

Récapitulatif du travail sur Transkribus

Lundi 8 juin 2020, J13

Jeudi 11 juin 2020, J14

Par Lucie Slavik, stagiaire pour le CRHXIX du 4 mai 2020 au 31 juillet 2020, trois jours par semaine.

Prise en main du projet

Avant de procéder au travail sur Transkribus, il a fallu prendre en main le dossier qui représente un fonds de plus de 2000 lettres, dont plus de la moitié sont écrites de la main de Le Play. Frédéric Le Play étant la figure majeure, celle autour de laquelle gravite toute la correspondance, il a paru prioritaire à l'équipe du Centre d'Histoire du XIXe siècle d'entraîner un modèle sur son écriture, et non sur celle de ses correspondants, ce qui pourra éventuellement se faire par la suite. On peut penser notamment à des correspondants qui ont beaucoup échangé avec lui, notamment Emmanuel Parent de Curzon (135 lettres MS_6062) ou Napoléon-Joseph Bonaparte (SIM Ms 6062_1.pdf) qui compte 94 lettres de sa main.

Il faudra déterminer si l'établissement d'un modèle serait un gain de temps pour les écritures d'autres mains, ou sinon continuer les transcriptions « manuelles », ce que nous ferons de toutes façons pour les correspondants de moindre importance, tel que Louis-Joseph Buffet (SIM Ms 6062_1.pdf) qui n'a que deux lettres, ou encore Lucien Brun qui n'en compte que 18.

Pour cela, une bonne connaissance du corpus est nécessaire, afin de ne pas naviguer à vue et faire les bons choix. A cet effet, j'ai réalisé deux tableaux excel, le premier afin de m'y repérer dans tout ce fond disparate, qui l'embrasse dans sa presque totalité, du moins pour ce qui m'a été transmis dans le cadre de mon stage en télétravail. Il s'intitule Inventaire_cor_LP.xlsx. Cependant, celui-ci étant trop détaillé et réalisé davantage pour mon usage personnel, j'ai réalisé un deuxième excel intitulé Inventaire_cor_LP_TRANSKRIBUS.xlsx qui recense les lettres sélectionnées pour mon stage et classées par ordre de priorité, de O (plus prioritaire) à 2 (moins prioritaire) selon la qualité des numérisations et des transcriptions, et l'importance des fonds ou quantité de lettres.

Transkribus, un pari

Transkribus est un outil utilisé par de nombreux projets, mais il n'est pas forcément approprié à tous les corpus. Une transcription « manuelle » nécessite, certes, beaucoup de temps et une relecture attentive, mais entraîner une machine telle que Transkribus nécessite également beaucoup de temps, et ceci toujours avec l'incertitude du résultat.

S'il y a plus de 90 % de taux de réussite (soit 10 % de taux d'erreur), la transcription pourra se faire sans trop de difficultés, mais la relecture restera longue et nécessaire. S'il y a 98 % de taux de réussite (donc 2 % de taux d'erreur), c'est à nous de voir si nous acceptons ce taux d'erreurs ou si nous préférons relire pour un rendu plus optimal, mais également plus coûteux en temps.

Par ailleurs, comme le soulignent les tutoriels mis en ligne pour initier à Transkribus, il faut «du temps pour explorer Transkribus et se familiariser avec son fonctionnement¹». C'est donc un réel investissement à prendre au début, avec l'incertitude du résultat et la possibilité d'un échec.

1 <http://regis-schlagdenhauffen.eu/wp-content/uploads/2018/01/Comment-utiliser-Transkribus-%E2%80%93-en-10-%C3%A9tapes-ou-moins.pdf>

Pour notre part, les résultats ont été dès le début encourageants. Sur 16 lettres, nous avons obtenu des scores de 16% d'erreur (donc 84 % de réussite) sur les caractères en entraînement, 26 en test, soit une lettre sur 6, et une lettre sur 4.

Nous avons donc continué cette aventure virtuelle, encouragée que nous étions par ces premiers retours.

Transkribus mode d'emploi

Le travail sur Transkribus s'est fait en plusieurs temps.

- Tout d'abord, il a fallu se familiariser avec l'outil, ce que nous avons fait notamment avec les tutoriels de l'EHESS, par Régis Schlagdenhauffen.

- Puis, il a fallu télécharger les manuscrits dans transkribus et mettre les transcriptions à chaque page, en se servant du travail déjà réalisé par les étudiants.

C'est un travail qui prend du temps mais s'avère n'être pas trop compliqué une fois que l'on s'est familiarisé avec l'outil. Pour voir quelles en sont les étapes, se reporter au tuto1 que j'ai fait afin d'assurer une transmission de savoir au CRHXIX.

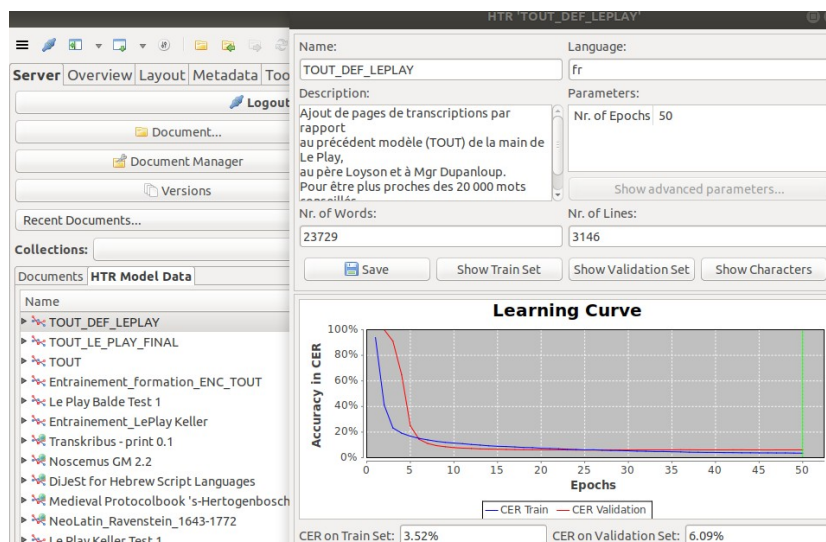
- Après le téléchargement des données, il faut entraîner le modèle, et pour cela prendre 90 % des lettres pour le train set, et 10 % pour le validation set. Voir le tuto2 à ce propos.

Dans l'idéal, il est conseillé de charger environ 20 000 mots, ce que j'ai fait.

- Le modèle étant entraîné, si le taux de réussite est suffisant, on peut passer à la transcription machine.

Bilan sur les côtés positifs et négatifs :

Une progression intéressante



La progression du modèle s'est avérée être riche de promesses.

Ainsi, en partant à 1849 mots, on obtenait :

CER ON TRAIN SET : 16,31 %

CER ON VALIDATION SET : 26,25 %

Puis, avec 1951 mots :

CER ON TRAIN SET:0,36 %

CER ON VALIDATION SET : 27,10 %

Avec 1977 mots :

CER ON TRAIN SET : 0,29 %

CER ON VALIDATION SET : 12,57 %

Avec 4207 mots :

CER ON TRAIN SET : 0,86 %

CER ON VALIDATION SET : 11,55 %

Avec 15928 mots :

CER ON TRAIN SET:2,58 %

CER ON VALIDATION SET : 6,40 %

Avec 18531 mots :

CER ON TRAIN SET : 2,91 %

CER ON VALIDATION SET : 7,83 % (même set de validation que le précédent)

Avec 23729 mots :

CER ON TRAIN SET:3,52 %

CER ON VALIDATION SET : 6,09 % (même set de validation que le précédent)

Cependant, comme souligné plus bas, certains de ces chiffres sont biaisés car je n'ai pas toujours utilisé le même set de validation, excepté pour les trois derniers.

Il y a certains chiffres fluctuants du fait de la part d'aléatoire en entraînement, peut-être dues à la qualité de l'image ou à un changement d'écriture (main qui vieillit).

Des résultats ambivalents

Ayant trouvé la progression de l'entraînement du modèle plutôt bonne quant aux chiffres, j'étais assez confiante pour l'appliquer sur des manuscrits non transcrits et voir le résultat du travail d'une petite dizaine de jours.

J'ai donc, à cette fin, mis de côté des manuscrits venant de fonds divers et variés, dans un dossier nommé E-echantillon_test_modele. Dedans se trouvent huit pages :

1-Une page est tirée du fonds « AD Haute Savoie », des lettres de Le Play à Joseph Despine. Or, ces numérisations sont d'une qualité très médiocre, on dirait presque que ce sont en réalité des photos floues. Le résultat transkribus a été mauvais pour cette page. Pas une ligne sans faute, à part le « Monsieur » et la signature de Le Play. *Néanmoins, il faut savoir que 5 % de CER, c'est une faute tous les 20 caractères, donc pour une moyenne de 4 lettres par mots, cela donne 1 faute tous les 5 mots au moins* (explication de Monsieur Clérice).

Je me l'explique par la mauvaise qualité de la numérisation d'une part, et d'autre part peut-être du fait que je n'ai entraîné pour le modèle aucune lettre de ce fonds.

2- Une page est tirée du fonds de Frédéric de Mercey à la BNF (BNF10). La qualité de la numérisation est plutôt bonne. Cependant, même si le résultat est meilleur que la précédente lettre, il reste encore beaucoup de fautes, et je me demande si le fait de les corriger ne prendrait pas plus de temps que de tout transcrire en partant de zéro. Pourquoi ce résultat décevant ? Là encore, je n'ai pas transcrit de lettres de ce fonds pour l'entraînement, mais je pense surtout que l'écriture de Le Play diffère en ce qu'elle est plus fine, plus penchée, finalement plus jeune ! et la machine ne l'a peut-être pas reconnue.

Il faudrait peut-être entraîner un modèle qui prenne en compte le changement d'écriture entre un Le Play jeune et un Le Play dont l'écriture s'est transformée avec l'âge, ou encore entraîner sur les deux périodes.

3- Une page est tirée de la bibliothèque publique universitaire de Genève « CD Le Play Antoine Savoye » : c'est une lettre de Le Play au père Hyacinthe Loyson. J'ai entraîné près de 90 pages de ce fonds. J'y ai trouvé malgré cela pas mal de fautes. D'autre part, la numérisation

comportant des nuances de gris assez foncées, la machine l'a pris pour de l'écriture : cela demandera beaucoup de temps de corriger ces fautes car cela matche à chaque fois une ligne qui n'en est pas une et cela croit y reconnaître de l'écriture. Il faut donc tout effacer et tout corriger : c'est une grande perte de temps (voir si parfois l'on peut supprimer par TR).



1-13 aes

1-14 pan

1-15

1-16 qu

1-17 Mon cher Père et ami

1-18 Dans une réunions d'amis qui se dévouent

1-19 à la réforme vocale. j'ai trouvée hier un clerc qui

Dans une réunion d'amis qui se dévouent
à la réforme sociale, j'ai trouvé hier un clerc qui
vous dira beaucoup. Son opinion est de vos amis,

4- Une page est tirée du fonds conservé au Château de Ligoure, de Le Play à son fils Albert. La transcription est assez propre, mais subsistent encore pas mal de fautes. Le Play forme parfois mal ses lettres, ce qui ne facilite pas la tâche, et pour l'œil humain, et pour la machine. Par ailleurs, même si les mots sont reconnus, les accents sont souvent absents. (Par exemple, il est écrit senat pour sénat). Voir si c'est potentiellement un problème de ligne.

5- Une page tirée du même fonds. Mêmes remarques. Beaucoup de fautes. Par ailleurs, Le Play a tendance à gommer les différences entre les majuscules et les minuscules. Souvent, ses « s » minuscules semblent être des « s » majuscules. En revanche, le « A » de « Albert » pourrait être une minuscule. D'où de nombreuses majuscules prises pour des minuscules.

6- Une page tirée du dossier « Keele », lettres de Le Play à Alfred Tylor. Le résultat est plutôt satisfaisant.

La machine bute face à certains obstacles : les ratures.

Je mets toujours une grande attention à ne pas com-
promettre mes amis.
L'athénæum du 22 octobre a publié une critique
de la Réforme Sociale. Bien qu'écrète par un partisan

3-20 seconde édition, je ne citerai votre nom que dans le

3-21 cas où vous m'en donneriez expressément l'autorisation

3-22 Je mets toujours une grande absensienà ne pas com-

3-23 promettre mes amis.

3-24 L'athénæum du 22 octobre à publié lirre critique

3-25 de la Réforme Sociale. Bien qu'écrète par un partisan

Cependant, le résultat est plutôt bon dans l'ensemble, et la bonne qualité de la numérisation joue probablement un rôle dans ce sens.

7- Cette page, qui est une lettre de Le Play à Hippolyte Taine (NAF 28420) est étonnamment réussie. C'est presque un sans faute, excepté le tampon qui a été matché et pris pour un « R ». Si tous les résultats étaient aussi excellents, je recommanderais transkribus sans hésitation aucune. Malheureusement, c'est parce que la lettre a été transcrite aussi pour le modèle, ce qui donne un résultat finalement biaisé.

8- Le résultat de cette page de Le Play à Peruzzi est plus que médiocre. Une des raisons de cet échec est la mauvaise délimitation des lignes : l'écriture de Le Play est à la verticale et la machine ne sait le reconnaître. Corriger ce genre d'erreurs prend à mon sens plus de temps que de transcrire directement sans passer par la machine.

14 mai 1881.
Mon cher Peruzzi,
Le général Galdini a présenté
hier à Mon ami Cheysson, M. le
Commandeur Luxali Membre de
votre parlement. Cheysson me dit
que votre collègue lui paraît à pré-
mices vue se rapprocher beaucoup de
M. Minghetti, qui m'a donné en 1867
un si utile concours pour la distri-
bution des récompenses accordées aux

2-13 votre parlement. Cheysson me dit

2-14 "

2-15 x 6

2-16 7

2-17 8 à

2-18 que votre collègue lui paraît à pré-

2-19 "

Au terme de ce premier résultat suite au modèle que j'ai entraîné, je ressens plutôt de la déception, par rapport au pourcentage de réussite qui me faisait espérer quelque chose de plus convaincant. Une question se pose : faut-il continuer à entraîner le modèle pour arriver à un résultat plus satisfaisant ou faut-il abandonner l'expérience ?

Les difficultés de ce travail se voient surtout :

- lorsqu'on rencontre des lignes mal délimitées par transkribus et qu'il faut tout supprimer et recommencer manuellement (cf. captures d'écran dans le pasàpas)

- les fautes rencontrées dans les transcriptions d'étudiants et qu'il faut corriger.

- les mots illisibles qu'on tente de déchiffrer.

- Un souci assez important est le fait que plusieurs pdf comprenant des centaines et des centaines de pages numérisées **sont bien trop lourds** et il faut les traiter avant de les mettre sur transkribus. J'ai essayé de pallier à cela en compressant les PDF puis en les divisant en JPG, mais ce n'est pas du tout une solution idéale, étant donné qu'on perd ainsi la qualité de l'image. Je n'avais pas le choix car je ne pouvais les diviser ou les transformer en JPG sur internet car ils étaient trop lourds. J'ai perdu plusieurs heures à cause de ce problème. Pour l'instant, je ne vois comme solution que la création d'un compte sur un site qui traite les pdf (et donc c'est payant si je ne me trompe) ou sinon la transcription des pdf sans passer par Transkribus, car pour beaucoup, on rencontre ce problème pour des écritures autres que celles de Le Play, je pense notamment aux manuscrits SIM, ou sinon pour l'écriture de Le Play, mais non de la correspondance (ce sont dans ce cas ses manuscrits de livres). De toutes façons, il va falloir trouver une solution rien que pour la mise en ligne, que l'on utilise ou non transkribus².

J'ai appris par la suite qu'il existe des outils sous linux pour diviser les PDF facilement. Il faudrait donc se renseigner là-dessus.

Comment gagner en temps :

- Utiliser les raccourcis de clavier (alt + tabulation) pour passer de la transcription word à transkribus.

- Faire les transcriptions directement dans Transkribus et non plus sur word afin de ne pas doubler le travail. Transkribus permet en plus à avoir le texte en regard et à s'y coller.

Mieux faire

On gagnerait en temps :

- si une relecture plus efficace était faite en amont (ce qui est aussi chronophage mais d'une autre manière. On séparerait ainsi l'aspect plus technique de transkribus de la pure correction).

- si tous les transkribus allaient à la ligne à chaque changement de ligne dans le manuscrit (transcription facsimilaire). Le gain de temps est alors considérable : il suffit de double-cliquer à chaque ligne, de copier/coller et le tour est joué.

- si les dossiers avaient un nom plus explicite dans transkribus. Ainsi ai-je chargé un dossier de la BNF, mais on ne sait qui écrit à qui avant de l'ouvrir, ce qui est regrettable. Plus de clarté dans le nommage serait important pour la suite, afin d'éviter une mauvaise compréhension des dossiers et donc une perte de temps (voir les propositions de solution dans le mémoire?)

² Une autre difficulté encore : les bugs car l'ordinateur, beaucoup sollicité, a parfois tendance à se bloquer, ce qui nécessite un à plusieurs arrêts forcés.

Une méconnaissance de l'outil et du vocabulaire précis de l'apprentissage machine amène à des confusions. Il est bon de se renseigner au préalable au lieu de se lancer tête baissée dans le travail. Pour ma part, la précision des termes employés et une meilleure connaissance de l'outil m'aurait aidé à être plus efficace.

Par exemple, dans le tuto que je suivais, il était mentionné qu'il fallait mettre une page de côté dans le set de validation, alors que j'ai appris par la suite qu'il fallait mettre 10 % des données dedans, ce qui n'est pas pareil, et qu'il valait mieux garder toujours le même set de validation pour mieux se rendre compte des progressions de l'apprentissage machine. Ne sachant pas cela, j'ai d'abord mis une page, ce qui était trop peu, et pas toujours la même, erreur que j'ai rectifiée pour les trois derniers entraînements de modèle. Les données sont donc un peu faussées.

Les outils sont en anglais et il est donc difficile de voir comment s'exprimer : doit-on reprendre les termes anglais de la plateforme ou les termes français des tutoriels. Il semble y avoir une sorte de flou autour des termes à employer.

De la même manière, j'ai fait plusieurs fois un entraînement de modèle, que je pensais être le dernier, d'où un nommage ambigu entre « TOUT », « FINAL », « DÉFINITIF » alors que ce n'est pas définitif.

Proposition de Monsieur Clérice : « Dans ce genre de situation, le *versioning* numérique est pratique : imaginons la possibilité de faire :

1.0 = Premier set de données, premier set de validation

1.1 = Mise à jour set de données, même set d'évaluation.

1.1.1 = Même set de données et d'évaluation mais réentraînement. Soit pour évaluer la variation, soit suite à des petites corrections typo ! »

Assurer la transmission

Afin de penser à la suite, nous avons pensé réaliser un tutoriel pour faciliter l'apprentissage de transkribus aux personnes de l'équipe du CRHXIX qui seront amenées à utiliser l'outil par la suite. Nous l'avons divisé en trois : la première partie est consacrée au téléchargement et à la préparation des données et à la transcription. La deuxième se concentre sur l'entraînement du modèle grâce aux données initialement préparées. La troisième aide à transcrire avec le modèle.

Récapitulatif sur les fonds utilisés

Afin d'assurer la continuité du projet, il est bon de voir quelles lettres ont été utilisées pour l'entraînement du modèle LE PLAY pour ne pas faire double travail.

Voici un petit récapitulatif des données que j'ai utilisé, parmi les fonds qui m'ont été transmis dans le cadre de mon stage en télétravail :

- Tout le dossier qui a été mis de côté pour la formation à Transkribus à l'École nationale des Chartes a été utilisé pour l'entraînement du modèle. Dossier C-Matériel formation Mars 2020 EdC, sous-dossiers Le Play Baldé, Le Play Keller, Le Play Sciences Po = 38 pages

- Dans Numérisations/BNF/BNF4/JPG : lettres de LP à Hippolyte Taine 6 pages. C'est sur cette correspondance que j'ai perdu du temps sur le papier strié avec des lignes faussement reconnues.

- Transcriptions/FLP_Transcriptions_étudiants/M12HI0365 - Pouvoirs, sociétés, enquêtes dans le monde occidental - MBrejonDeLavergnée-Lettres transcrites Le Play-120177/PIERRE DULIN_21623_assignmentsubmission_file_/LP à Charles de Ribbe

13 pages sur 15 de Le Play à Charles de Ribbe.

- Numérisations/Peruzzi : 3 pages sur 33.

- Numérisations/BNF/BNF9/BNF9enJPG : 12 pages sur 43. Lettres de Le Play à Monseigneur Félix Dupanloup.

-Numérisations/CD Le Play Antoine Savoye/73 pages sur 117. Lettres de Le Play au père Hyacinthe Loyson.

=> 145 pages, 23 729 mots

Pour le set de validation, j'ai sélectionné quelques textes de plusieurs fonds afin d'embrasser un large champ d'écriture, dans le dossier nommé D-Set_validation_le_play (ou set_entrainement_leplay dans transkribus). On y trouve des lettres tirées du dossier composé pour la formation à l'École nationale des Chartes, des lettres de Le Play au père Loyson, à Mgr Dupanloup, et bien d'autres encore.

Continuer ?

Face à ces résultats, une question se pose. Est-il bon de continuer avec Transkribus ? L'expérience s'avère-t-elle être positive ou négative ? Encourageante ou dissuasive ?

N'étant pas une spécialiste, je ne saurais trancher franchement en faveur de l'un ou de l'autre. Je dirais que pour que l'utilisation de Transkribus soit plus convaincante, il faudrait absolument avoir des numérisations de bonne qualité, et non des numérisations floues ou des photos³.

Les lettres qui sont écrites sur un papier qui est légèrement strié entraînent des incompréhensions de la machine. Les numérisations en noir et blanc (comme celles de Le Play à Loyson) qui comportent des nuances de gris prêtent également à confusion.

Pour une meilleure efficacité de Transkribus, il faudrait avoir une matière première plus unifiée et de meilleure qualité dans l'ensemble⁴.

Mais n'est-ce pas là utopique ?

L'écriture de Le Play pose aussi pas mal de soucis. Je pense que nous n'arriverons pas à du 98 ou 99 % (à voir!). Les problèmes majeurs restent :

- les ratures

- les lettres non accentuées alors qu'il le faudrait (senat pour sénat)

- les lettres mal formées, les z qui ressemblent à des r, les s qui ressemblent à des r. Déjà pour l'œil humain, certains mots sont parfois illisibles. Nombre de transkribus le font remarquer.

- l'évolution de son écriture, entre sa jeunesse et sa vieillesse. Il faudrait entraîner plus de lettres de jeunesse. Ou sinon avoir une meilleure connaissance du fonds en faisant un point sur le nombre de lettres par année, la proportion de lettres écrites dans la jeunesse et celles de l'âge mûr. Selon ce résultat (s'il y a une majorité de lettres de l'âge mûr, ce que je pense et espère) ce serait un gain de temps : il n'y aurait pas besoin d'entraîner beaucoup de lettres de jeunesse je pense.

On pourrait dire que la moitié des lettres transcrites avec le modèle comportent pas mal de fautes, et qu'il est plus long de corriger les fautes, reprendre la transcription, que de faire une transcription en partant de rien.

3 Proposition de Monsieur Clérice : « Il faudrait aussi voir, mais vous n'en avez pas les moyens, les résultats avec d'autres outils à partir des mêmes données d'entraînement.

Deux phases mélangées dans transkribus : la reconnaissance de zone et la reconnaissance de texte. Il me semble que dans votre cas, la reconnaissance de zone pose de « gros » problèmes. On pourrait imaginer que d'autres outils seraient plus efficaces là-dessus (du côté de Kraken et e-scriptorium, mais encore plutôt instable) »

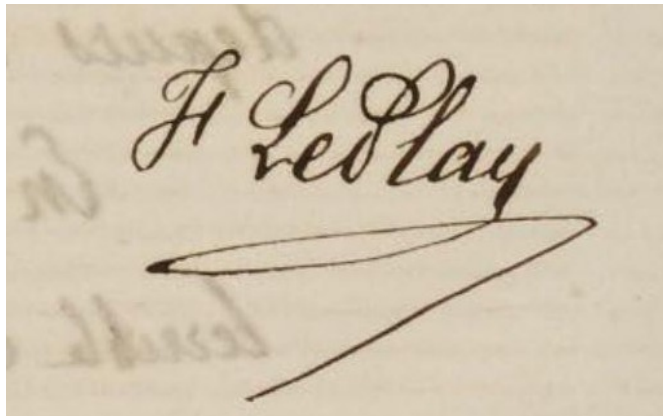
4 Remarque de Monsieur Clérice : « Cependant, si cela demande plus de données, cette variation de qualité permet aussi d'entraîner un modèle BEAUCOUP plus robuste. Ce n'est pas parfait mais le modèle se fait à cette difficulté. Plus lentement, c'est tout. »

Pour ma part, je serais pour tenter d'entraîner encore 5 000 ou 10 000 mots sur un corpus assez large et englobant, et voir si le modèle est meilleur.

Si ce n'était pas le cas, je n'insisterais pas forcément pour le modèle, mais Transkribus pourra toujours être utilisé comme interface de transcription.

Pour les écritures autres que celle de Le Play, il faudrait choisir vraiment celles qui ont plus de cent lettres et dont l'écriture est assez lisible, sinon je pense que c'est plutôt une perte de temps.

En général, je serais pour continuer les transcriptions de façon manuelle, au moins pour les écritures autres que celles de Le Play.

A photograph of a handwritten signature in dark ink on aged, slightly yellowed paper. The signature reads "H. Le Play" in a cursive script. Below the name is a large, elegant, horizontal flourish that loops back under the name. The background shows faint, ghosted impressions of other handwriting from the reverse side of the page.