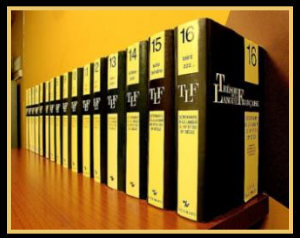
ANALYSE ET TRAITEMENT INFORMATIQUE DE LA LANGUE FRANÇAISE

« LGeRM : un outil de gestion des états anciens du français »

gilles.souvay@atilf.fr

Séminaires du CENTAL
17 février 2017






www.atilf.fr

cnrs dépasser les frontières

UNIVERSITÉ DE LORRAINE

Plan de l'exposé

- **DICTIONNAIRE ET VARIATION GRAPHIQUE**
- PRÉSENTATION DU LEMMATISEUR LGeRM
- ÉVOLUTIONS
- OUTIL GLOSSAIRE

Dictionnaire et variation

- Comment trouver un mot dans le dictionnaire
  - dictionnaire langue contemporaine vs dictionnaire en diachronie
    - variation morphologique
    - variation graphique : norme / absence de norme
- Problème pour l'utilisateur
  - spécialiste ou pas de la langue médiévale
- Quelle entrée ? (quel lemme ?)
  - *destroict*, *ameroyent*, *acoremens*, *polra*, *menra*

Dictionnaire et variation

- Choix de la graphie du lemme
  - problème aussi pour le rédacteur
    - agnel, aigneau, agneau
    - limites du choix de moderniser
      - cohérence dans une famille, mots disparus
- DMF : dictionnaire électronique
  - dans sa consultation
  - dans sa conception
- Taper la forme telle que rencontrée dans le document, le DMF fera des propositions

Exemple de forme atypique

Formulaire

embache | Rechercher | Effacer | - options

- [x] lemmatiser
- [ ] développer une graphie connue
- [x] trace **LGeRM**
- [ ] attestation dans les corpus textuels
- [x] analyse dans la *Base de Graphies Verbales*
- [x] afficher les dictionnaires cités

Saisir un mot ou une forme sans se préoccuper des entrées du DMF : des propositions s'afficheront.

La recherche porte sur les variantes graphiques connues du lemmatiseur.

Résultat de la recherche

La forme *embache* est connue du lemmatiseur avec l'analyse suivante :

EMBATTRE, verbe | structure | sans exemple | complet | textes

[TL : *embatre* ; GD : ***embatre*** ; AND : ***enbatre1*** ; DÉCT : ***embatre*** ; FEW I, 293a ***battuere*** ; TLF : ***embat(t)re***]

Plus d'hypothèses

BGV

2 attestations dans la **Base de Graphies Verbales**

http://www.atilf.fr/bgv/

embache	embatre	subjonctif présent 3	TL
embache	embatre	subjonctif présent 3	Gdf

La variation graphique médiévale

- PON

PONDRE Voir diachronie dans FRANTEXT

	D	L	I	P	BFM	7FMR	NCA	DÉCT	BGV	PIZ	OFP	XVIe	IMP	
pon	-	-	-	2	-	-	112	-	1	-	-	-	-	PION1 (3)
pond	1	-	-	2	-	2	-	-	6	-	3	4	-	
pondans	1	-	-	-	-	-	-	-	-	-	-	-	-	PONDANT (2)
ponde	1	1	-	-	-	-	-	-	3	-	-	-	-	POINDRE (3)
pondent	-	-	1	-	-	-	-	-	3	-	-	2	1	
pondera	-	-	1	-	-	2	-	-	-	-	-	-	2	
ponderons	1	1	-	1	-	-	-	-	-	-	-	-	-	PONDÉRON (2)
pondez	3	3	-	-	-	-	-	-	2	-	-	-	-	PONDE (2)
pondre	5	2	4	1	1	26	1356	-	11	-	-	21	3	
pondus	-	-	1	-	-	2	-	-	2	-	-	5	4	
pone	-	-	1	-	-	1	1	-	1	-	-	2	1	PONDRE (2)
ponent	-	-	2	2	10	6	-	-	2	-	-	6	1	PONANT (3)
ponge	1	-	-	-	-	-	-	-	1	-	-	-	-	PONDRE (2)
ponis	-	-	1	-	-	1	-	-	1	-	-	-	-	
ponnes	1	1	-	1	-	-	1	-	-	-	-	-	-	PONDRE (3)
ponnu	-	-	4	-	-	4	-	-	5	-	-	-	-	
pons	4	3	99	27	12	134	44	2	8	-	78	11	3	PION1 (4)
ponse	4	2	2	4	-	4	-	-	1	-	-	-	-	PONCE (3)
pont	45	42	711	281	62	788	354	146	10	3	925	234	298	POINDRE (3)
ponte	-	-	2	-	-	10	-	-	6	-	-	10	-	
pos	12	5	60	30	8	68	62	5	9	10	38	1	-	PONDRE (6)
post	5	3	43	9	5	94	24	-	17	-	-	18	8	PONDRE (2)
poste	36	16	17	19	-	26	7	-	6	1	1	111	230	POESTE (4)
puins	-	-	-	3	10	-	22	-	1	-	-	-	-	
puns	3	-	-	-	-	-	1	-	2	-	-	-	-	PON (2)
pus	1	1	3	1	1	4	174	-	17	1	1	7	116	PONDRE (3)
	D	L	I	P	BFM	7FMR	NCA	DÉCT	BGV	PIZ	OFP	XVIe	IMP	

26 formes du lemme attestées

Extrait du lexique morphologique LGeRM

Plan de l'exposé

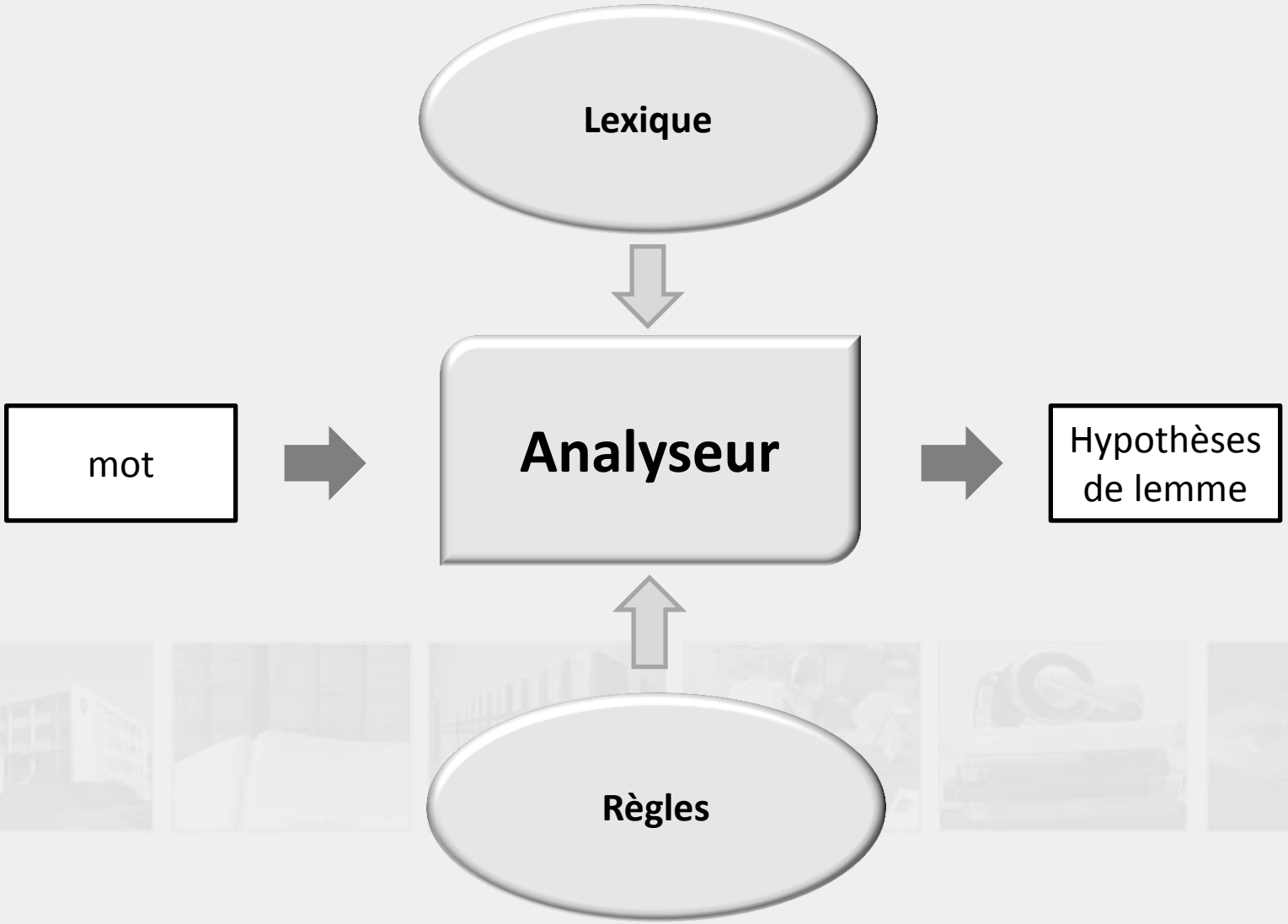
- DICTIONNAIRE ET VARIATION GRAPHIQUE
- **PRÉSENTATION DU LEMMATISEUR LGeRM**
- ÉVOLUTIONS
- OUTIL GLOSSAIRE

Le lemmatiseur LGeRM

- LGeRM
  - **L**emmes **G**raphies **e**t **R**ègles **M**orphologiques
  - http://www.atilf.fr/LGeRM

- Évolution de l'outil
  - 1986 : graine (DEA Informatique, LORIA+URFA)
  - 2001-2003 : Dictionnaire du Moyen Français
  - 2004 : CILPR Aberystwyth, Pays de Galles
  - 2009 : TAL volume 50
    http://www.atala.org/LGeRM

Architecture



Analyseur

Algorithme

```
si la graphie est dans le lexique alors
    Proposer l'analyse
sinon
  tantque conditions faire
      Appliquer les règles pour trouver une graphie connue
  fait
finsi
```

Conditions

- stratégie de gestion des formes produites
- mécanisme d'arrêt

Le lexique

- Liste de triplet (forme, lemme, étiquette)
  - (amer, AIMER, verbe)
  - (amer, AMER, adj.)
  - (amera, AIMER, verbe)

- Construction du lexique
  - Lexique initial issu des exemples du DMF, lemmes et étiquette DMF
  - de collaborations
  - enrichi à partir des corpus FRANTEXT
  - janvier 2017 environ 954 000 entrées.

Les règles 1/5

- Règle (morphologique) : morphologie et variation graphique
- Structure générale d'une règle
  - `Si` *conditions* `alors` *action* `finsi`
- Conditions sur les graphèmes du mot
  - en finale, en initiale
  - précédé de, suivi de : une lettre, liste de lettres, d'une consonne, d'une voyelle, sauf …
- Conditions sur le lemme
- Conditions sur le succès de la règle

Les règles 2/5

- Règles sur la flexion verbale
  - finale : retrouver l'infinitif
  - transformation de la finale : autre personne
    - si (en finale) alors RONT → RA finsi
    - *menront* → *menra*, MENER
  - autre transformation de la finale
    - si (en finale) et (précédé de [D,T,V]) alors ERAI → RAI finsi
    - *ponderai* → *pondrai*, PONDRE
  - ou cas inverse
    - *menront* → *meneront*

Les règles 3/5

- Flexion non verbale
  - si (en finale) alors ES → EFS finsi — *nes*→ *nefs*, NEF
- Modernisation/archaïsation
  - Y → I — *fayre*→ *faire*, FAIRE
- Équivalence graphique
  - C→ SS — *mesfacent* → *mesfassent*, MÉFAIRE
- Agglutination adverbe, pronom, élément formant
  - *tresadvisé* → *advisé*, TRÈS +AVISÉ
- Variantes régionales
  - OUN → ON — *mount*→ *mont*, MONT

Les règles 4/5

- Le système comporte environ 6 500 règles
  - 200 règles initiales de 1986
  - ajout de la flexion verbale en 2001
  - confrontation au DMF et aux corpus textuels
  - ¾ pour la flexion verbale et ses variations
- Pallier les lacunes du lexique
  - On n'aura jamais toutes les variations possibles d'un mot dans le lexique
  - la variation graphique est contenue dans les règles

Les règles 4/5

- Décrire la variation/flexion du lemme CONNAISSANCE

[c|k|q][o|oi|e|oei][n|nn|gn|ngn][oi|ai|i|ioi|e|oe][s|ss|sc|sç|ç|c][i]?[en|an|ã|ẽ][s|ss|c|sc|ç|ch][e][sz]?

cognescence cognissance cognissanche
cognoeissance cognoiscences cognoisçance conaisanche
congnioissance congnoessance
connissanche conoissances cougnoissance…

- 54 formes attestées dans nos corpus médiévaux

Plan de l'exposé

- DICTIONNAIRE ET VARIATION GRAPHIQUE
- PRÉSENTATION DU LEMMATISEUR LGeRM
- **ÉVOLUTIONS**
- OUTIL GLOSSAIRE

Outil glossaire

- 2007- : préparation du glossaire d'un texte
  - projet franco-britannique : Christine de Pizan (Édimbourg) et Froissart (Sheffield/Liverpool)
  - intégration du lemmatiseur dans un environnement
    - texte source en XML/TEI
    - accès au texte en continu, par mot, par lemme
    - correction de l'édition, désambiguïsation, glose
    - export des résultats
  - http://www.atilf.fr/LGeRM/glossaire/

Couverture diachronique plus large

- 2010 : adaptation à la langue du XVIIe
  - projet européen IMPACT (IMProving ACcess to Text) : océrisation et interrogation des fonds anciens des bibliothèques dans chacune des langues des partenaires
  - produire un lexique pour l'océrisation et un lexique pour l'interrogation
  - archaïsation du lexique moderne et projection sur un corpus textuel
    - Morphalou (entrées du TLF et flexions)

Couverture diachronique plus large

- Morphologie/variation de nature différente
  - ajout de lettres étymologiques : *havons poincter*
  - règles typographiques : *cinquiême à/a*

- Capable de traiter une transcription diplomatique
  - u/v i/j barre de nasalisation s long
  - *ſciẽce, neceßité, comẽ*

Intégrer LGeRM dans son application

- 2010 : service web
  - http://www.atilf.fr/dmf/definition/*aimer*
    http://www.atilf.fr/dmf/morphologie/*amer*
  - http://www.atilf.fr/dmf/LGeRM?w=aimer

```
<LGeRM>
  <orth>aimer</orth>
  <list>
    <w n="0"> <!-- no rule applied, the word is known -->
      <lem>AIMER</lem>
      <pos>verbe</pos>
    </w>
  </list>
</LGeRM>
```

```
<LGeRM>
  <orth>hamour</orth>
  <list>
    <w n="1"> <!-- 1 règle appliquée -->
      <lem>AMOUR1</lem>
      <pos>verbe</pos>
    </w>
    <w n="2"> <!-- 2 règles appliquées -->
      <lem>AIMEUR</lem>
      <pos>verbe</pos>
    </w>
  </list>
</LGeRM>
```

```
<LGeRM>
  <orth>qqqqq</orth>
  <list n="0"/>
</LGeRM>
```

Les lexiques morphologiques LGeRM

- Impulsion projet ANR/DFG PRESTO
- 2013 : distribution des lexiques morphologiques
  - LGeRM médiéval (juin 2016)
    - optimisé pour 1300-1500
    - 88 039 lemmes et étiquettes DMF
    - 951 452 entrées / 192 607 attestées FRANTEXT
  - LGeRM XVIIe (2013) / LGeRM PRESTO
    - optimisé pour 1550-1700
    - 89 754 lemmes et étiquettes TLF
    - 2 959 371 entrées / 116 161 attestées (3,9%)

Les lexiques morphologiques dans FRANTEXT

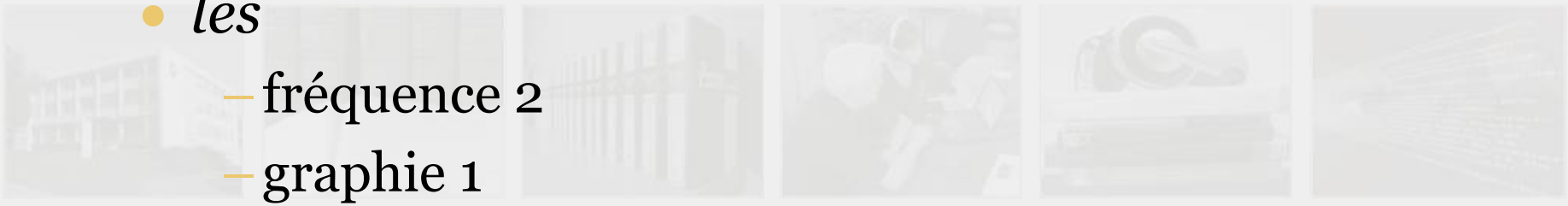
- 2013 : gestion de différents états de langue dans FRANTEXT

▶[12]	6225	Symon, le juste et le cremeu. - Item, où fut rosty l'	aigniel		
▶[13]	6225	lieu où Nostre Seigneur mengea avecq ses appostres l'	aigniel		
▶[14]	7006	son filz. Et ce meismes experiment elle avoit fait d'un	aignel		
▶[15]	7006	; et fut tres piteux et se jouoit aux lions comme a	aigneaulx		
▶[16]	7006	solennité dicte Pasque en la quelle ilz mengoient l'	aignel	de Pasques atout le pain aliz, comme il appert en Exode	Zoom
▶[18]	6203	beste en char, en os et en sanc, tout ainsi comme vn petit	aigniel	senz laine, si que on mangue et le fruit et la beste. Et	Zoom
▶[19]	9998	compaignie a ses filles. clxxvii. De l'arbre qui est dit	aignel	chaste. clxxviii. Des pastures d'Egypte. clxxix.	Zoom
▶[20]	9998	bestes, il separoit les poulains des vielz chevaus et les	aigniaus	des grans brebis. Et il ot i. frere qui ot non Jubal,	Zoom
▶[21]	8203	appartenans audit prieur, assavoir est de laines, d'	aigneaux	, de pourceaux, d'oisons, de chanvres, de lin et de	Zoom
▶[22]	0806	et puis; Qu' onques turtre ne turterelle,	Aingnaus	, coulons, ne coulombelle, Damoiselle, ne pucelette	Zoom
▶[23]	6701	anel gisant, lequel y getera en la mer. Et tantost que li	aigneaux	sera departi de moy, la nef s' en pourra aler saine et	Zoom

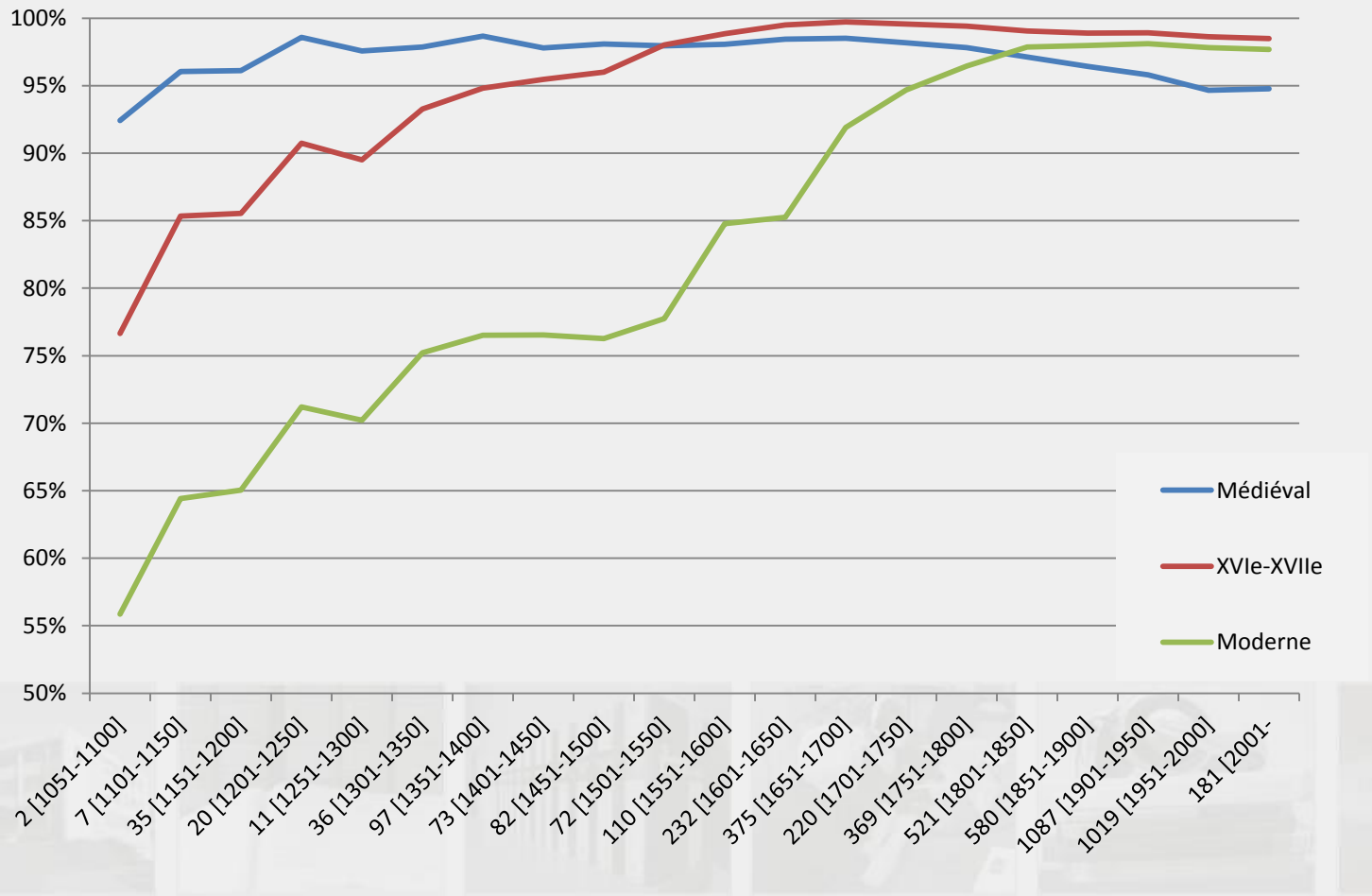
▶ [1]	6222	Sainct Estienne et Abibon. Lieu où fut rôti l' **agneau**	pascal. - A main dextre, et environ le coing de l'	Zoom
▶ [4]	6222	son chief entier, ses deux mains ou en l'une y a ung **agneau**	d'or et une pierre rouge, l'os d'une espaulle, et	Zoom

Taux de couverture des lexiques

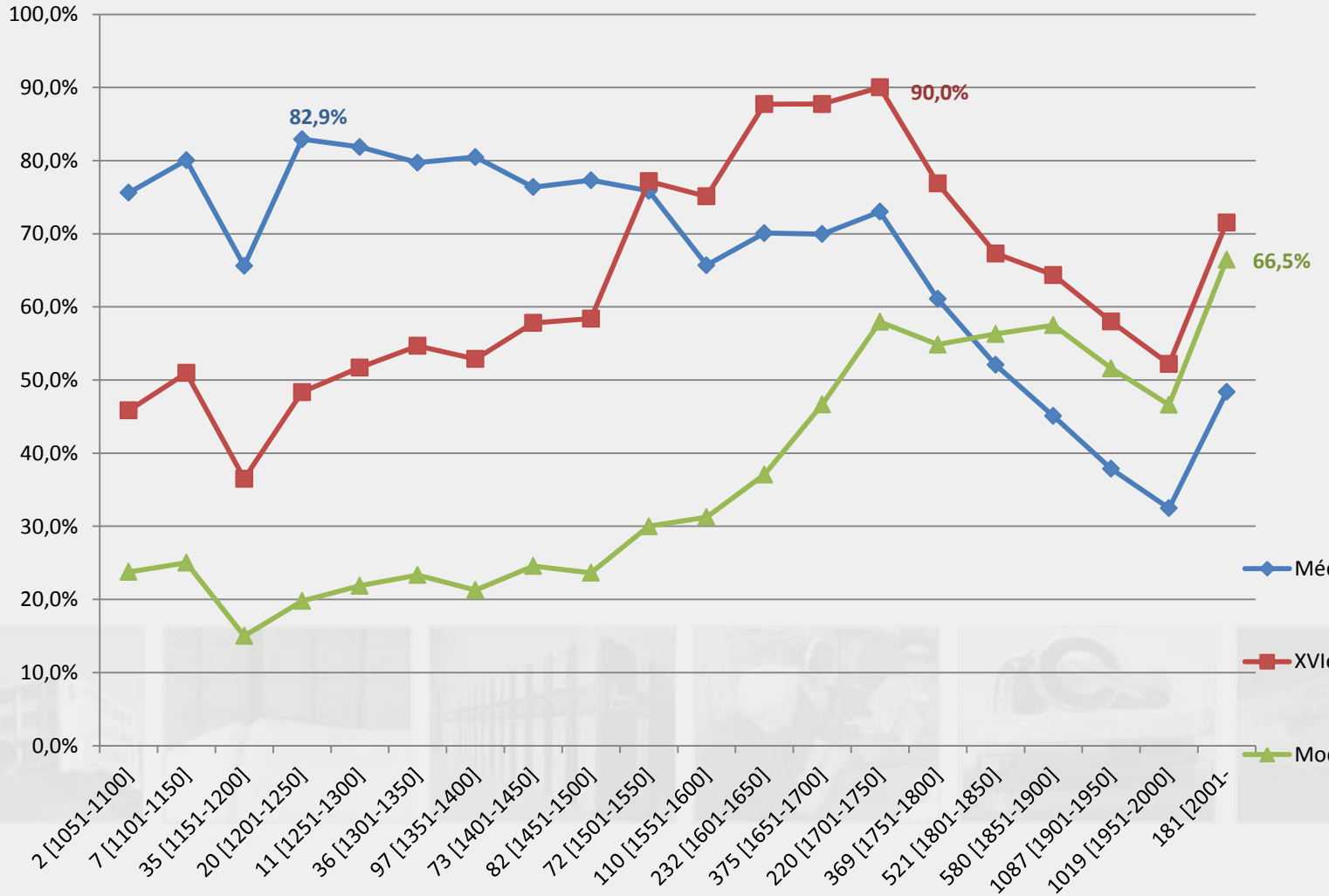
- Taux de couverture des lexiques :
  - quel pourcentage de mots du corpus sont accessibles
  - février 2017
- Distinction : fréquence / graphies
  - *les chas et les soris.*
  - *5 mots 4 graphies*
  - *les*
    - fréquence 2
    - graphie 1



Taux de couverture des lexiques : fréquence



Taux de couvertures des lexiques : graphies



Plan de l'exposé

- DICTIONNAIRE ET VARIATION GRAPHIQUE
- PRÉSENTATION DU LEMMATISEUR LGeRM
- ÉVOLUTIONS
- **OUTIL GLOSSAIRE**

Outil glossaire

- Outil collaboratif en ligne
  - utilisateurs, projets, droits
- Outil en cours de développement
- Étapes du travail
  - lemmatisation
  - étude des résultats de la lemmatisation
  - exploitation des résultats
  - diffusion des résultats

Outil glossaire

- 3 modes de consultation
  - texte en continu
  - accès aux formes
  - accès aux lemmes
- 2 modes d'affichages
  - liste d'attestations
  - concordancier

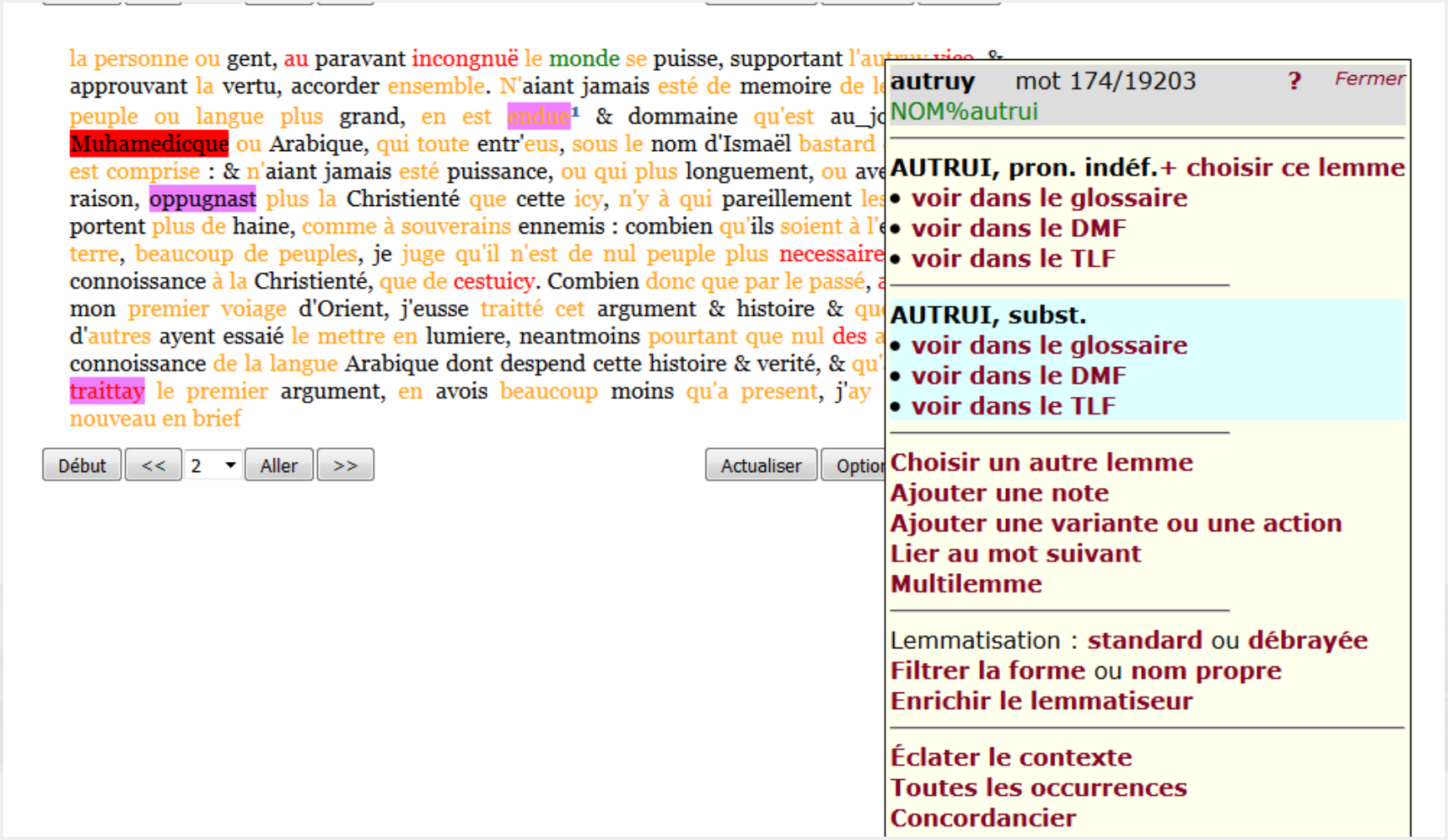
Lemmatisation

- Lemmatisation hors ligne par l'informaticien
  - exécutable intégré dans une chaîne de traitement
  - lemmatisation hors contexte
  - outil de désambiguïsation
- Mise en forme du texte
  - XML/TEI vs format tabulaire
  - segmentation
- Choix de l'état de langue, étiquettes

Étude des résultats

- Contrôle des résultats
  - erreurs de transcription/océrisation
  - mots absents du lexique (forme ou lemme)
  - règles absentes
  - nouvelle(s) lemmatisation(s)

Étude des résultats



Étude des résultats

1	Muhamedicque NAM% Frantext **mot 211**	mot inconnu +	, en est endue[1] & dommaine qu'est au_jourd'huy la **Muhamedicque** ou Arabique, qui toute entr'eus, sous	**2**
2	JESUSCHIST NAM% Frantext **mot 1768**	mot inconnu +	estre servi qui jamais sera. Les Ismaëlites confessant que **JESUSCHIST** receu des Christiens est le Messie, promis aux	**8**
3	tresmãuvais VER:impf% Frantext **mot 3728**	mot inconnu +	Ismaël feust chassé, & desherité, Abraham le trouva **tresmãuvais** & tresdolent, vaincu du divin commandement, luy	**17**

19	ignoranrans NOM% Frantext **mot 13318**	mot inconnu +	, & de celle de Grace, estant premierement par **ignoranrans**, ou mauvais Payens, Juifs ou Christiens,	**58**
20	baucoup NOM% Frantext **mot 13507**	mot inconnu +	l'Orient & Midy sous l'alCoran, lequel avec **baucoup**[3] d'autres livres est en telle authorité, comme	**59** **Note** : erreur LGeRM

Étude des résultats

3	traittay VER:cond% Frantext **mot 373**	**TRAITER**, verbe 1 10 + **TRACTER**, verbe 2 20 + **TROTTER**, verbe 2 20 +	cette histoire & verité, & qu'alors que je **traittay** le premier argument, en avois beaucoup moins qu'	**2**
9	tresraisonnable ADJ% Frantext **mot 690**	**RAISONNABLE**, adj.1 10 + TRÈS, +	sage donne raison de toutes ses actions, & plusque **tresraisonnable**, que toutes actions durables soient faittes avec l'	**4**
140	Historiografe NOM% **mot 9069**	**HISTORIOGRAPHE**, subst. 1 10 +	par les Ismaëlites confermés : tellement que comme escrit Giafer **Historiografe** Arabe duquel l'histoire Arabique est en Baviere chez	**40**

Étude des résultats

- Choix du lemme
  - choix manuels
  - outils externes
    - XVIe-XVIIe
      - TreeTagger + paramètres modernes +modernisation à la volée
      - TreeTagger + paramètres PRESTO
- Notes

Étude des résultats

<< A B C D E F G H I J K L M N O P Q R S T U V >> Options
W Z Accès direct à...

ABAISSÉ, adj.	1	1	abbaissé
ABAISSER, verbe	2	2	abbaisser abbaissé
ABANDONNÉ, adj.	7	2	abandonné abandonnés
ABANDONNÉ, subst.	7	2	abandonné abandonnés
ABANDONNER, verbe	9	3	abandonner abandonné abandonnés
ABÂTARDI, adj.	21	7	abastardi abastardie abbastardi abbastardie abbastardies abbastardis abbastardy
ABÂTARDIR, verbe	21	7	abastardi abastardie abbastardi abbastardie abbastardies abbastardis abbastardy
ABBÉ, subst.	1	1	abbé
ABCÈS, subst.	3	1	assés
ABOLIR, verbe	5	3	aboli abolie aboly

Étude des résultats

ABAISSER, verbe		2 formes 2 attestations
abbaisser 1 att.		
1	, neantmoins le bien qu'ils ont fait tant en **abbaisser** l'orgueil des faux Christiens, Juifs, &	**mot 15302** --- **66**
abbaissé 1 att.	**ABAISSER**, verbe | **ABAISSÉ**, adj.	
1	pechés, de Dieu laissé, & son ange premierement **abbaissé**, jamais ledit ange ne laisse ledit peuple,	**mot 7772** --- **34**

Exploitation des résultats

- Finalisation des résultats :
  - prise en compte des informations de désambiguïsation

- Exploitation
  - statistiques
  - interrogation : attestations/concordancier
  - gloses, article au format DMF, export texte

Diffusion des résultats

- Édition électronique en ligne
  - interrogation : attestations/concordancier
  - glossaire en ligne

Démonstration

- Accès à l'outil :
  - http://www.atilf.fr/LGeRM/glossaire
- Demande de création de compte
  - http://www.atilf.fr/LGeRM/compte
- lgerm@atilf.fr

Perspectives

- Distribution du lemmatiseur
  - redéfinir les entrées/sorties
  - optimisation
  - multi plate-forme
  - gestion du lexique
- Donner la possibilité de déposer son texte et lemmatiser
- Adaptation du lexique au texte