Traitements numériques pour l'analyse du changement linguistique

Journée d'études « Philologie computationnelle : au delà de l'encodage du texte »

Lucence Ing

Centre Jean Mabillon, École des chartes – PSL

2 décembre 2021

1. Introduction

Disparitions lexicales en diachronie

➥ étudier les **disparitions lexicales** entre ancien français et moyen français ; essayer de voir si des mouvements systématiques se dessinent

Au sein d'un même texte, le *Lancelot* en prose. Comparaison principale basée sur **deux témoins** :

- **Ao** : BnF français 768 (premier tiers XIII^{e}), d'après l'édition de E. Kennedy
- **Ez** : Mazarine, Inc. 491 (édition *princeps* du *Lancelot*, parue à Rouen chez Jean et Gaillard le Bourgeois en 1488)

Témoins complémentaires (sur quelques parties) :

- Rennes BM, ms 255 (premier tiers XIII^{e}, scripta centrale)
- Bristish Library, Add. 10293, d'après éd. de A. Micha (ca. 1316, pic.)
- BnF, fr. 16999 (milieu XIV^{e}, Paris)
- BnF, fr. 113-4 (dernier quart XV^{e}, Poitiers)

Le texte du *Lancelot*

Le texte

- anonyme, premier tiers du XIII^{e} siècle
- **partie centrale** du *Lancelot-Graal*
- un texte **très diffusé** jusqu'à la fin du XVI^{e} siècle

Les **versions**

- version longue et version courte
- version de Londres et version de Paris à l'intérieur de la version longue
- contaminations diverses

La tradition du texte

Quant à vouloir fixer strictement la place des manuscrits les uns par rapports aux autres et ***dessiner un de ces beaux arbres généalogiques*** *dont s'ornent les éditions critiques, il* ***n'y faut pas songer.***
(A. Micha, « La Traduction manuscrite du *Lancelot* en prose », 1964)

Nos témoins

- *Lancelot* traditionnellement divisé en trois parties : le **Galehaut**, la Charette et l'Agravain
- les versions commencent à diverger lors l'épisode du deuxième voyage en Sorelois (fin du Galehaut)
- notre étude porte sur la **partie similaire** du *Lancelot*

La question de l'obsolescence des mots

Études sur le lexique

- ce qui est souvent étudié : les **apparitions**
- surtout dans l'**étude diachronique** de la langue de l'**ancien au moyen français**

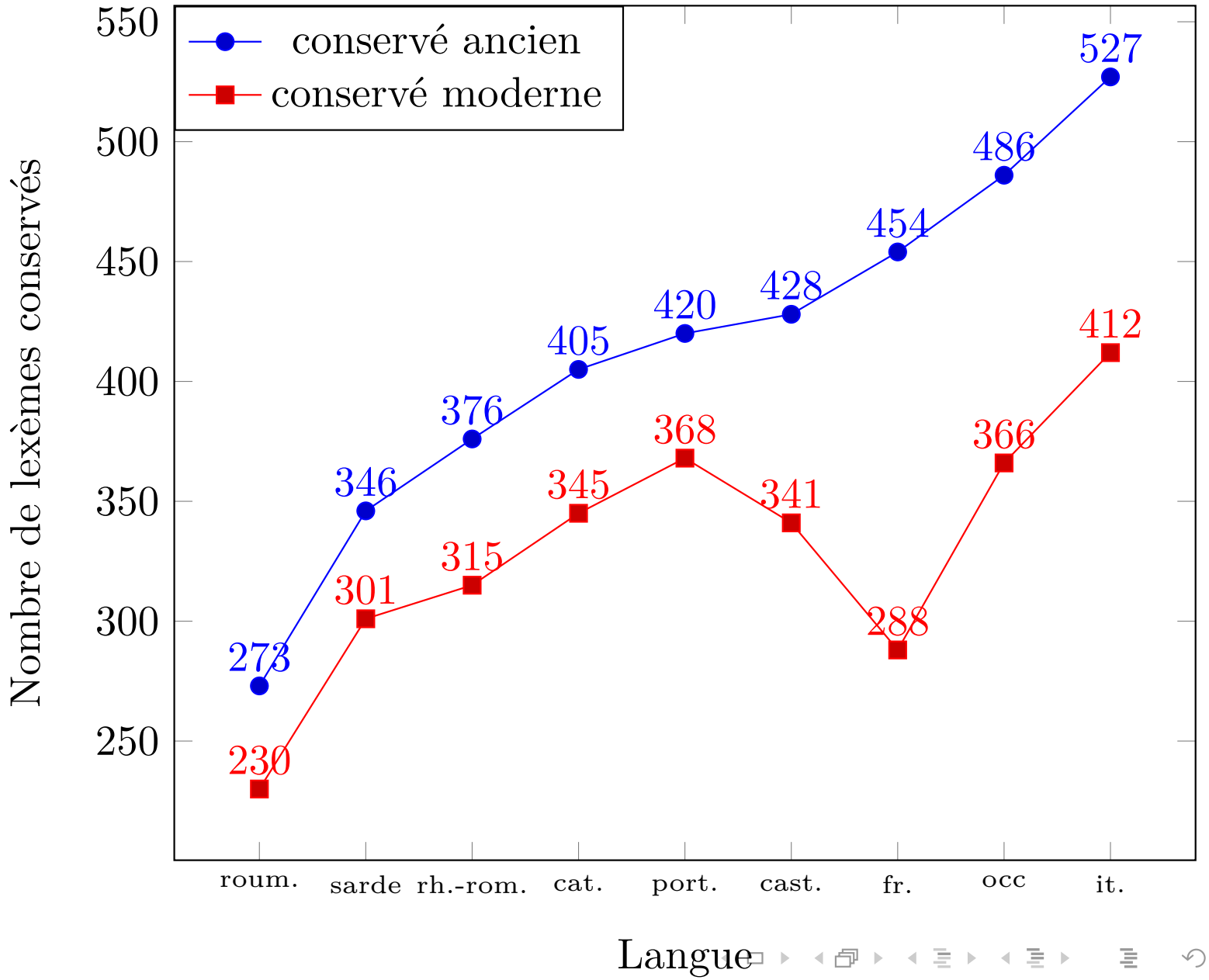
Comme une mort fait moins plaisir qu'une naissance, il n'y a que peu de lexicographes qui essaient d'indiquer avec une précision comparable [à celle donnée pour les apparitions] la date où ces mots-là ont cessé de circuler.

C. Wittlin, « Qu'est-ce qui a tué *ocire* ? », 1989

Les **disparitions**

- un des huit mécanismes du **changement linguistique** (C. Marchello-Nizia, 2006)
- un « **épiphénomène** » (C. Badiou-Monferran, 2008) **?**
- quelques études : C. Badiou-Monferran & et T. Verjans, 2015 ; M. Glessgen, 2008 ; S. Dworkin, 2011 ; W. de Mulder et al., 2020

Les disparitions de lexèmes dans les langues romanes d'après A. Stefenelli (1992)



Les disparitions : approche typologique

Classification des facteurs	Facteurs de faiblesse	Type de disparition	Statut de la concurrence	Qualité du remplacement
formel	faiblesse phonétique	disparition totale	concurrence sémantique	remplacement par un lexème
formel	homophonie	disparition totale, partielle ou marginalisation	concurrence sémantique	remplacement par un lexème
formel	faiblesse phonotactique	disparition totale	concurrence sémantique	remplacement par un lexème
formel	complexité morphologique (isolement morpho-syntaxique)	disparition totale, partielle ou marginalisation	concurrence sémantique et formelle	remplacement par un lexème
formel (sémantique)	isolement morpho-lexical	disparition totale, partielle ou marginalisation	concurrence sémantique et formelle	remplacement par un ou plusieurs lexème(s)
sémantique	faiblesse sémantique	disparition totale, partielle ou marginalisation	concurrence sémantique	remplacement par un ou plusieurs lexème(s)
sémantique	polysémie excessive	disparition totale, partielle ou marginalisation	concurrence sémantique	remplacement par un ou plusieurs lexème(s)
extra-linguistique	faiblesse référentielle	disparition totale ou marginalisation	pas de concurrence	pas de remplacement
extra-linguistique	tabou	disparition totale	concurrence sémantique	remplacement par un lexème
extra-linguistique	concurrence avec des lexies plus prestigieuses	disparition totale	concurrence sémantique	remplacement par un lexème

L'intérêt de la recherche

Lexique et **évolution des mentalités**

Il semble qu[e ceux qui ont étudié l'histoire du vocabulaire français] ne voient la vie et le mouvement que dans la naissance de néologismes, tandis qu'en vérité le registre des décès n'est pas moins important quand on veut caractériser une époque. Et avant tout, il faudrait établir les rapports entre les mots qui s'en vont et ceux qui prennent leur place. Ici tout est encore à faire. Et pourtant ***l'histoire de la civilisation tout entière s'y reflète****.*

W. von Wartburg, *Évolution et structure de la langue française*, 1934

Étudier l'évolution **au sein d'un même texte**

- des **passages identiques** ou similaires
- la **modernisation** du texte par des locuteurs de la fin du XV^e^ siècle : le **sentiment des locuteurs** de l'obsolescence
- complexité de la question : **archaïsmes** comme « **gages de littérarité** » du texte (D. Capin, 2004 ; M. Colombo-Timelli, 2017)

Identifier les lexèmes disparus et les analyser

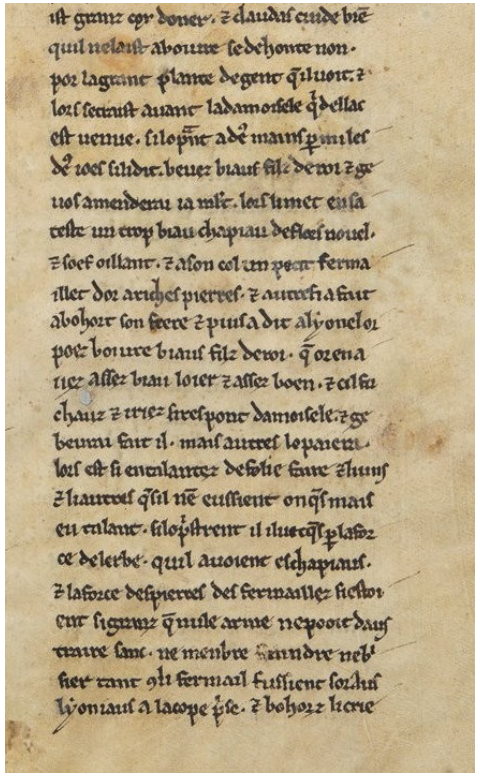
bëer

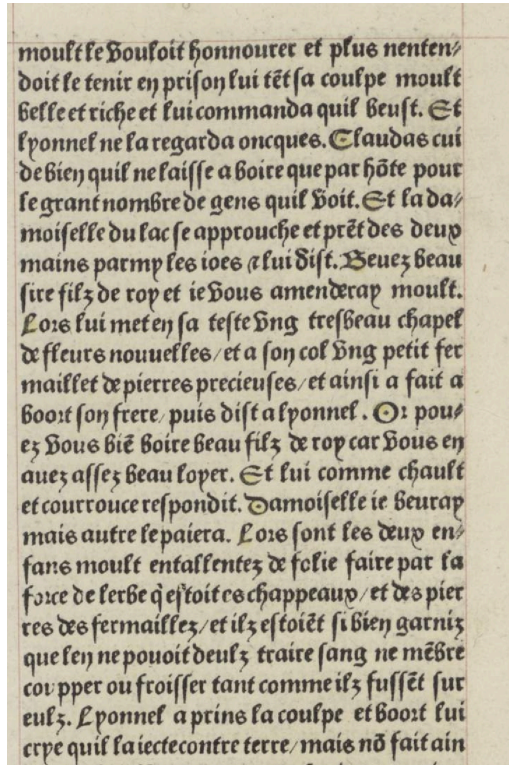
Onques par amors n'avoit amé c'une foiee et qant l'an li demandoit por quoi il avoit aimors laissiees, si disoit que por ce qu'il baoit a vivre longuement.

Oncques n'avoit aymé fors une fois et quant on lui demandoit pourquoy il avoit amours laissees, il disoit qu'il desiroit vivre longuement.

... ele ot fait grant partie de ce que ele baoit affaire, si fu mout liee et petit prisa lo cop...

... elle eust fait grant partie de ce qu'elle vouloit faire, elle fut moult joyeuse et peu prisa le coup...





fronchier
FEW, x, 470b : RONCARE

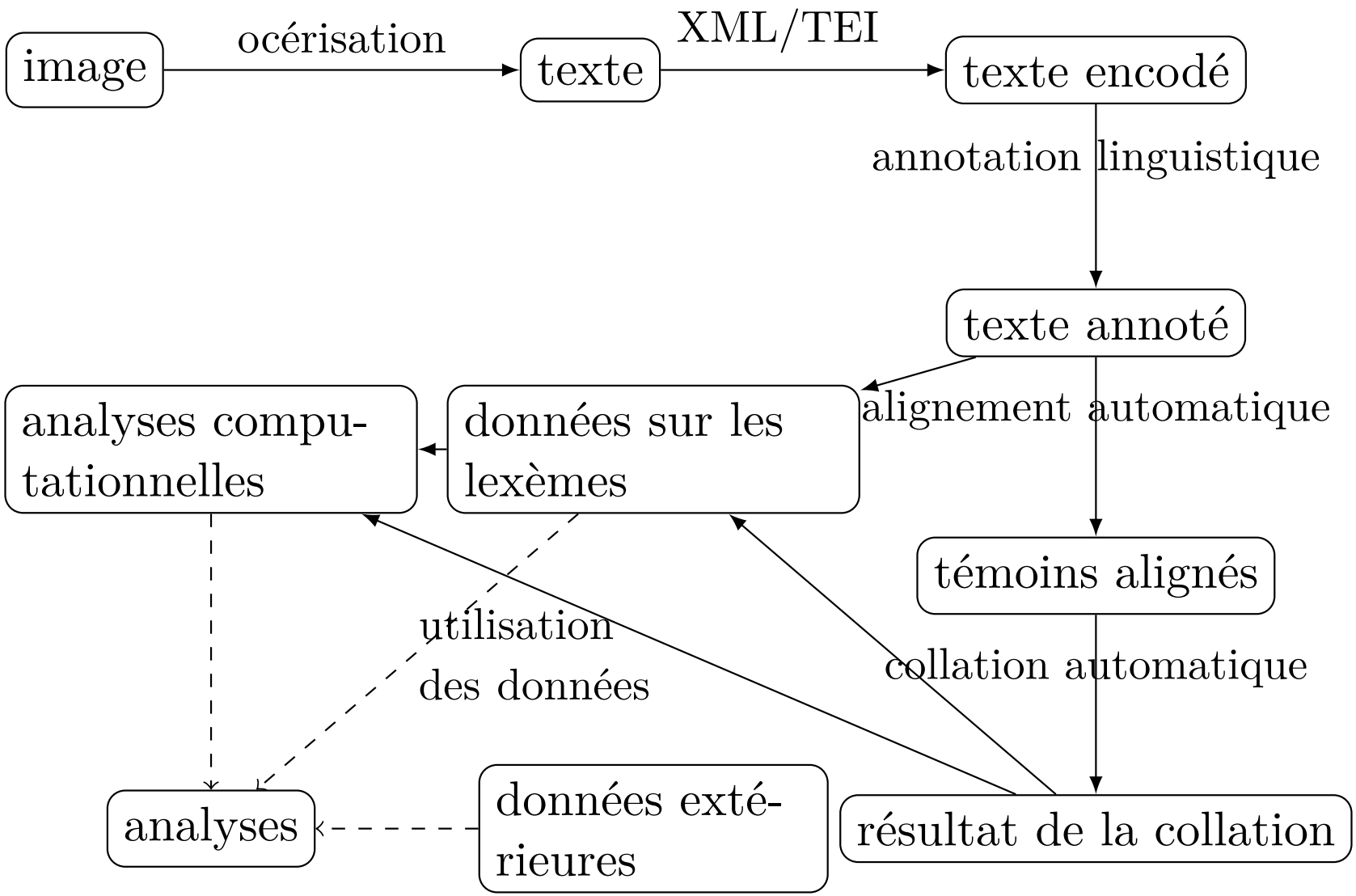
Ao : 1 ; Ez : 0

"souffler bruyamment" (CoincyI18K – PercefR2)

*… et **fronchoit** del nes an sa grant ire autresin com uns chevax.*
*… et **rouffloit** du nez en sa grant ire ainsi comme ung cheval.*
(9-18)

Lexème qui remplace : *ronfler*, (EneasS2 – Cresp 1637 ; RONFL-)

La chaîne de traitements



2. Structurer les données

Océrisation

L'océrisation

- **OCR** : *Optical Character Recognition* ou reconnaissance optique de caractères
- **transcription automatique** de caractères imprimés

Exemple d'interface de correction d'OCR

Fonctionnement de l'océrisation

Apprentissage profond

- un modèle…
- … qu'on **entraîne** sur des données pré-corrigées…
- … qu'on **applique** sur d'autres données…
- … qu'on corrige…
- … dans une sorte de boucle vertueuse d'**amélioration du modèle**

Utilisation d'`ocropy`, maintenant dépassé

Encodage XML/TEI (1)

Encodage (semi-)automatique

tokénisation

```
<w xml:id="Ao_w_0135673">il</w>
<w xml:id="Ao_w_0135674" rend="aggl">n</w>
<w xml:id="Ao_w_0135675">eussient</w>
```

régularisations (*s* longs, lettres ramistes, etc.)

```
<w xml:id="Ao_w_0005127"><choice><orig>ſ</orig>
<reg>s</reg></choice>i</w>
<w xml:id="Ao_w_0005128">a<choice><orig>u</orig>
<reg>v</reg></choice>oient</w>
<w xml:id="Ao_w_0005129">a</w>
<w xml:id="Ao_w_0005130">fe<choice><expan><ex>m
</ex></expan><abbr><am>&#x0303;</am></abbr>
</choice>mes</w>
```

éléments de bibliographie matérielle (foliotation, titres courants)

```
<pb n="a.i.r" source="#Ez"/>
<cb n="a"/>
<lb facs="bbox 1360 60 1760 135" break="yes"/>
<fw type="header" place="top">
<w xml:id="Ao_w_0005108">La</w>
<w xml:id="Ao_w_0005109">premiere</w>
<w xml:id="Ao_w_0005110">partie</w>
</fw>
```

Encodage XML/TEI (2)

Encodage manuel

divisions en chapitre

```
<div type="chap" rend="div" n="029"
xml:id="Ez_025b" corresp="Ao_029">
....
</div>
```

```
<w xml:id="Ez_w_0010495">ma<choice><sic>m</sic>
<corr resp="#LI" type="coq">n</corr></choice>
iere</w>
```

corrections ponctuelles sur l'incunable

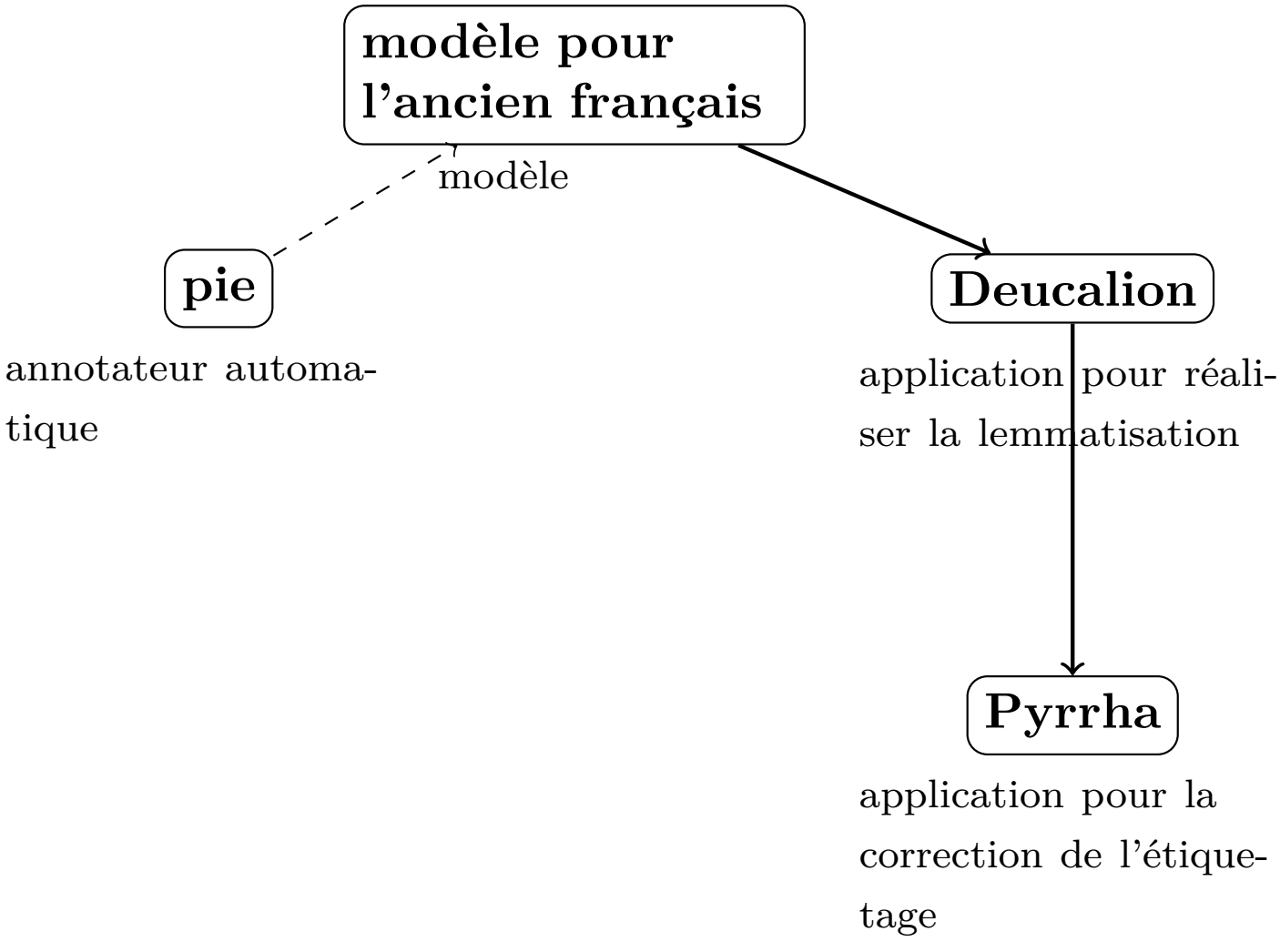
```
<w xml:id="Ez_w_0072760">
<choice><sic>di</sic>
<corr resp="#LI" type="repetLigne"/>
</choice>
<lb source="#Ez" break="no" rend="ind"
facs="bbox 1592 2300 2549 2367"/>dieu
</w>
```

Annotation linguistique : outils

Annotation et correction

- annotateur automatique : **pie** (E. Manjavacas, T. Clérice, M. Kestemont)
  https://doi.org/10.5281/zenodo.4572585
- interface Deucalion (T. Clérice)
  https://dh.chartes.psl.eu/deucalion/
- **modèle pour l'ancien français** entraîné à l'École nationale des chartes (J.-B. Camps et al.)
- correction avec l'interface Pyrrha (T. Clérice, J. Pilla et al.)
  https://doi.org/10.5281/zenodo.5144781

Schéma des outils d'annotation linguistique



Annotation linguistique : principes

Étiquettes retenues pour le projet : **lemmes** et **partie du discours** (POS, *part-of-speech*)

Modèle pour l'ancien français

lemme dictionnaire Tobler-Lommatzsch
`https://www.ling.uni-stuttgart.de/institut/ilr/toblerlommatzsch/work/workfr.htm`

POS en suivant la documentation Cattex
`http://bfm.ens-lyon.fr/IMG/pdf/Cattex2009_manuel_2.0.pdf`

Ao	ses	hiaumes	fu	toz	fanduz
Ez	son	heaume	fut	tout	fendu
lemma	son4	heaume	estre1	tot	fendre
pos	DETpos	NOMcom	VERcjg	ADVgen	VERppe

Annotation dans le fichier XML

```
<w xml:id="Ao_w_0135673" lemma="il" pos="PROper">il</w>
<w xml:id="Ao_w_0135674" lemma="ne1" pos="ADVneg">n</w>
<w xml:id="Ao_w_0135675" lemma="avoir" pos="VERcjg">eüssient</w>
<w xml:id="Ao_w_0135676" lemma="mie" pos="ADVneg">mie</w>
<w xml:id="Ao_w_0135677" lemma="vëoir" pos="VERppe">veü</w>
<w xml:id="Ao_w_0135678" lemma="le" pos="DETdef">lo</w>
<w xml:id="Ao_w_0135679" lemma="lïon" pos="NOMcom">lion</w>
<w xml:id="Ao_w_0135680" lemma="en1" pos="PRE">en</w>
<w xml:id="Ao_w_0135681" lemma="le" pos="DETdef">l</w>
<w xml:id="Ao_w_0135682" lemma="aigue" pos="NOMcom">eive</w>
<w xml:id="Ao_w_0135683" lemma="mais1" pos="CONcoo">mais</w>
<w xml:id="Ao_w_0135684" lemma="làsus" pos="ADVgen">laïssus</w>
<w xml:id="Ao_w_0135685" lemma="en1+le" pos="PRE.DETdef">el</w>
<w xml:id="Ao_w_0135686" lemma="ciel" pos="NOMcom">ciel</w>
<w xml:id="Ao_w_0135687" lemma="car" pos="CONcoo">Car</w>
<w xml:id="Ao_w_0135688" lemma="le" pos="DETdef">li</w>
<w xml:id="Ao_w_0135689" lemma="ciel" pos="NOMcom">ciaus</w>
<w xml:id="Ao_w_0135690" lemma="estre1" pos="VERcjg">est</w>
<w xml:id="Ao_w_0135691" lemma="siecle" pos="NOMcom">siegles</w>
<w xml:id="Ao_w_0135692" lemma="pardurable" pos="ADJqua">pardurables</w>
```

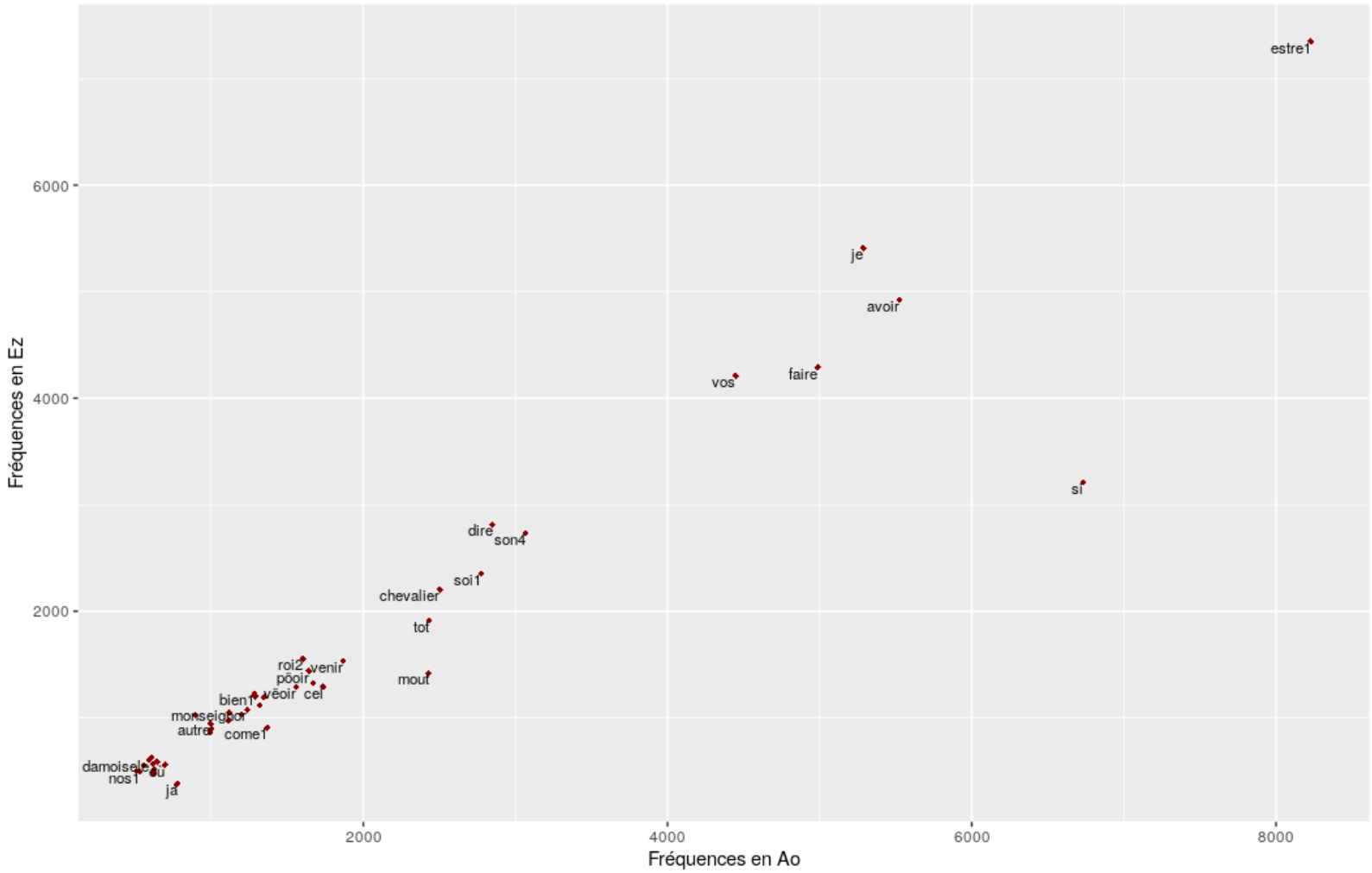
La force des lemmes

Pourquoi lemmatiser ?
→ obtenir des **formes stables**

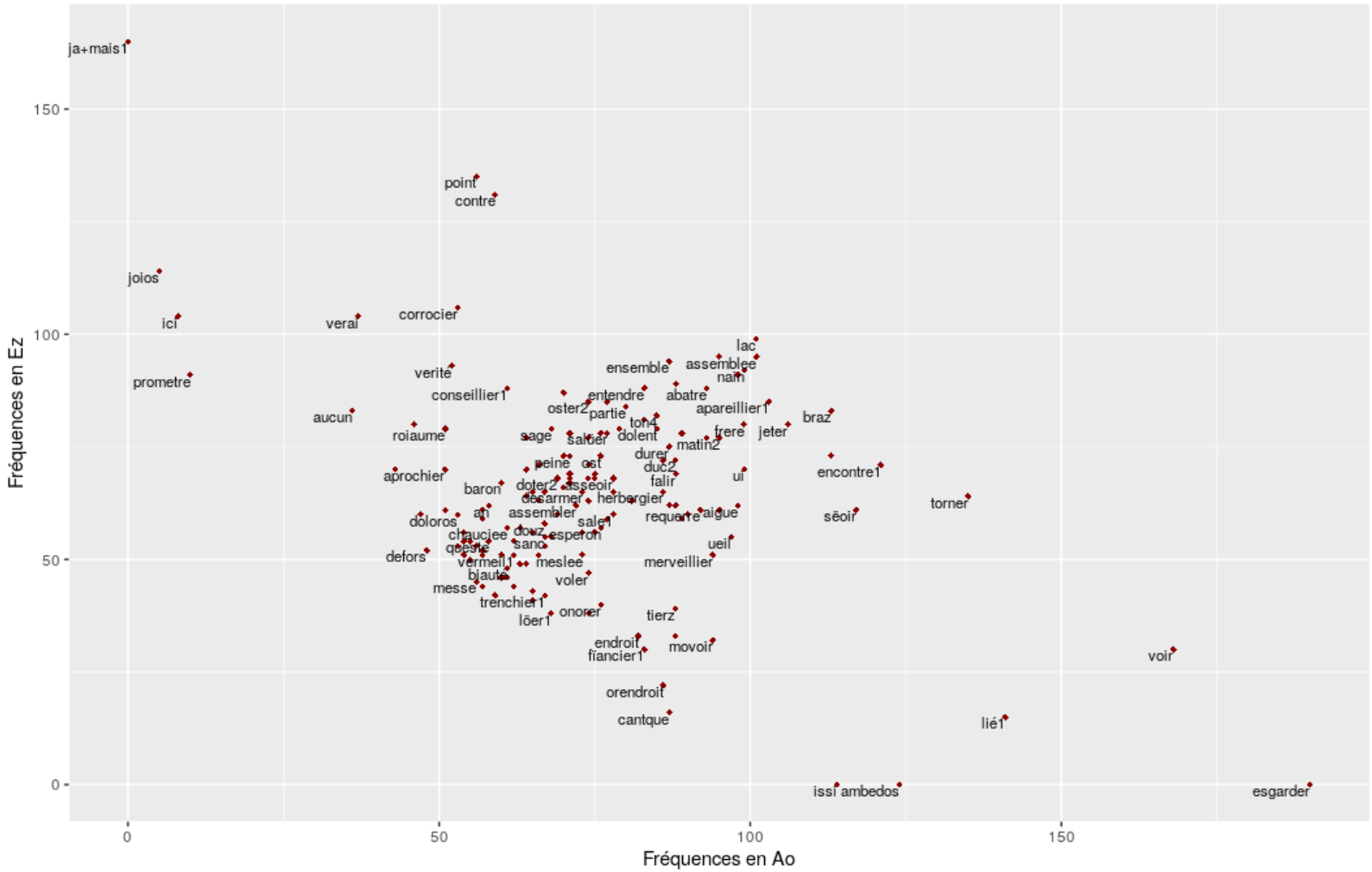
10 formes pour **enfant** dans notre corpus : *enfans*, *anfanz*, *enfant*, *anfant*, *anfes*, *enfes*, *anffans*, *enffant*, *amfant*, *enfanz*

- faire des calculs de **fréquence**
- pouvoir rechercher des **usages** de manière simple
- permettre un **alignement** sur des **formes identiques**

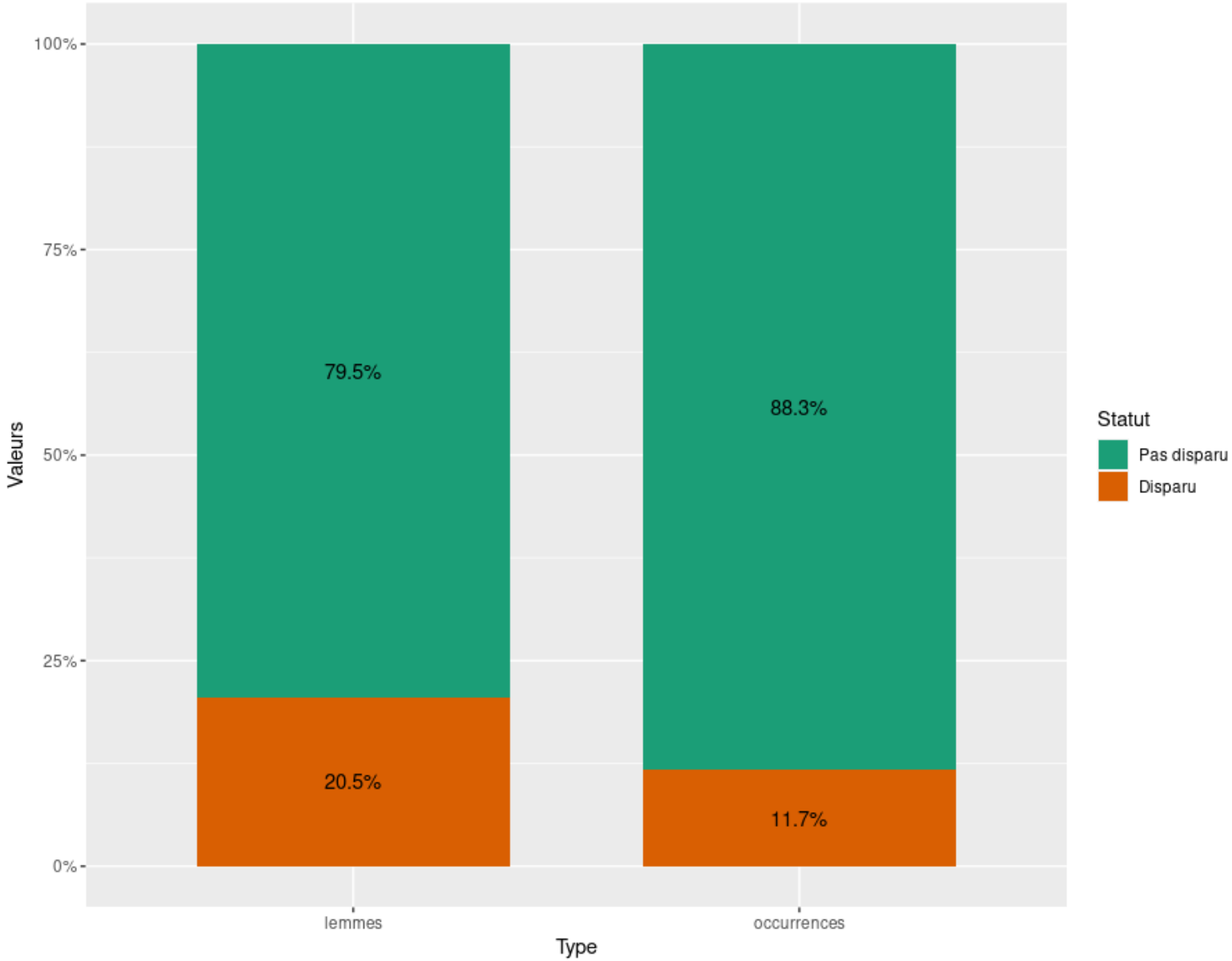
Répartition des lemmes à plus de 1000 occurrences



Répartition des lemmes entre 100 et 200 occurrences



Les lemmes et leur nombre d'occurrences



Les lexèmes qui disparaissent (1)

lexèmes	total	Ao	Ez
acesmer	10	10	0
delaiier1	10	10	0
estovoir1	12	12	0
consirer	13	13	0
encui	13	13	0
onoreement	14	14	0
sospecier	14	14	0
main1	15	15	0
espoir2	17	17	0
aerdre	18	18	0
membrer2	21	21	0
autretant	28	28	0
auques	31	31	0
sen2	35	35	0
mar	36	36	0
o4	38	38	0
peçoiier	40	40	0
desor	41	41	0
neporcant	88	88	0
si+il	90	90	0
ambedos	120	120	0
je+il	150	150	0
esgarder	172	172	0

Les lexèmes qui disparaissent (2)

Fréquences à 1

lexèmes	total	Ao	Ez
encomencier	19	18	1
covine	27	26	1
dalez	36	35	1
irier	37	36	1
bëer	44	43	1
ester	49	48	1
covent1	61	60	1
remanoir	201	200	1

Fréquences peu élevées

lexèmes	total	Ao	Ez
hardement	16	14	2
trestot	25	23	2
guenchir	26	24	2
conrëer	32	28	4
emprendre	44	40	4
vis1	56	52	4

Différences importantes de fréquences

lexèmes	total	Ao	Ez
lié	132	120	12
siecle	96	66	30
fiancier1	111	81	30
traire	242	174	68
durement2	412	317	95
rien	520	332	188

Les lexèmes qui apparaissent

lexèmes	total	Ao	Ez
ja+mais1	147	0	147
lequel	134	0	134
tellement	65	0	65
enmener	59	0	59
prisonier	40	0	40
sinon	36	0	36
depuis	31	0	31
afin2	26	0	26
emprès	26	0	26
nonportant	26	0	26
cause	25	0	25
incontinent	25	0	25
fraper	22	0	22
portant	22	0	22
desormais	18	0	18
pareillement	17	0	17
desplaire	16	0	16
rencontrer	15	0	15
neportant	14	0	14
estat	13	0	13
remercïer	13	0	13
vaillantment	13	0	13
onorablement	12	0	12

lexèmes	total	Ao	Ez
esmerveillier	25	1	24
persone	25	1	24
tirer	51	2	49
emporter2	32	2	30
merveillosement	16	2	14
joios	103	5	98
corage	30	7	23
sovenir	29	7	22
ici	110	8	102

lexèmes	total	Ao	Ez
covenant	72	14	58
aussi	178	22	156
demorer	333	96	237

Les disparitions de lexèmes

Les lexèmes qui disparaissent :

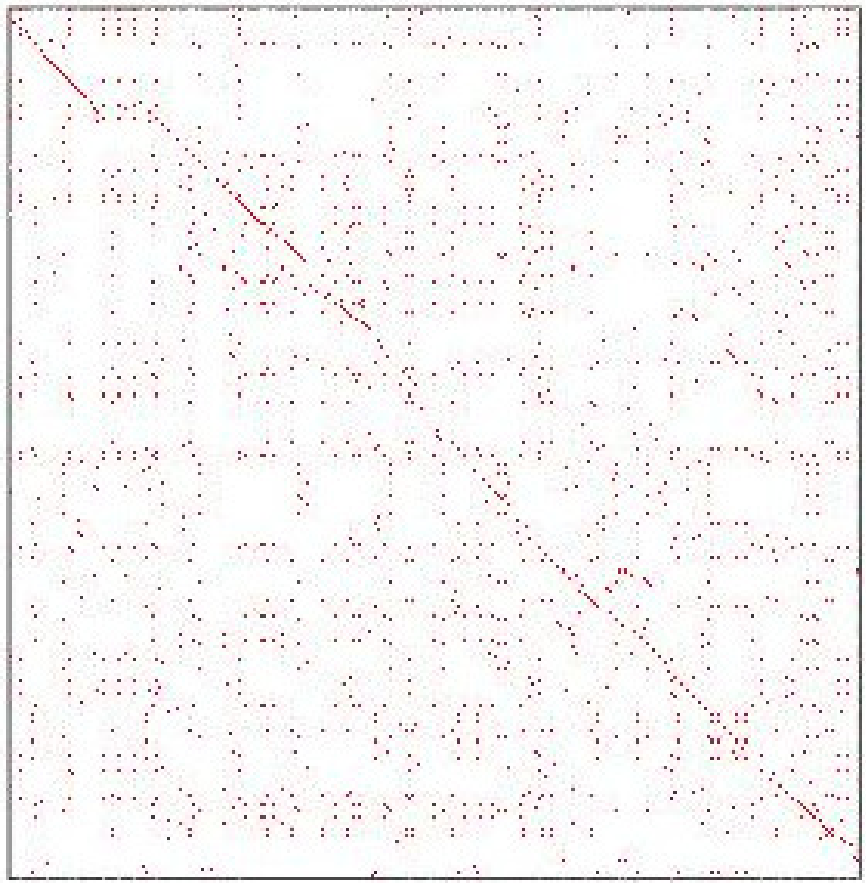
- **disparition totale** (fréquence tombe à 0)
- disparition **presque totale** (fréquence à 1-2)
- disparition **partielle** (chute de fréquence)

Distinguer :

- ce qui se passe à l'**intérieur des témoins retenus**
- ce qui se passe dans les autres textes / les **autres registres de langue**
- les lexèmes qui survivent dans des **lexiques spécialisés**
- les lexèmes qui survivent dans les **dialectes**
- l'emploi d'**archaïsmes** dans le témoin Ez

3. Alignement et collation

Aligner les témoins



Matrice représentant l'alignement des témoins sur le chapitre 5. Chaque point rouge représente un valeur de distance à 0 (tokens identiques)

Matrice de distance

	cant1	le	chevalier	desireter	öir	le	novele
le	5	0	7	8	3	0	4
conte1	3	5	7	7	6	5	5
dire	5	3	8	5	2	3	5
que4	5	3	8	8	4	3	5
cant1	0	5	7	8	5	5	6
le	5	0	7	8	3	0	4
chevalier	7	7	0	7	7	7	6
desireter	8	8	7	0	7	8	7
öir	5	3	7	7	0	3	6
le	5	0	7	8	3	0	4
novele	6	4	6	7	6	4	0
Montlair	6	7	7	8	6	7	6

Exemple de la matrice de distance, sur le début du chapitre 5

J.-B. Camps, E. Spadini, L. Ing, FALCON, “Collating Medieval Vernacular Texts. Aligning Witnesses, Classifying Variants”, `https://hal.archives-ouvertes.fr/hal-02268348`

Pourquoi pré-aligner ?

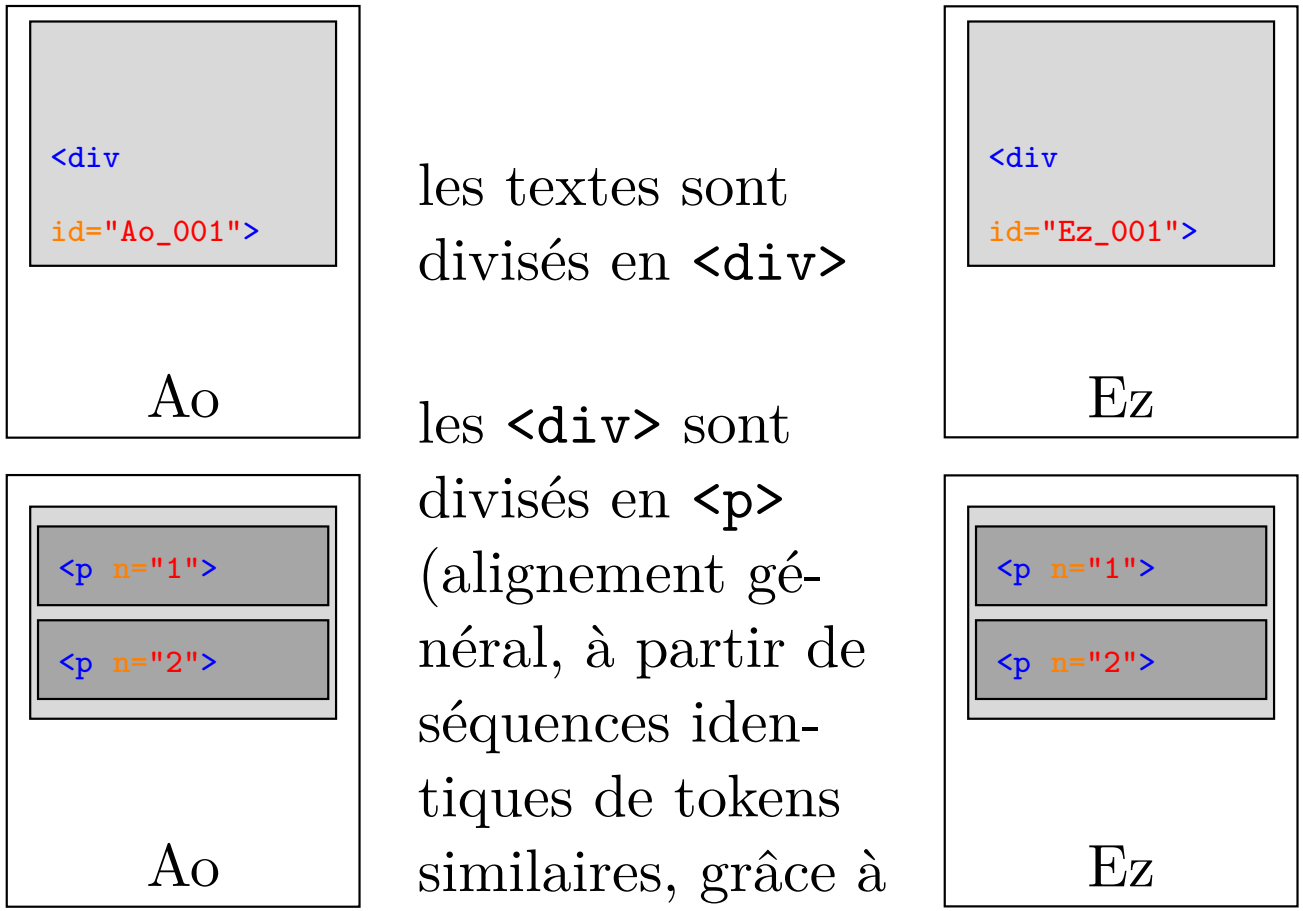
Ao

Or te dirai don qui cele flors est et coment ses conseuz te sauvera. Cele flors est flors de totes les autres flors. De ce cele flor nasqui li fruiz de qoi totes choses sont sostenues. C'est li fruiz don li cors est sostenuz et l'ame païe. C'est li fruiz qui saola les cinc mile homes en la praerie, qant les doze corboilles furent anplies del reillié. Ce est li fruiz par coi li pueples Israel fu sostenuz quinze anz es desserz la ou li om, ce dit l'Escripture, manja lo pain as angles. Ce est li fruiz par coi Josep de Barimathia et si compaignon furent sostenu qant il s'an venoient de la terre de promission an ceste estrange païs par lo comendement Jhesu Crist et par son conduit. Ce est li fruiz don Sainte Eglise est repaüe chascun jor. Ce est Jhesu Criz, li Filz Deu. C'est la flors de cui doiz avoir lo consoil et lo secors se tu ja mais l'as. Ce est sa douce Mere, la glorieuse Virge, don il nasquié contre acostumance de nature. Cele dame est a droit apelee Flors car nule fame ne porta onques anfant devant li ne aprés, qui par charnel asemblement ne fust ançois desfloree. Mais ceste haute dame fu virge pucele et avant et aprés c'onques la flor de son pucelaige ne perdi. Bien doit dons estre apelee Flors de totes autres flors, qant ele garda sa glorieuse flor saigne et antiere la ou totes les autres flors perissent,ce est au concevoir et an l'anfanter, et qant de lui nasquié li Fruiz qui done vie a totes choses.

Ez

La fleur, c'est la mere Jesuchrist, ou le Pere et le Filz et le Saint Esperist se aombra sans lui corrompre sa virginité et la fleur qui est en elle.

Alignement



les textes sont divisés en `<div>`

les `<div>` sont divisés en `<p>` (alignement général, à partir de séquences identiques de tokens similaires, grâce à `text-matcher`)

J. Reeve, https://doi.org/10.5281/zenodo.3937738

Fonctionnement de l'alignement (1)

```
<div xml:id="Ao_051" corresp="Ez_036"> ...
<w xml:id="Ao_w_0140584" lemma="faire"
pos="VERcjg">fait</w> <w xml:id="Ao_w_0140585"
lemma="ainz" pos="ADVgen">ainz</w>
<w xml:id="Ao_w_0140586" lemma="il"
pos="PROper">l</w> <w xml:id="Ao_w_0140587"
lemma="amer1" pos="VERcjg">aimme</w>
```

`text-matcher` : identification des séquences similaires ; ici, la chaîne est identique en Ao et Ez

ainz il amer1

transformation en `json` et ajout de la position du premier caractère

```
{div : [{'id' : 'Ao_051', 'corresp' :
'Ez_036', 'tokens' : [{'text' : 'fait', 'id'
: 'Ao_w_0140584', 'lemme' : 'faire', 'pdd' :
'VERcjg', 'pos' : '1932' } {'text' : 'ainz',
'id' : 'Ao_w_0140585', 'lemme' : 'ainz', 'pdd' :
'ADVgen', 'pos' : '1937' }]}]}
```

Fonctionnement de l'alignement (2)

`text-matcher` : récupération des premiers caractères des chaînes qui se correspondent 1937

Ez : 1468

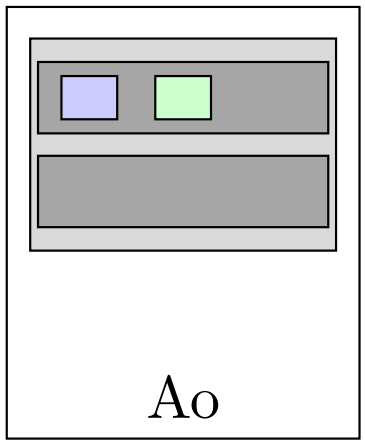
correspondance entre premier caractère et identifiant

```
{div : [{'id' : 'Ao_051', 'corresp' :
'Ez_036', 'tokens' : [{'text' : 'fait', 'id'
: 'Ao_w_0140584', 'lemme' : 'faire', 'pdd' :
'VERcjg', 'pos' : '1932' } {'text' : 'ainz',
'id' : 'Ao_w_0140585', 'lemme' : 'ainz', 'pdd' :
'ADVgen', 'pos' : '1937' }]}]}
```

constitution du doc XML avec des marques de `<p>` avant les mots correspondants

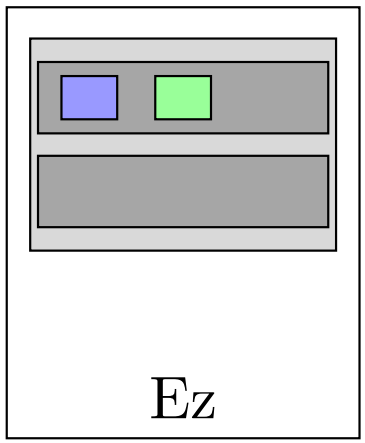
```
<div xml:id="Ao_051" corresp="Ez_036"> ...
<w xml:id="Ao_w_0140584" lemma="faire"
pos="VERcjg">fait</w> <p n="11"/>
<w xml:id="Ao_w_0140585" lemma="ainz"
pos="ADVgen">ainz</w> <w xml:id="Ao_w_0140586"
lemma="il" pos="PROper">l</w> <w
xml:id="Ao_w_0140587" lemma="amer1"
pos="VERcjg">aimme</w>
```

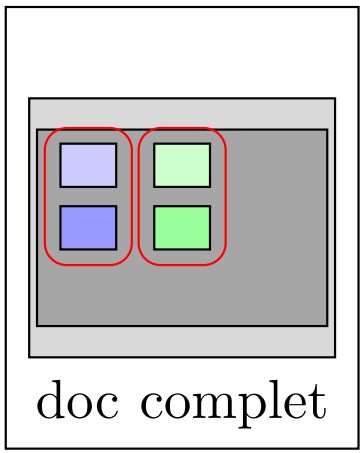
Collation



la collation automatique produit un alignement mot à mot

il est fait sur les lemmes grâce à Collatex





Fonctionnement de la collation (1)

La **collation**

- étape dans l'**édition d'un texte** : repérer les différentes variantes des témoins d'un texte afin d'en établir la tradition textuelle
- ici, la collation ne sert pas à l'établissement d'une édition, mais à la **comparaison de nos deux témoins**

Collatex

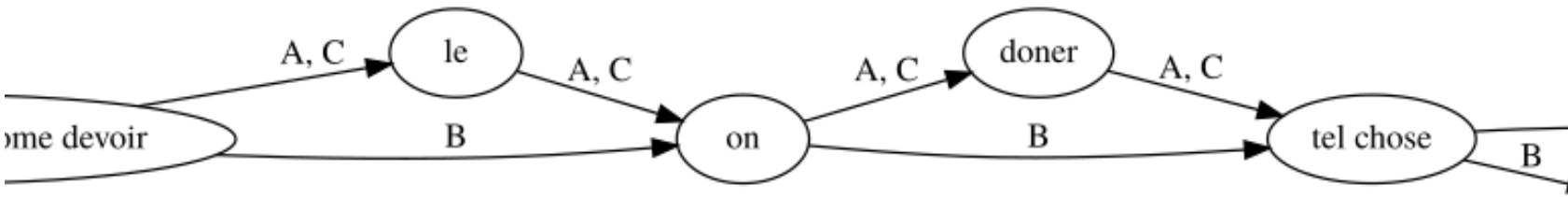
- un outil de **collation automatique**
- permet d'aligner les tokens et d'identifier similarités et différences entre les témoins
- l'alignement est **basé sur l'identification de chaînes de caractères similaires** dans deux témoins ou plus
- à cause de la **variance** en français médiéval, la collation est réalisée sur les **lemmes**

Fonctionnement de la collation (2)

Modèle Gothenburg

Modèle défini en 2009, composé de **quatre étapes** :

1. tokénisation
2. alignement (à partir d'un témoin, témoin après témoin)
3. détection des transpositions
4. visualisation



`https://collatex.net/`
`http://interedition.github.io/collatex/pythonport.html`

Exemple de collation (vue table)

Ao	**Ez**
Li	-
contes	-
dit	-
que	-
qant	Quant
li	le
chevaliers	chevalier
deseritez	desherité
oï	ouyt
les	les
noveles	nouvelles
-	de
Monlair	Moncler
lo	-
chastel	-
qui	qui
pris	prins
estoit	estoit
et	et
il	il
vit	vit
Claudas	Claudas

Exemple de collation (vue XML)

```
<app type="graph">
<rdg wit="#Ao">
<w lemma="öir" xml:id="Ao_w_0009258" pos="VERcjg">oï</w></rdg>
<rdg wit="#Ez">
<w lemma="öir" xml:id="Ez_w_0013454" pos="VERcjg">ouyt</w></rdg>
</app>
<app type="absVar">
<rdg wit="#Ao">
<w lemma="le" xml:id="Ao_w_0009259" pos="DETdef">les</w></rdg>
<rdg wit="#Ez">
<w lemma="le" xml:id="Ez_w_0013455" pos="DETdef">les</w></rdg>
</app>
<app type="graph">
<rdg wit="#Ao">
<w lemma="novele" xml:id="Ao_w_0009260" pos="NOMcom">noveles</w></rdg>
<rdg wit="#Ez">
<w lemma="novele" xml:id="Ez_w_0013456" pos="NOMcom">nouvelles</w></rdg>
</app>
<app type="leconIsolee" corresp="#Ez">
<rdg wit="#Ez">
<w lemma="de" xml:id="Ez_w_0013457" pos="PRE">de</w></rdg>
</app>
<app type="graph">
<rdg wit="#Ao">
<w lemma="Montlair" xml:id="Ao_w_0009261" pos="NOMpro">Monlair</w></rdg>
<rdg wit="#Ez">
<w lemma="Montlair" xml:id="Ez_w_0013458" pos="NOMpro">moncler</w></rdg>
</app>
```

Les erreurs de collation

chap	n°p	id. l	préc. Ao	préc. Ez
divColl62	122	Ao_w_0232114-Ez_w_0172535	que vos an vandroiz par moi sanz autre bessoigne	damoiselle que vous en revendrez par moy sans autre affaire
divColl63	41	Ao_w_0242661-Ez_w_0182317	por vostre bessoigne Mais ge ne la cuidai pas avoir	pour vostre besongne et je ne la cuidoie pas cy

l. Ao	l. Ez	suiv. Ao	suiv. Ez
anprandre	entreprendre	Oïl fait il se ce n estoit afaires dom ge	Oy dit il se se n est chose donc je
amprise	endroit	si a droit com ge ai Mais or sai	avoir entreprinse comme j ay mais ore scay je

4. Études des lexèmes

Les lexèmes intéressants

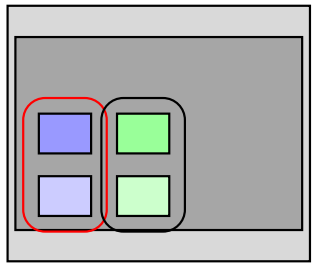
- les lexèmes dont les **fréquences sont à 0**
- les lexèmes dont les **fréquences chutent** : comment les repérer ?

→ utilisation de l'indice de spécificité de TXM

Unité	**F**	**f_Ao**	**score_Ao**	**f_Ez**	**score_Ez**
bëer	44	43	9.8976	1	-9.8976
totevoies	80	70	9.4043	10	-9.4043
vis1	56	52	9.3297	4	-9.3297
dalez	36	35	7.8969	1	-7.8969
pensé	40	38	7.676	2	-7.676
traire	242	174	7.4069	68	-7.4069

Indice de spécificité : « calcul de la probabilité du fait que l'événement apparaisse autant de fois qu'on l'observe effectivement dans la partie (soit f_obs) ou plus fréquemment encore à concurrence de la taille de la partie (en suivant [une] loi hypergéométrique) »
`http://textometrie.ens-lyon.fr/html/doc/manual/0.7.9/fr/manual43.xhtml`

Suite de la chaîne



doc complet

```
<xsl:template
match= "row">
 <xsl:element
name= "fiche">
```

pour chaque lemme pertinent, récupération données

```
<fiche
xml:id=
"fronchier">
```

fiches automatiques

```
<datePremAtt>
RoisC
</datePremAtt>
```

fiches structurées

```
<row> <lemme>
fronchier</row>
</lemme>
```

csv vers XML

données extérieures (dictionnaires, corpus divers)

Analyses des lemmes

Plusieurs types d'analyses **en fonction des fréquence**s des lemmes :

- grammèmes : **rapide analyse**, globale
- lexèmes présents moins de 4 fois en Ao : présentation des **contextes**
- lexèmes présents au moins 4 fois en Ao : **fiche lexicale**

Pour les lexèmes présents au moins 5 fois en Ao : si 50 occurrences ou moins en Ao, analyse de chaque occurrence.

Structuration XML des fiches

```
<fiche xml:id="fronchier">
<lemme>fronchier</lemme>
<freqAo>1</freqAo>
<freqEz>0</freqEz>
<definition>souffler bruyamment</definition>
<datePremAtt>CoincyI18K</datePremAtt>
<dateDernAtt>PercefR2</dateDernAtt>
<source>Mats-DMF</source>
<etymon>roncare</etymon>
<sourceEtymon>\textsc{x}, 470b</sourceEtymon>
<contexte  ref="#divColl9-18" type="remplacement"
xml:id="divColl9-18-ctxt-001">
<contexteAo ref="#Ao_w_0018755">\dots{} et <bold>fronchoit</bold> del
nes an sa grant ire autresin com uns chevax.</contexteAo>
<contexteEz ref="#Ez_w_0021715">\dots{} et <bold>rouffloit</bold> du
nez en sa grant ire ainsi comme ung cheval.</contexteEz>
</contexte>
<ficheCorresp xml:id="ronfler">
<lemme>ronfler</lemme>
<definition>souffler bruyamment</definition>
<datePremAtt>EneasS2</datePremAtt>
<dateDernAtt>Cresp 1637</dateDernAtt>
<source>Mats-FEW</source>
<sourceEtymon>\textsc{x}, 470b</sourceEtymon>
<etymon>ronfl-</etymon>
```

Présentations succinctes des lexèmes peu fréquents

fronchier
FEW, x, 470b : RONCARE

Ao : 1 ; Ez : 0

“souffler bruyamment” (CoincyI18K – PercefR2)

*… et **fronchoit** del nes an sa grant ire autresin com uns chevax.*
*… et **rouffloit** du nez en sa grant ire ainsi comme ung cheval.*
(9-18)

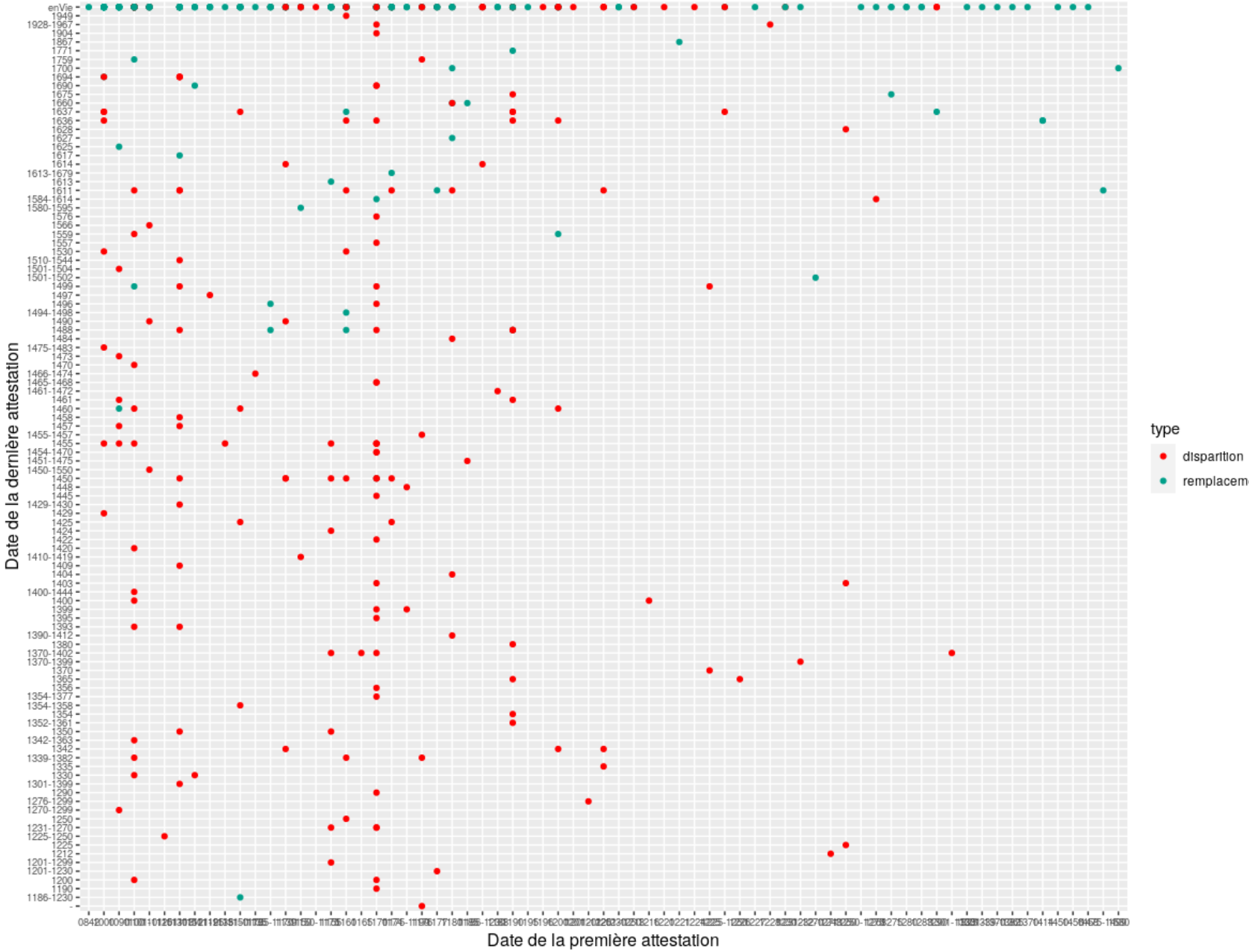
Lexème qui remplace : *ronfler*, (EneasS2 – Cresp 1637 ; RONFL-)

Analyses des lemmes plus fréquents

1. **analyse des exemples** :
   - remplacement **systématique** ? *onoreement* → *onorablement*
   - **conservation partielle** en Ez ? syntagme particulier (*enmy le piz*)
2. **analyse interne** :
   - mots remplaçants existent-ils en Ao ?
   - si oui, distinction ? *damagier*, qui remplace *estoutoiier*, est déjà utilisé en Ao
3. **analyse externe** : utilisation des données issues
   - des **corpus en ligne** (Frantext, DMF)
   - des dictionnaires de langue

   pour essayer de cerner l'évolution des lexèmes
   estoutoiier dans sa forme participiale entre en hononymie avec *estoutie*, "action téméraire", formes opposées sur le plan sémantique
4. proposition de **typologie** et de **facteurs**
   Typologie : Marginalisation ; survivance dans des lexiques spécialisés
   Facteurs supposés : Homonymie avec un lexème fréquemment utilisé en corrélation ; spécialisation dans deux sens antagonistes ; existence d'un autre lexème de même sens, plus marqué d'un point de vue phonétique

Visualisation des lemmes par dates



5. Approches computationnelles

Deux approches utilisées

- le **Topic Modeling**, qui permet de déterminer un certain nombre de **thèmes au sein d'un corpus**, et de relier ces thèmes aux différentes occurrences d'un lexème
- le **Word Embeddings**, ou plongement de mots, qui permet d'obtenir un « espace sémantique » des textes, en se basant sur la représentation de **chaque mot comme un vecteur au sein d'une matrice**

Des approches qui permettent de parler du **sémantisme** des lexèmes, basées sur le principe qu'un mot prend sens en contexte. J. Rupert Firth (1957) : « ***You shall know a word by the company it keeps*** ».

Le Topic Modeling

Interêts

- déterminer un **nombre de thèmes** qui structurent le texte
- permet de dessiner des **espaces sémantiques**, sans avoir à assigner manuellement ces espaces (P. Schöch, 2012)
- assigner un **thème à chaque occurrence** : est-ce que certains thèmes sont davantage concernés que d'autres par le processus d'obsolecence ?

→ tout document textuel est constitué de thèmes qui le structurent, et chacun de ces thèmes se caractérise par les mots qui permettent son expression

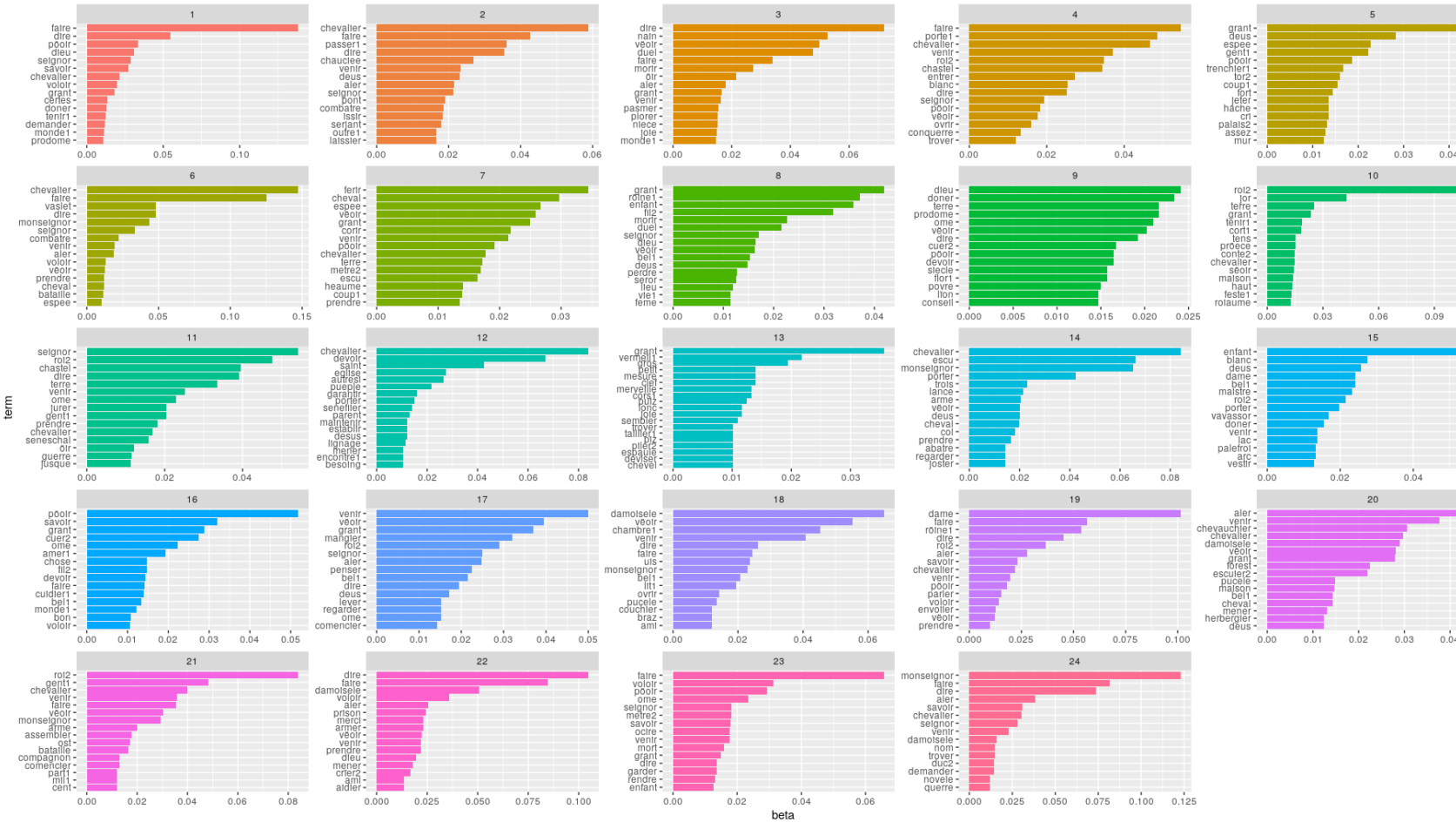
Fonctionnement et application

Principes de fonctionnement

- technique **non-supervisée** d'apprentissage machine
- déterminer des **ensembles de mot**s pour chaque document, basé sur la reconnaissance de séquences identiques d'utilisation
- **méthode probabiliste**, algorithme LDA (Latent Dirichlet Allocation, D. Blei et al., 2003)

- utilisation de `mallet` dans R
  `https://rdrr.io/cran/mallet/man/mallet-package.html`
- le paramètre **"nombre de thèmes" est à assigner préalablement**
- tranches 500 mots et 24 *topics*
- il faut ensuite aller voir ce que chacun des thèmes contient pour comprendre ce qu'il englobe

Les thèmes qui structurent les textes



Quelques thèmes

Topic 1 : **la description des combats**
grant, espee, coup1, vëoir, ferir, escu, sanc, trenchier1, heaume, hauberc…
mesler, esbäir, midi…

Topic 5 : **les rapports de féodalité**
chastel, seignor, chevalier, terre, dire, gent1, venir, prendre, ome secorre1…
marche1, guerre, jurer, passer1, joster…

Topic 7 : **le rapport à la signifiance**
chevalier, devoir, saint, duel, eglise, porter, terme, mal1, pueple, senefiier…
letre, aventure, biere, escriture, maufaitor…

Topic 8 : **la guerre**
faire, voloir, seignor, dire, pöoir, savoir, prendre, ome, venir, ocire, tenir1,
mort, roi2… prison, defendre, garder, morir, gent1, conseil, terre, träison,
cité, garantir, pais, päis baron, destruire…

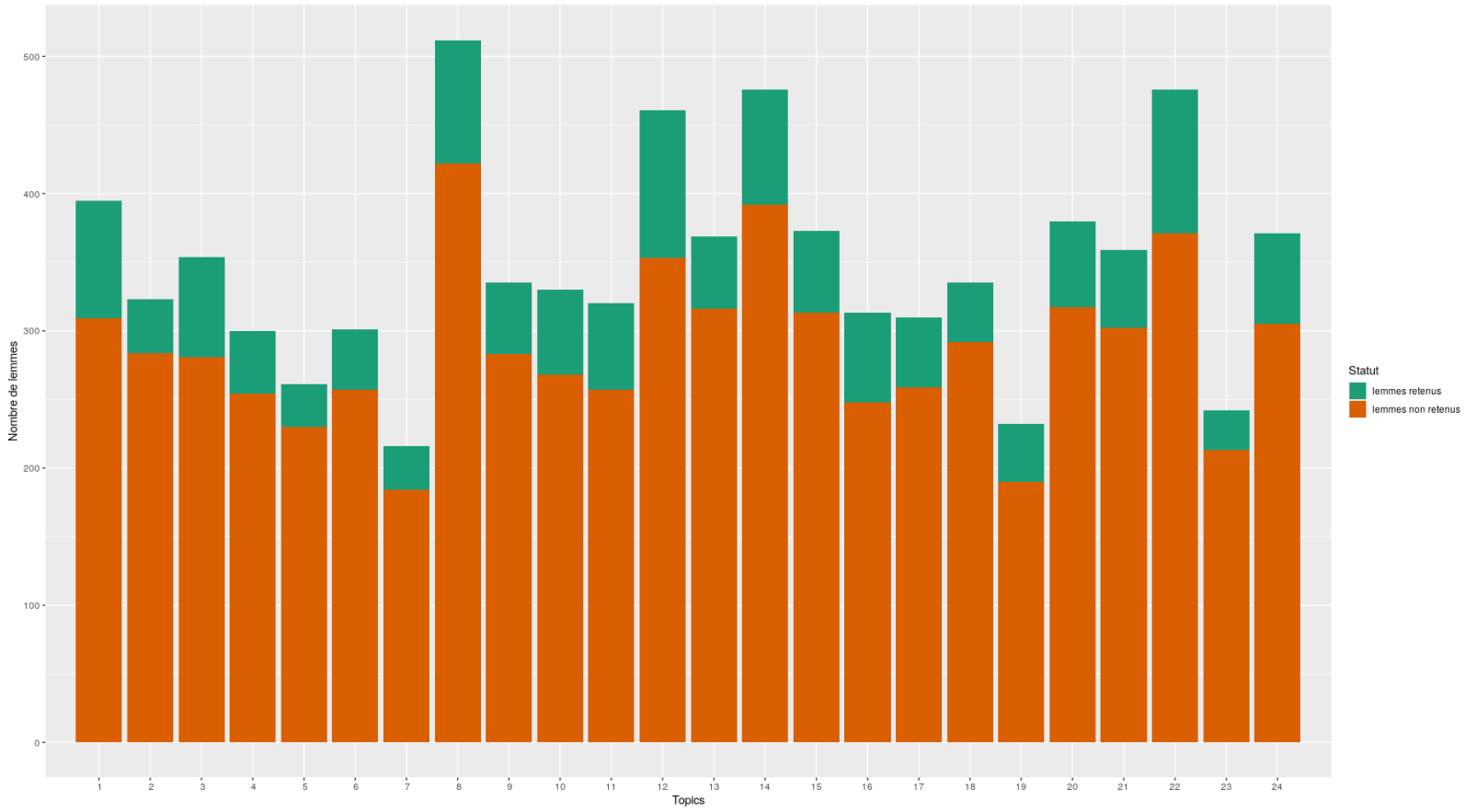
Topic 12 : **les actions pendant les combats**
chevalier, cheval, ferir, metre2, venir, corir, faire, espee, glaive, pöoir…
defendre, senestre, retorner, ocire…

Topic 14 : **les valeurs de la prodomie**
pöoir, grant, ome, faire, doner, bel1, cuer2, chose, devoir, bon, monde1,
chevalier…… prodome, onor, droit, chevalerie…

Topic 22 : **les sentiments**
amer1, grant, venir, voloir, rien, seignor, joie, certes, röine1, conoistre…
amor, ha !, ami, plorer…

Topic 23 : **les repos à la cour du roi Arthur**
faire, chevalier, vaslet, seignor, grant, monseignor, venir, roi2, bel1, espee,
ome, envoiier… lit1, ostel, desarmer, mangier, sale1, messe, feste1…

Mots obsolescents par thème



Word Embeddings

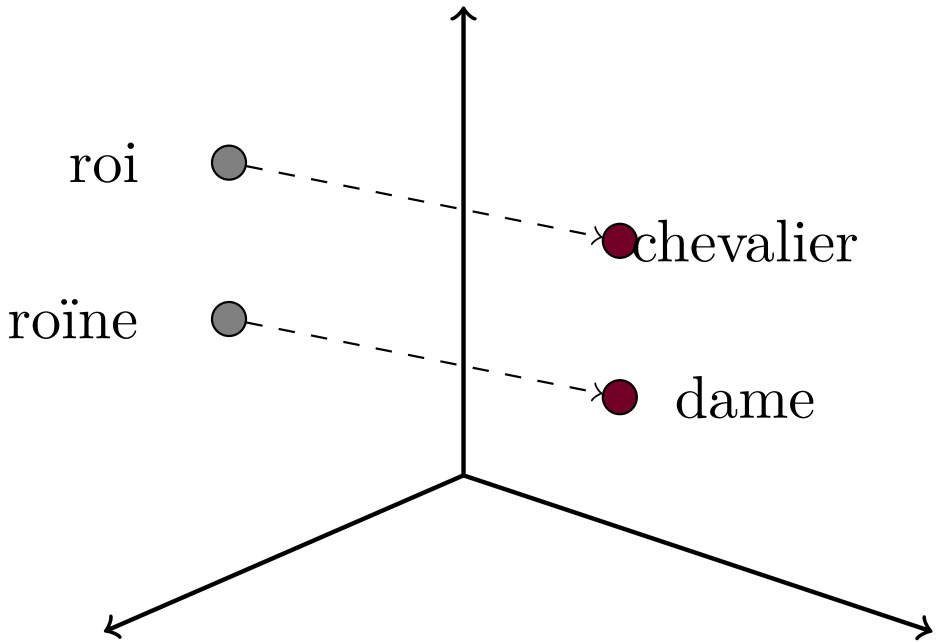
Principe de fonctionnement

- le **texte** est représenté sous une forme de **matrice**
- la **représentation d'un mot** se fait à l'intérieur d'un **vecteur** : le mot est un vecteur de chiffres
- ce vecteur est composé de valeurs établies **en fonction de la co-occurrence** du lexème avec les autres termes

chevalier	-1.5488147	-1.117865	-0.13586469	-0.9344863	1.6124617
roi2	-3.944645	-1.7473782	0.25746804	-2.7548037	2.2756793
dame	0.6642276	-0.61475843	0.00583692	1.5918058	2.1516817

Représentations

Les **mots avec des vecteurs similaires sont des mots qui apparaissent dans des contextes similaires** donc qui ont des sens similaires



Les modèles qui représentent bien les textes peuvent résoudre des équations du type : `roi + dame - chevalier ?`
→ **`roïne`**

Utiliser des plongements de mots ?

De nombreuses recherches actuelles

- depuis `word2vec` (Mikolov et al. 2013)
- jusqu'aux **représentations des mots en contexte** comme BERT (Delvin et al., 2019) ; CamemBERT pour le français (Martin et al., 2020) pour palier aux problèmes que posent les **lexèmes polysémiques**

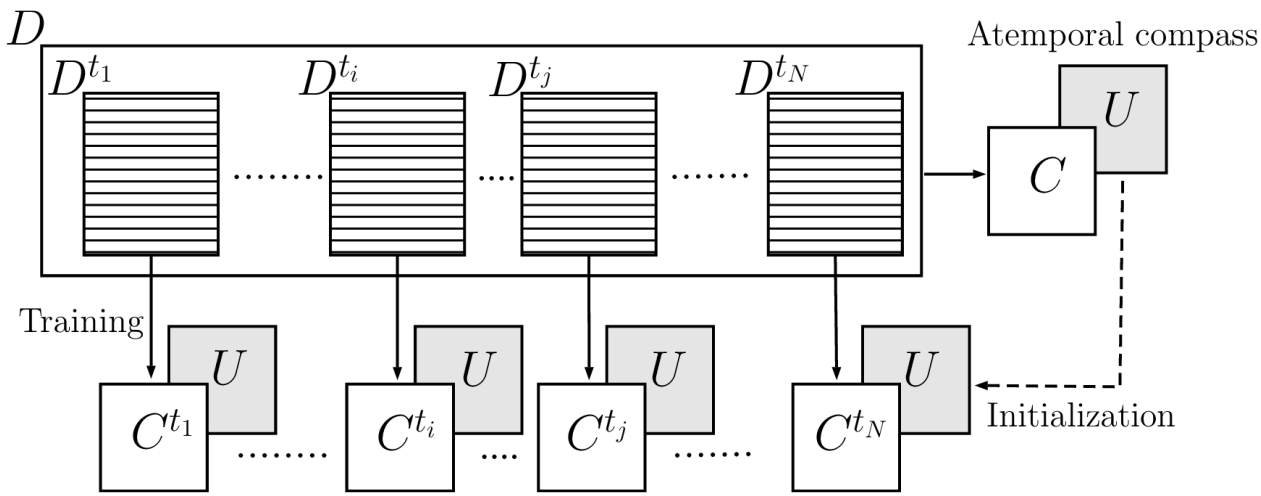
Représenter le changement sémantique

- si **un seul modèle** est appliqué sur différentes parties d'un corpus, on obtient **un système de référence différent** pour chacune
- de nombreux travaux sur la question depuis D. Hamilton et al., "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change" (2016)

Temporal Word Embeddings

◆ TWEC (Temporal Word Embeddings with a Compass, Di Carlo et al., 2019)
`https://github.com/valedica/twec`

- construit à partir de `word2vec`
- on entraîne d'abord sur le **corpus entier**, puis sur les **tranches**
- immobilisation de l'une des couches de l'architecture CBOW pour avoir une concordance



source : TWEC

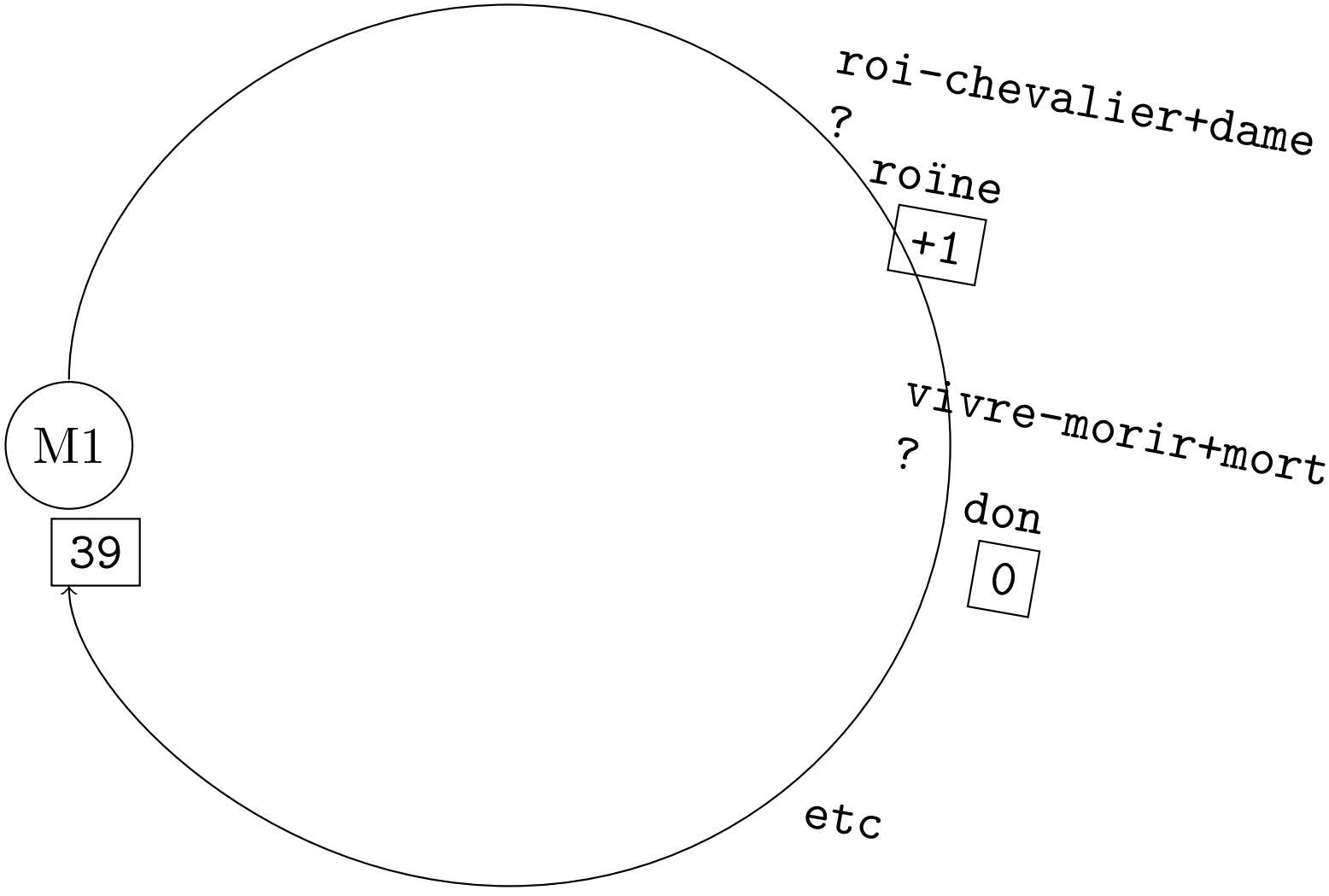
Méthode d'entraînement et d'évaluation

Méthode (toute personnelle) pour **choisir un modèle**

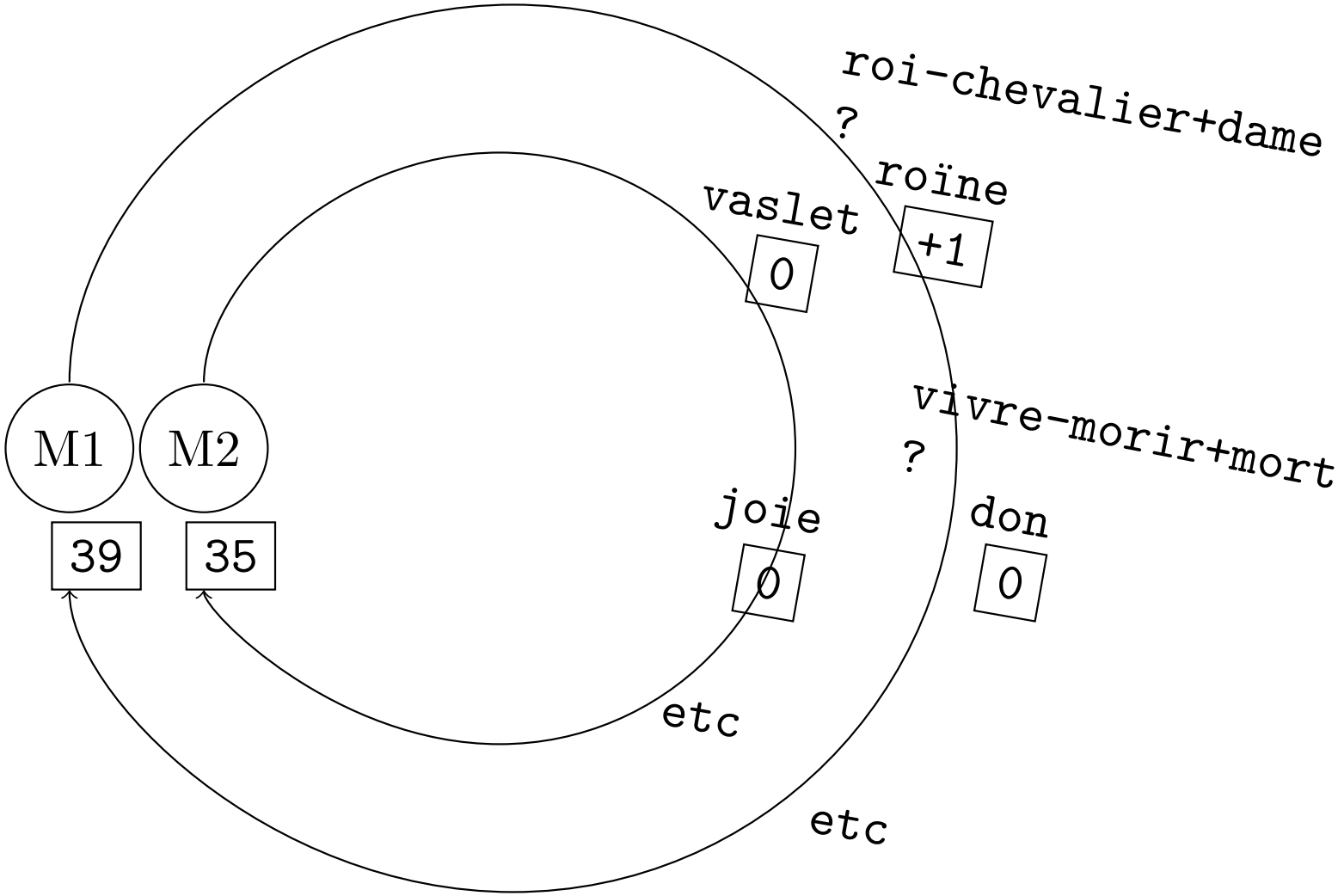
- objectif : création du modèle qui r**eprésente le mieux l'espace sémantique** des textes
- moyen : utilisation des **résultats des équations données**
- méthode : une **boucle** est créée. À chaque boucle, un nouveau modèle est créé et **doit résoudre un certain nombre de questions** déterminées à l'avance. S'il répond correctement il gagne un point.

Le **meilleur modèle** est celui qui a **le plus de points**.

Word Embeddings : le modèle



Word Embeddings : le modèle



Exploitations

Évolution sémantique

- les mots les plus similaires de **durement2** en Ao : *blecier, afeblir, vigorosement, doucement, irier, angoissos, estoner, esfrëer, vistement, sospirer, doloir, visage, estordire, comencier, eschine, menacier, esbäir*
- les mots les plus similaires de **durement2** pour Ez : *blecier, afeblir, estordir, roidement, plaissier3, saignier, asprement, detrenchier, escrois1, larme, froit1, chëoir, contreval, tomber, tellement*

→ l'emploi dans les scènes de combat est conservé
→ en Ao, associé à l'expression des sentiments ; en Ez, association avec l'intensité

Remplacement ?

- les mots les plus similaires de **peçoiier** en Ao : *rompre, brisier, decoper, lancier3, estoner, abatare, voler, hurter, lance, sachier2, parchëoir, aprester, detrenchier*
- les mots les plus similaires de **rompre** en Ez : *brisier, detrenchier, abatre, depecier, lance, arrachier, percier, covrir, escrever, parchëoir, voler, rüer, desrompre*

→ *peçoiier* disparaît totalement ; pour voir ce qui le remplace, on regarde en Ez les similarités avec le lexème qui est le plus similaire en Ao

Intérêts et limites de ces pratiques

Intérêts

- représenter tout ce qui a trait au **changement sémantique**
- avoir un **point de vue global** sur les lexèmes
- avoir aussi des données **sur chacun des lexèmes**
- avoir des informations précises sur les lexèmes **au sein du corpus étudié**
- obtenir ces informations sans avoir besoin de corriger des données

Limites

- méthodes **peu efficaces sur les lexèmes peu fréquents** — qui représentent une partie importante des lexèmes retenus pour étude — puisqu'elles sont basées sur la représentation de leurs occurrences en contexte
- méthodes qu'il faut évaluer (évaluation chronophage)

Implémentation des données

Lemme Ao	Disparition	statut remplac.	Lemme remplac.	Statut l. r. en Ao	relation etym.	Fr. l. 1
esgarder	T	RTT	regarder	TE	ME	172 – 0

Fr. l. 2	Simil. Ao 1	Simil. Ao 2	Simil. Ez 1	Simil. Ez 2	Topic l. 1	Topic l. 2
129 – 243	regarder, sëoir, luisir, envoleper, cortoisement	esgarder, arester, sospirer, esbäir, sorire	NA	avaler, esbäir, comencier, sospirer, ovrir	20,18,11, 22	15,22,11, 6

6. Conclusions et perspectives

Les disparitions : quelques pistes

- un seul facteur ne suffit pas : il y a un **rapport entre forme et sémantisme**
- **redistribution des formes**
- **vers une simplification** ? Un lexème qui remplace plusieurs autres ; peut-être, dans certaines parties du lexique, dans ce type de texte particulier

Encore de nombreuses questions

- voir si une **systématicité** se dessine, aussi du point de vue **sémantique**
- les mots peu employés disparaissent : concurrence synonymique ou **peu de connaissances** sur ces lexèmes ?
- **modéliser** les lexèmes difficiles comme les **archaïsmes**

Les approches numériques

Les outils numériques sont **stimulants** car ils permettent :

- de traiter de manière **systématique** des données
- de traiter **une grande masse** de données
- d'avoir une **nouvelle approche** sur les sujets (et d'avoir des nouvelles pistes de recherche sur les objets (?))
- d'adopter une **démarche scientifique** (résultats reproductibles, démarche transparente)
- de **partager** les données (réutilisation des données rendue possible)

Les outils numériques demandent **beaucoup de travail** :

- beaucoup de données !
- l'**harmonisation** des données, nécessaire pour permettre leur traitement, prend du temps
- l'apprentissage et la **maîtrise des outils numériques** sont de longs processus
- la mise en place d'une **chaîne de traitements** est longue
- la composition des scripts prend du temps
- la **post-correction** des données est très chronophage
- l'**évaluation** des outils utilisés prend du temps

Merci pour votre attention !

Gallica/BnF, français 118, f. 223r