Création de modèle(s) HTR pour les documents médiévaux en ancien français et moyen français entre le Xe -XIVe siècle

Séance 3 : L'allographie, entre besoins scientifiques et pragmatiques. Comment modéliser et optimiser les données d'entraînement pour l'HTR (I) ?

J.-B. Camps, F. Duval, A. Pinche

14 décembre 2021

1. L'allographie, entre besoins scientifiques et pragmatiques
2. Comment signaler les corrections dans le manuscrit ?
3. L'hyphénisation
4. Diacritiques médiévaux
5. La ponctuation médiévale

L'allographie, entre besoins scientifiques et pragmatiques

- Transcription « allographétique » (*graphetic* en anglais) qui vise à donner accès à toutes les formes de chaque lettre ou signe
- Transcription « graphématique » (*graphemic* en anglais) qui préserve la suite des lettres et réduit chaque forme à son sens dans un système alphabétique

Nous reprenons ici les terminologies utilisées par Dominique Stutzmann, « Paléographie statistique pour décrire, identifier, dater... Normaliser pour coopérer et aller plus loin ?», Franz Fischer, Christiane Fritze, Georg Vogeler (eds), *Codicology and Palaeography in the Digital Age 2*, pp.247-277, 2011, halshs-00596970

1 L'allographie, entre besoins scientifiques et pragmatiques

2 Comment signaler les corrections dans le manuscrit ?

3 L'hyphénisation

4 Diacritiques médiévaux

5 La ponctuation médiévale

Classer les types de corrections : Mise en page ou signes diacritiques ?

Les catégories suivantes :

- glose marginale
- glose encadrante
- marginalia
- note

peuvent être signalées grâce à la segmentation du document en utilisant, par exemple, la zone *Margin* de SegmOnto. Les gloses interlinéaires peuvent être signalées par un type de ligne particulier comme le type *Interlinear* de SegmOnto.

Classer les types de corrections : mise en page ou signes diacritiques ?

Comment indiquer les corrections quand le texte est barré, exponctué ou que l'encre a été modifiée ? Comment signaler un blanc en attente de correction ? Un signe non imitatif pourrait être utilisé pour signaler les zones de correction suivante :

- exponctuation
- texte souligné
- biffure/rature (quand le texte est lisible)
- Proposition de signe : © (COPYRIGHT SIGN, 00A9)

Classer les types de corrections : mise en page ou signes diacritiques ?

- Comment signaler un blanc en attente de correction ?
- Comment signaler un morceau de texte qui a été rendu illisible par un grattage ?
- Nous proposons d'utiliser le signe UTF-8 : « ¬ » (NOT SIGN, 00AC)

Signes fonctionnels et signes de renvois

Dans les manuscrits, les ajouts peuvent être signalés :

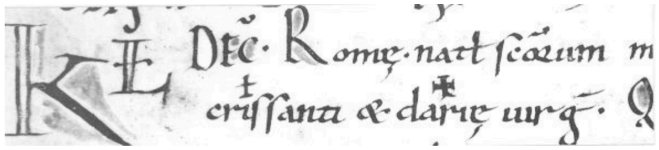
- par une mise en page différente (vu précédemment)
- par des signes fonctionnels qui permettent de faire le lien entre la zone de texte principal concernée et l'ajout

Il peut être très utile de les signaler pour le post-traitement des prédictions et faire, ainsi, un matching automatique du texte pour mettre les ajouts ou les notes au bon endroit.

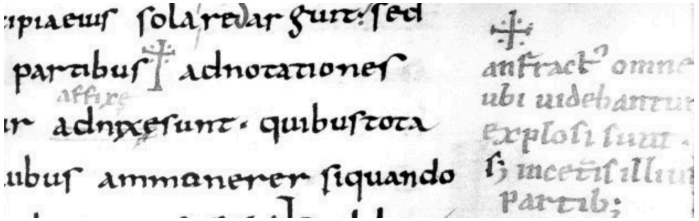
Signes de renvois

Dans le cas simple d'utilisation d'une lettre ou d'un chiffre en guide de signe de renvoi, nous proposons de transcrire le caractère tel quel.
Voici une liste rapide des autres signes de renvois (établie à partir de Codicologia et du *Vocabulaire codicologique* de Denis Muzerelle)

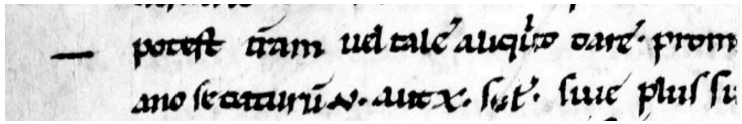
- Croisette



- Astérisque

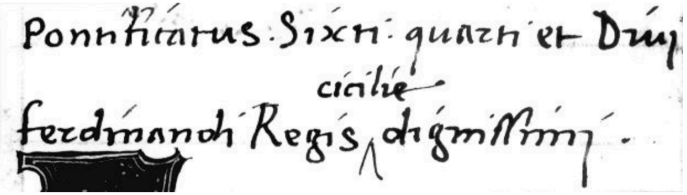


- Obèle



Signes de renvois

- Caret



- Manicule

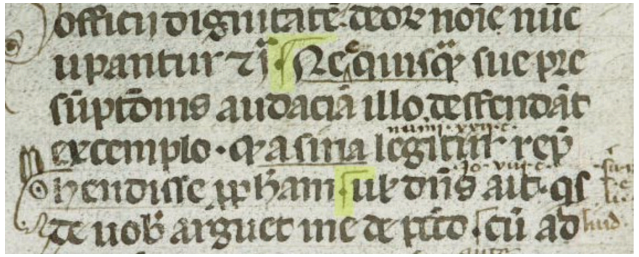
Signes de renvois

Vers des propositions de solutions :

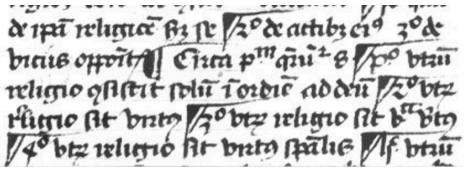
- Les croisettes et astérisques pourraient être signalés par le signe * (U+002A)
- Que faire des obèles ?
- Les carets pourraient être signalés par le signe ˄ (chevron d'insertion, U+2038)
- Les manicules pourraient être signalées par le signe → (Flèche vers la droite, U+2192) ou par le signe Unicode ☞ (manicule, U+261E)

Signes fonctionnels

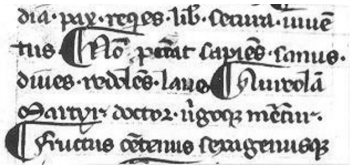
- Gamma capitulaire



- Crochet alinéaire



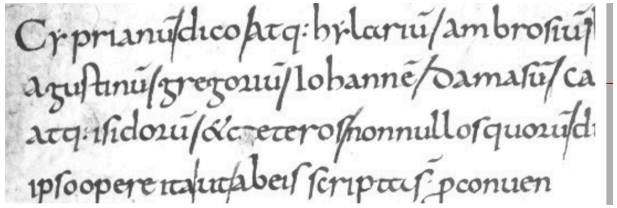
- Pied-de-mouche



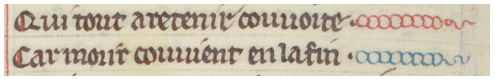
Tous ces signes étant assez semblables et ayant la même fonction, ils pourraient être représentés par un même signe ⁋ (pied-de-mouche réfléchi, U+204B), qui évite d'utiliser le signe ⁋ de la zone privée MUFI.

Signes fonctionnels

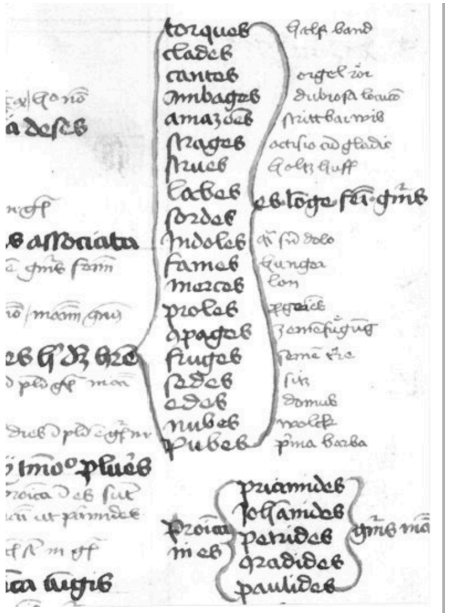
- Diastole



- Bout de ligne décoré

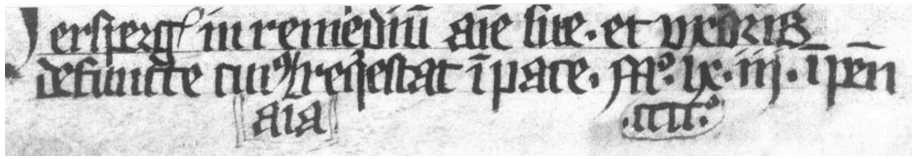


- feston et accolade

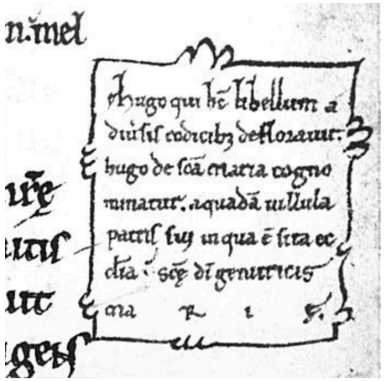


Signes fonctionnels

- Circonduction



- Cartouche



1 L'allographie, entre besoins scientifiques et pragmatiques

2 Comment signaler les corrections dans le manuscrit ?

3 L'hyphénisation

4 Diacritiques médiévaux

5 La ponctuation médiévale

L'hyphénisation

- Faut-il noter les informations sur l'hyphénisation ?

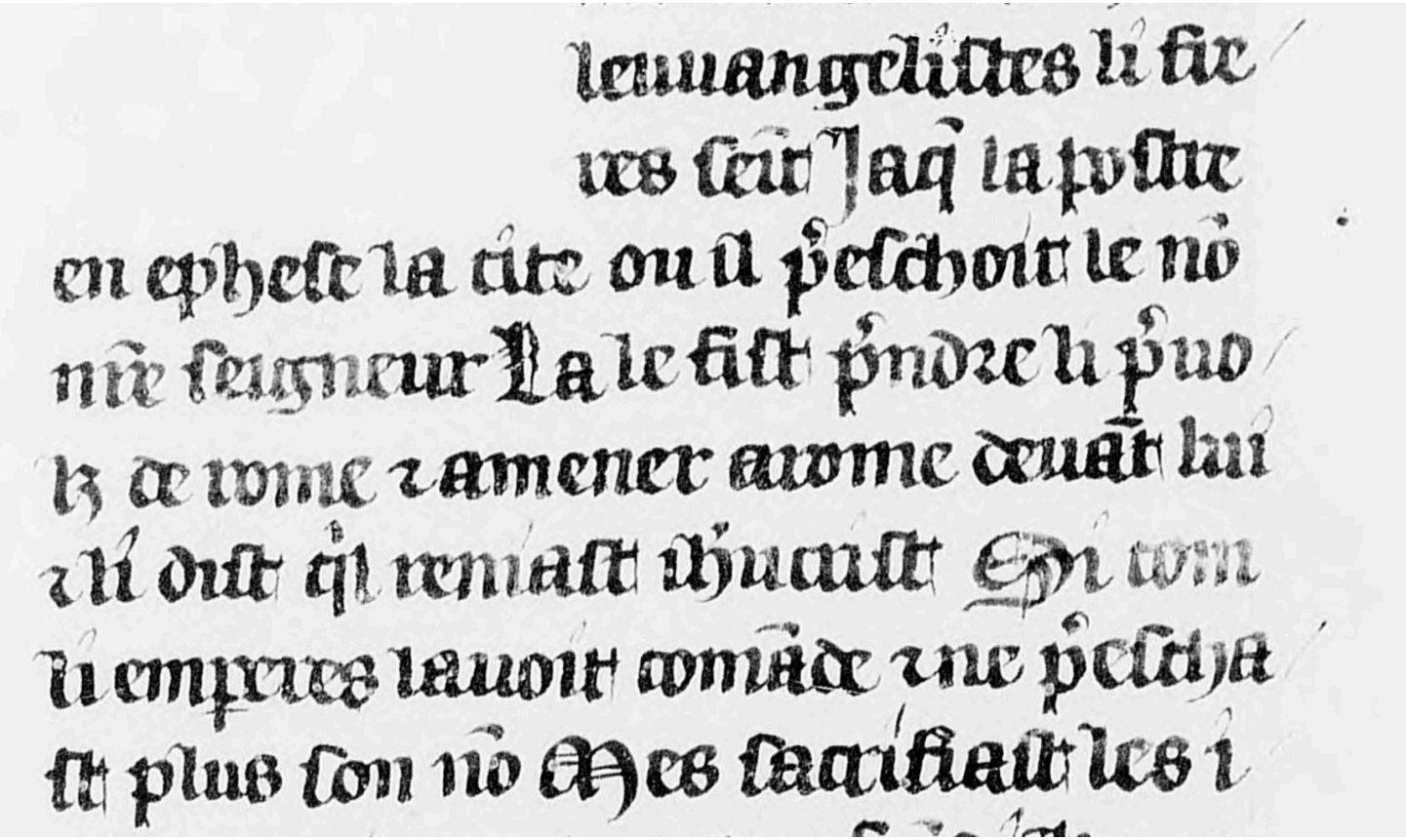


FIGURE – Hyphénisation, ms. BnF fr. 411

L'hyphénisation

- Si oui, comment ?

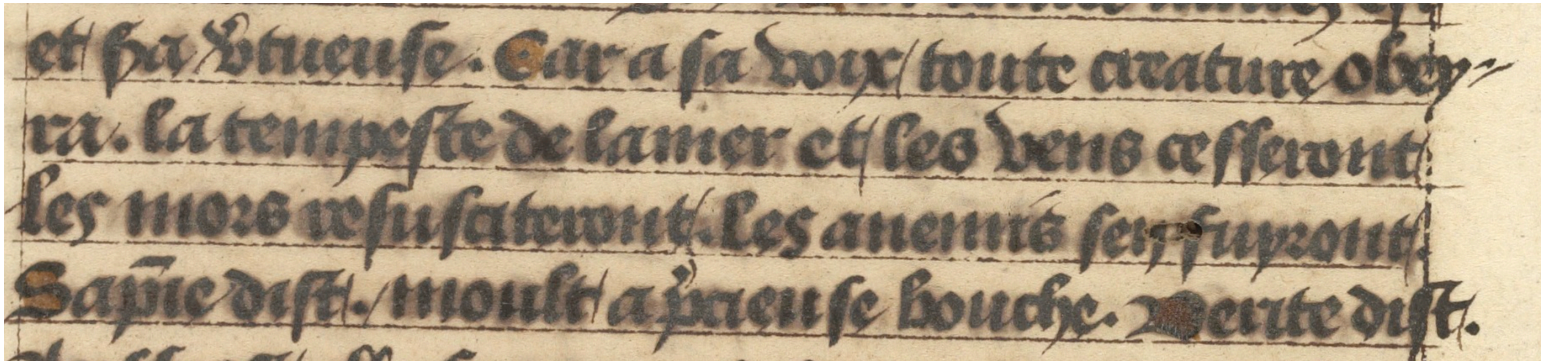


FIGURE – Hyphénisation, bibliothèque de l'université de Pennsylvanie 660 ; fol.88

1 L'allographie, entre besoins scientifiques et pragmatiques

2 Comment signaler les corrections dans le manuscrit ?

3 L'hyphénisation

4 Diacritiques médiévaux

5 La ponctuation médiévale

Pointage des i

- Faut-il noter les informations sur le pointage des i ?

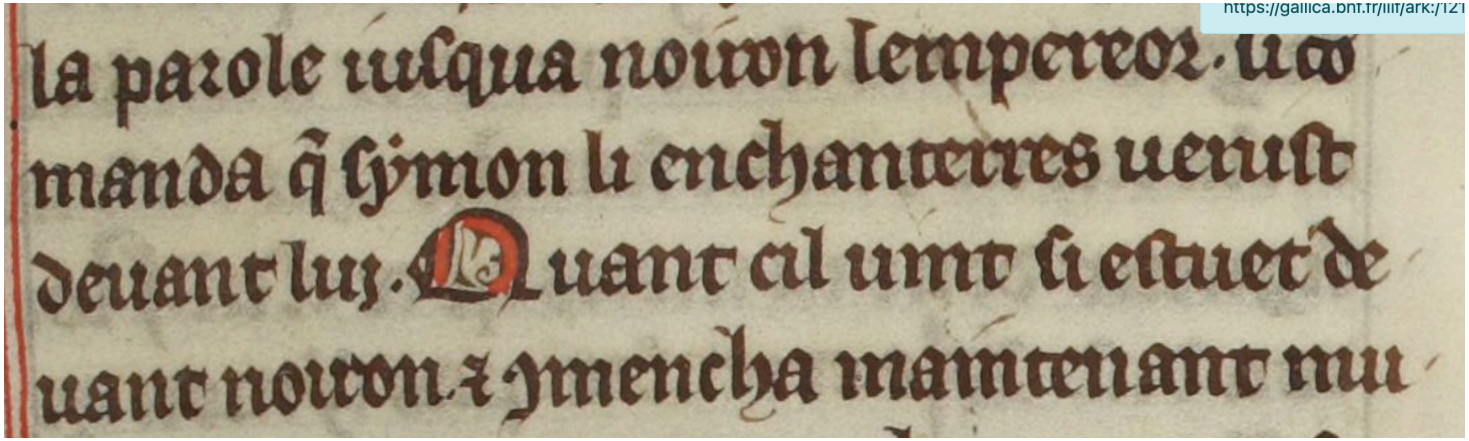


FIGURE – Pointage des i, ms. BnF fr. 412

- Si oui, comment ?

voyelles accentuées

- Faut-il noter le e accentué ?

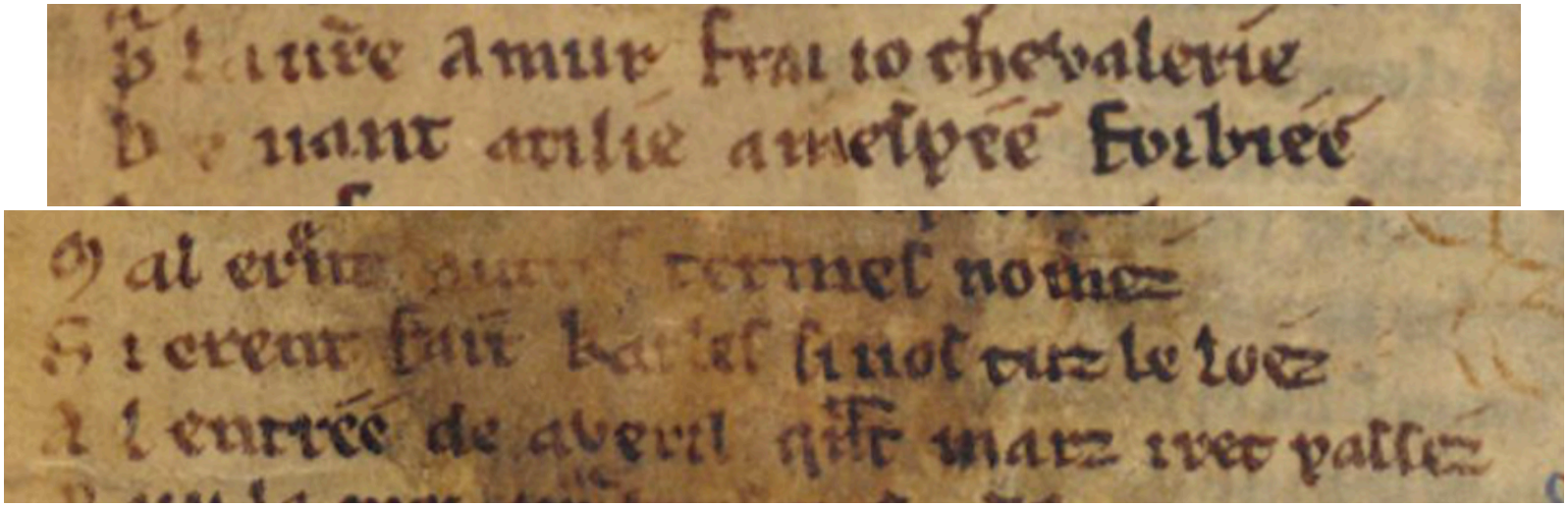


FIGURE – e accentué, fragments d'anciens manuscrits de poésie française, BnF NAF 5094

- Si oui, comment ?

e caudata

- Faut-il noter le e cédille ?

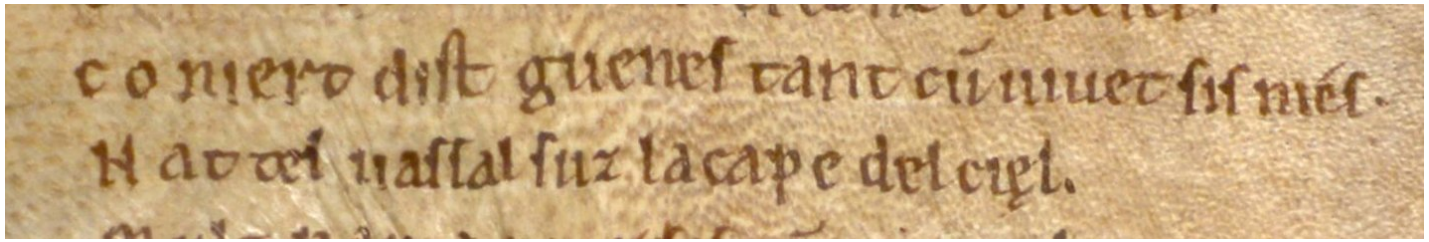


FIGURE – e cédille, Bodleian Library, Oxford, ms. Digby 23b

- Si oui, comment ?

1 L'allographie, entre besoins scientifiques et pragmatiques

2 Comment signaler les corrections dans le manuscrit ?

3 L'hyphénisation

4 Diacritiques médiévaux

5 La ponctuation médiévale

La ponctuation médiévale

Comment représenter la ponctuation ?

- Point : "."
- Point bas, médian : " ▪ "
- Point haut : "˙" (F1F8, zone privée MUFI)
- ponctus elevatus : "ʔ" (F1E0, zone privée MUFI)
- Punctus circumflexus : "ˀ" (F1F5, zone privée MUFI)
- Point-virgule : " ;"
- Deux points superposés : " :"
- Virgula : "ˏ" (F1F7, zone privée MUFI)