Contenu de la documentation

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 été privilégiée par le projet Lectaurep [1]).

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 volume [2] présente une grande variété de main s'enchaînant fréquemment les unes aux autres ;
2. Le deuxième volume [3] 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 Typer 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. Chiffoleau a formulé une ontologie pour les régions et lignes des écrits de correspondance en langue française pour le XXe siècle [7] :

Cetaines 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 Chiffoleau, *Correspondence : Guidelines*, DAHN Project, 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 :*

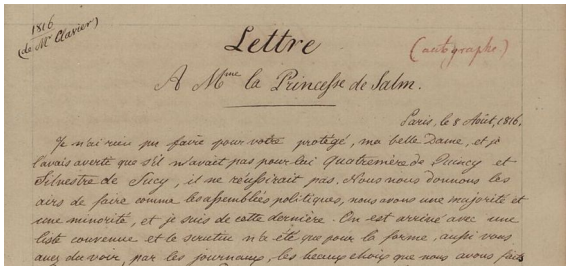

Figure 1.1 – Rubrique "autographe".

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]
— **Postscritp** (yellow): *cela repmplacerait 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 ci-dessous propose une mise en oeuvre de ce typage des régions.

1.3.2 Typer les 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 DAHN [10].

J'envisage plutôt **ne pas encoder ces éléments dans la phase d'HTR**, et de ne les aborder que la phase d'édition. Il sera de toute façon necessaire, lors de la reprise manuelle de l'édition TEI, de suivre la reproduction du manuscrit à éditer. En outre, introduire des caractères tels que £, €, etc. dans la transcription génèrerait du bruit dans l'entraînement du modèle HTR et imposerait une phase de nettoyage pour les réutilisations éventuelles des vérités de terrain.

En somme, il s'agirait de **transcrire tout ce qui est lisible** (y compris les lettres biffées), en privilégiant le dernier état du texte dans le cas où la correction a été superposée à la première couche d'écriture.

1.4 Mise en oeuvre de la reconnaissance d'écriture

Dans le cadre de son stage, H. Souvay a initié l'entraînement d'un modèle HTR à partir d'un petit volume de vérités de terrain [11]. La méthodologie employée était la suivante :

> Nous avons décidé de tenter la transcription automatique sur un sous-ensemble du corpus composé de copies de lettres compilées dans des recueils. […] Les mains sont relativement constantes dans le temps dans ce sous-ensemble contrairement au reste du corpus. Même proches, ces mains demeurent différentes. Nous avons donc opté pour l'entraînement d'un modèle multi-mains, c'est à dire un modèle non-spécialisé capable de transcrire plusieurs mains représentées dans le corpus d'entraînement [12]

Il s'agit dans un premier temps d'augmenter le volume des vérités de terrain pour améliorer les performances du modèle entraînés par H. Souvay.

10. F. Chiffoleau, *Few Tips for Reading the Text Files*, DAHN Project, URL : `https://github.com/FloChiff/DAHNProject/tree/master/Project%20development/Texts` (visité le 11/04/2022).
11. Hippolyte Souvay, *La Correspondance de Constance de Salm (1767-1845) : Rapport de Stage*, rapport de stage de seconde année de master Humanités numériques et computationnelles, École nationale des chartes-Institut historique allemand à Paris, 2021.
12. *Ibid.*, p. 6-7.

1.5 Automatiser la correction des prédictions

Nous avons suivi la démarche explicitée dans la documentation du projet DAHN [13] et proposé quelques développements aux scripts issus de ce projet.

1.5.1 Analyser les mots

Nous avons appliqué le script d'analyse de mots spellcheck-texts-PAGEXML.py à nos prédictions HTR. Les corrections sont plus nombreuses sur des prédictions HTR que sur des prédictions OCR, surtout avec un modèle encore peu entraîné. La correction est donc un travail conséquent. En outre, il faut veiller à ne produire que des corrections dépourvues d'ambiguïtés et applicable en toutes circonstances. Si le modèle lit "celle" pour "cette", seule une correction manuelle peut y remédier ; le risque du dictionnaire est de remplacer automatiquement ailleurs des prédictions justes par le terme trouvé. Il ne faut pas oublier que le remplacement des mots par le dictionnaire est indépendant du contexte du mot en question.

Afin de faciliter la correction des dictionnaires généré par le script pour chaque page (chaque proposition de correction doit en effet être contrôlée), on a développé le script pour afficher le contexte du mot et en conserver la mémoire, ce qui limite les allers-retours entre le dictionnaire à corriger et l'image ou la prédiction d'origine ; le contexte est en effet déterminant pour valider ou modifier une correction proposée automatiquement.

Dans le but d'optimiser la performance de l'analyse des mots on a développé une fonction appelée get-lemmes, qui fouille les vérités terrains déjà constituées et permet de donc de valider rapidement les mots déjà rencontrées dans le traitement de la correspondance de CdS avant de devoir lancer leur recherche dans un dictionnaire généraliste plus vaste.

Les corrections précédemment validées sont elle aussi mobilisées lors de l'analyse des prédictions, ce qui permet non seulement de réexploiter facilement des corrections déjà effectuées, mais aussi d'écarter des formes qui auraient précédemment identifiées comme ambiguës (pouvant être résolues de plusieurs façons selon le contexte, et donc à ne pas corriger automatiquement).

13. F. Chiffoleau, *How to Do a Post-OCR Correction for TEXT Files*, DAHN Project, 8 avr. 2022, URL : `https://github.com/FloChiff/DAHNProject/blob/8df8dfc6053a7dd57a6c5510d1e56bb336ce1d04/Project%20development/Documentation/Post-OCR%20correction%20for%20TEXT%20files.md` (visité le 11/04/2022).

1.5.2 Constituer un dictionnaire personnalisé

1.5.3 Appliquer le dictionnaire personnalisé aux prédictions HTR

Les corrections s'avérant particulièrement nombreuses, le script text-correction-XML.py écrit par F. Chiffoleau a dû être perfectionné afin de procéder à une tokénisation des mots avant de mobiliser les entrées du dictionnaire pour corriger les formes erronées présentes dans le texte.

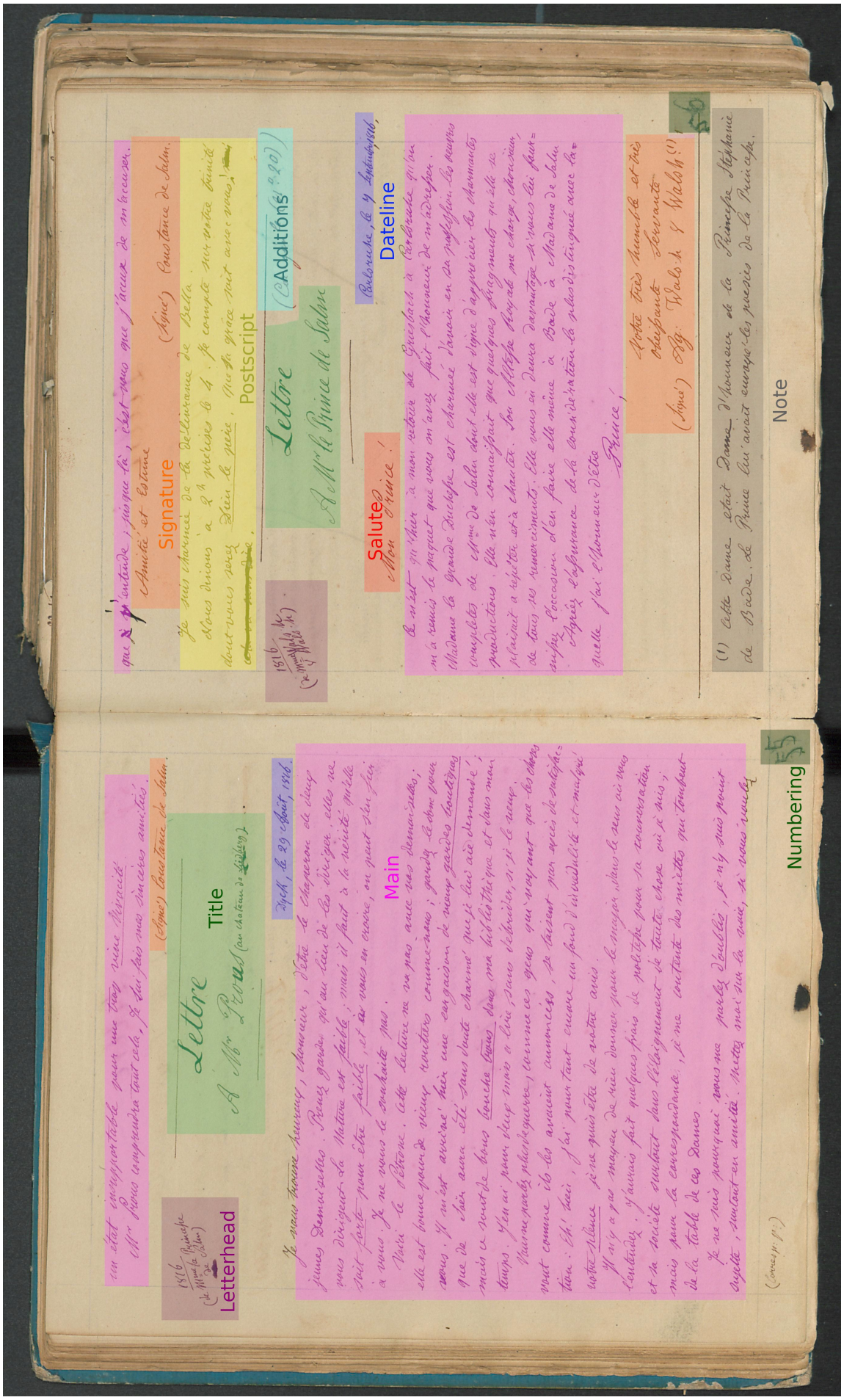

FIGURE 1.2 – Exemple de typage des zones de texte sur une double page.