Experiments with Collatex

Collation and the categorisation of variants

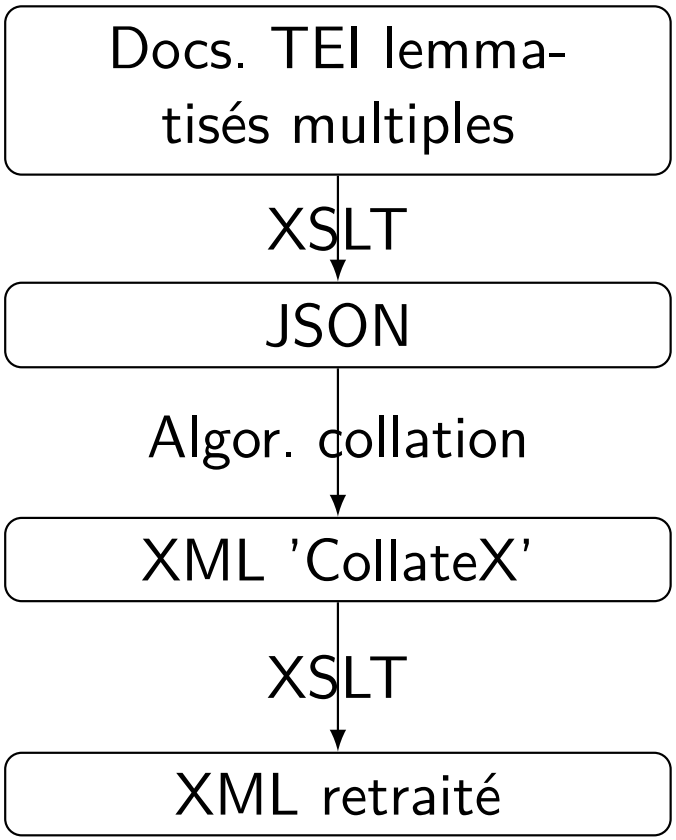
Jean-Baptiste Camps Lucence Ing

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

17 janvier 2019

Outline

Traitement en Python avec CollateX



XML

```
<w lemma="il"
type="PROper|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": "PROper",
  "morph": "PERS.=3|NOMB.=s|GENRE=m|C
}
```

Sorties de Collatex

Table d'aligment

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="VERcjg" wit="#A"
        xml:id="Ao_w_008131">savroit</rdg>
    <rdg ana="MODE=ind|TEMPS=fut|PERS.=3|NOMB.=s" lemma="savoir" type="VERcjg" 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>
```

Outline

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*

Typologie : remarques et questions

Typologie Camps (en cours d'élaboration)

Remarques

- ces typages peuvent être (majoritairement) inférés à partir de l'annotation linguistique.
- seule la distinction *derivational* / *synonymism* / *semantic* demande une instrumentation un peu plus avancée.

Questions

- Pour l'instant, les cat. de lieux variants sont graduées. Faut-il distinguer, par ex., une catégorie même PdD, lemme différent ?
- cette typologie est très adaptée à des variantes sur un seul mot, qui seront majoritaires vu la granularité typologique. Comment typer les variantes plus macrostructurelles d'un même type (plusieurs lemmes différents de suite) ? substantive ?

Cause de variation (rdg/cause)

Typologie Camps (très provisoire)

Involontaires

Erreurs de lecture

confusion paléographique lettres, ponct., abr. mal lues

saut haplologie, saut du même au même, homéotéleute/-archie...

répétition dittographie, ...

Erreurs linguistiques

graphologique mauvaise compr. d'une abr., segmentation,

phonologique homo-/parophonie,...

contextuelle parallélismes, accord, ...

Indéterminables

neutralisation dé-régionalisation, modernisation ;

banalisation noms propres, mots rares,...

Volontaires

interpolation remplissage d'un blanc/lacune, amplification, glose ...

abrégement

réécriture stylistique, métrique, thématique...

(fausse-)correction fautes critiques.

Cause de variation : remarques et problèmes

Typologie Camps (très provisoire)

Nécessité de faire un retour sur Havet 1911, Marichal 1956 et 1971, Guyotjeannin et al. 2001 (ainsi que Willis, 1972 ; West, 1973).

Cause et source

En réalité, faut-il encore distinguer cause et source, qui peuvent être deux choses différentes ? Par ex.,

- saut et proximité de séquences voisines ?
- Mauvaise résolution d'une abréviation et archaïsme graphique de la source ?
- Réécriture d'un passage et sens défectueux de la

Fautes à histoire

Peut-être faut il traiter différemment fautes de 1re génération et "à histoire" (de 2e, 3e ...ne génération).

Ex. (Havet, § 1521), metus / motus,
m*eo*tus > me otius > meo totius

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="envïos" 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