

Contenu de la documentation

1	HTR	3
1.1	Problématique	3
1.2	Choisir un corpus d'entraînement	3
1.2.1	Main 1	4
1.2.2	Écriture de Constance de Salm (CdS)	4
1.3	Segmentation et typage des zones d'écriture	4
1.3.1	Typier les régions d'écriture	4
1.3.2	Typier les lignes d'écriture	5
1.3.3	Phénomènes graphiques particuliers	5

Chapitre 1

HTR

1.1 Problématique

Quatre à cinq mains différentes ont été repérées jusqu'à présent dans la correspondance de CdS. Cette variété d'écritures peut sérieusement entraver les performances d'un modèle de reconnaissance.

Deux pistes méthodologiques se dessinent :

1. Rassembler dans un premier temps des lettres qui sont de la même main, pour voir quels sont les résultats du modèle qu'H. Souvay a commencé à entraîner ;
2. Reprendre un modèle déjà entraîné à travailler sur plusieurs mains ; c'est l'option qui a été privilégiée par le projet Lectaurep¹).

1.2 Choisir un corpus d'entraînement

Les recueils de lettres constituent la part du corpus la plus normée sur le plan paléographique. La distribution des mains y est variable selon les tomes :

1. Le premier volume² présente une grande variété de main s'enchaînant fréquemment les unes aux autres ;
2. Le deuxième volume³ présente en revanche une meilleure cohérence paléographique : la même main peut se suivre sur un bon nombre de pages consécutives, facilitant l'entraînement d'un modèle sur une écriture particulière. En partie utilisé par H. Souvay pour ses tests, nous avons repris ce volume pour constituer un premier sous-corpus paléographiquement cohérent.

1. Alix Chagué, *Création de modèles de transcription pour le projet LECTAUREP #2*, Lectaurep : l'intelligence artificielle appliquée aux archives notariales, URL : <https://lectaurep.hypotheses.org/488> (visité le 05/04/2022).

2. Constance de Salm, *Correspondance générale, seconde copie, 1er volume, 1785-1814*, URL : <https://constance-de-salm.de/archiv/#/document/11215> (visité le 11/04/2022).

3. Id., *Correspondance générale, seconde copie, 2ème volume, 1815-1821*, URL : <https://constance-de-salm.de/archiv/#/document/11216> (visité le 11/04/2022).

1.2.1 Main 1

Nous avons établi une liste de 47 images (soit 47 doubles pages) au sein du 2e volume attestant une écriture homogène que nous dénommons *Main 1*. Nous avons pour cela arbitrairement découpé les lettres afin de ne travailler que sur un seul type d’écriture, sachant que les changements de main interviennent souvent en milieu de page. Quelques corrections de la main de CdS apparaissent ponctuellement ⁴.

1.2.2 Écriture de CdS

Le site ne publie aucune lettre originale de la main de CdS, mais 52 brouillons (*Entwurf*) ⁵.

Entraîner un modèle de reconnaissance sur cette écriture supposerait un travail délicat de transcription pour une écriture particulièrement cursive (compter environ deux semaines pour disposer d’une bonne vingtaine de pages), mais l’investissement peut en valoir la peine.

1.3 Segmentation et typage des zones d’écriture

Nous avons procédé à une première expérience de transcription sur le sous-corpus *Main 1* avec le logiciel e-Scriptorium installé localement.

1.3.1 Typier les régions d’écriture

Le typage est utile en ce qu’il permet de traiter de manière différentielle des régions et des lignes selon leur type, afin de les affecter à des éléments distincts de l’arborescence XML-TEI qu’il faudra construire.

Il faut donc réfléchir aux besoins de cette transformation vers le format TEI. Les *Guidelines* de l’édition de correspondance du projet DAHN permettent de guider cette réflexion ⁶. Par ailleurs, F. Chiffolleau a formulé une ontologie pour les régions et lignes des écrits de correspondance en langue française pour le XXe siècle ⁷ :

Certaines régions pourraient être directement appliquées :

- **Main** (pink)
- **Title** (green)

4. Des reproductions en qualité réduite de ces images ont été placées dans le dossier `./htr/img/main1bd`

5. Le dépouillement se trouve dans le fichier `./htr/mains/brouillonsCDS.md`

6. Floriane Chiffolleau, *Correspondence : Guidelines*, 10 janv. 2022, URL : <https://github.com/FloChiff/DAHNProject/blob/8df8dfc6053a7dd57a6c5510d1e56bb336ce1d04/Correspondence/Guidelines/Documentation-Correspondance.pdf> (visité le 07/04/2022).

7. Id., *[Correspondance En Langue Française, XXe s.]* SegmOnto, 10 déc. 2021, URL : https://github.com/SegmOnto/examples/tree/main/sources/lettre_fr_XXe (visité le 07/04/2022).

- **Signature** (orange)
- **Letterhead** (purple)
- **Numbering** (dark green)
- **Salute** (red)
- **Dateline** (dark blue)

Il pourrait être pertinent de modifier l'usage de :

- **Additions** (turquoise): *cette catégorie est utilisée ailleurs dans Segmonto, pour les documents administratifs* ⁸ ; elle intervient dans le traitement du document postérieurement à sa rédaction. Cette pertinence reste à confirmer. Cette catégorie pourrait également s'appliquer aux rubriques :

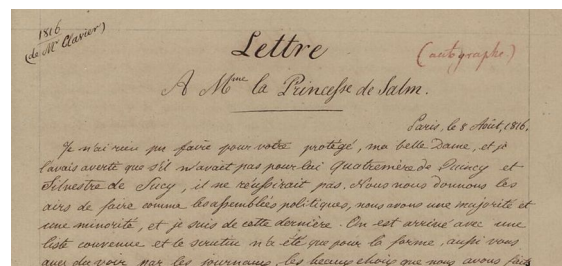


Figure 1.1 – Rubrique "autographe".

Il pourrait être pertinent de reprendre ou de créer d'autres concepts :

- **Note** (turquoise): *pour les notes infrapaginales (utilisé dans SegmOnto pour les imprimés* ⁹
- **Postscript** (yellow): *cela remplacerait le rôle à l'origine assigné à Additions. J'opterais bien pour le jaune car il ne va pas me servir par ailleurs, et qu'on ne risque guère d'avoir un tampon proche du post-scriptum.*

La figure 1.2 propose une mise en oeuvre de ce typage des régions.

1.3.2 Typage des lignes d'écriture

Les types de lignes dont on propose l'utilisation sont :

- **Main**
- **Verse** : *les passages en vers sont relativement nombreux*
- **Correction** : catégorie existant par défaut dans e-Scriptorium, elle s'appliquerait uniquement pour les corrections appliquées dans l'interligne.

1.3.3 Phénomènes graphiques particuliers

CdS a corrigé certains mots de sa main :

8. A. Chagué, [Documents Administratifs, XIXe s.] SegmOnto, 10 déc. 2021, URL : https://github.com/SegmOnto/examples/tree/main/sources/administratif_XIXe (visité le 07/04/2022).

9. [Imprimés], SegmOnto, 10 déc. 2021, URL : https://github.com/SegmOnto/examples/tree/main/sources/prints/BnF_cb120117553 (visité le 07/04/2022).

- En rayant une lettre, un mot ou plusieurs mots, ou bien en réécrivant par dessus le texte. Dans de nombreux cas cela consiste en une simple lettre barrée ; le typage de la ligne demanderait alors beaucoup d’effort pour un résultat minime ;
- En réécrivant dans l’interligne : il est alors pertinent d’utiliser le type de ligne e-Scriptorium **Correction**.

Un ensemble de solutions d’encodage des corrections a été proposé dans le cadre du projet DAHN¹⁰.

10. F. Chiffoleau, *Few Tips for Reading the Text Files*, GitHub, URL : <https://github.com/FloChiff/DAHNProject/tree/master/Project%20development/Texts> (visité le 11/04/2022).

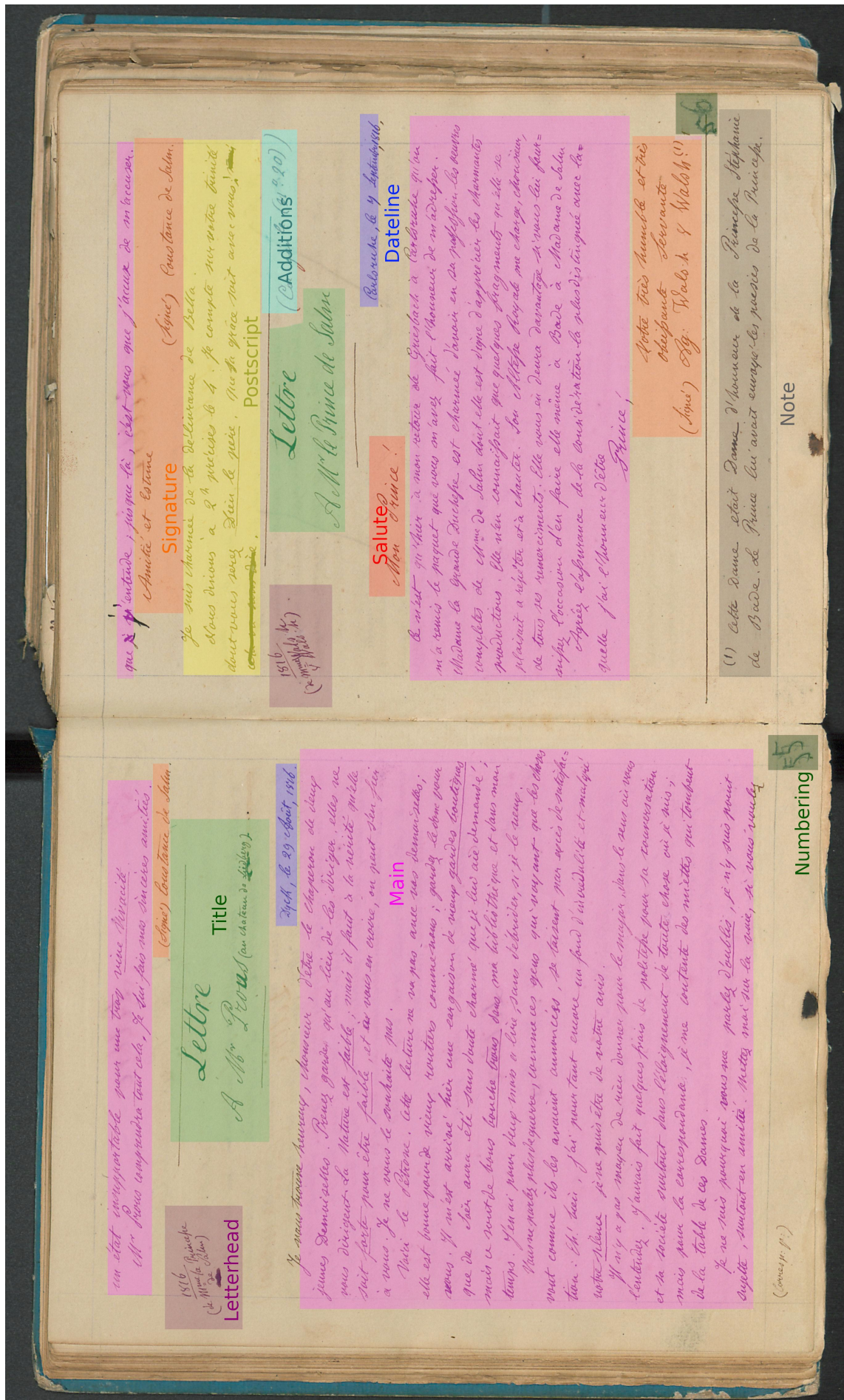


FIGURE 1.2 – Exemple de typage des zones de texte sur une double page.