



**HAL**  
open science

## Collation assistée par ordinateur de témoins de textes en ancien français

Jean-Baptiste Camps, Lucence Ing

► **To cite this version:**

Jean-Baptiste Camps, Lucence Ing. Collation assistée par ordinateur de témoins de textes en ancien français : défis et perspectives nouvelles. 2018. halshs-02023936

**HAL Id: halshs-02023936**

**<https://shs.hal.science/halshs-02023936>**

Preprint submitted on 18 Feb 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

# Collation assistée par ordinateur de témoins de textes en ancien français : défis et perspectives nouvelles

Jean-Baptiste Camps    Lucence Ing

Centre Jean-Mabillon  
École nationale des chartes  
Université Paris Sciences & Lettres  
jbcamps@hotmail.com  
lucence.ing@chartes.psl.eu

Anciens textes, nouveaux outils :  
des corpus alignés aux éditions multiples  
Colegio de España, 19 octobre 2018  
Atelier « Romanités Numériques »  
Laboratoire d'Études Romanes, Université Paris 8

# Enjeux principaux

## Constats sur la collation

- une étape importante dans le processus éditorial, qui ouvre la porte à une variété d'analyses ;
- un travail souvent long, complexe et fastidieux.

## Vers une collation assistée par ordinateur ?

Des algorithmes existent pour tenter d'automatiser ce processus, **mais...**

- inadaptés aux langues vernaculaires médiévales ;
- efficacité décroît fortement quand la variation ou le nombre de témoins augmente.

## Approche

- Penser la collation comme étape d'une chaîne de traitement global ;
- décomposer en unités de traitement plus aisées à accomplir.

# Outline

- 1 Introduction
  - La collation, pourquoi ? comment ?
  - Difficultés particulières des textes romans médiévaux
- 2 Visualiser l'alignement de deux versions avec une carte de chaleur
  - Matrice des distances d'édition
  - Élaboration de la carte de chaleur
  - Cas
  - Un outil suffisant ?
- 3 Structuration des données et collation automatique
  - Premières étapes
  - Collation automatique
  - Exploitation avancée des variantes
- 4 Conclusion

# La collation : pourquoi ?

## DÉFINITION

F. Duval, *Les Mots de l'édition de textes...*

*Comparaison des leçons de deux témoins ou plus.*

M. Buzzoni, C. Macé, *Parvum lexicon stemmatologicum*

*[Collation] can be defined as the comparative examination of the witnesses in order to determine the variant readings in all witnesses.*

Une phase importante de l'édition critique

- ① *recensio* des manuscrits ;
- ② *collatio* du texte des différents témoins ;
- ③ classement des manuscrits et établissement de leur généalogie, rendue par un stemma ;
- ④ établissement du texte, et *emendatio*.

*Faire de l'ordre dans la foule des variations que l'on observe dans les témoins.*

# La collation : processus

## Au lieu variant *v* du texte

*A* : ...maint el cuer...

*B* : ...as iex...

*D* : ...es iex...

*H* : ...maint el cuer...

*I* : ...maint en cuer...

*J* : ... et maint en cuer...

*O* : ...as iex...

## Modélisation

DOB :

D : es | OB : as

iex

AHIJ :

AHI :  $\emptyset$  | J : et

**maint**

IJ : **en** | AH : el

**cuer**

Entrée d'apparat (Richart de Fournival,  
*Li Bestiaires...*, éd. C. Segre)

maint en cuer] es (as OB) iex DOB

maint (et maint J) en (el AH) cuer *coet.*

# Comment ?

## Approche traditionnelle

- 1 choisir un témoin de référence (*exemplaire de collation*) ;
- 2 définir le type de variation que l'on veut retenir ;
- 3 sur une grande feuille, noter, par rapport à l'ex. de référence, ce que l'on voit dans les autres témoins ;
- 4 classer, éditer avec un appareil de variante, jeter la feuille...

## Approche computationnelle

- 1 faire l'acquisition du texte de tous les témoins (HTR, OCR, transcription...) ;
- 2 aligner et collationner avec l'aide d'un algorithme ;
- 3 retranscrire les données dans un modèle et une implémentation donnée (ex., TEI) ;
- 4 poursuivre le processus éditorial, tout en conservant l'ensemble du processus de collation.

# Approche computationnelle

## Avantages

- transparence ;
- reproductibilité ;
- pas de perte de données ;
- réutilisations.

## Inconvénients

- lourdeur ;
- complexité ;
- pas de perte de données ;
- résultats encore en partie insatisfaisants.

## Tradizione attiva (Vàrvaro)

### Chanson d'Otinel

B ke Guenes les traï od la salvage gent  
 A que li fel Guennes, le cuvers sodiant,  
 les i vendi, ce sevent li auquant,  
 cel jor meismes qu'il furent combatant.

### Chrétien de Troyes, *Chev. au lion*, v. 38

H Que toz jorz durra li renons  
 P En tant qu'i nomment des boins les nons  
 V Que toz jors vivera lor nons  
 F Que tos jors mais durra ses nons  
 G Q'au mains touz jorz vivra ses nons  
 A Que tos jors mais dura ses nons  
 S Que tous jours mais dura ses nons  
 R C'al mains tous tans vivra ses nons

# Variation graphique

forme	occurr.	forme	occurr.
cheval	785	ceux	10
cheual	375	cevax	10
chevaus	248	ceuaus	9
ceval	98	chiuau	9
chevax	92	cheuax	8
chevals	84	kevaus	6
ceual	66	chevau	5
cheuaus	65	cevax	3
chival	34	chivals	3
chevaux	30	cheuas	2
chivaus	27	keval	2
cheuax	23	chaval	1
chiual	23	chavaux	1
cevaus	19	cheua	1
chevas	19	cheualx	1
cheuals	14	cheuau	1
cevals	12	chevalx	1
chiuaus	11	chiuals	1

Identique ou non ?

Cait del fuere

Chiet dou fuerre

Kiet du feurre

Évident pour le philologue... mais  
pour l'algorithme ?

## Les corpus

*Chanson d'Otinel* chanson de la fin du XII<sup>e</sup> siècle ; un témoin continental (Reg. lat. 1616 ; Saint-Brieuc, 1317 ; traits de l'Est) ; deux témoins anglo-normands (fragm. NAF 5094, XII<sup>ex</sup>-XIII<sup>inc</sup> ; ms. Bodmer 168, XIII<sup>3/3</sup>) ; nombreuses traductions médiévales (anglaises, galloise, norroises, italienne ; traces d'une diffusion ibérique).

*Chevalier au lion* de Chrétien de Troyes, à partir des transcriptions alignées de Pierre Kunstmann et Kajsa Meyer. Plus de 13 témoins.

*Lancelot en prose* Deux témoins : Ao, daté du premier tiers du XIII<sup>e</sup> siècle, et Ez, incunable imprimé en 1488, servant une étude sur la disparition lexicale en diachronie.

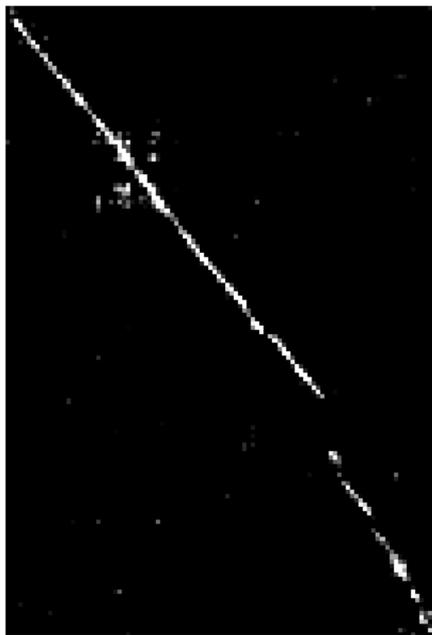
# Outline

- 1 Introduction
  - La collation, pourquoi ? comment ?
  - Difficultés particulières des textes romans médiévaux
- 2 Visualiser l'alignement de deux versions avec une carte de chaleur
  - Matrice des distances d'édition
  - Élaboration de la carte de chaleur
  - Cas
  - Un outil suffisant ?
- 3 Structuration des données et collation automatique
  - Premières étapes
  - Collation automatique
  - Exploitation avancée des variantes
- 4 Conclusion

# Matrice des distances de Levenshtein

Tém. 1 \ Tém. 2	Cant	a	este	lancelot	en	la	garde
Or	4	2	4	8	2	2	4
dit	3	3	3	7	3	3	5
li	4	2	4	7	2	1	6
contes	4	6	4	6	5	6	5
que	4	3	3	7	3	3	4
tant	1	3	4	5	3	3	4
a	3	0	4	7	2	1	4
esté	4	4	1	7	3	4	5
Lanceloz	6	7	7	2	7	7	6
en	3	2	3	7	0	2	5
la	3	1	4	6	2	0	4
garde	4	4	4	6	5	4	0

# Image-texte dans ImageJ



**FIGURE** – Carte de chaleur binarisée obtenue à partir des données de la matrice des distances entre les deux témoins

# Carte de chaleur de la matrice après traitement

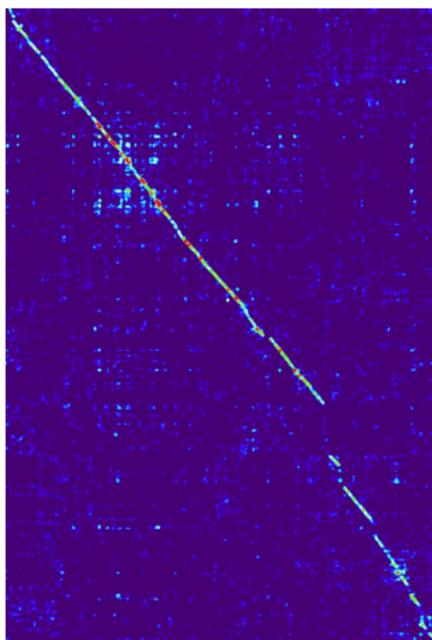


FIGURE – Alignement de deux témoins du *Lancelot en prose*

# La visualisation : intérêts et limites

- 1 un outil intéressant :
  - une **première approche** des données
  - un **point de vue global** qui permet de dégager la structure textuelle qu'entretiennent des témoins

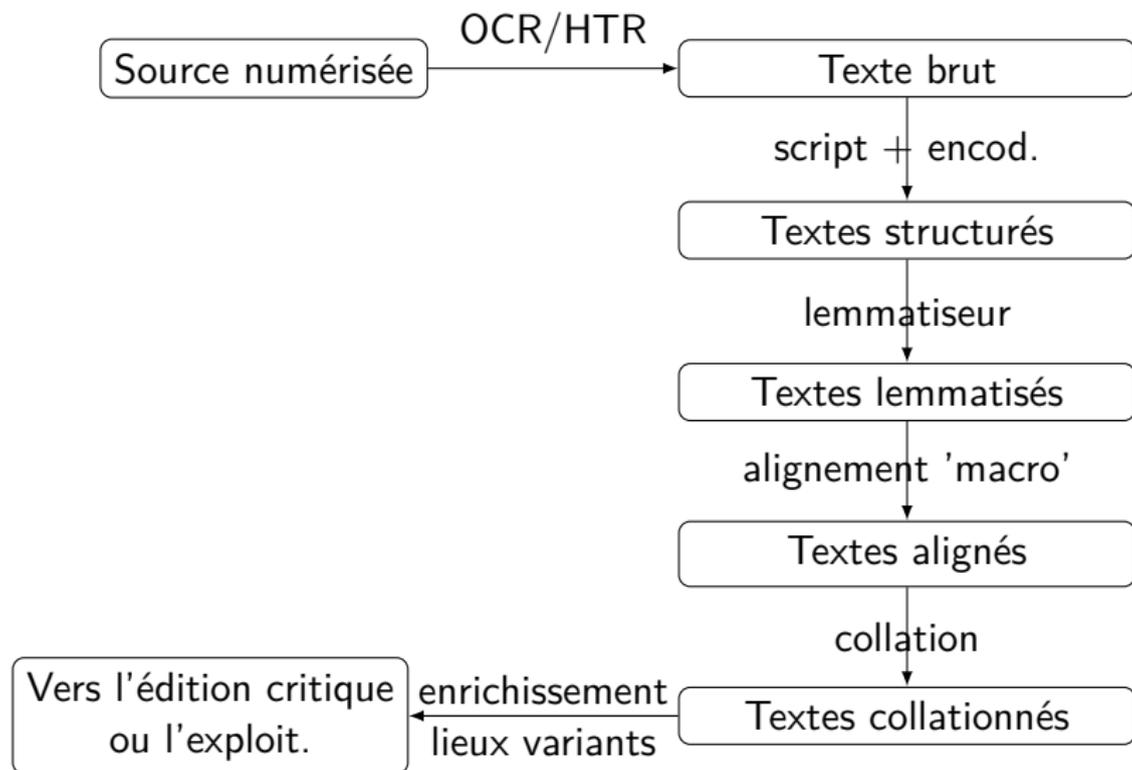
# La visualisation : intérêts et limites

- 1 un outil intéressant :
  - une **première approche** des données
  - un **point de vue global** qui permet de dégager la structure textuelle qu'entretiennent des témoins
- 2 mais un outil insuffisant :
  - **seuls deux témoins** peuvent être traités en même temps
  - l'**exploitation** des données textuelles à partir de cette visualisation n'est **pas possible**

# Outline

- 1 Introduction
  - La collation, pourquoi ? comment ?
  - Difficultés particulières des textes romans médiévaux
- 2 Visualiser l'alignement de deux versions avec une carte de chaleur
  - Matrice des distances d'édition
  - Élaboration de la carte de chaleur
  - Cas
  - Un outil suffisant ?
- 3 Structuration des données et collation automatique
  - Premières étapes
  - Collation automatique
  - Exploitation avancée des variantes
- 4 Conclusion

# La chaîne de traitement



# En amont de la collation...

## Étapes

- OCR/HTR (OCRopy, Kraken, Transkribus...);
- post-correction (PoCoTo; relecture);
- encodage selon le modèle TEI (script; manuellement);
- lemmatisation (Pandora, Lemming);

## OCRopy

- réseaux récurrents LSTM;
- logique d'apprentissage : préparation des données, entraînement d'un modèle, application, correction...

book/inc1-100/010040.bin.png

**droit a monfeigneur gauuain et lui a eulz, et**

droit a monfeigneur gauuain et lui a eulz, et

book/inc1-100/010041.bin.png

**cuida que brehin fift ainfi mais non fift. Lun**

cuida que brehin fift ainfi mais non fift. Lun

book/inc1-100/010042.bin.png

**des cheualiers fiert meffire gauuain en lefcu**

des cheualiers fiert meffire gauuain en lefcu

book/inc1-100/010043.bin.png

**fi fort que fa lance volle en pieces. Et meffire**

fi fort que fa lance volle en pieces. Et meffire

book/inc1-100/010044.bin.png

**gauuain lattaît de telle force quil le porte a ter**

gauuain lattaît de telle force quil le porte a ter

FIGURE – Fichier html à éditer

# Notre approche de la collation

Pour pallier aux difficultés posées aux algorithmes par les particularités des textes vernaculaires, notre approche se décompose comme suit :

- alignement 'macro', des paragraphes ou des vers entre les témoins ;
- collation faite sur les lemmes (et non pas les formes) ;
- récupération de l'information sur les formes, et restructuration de la collation en fonction ;
- annotation des lieux variants.

# Alignement 'macro'

## Situation actuelle

- pour l'instant, à la main ;
- nouvelle approche en test : utilisation d'un algorithme calculant des similarités entre passages, de type 'text reuse'.

## Un outil possible, **Iteal** (Jänicke and Wrisley 2017)

- chaque témoin,  $A, B$  divisé en unités de structures (lignes),  $A_1, \dots, A_n$  et  $B_1, \dots, B_n$ ,
- calcul de distance entre chaque couple possible  $\{A_i, B_j\}$ 
  - $n$ -grams,
  - continus ou non,
  - approximatifs (Relative Edit Distance,  $RED(A, B) = \frac{2 \times LevDist(A, B)}{|A| + |B|}$   
(défaut :  $RED \leq 0,5$ ).
- rapprochement des vers qui dépassent le seuil fixé.

# Premiers résultats avec Iteal

interactive text  
edition alignment

Data Source:

Edition 1:

Edition 2:

---

String Similarity: 70%

MinCoverage: 40%

NgramMin: 3

BrokeNgrams?

 no  yes

Read Direction

 ltr  rtl  
(only display)

Cut Text

 no  yes  
(only display)

---

Lancelot

AF\_travail.txt

MF\_travail.txt

AF\_travail.txt

Comment fait messire Yvains biax dolz ...  
Sire fait li vallez qe ne serai or mon...  
Certes fait il mout volentiers  
Messire Yvains s an vait au roi si li dit  
Sire vostres vallez vos mande par moi...  
Li qex vallez fait li rois  
Sire fait il li vallez qui arsoir vos f...  
A ces paroles vint la reine parmi la sa...  
Comment fait li rois si velt ja estre ch...  
Voire fait il demain el jor  
Oëz Gauvains fait li rois de nostre va...  
Qui est fait la reine cil vallez  
Qui dame fait messire Yvains c est tre...  
Lors li conte comment il avoit esté ...  
Comment fait la reine ersoir vint a co...  
Voire dame fait messire Yvains car il ...

MF\_travail.txt

Messire yvain va au roy et lui dist  
Sire vostre varlet veult et vous prie ...  
Le quel varlet fait le roy  
Vostre varlet de er soir qui veult ja es...  
Certes fait messire gauvain il a grant ...  
Comment fait le roy veult il ja estre c...  
Oy sire fait il demain au jour  
En verite fait le roy il le sera  
Qui est ce varlet fait la royne  
Qui il est fait messire yvain C est le...  
Lors lui compte comment il avoit est...  
Comment fait la royne er soir vint a ...  
Certes fait messire yvain il en a trop ...  
Je le verroie volentiers fait la royne  
Dame fait le roy vous le verrez le mie...  
Lors dist a messire yvain qu il le voi...

J.B. Camps, L. Ing (ENC)

Collation assistée par ordinateur

Rom. num., 19 oct. 2018

21 / 31

# CollateX (Dekker et al., v. 2.2.)

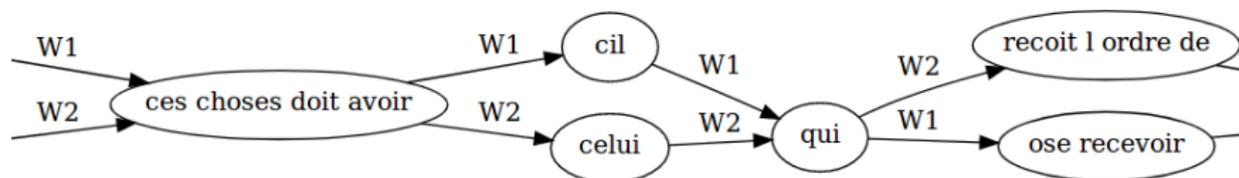


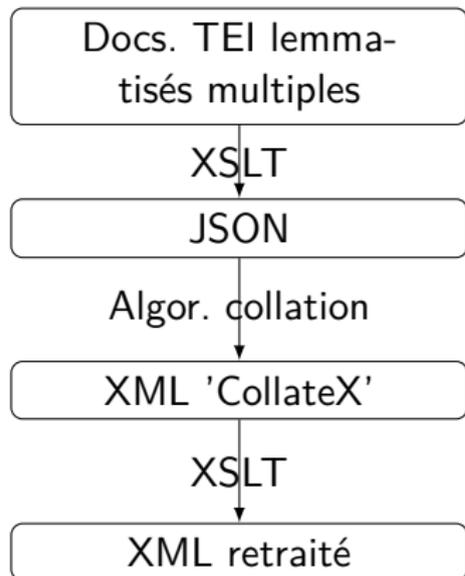
FIGURE – Exemple de graphe issu de CollateX

- outil de collation assistée par ordinateur ;
- matrice en n-dimensions des textes ;
- alignement des séquences identiques ;
- cherche les 'diagonales' les plus longues possibles ;
- modèle graphe (DAG).

## Modèle de Göteborg :

- tokénisation
- alignement
- détection des transpositions
- visualisation

# Traitement en Python avec CollateX



## XML

```

<w lemma="il"
type="PR0per | PERS.=3 | NOMB.=s | GENRE=m
| CAS=i" xml:id="w_A_002">lui</w>
  
```

## JSON

```

{
  "t": "lui",
  "xml:id": "w_A_002",
  "lemma": "il",
  "POS": "PR0per",
  "morph": "PERS.=3 | NOMB.=s | GENRE=m | CAS=i"
}
  
```

# Sorties de Collatex

## TABLE D'ALIGNMENT

A	B
a	a
si	si
preudome	preudomme
com	comme
vos	vous
iestes	estes
ne	-
doi	-
ge	je
-	ne
-	dois
pas	pas
mon	mon
non	nom
celer	celer
et	-
gel	-
vos	-
dirai	-

CollateX propose des sorties multiples :

Table d'alignement

Graphe de variantes

JSON

XML

XML/TEI

TEI

`<app>`

```
<rdg wit="#A #H #M #P #S #V">lui</rdg>
```

```
<rdg wit="#F">li</rdg>
```

```
<rdg wit="#G #R">soi</rdg>
```

`</app>`

# Erreurs produites par l'alignement automatique

```

<app>
  <rdg ana="MODE=con|PERS.=3|NOMB.=s" lemma="savoir" type="VERcjpg" wit="#A"
    xml:id="Ao_w_008131">savroit</rdg>
  <rdg ana="MODE=ind|TEMPS=fut|PERS.=3|NOMB.=s" lemma="savoir" type="VERcjpg" wit="#B"
    xml:id="Ez_w_006332">saura</rdg>
</app>
<app>
  <rdg ana="MORPH=empty" lemma="enseignier" type="VERinf" wit="#A"
    xml:id="Ao_w_008132">enseignier</rdg>
</app>
<app>
  <rdg ana="DEGRE=c" lemma="mieus" type="ADVgen" wit="#A" xml:id="Ao_w_008133"
    >miauz</rdg>
  <rdg ana="DEGRE=c" lemma="mieus" type="ADVgen" wit="#B" xml:id="Ez_w_006333"
    >mieulx</rdg>
</app>
<app>
  <rdg ana="MORPH=empty" lemma="de" type="PRE" wit="#A" xml:id="Ao_w_008134">de</rdg>
  <rdg ana="MORPH=empty" lemma="enseignier" type="VERinf" wit="#B"
    xml:id="Ez_w_006334">enseignier</rdg>
</app>
<app>
  <rdg ana="MORPH=empty" lemma="quel" type="CONsub" wit="#B" xml:id="Ez_w_006335"
    >que</rdg>
</app>

```

# Questions de modélisation

## Principes

- **un lieu variant est la plus grande unité de co-variation d'un même type.**
- dans l'annotation des variantes, distinguer ce qui caractérise la **relation** entre plusieurs variantes, ou la variante en elle-même ;
- distinguer en outre ce qui concerne le type de variation, la cause et la source.

# Typologie

app/@type

**1. même lemme, partie du discours & morph., forme différente :**

**graphic** diatopic, diachronic, ... , ex. *chivalier/chevalier*

**2. même lemme & partie du discours, morph. différente :**

**flexional** verbal, nominal,... ex. *chivalier/chivaliers*

**3. même lemme, partie du discours différente :**

**morphosyntactic** substantivation, ... Ex. *mangier/(li) mangier(s)*

**4. lemme différent :**

**derivational** prefix, suffix,... ex. *creanter/acreanter*

**synonymism** synonymes, hypero-/hyponymes, cohyponymes,  
holo-/méronyme (paronymes?). Ex., *chevalier/baron*

**semantic** nonsense, equipollent, *difficilior/facilior*... Ex. *chevalier/charete*

# De la sortie de CollateX à des documents enrichis

## Sortie CollateX

```
<app>
  <rdg wit="#A">venimeus</rdg>
  <rdg wit="#F">venimeus</rdg>
  <rdg wit="#P">enuious</rdg>
  <rdg wit="#R">venimex</rdg>
</app>
```

## SORTIE TRAITÉE (PROVISOIRE)

```
<app type="substantive">
  <rdg wit="#A #F #R" lemma="venimos" POS="NOMcom">
    <app type="graphic">
      <rdg wit="#A #F">venimeus</rdg>
      <rdg wit="#R">venimex</rdg>
    </app>
  </rdg>
  <rdg wit="#P" lemma="envios" POS="VERppe"
    morph="NOMB.=p|GENRE=m|CAS=r">enuious</rdg>
</app>
```

# Exploitation des variantes

## Les textes après la collation

- annotés linguistiquement
- collationnés au mot près
- dont les variantes sont précisées

## Fin de la chaîne de traitement

- possibilités d'édition numérique
- études linguistiques

# Outline

- 1 Introduction
  - La collation, pourquoi ? comment ?
  - Difficultés particulières des textes romans médiévaux
- 2 Visualiser l'alignement de deux versions avec une carte de chaleur
  - Matrice des distances d'édition
  - Élaboration de la carte de chaleur
  - Cas
  - Un outil suffisant ?
- 3 Structuration des données et collation automatique
  - Premières étapes
  - Collation automatique
  - Exploitation avancée des variantes
- 4 Conclusion

## Pour finir...

### Bilan d'étape

- une chaîne de traitement encore à l'état de prototype ;
- qui devra être complétée pour diminuer la part d'intervention manuelle ;
- mais des résultats améliorés sur la collation.

### Remerciements

Un grand merci à Daniel Stockholm (EPHE | PSL) pour les cartes de chaleur.

## Pour finir...

### Bilan d'étape

- une chaîne de traitement encore à l'état de prototype ;
- qui devra être complétée pour diminuer la part d'intervention manuelle ;
- mais des résultats améliorés sur la collation.

### Remerciements

Un grand merci à Daniel Stockholm (EPHE | PSL) pour les cartes de chaleur.

**Merci de votre attention !**