same that 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>
    ne <app type="graphic" xml:id="VL_3686.3">
        <rdg wit="#H #P #V #F #G #A #S #R">fu</rdg>
        <rdg wit="#M">fut</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>
```

Using the xslt stylesheet provided this could then be converted, generating for instance the outcome presented in table 2.

	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

Table 2: Result from the automated conversion into the data format used by the package (all types of variation retained)

In this example, we make the assumption that the user chose to retain all types of variations. By default, only variant locations labelled as substantive are retained in the transformation, but it can be configured.

To facilitate quick experimentation, and reproducibility of procedures, we ship the package with various datasets, based on artificial or historical traditions. The datasets are provided in .rda format, as part of the package. They consist of tabular data, in the format used for analysis, and represent the selection of significant variant locations as used in [3]. The main datasets are *Fournival* (historical, [15]) and *Parzival* (artificial, [19]). Documentation and reference of datasets are included in the package manual.

3 Features of the package

To implement our package, we chose to use the open-source environment for statistical analysis R [16]. R is a well-established software, with a strong community, allowing us to build on a wide range of built-in or community-developed functions. Furthermore, the choice of R makes it very easy to fully document the procedures, to include both commands and datasets, and to distribute the package. We adopted a copyleft license, the GNU General Public License v.3 [10].

In its current state, the package mainly consists in two sets of features. The first set is dedicated to exploratory analysis of a textual tradition, the second to the building of a stemma.

In the exploratory set of functions, we first implemented a procedure (the function `PCC.conflicts`) to identify contradictions in the variant locations' genealogical configurations, comparing their readings two by two. The underlying intuition is simple: if a variant location is in conflict with a relatively large number of variant locations, it can be considered as unreliable, and ruled out of the further computations. On the other hand, variant locations in conflict with an unreliable variant location can be considered as potentially reliable. The rest of the procedure will help the researcher to assess which variant locations are the source of the contradictions, through visualisation or computations.

To analyse this phenomenon, we represent the set of conflicting variant locations as dots (or "nodes") on a graph. When there is a conflict between two variant locations, we draw an undirected link (or "edge") between them. The user can then view the conflicts between variant locations as a network, that can be plotted and mathematically analysed. Current visualisation uses the classical Fruchtermann-Rheingold [9] spatialisation.

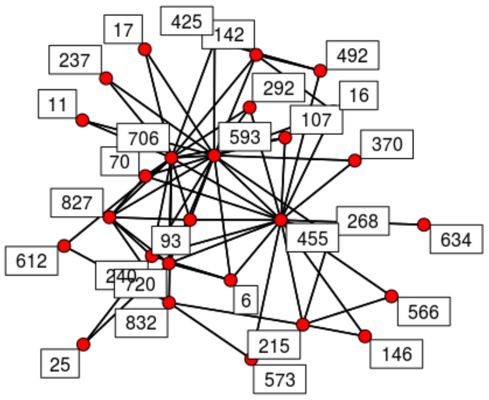
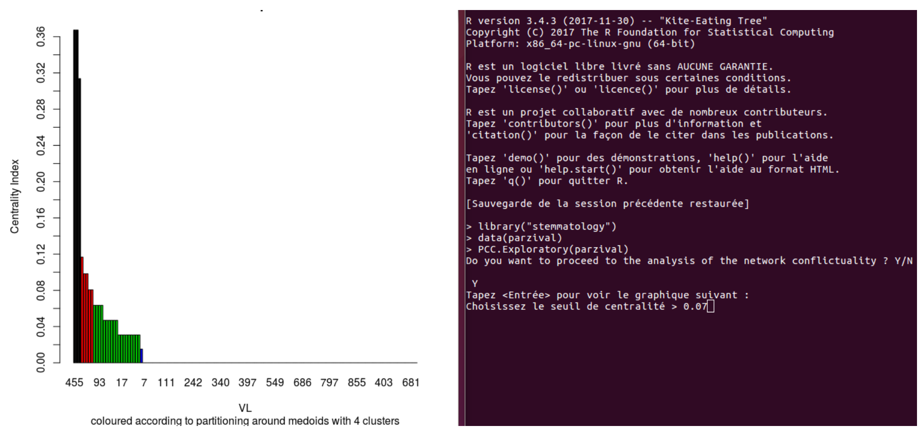


Figure 1: Graph of the conflicts between the variant locations in the *Parzival* dataset

The user is then guided in determining the level of conflictuality that seems acceptable in his corpus. The "degree centrality" [7] of the various nodes is computed and displayed. The nodes are clustered according to the value of their centrality. The higher is the value, the more uncertain is the interest of the associated variant location.

This step is meant to help the user to have an intuition about the "conflictuality" acceptable between variant locations. He is then invited to give the level he deems



```
R version 3.4.3 (2017-11-30) -- "Kite-Eating Tree"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R est un logiciel libre livré sans AUCUNE GARANTIE.
Vous pouvez le redistribuer sous certaines conditions.
Tapez 'license()' ou 'licence()' pour plus de détails.

R est un projet collaboratif avec de nombreux contributeurs.
Tapez 'contributors()' pour plus d'information et
'citation()' pour la façon de le citer dans les publications.

Tapez 'demo()' pour des démonstrations, 'help()' pour l'aide
en ligne ou 'help.start()' pour obtenir l'aide au format HTML.
Tapez 'q()' pour quitter R.

[Sauvegarde de la session précédente restaurée]

> library("stemmatology")
> data(parzival)
> PCC.Exploratory(parzival)
Do you want to proceed to the analysis of the network conflictuality ? Y/N

 Y
Tapez <Entrée> pour voir le graphique suivant :
Choisissez le seuil de centralité > 0.07
```

Figure 2: User interface of the package: visualising the centrality of the nodes

bearable (`PCC.overconflicting` function), and the variants generating too many contradictions according to this setting are displayed.

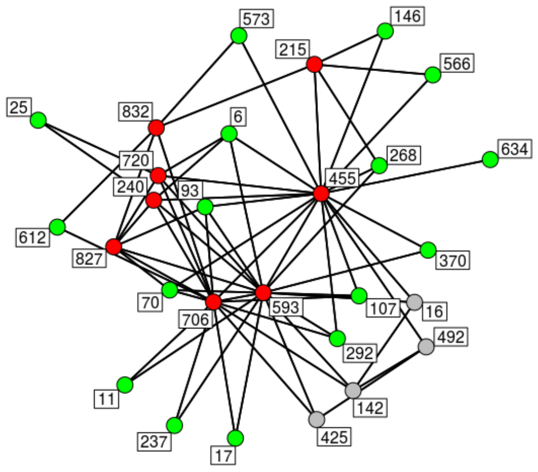
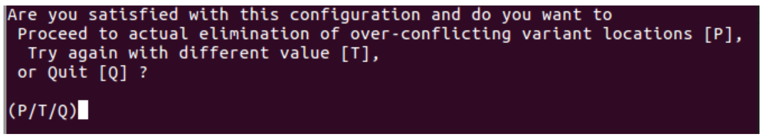


Figure 3: Overconflicting variants isolated - in red - in the *Parzival* dataset

`PCC.elimination` eventually gets rid of those variant locations, under the user's supervision.



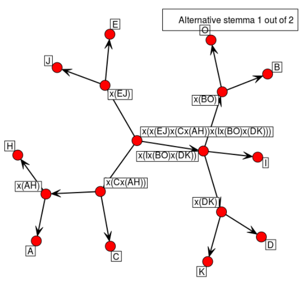
```
Are you satisfied with this configuration and do you want to
 Proceed to actual elimination of over-conflicting variant locations [P],
  Try again with different value [T],
 or Quit [Q] ?

(P/T/Q)
```

Figure 4: User interface of the package: assisted selection of variants

If contamination is suspected, the function `PCC.contam` can be called. It removes a witness from its calculations, and computes the number of conflicts between variant locations remaining without it. The function repeats the same computation for every witness in the textual tradition. If removing one specific witness induces a significant decrease in conflictuality between variant locations, the function offers to label this manuscript as plausibly contaminated.

At the end of this selection process, some uncertainty can remain about several variant location. Yet, in the event of algorithmically undecidable situations, the user should not be stuck. The function (`PCC.equipollent`) thus allows to create different databases corresponding to the competing configurations. From then on, these databases can be studied separately, and might result in different plausible stemmata.

The second set of functions (called by the general function `PCC.stemma`) allows the user to build one or several stemmata, depending on the input. For the construction of the stemma, only one method is implemented as of now, relying on the transformation of the common error method into a disagreement-based algorithm, formulated by Poole [13, 14]. As such, it is mainly based on disagreements opposing at least two witnesses to at least two others. The algorithm proceeds step by step, first assessing groups (`PCC.buildGroup`), then reconstructing or identifying their model, and then restarting from step one after eliminating the *codices descripti* from the database. It keeps going until there are less than four witnesses in the database.


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)))
x(Ix(BO)x(DK))
I
H
x(AH)
x(Cx(AH))
x(DK)
A
C
D
K

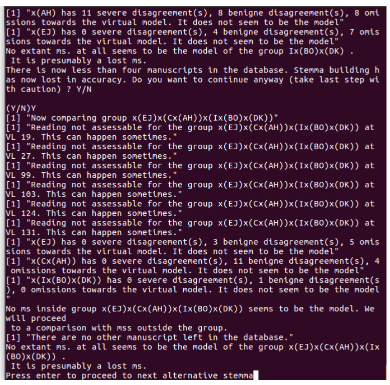

```
[1] "x(AH) has 11 severe disagreement(s), 8 benigne disagreement(s), 8 omi
ssions towards the virtual model. It does not seem to be the model"
[1] "x(EJ) has 0 severe disagreement(s), 4 benigne disagreement(s), 7 omis
sions towards the virtual model. It does not seem to be the model"
No extant ms. 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
```

Figure 5: User interface of the package: building the stemma (*Fournival* dataset)

Even though the algorithm can compute a plausible configuration, its default setting incites the expert to make is own decision regarding the very top of the stemma, which is the most delicate to assess [2]. Warned about the loss of reliability of the algorithm's computations, the user can however demand to display its final results.

4 Further developments

This first version of the package calls for new developments and improvements. Some minor revisions to improve the different visualisations are undergoing. To highlight the variant locations' different levels of conflictuality, the default mapping could be changed to a circular [5] or radial axis layout, where nodes could be ordered by decreasing centrality. Various mappings could be accessible to the user via a dedicated option. Other visual options could also be implemented for the *stemma codicum*. Obtaining a clearer, more customizable, ready for publication graph, is one of our short-term objectives.

More importantly, we are in the process of implementing other functions. One of them could be dedicated to a better detection of contamination, with procedures such as "cardiograms" [4, 20]. In the future, we also would like to offer an easy access to a large set of existing methods for stemma-building, to facilitate comparisons or benchmarks. The availability of the open source code and its online repository make it easy for others to help us complete our goal and contribute to this software's development.

References

[1] Andrews, Tara, Gershoni, Ido, Imhof, Ramona, Kaufmann, Sascha, Schaerer, Jakob, Studer, Thomas, & Zumbrunn, Severin, s.d., "Efficient Stemmatology: a Graph Database Application in the Digital Humanities", `http://home.inf.unibe.ch/~tstuder/papers/StemmaRest.pdf`.

[2] Bédier, Joseph, 1928, "La tradition manuscrite du *Lai de l'ombre* : réflexions sur l'art d'éditer les anciens textes", *Romania*, 54, p. 161-196 and 321-356.

[3] Camps, Jean-Baptiste & Cafiero, Florian, 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).

[4] 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.

[5] Doğrusöz, Uğur, Belviranli, M., & Dilek, A., 2012, "CiSE : A circular spring embedder layout algorithm", *IEEE Transactions on Visualization and Computer Graphics*, DOI : 10.1109/TVCG.2012.178 .

[6] Duval, Frédéric, 2015, *Les mots de l'édition de textes*, Paris (Magister).

[7] Sabidussi Gert, 1966, “The centrality index of a graph”, *Psychometrika*, 31, p. 581–603.

[8] Froger, Jacques, 1968, *La critique des textes et son automatisation*, Paris (Initiation aux nouveautés de la science).

[9] Fruchterman, Thomas M. J., & Reingold, Edward M., 1991, “Graph Drawing by Force-Directed Placement”, *Software – Practice & Experience*, 21-11, p. 1129–1164, DOI : 10.1002/spe.4380211102.

[10] GNU General Public License, version 3, 2007, Free Software Foundation, `http://www.gnu.org/licenses/gpl.html`

[11] Heikkilä, Tuomas, & Roos, Teemu, 2016, “Thematic Section on Studia Stemmatologica”, *Digital Scholarship in the Humanities* 31-3, p. 520-22, DOI : 10.1093/llc/fqw038.

[12] *Parvum lexicon stemmatologicum - PLS - HIIT Wiki*, ed. Caroline Macé & Philipp Roelli, 2015, `https://wiki.hiit.fi/display/stemmatology/Parvum+lexicon+stemmatologicum`.

[13] Poole, Eric, 1974, “The Computer in Determining Stemmatic Relationships”, *Computers and the Humanities*, 8-4, p. 207-16.

[14] Poole, Eric, 1979, “L’analyse stemmatique des textes documentaires”, in *La pratique des ordinateurs dans la critique des textes*, Paris, p. 151-161.

[15] Richart de Fornival, 1957, *Li Bestiaires d’Amours di maistre Richart de Fornival e li response du bestiaire*, ed. Cesare Segre, Milano & Napoli.

[16] R Development Core Team, 2014, *R : A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, `http://www.R-project.org`.

[17] Roos, Teemu, & Heikkilä, Tuomas, 2009, “Evaluating methods for computer-assisted stemmatology using artificial benchmark data sets”, *Literary and Linguistic Computing*, 24-4, p. 417-433.

[18] Schmidt, Desmond, & Colomb, Robert, 2009, “A data structure for representing multi-version texts online”, *International Journal of Human-Computer Studies*, 67-6, p. 497-514, DOI : 10.1016/j.ijhcs.2009.02.001.

[19] Spencer, M., Davidson, E. A., Barbrook, A. C., & Howe, C. J., 2004, “Phylogenetics of artificial manuscripts”, *Journal of Theoretical Biology* 227, p. 503–11, DOI : 10.1016/j.jtbi.2003.11.022.

[20] Wattel, E., & Van Mulken, M.J.P., 1996, “Shock Waves in Text Traditions, Cardiograms of the Medieval Litterature”, in *Studies in Stemmatology*, ed. P. Van Reenen, M.J.P. Van Mulken, Amsterdam, p.105-121.