Tutoriel Transkribus
CRHXIX

1ère étape, créer des données d'entraînement pour le moteur HTR

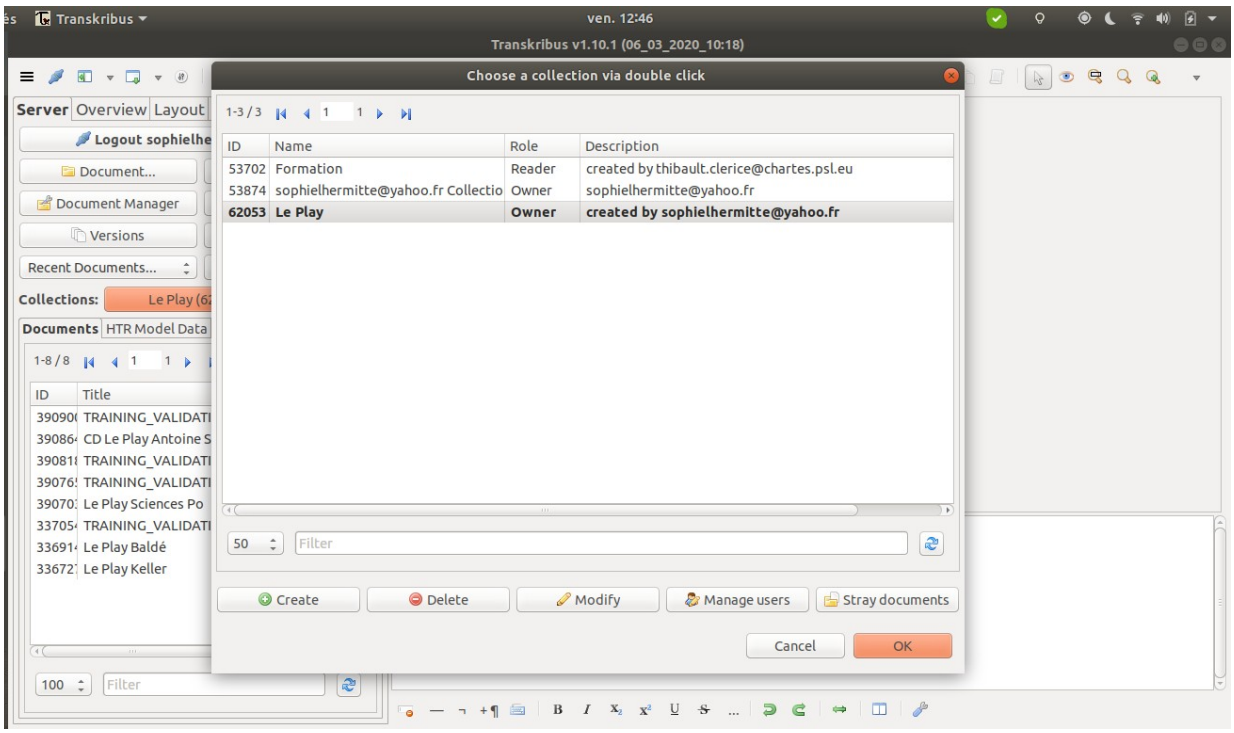
Par Lucie Slavik, stagiaire, en stage filé trois jours par semaine, 4 mai 2020 – 31 juillet 2020
Tutoriel réalisé le vendredi 22 mai 2020
Mis à jour le 8 juin 2020.

Matériel :
Exemple réalisé à partir des numérisations et transcriptions des lettres de Le Play à Mgr Félix Dupanloup (16 lettres transcrites par Clara Forcier et relues par Margaux Faure).

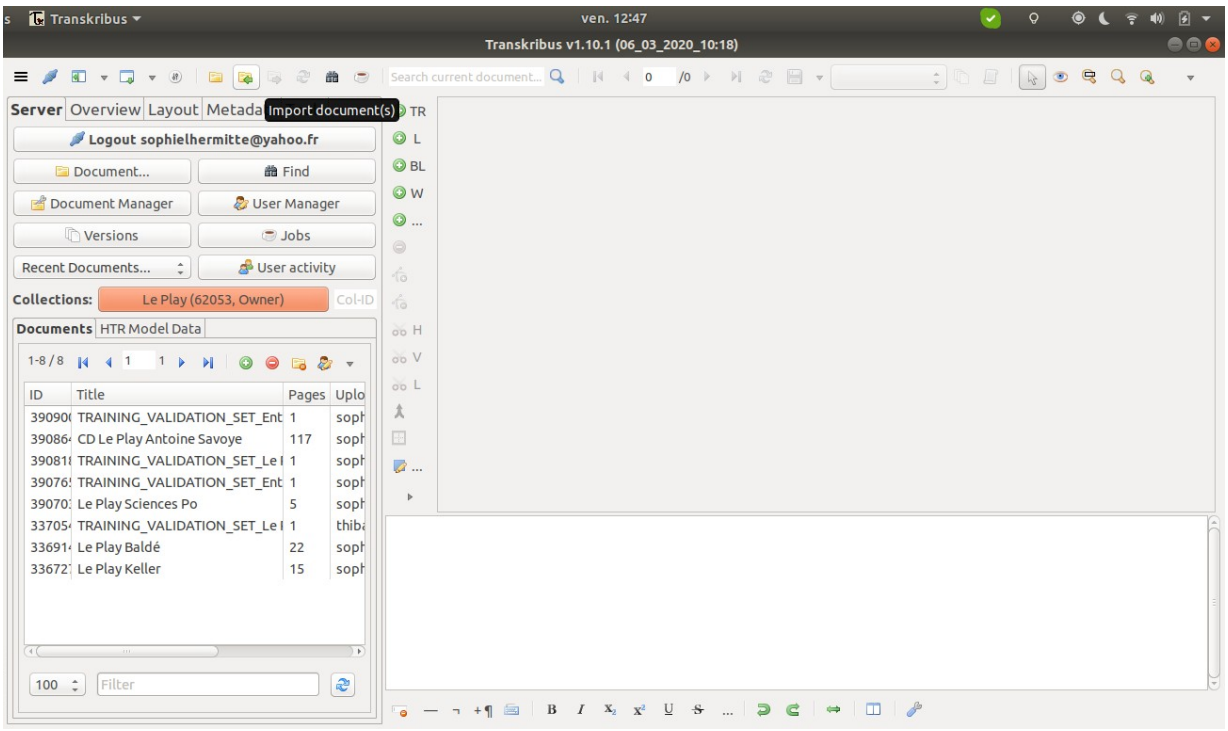
Inspiration :
Tuto de base qui a servi pour la présente adaptation au CRHXIX : Comment_utiliser_Transkribus_–_en_10_étapes_ou_moins_with_Screenshots.pdf, disponible sur internet : http://regis-schlagdenhauffen.eu/wp-content/uploads/2018/01/Comment-utiliser-Transkribus-%E2%80%93-en-10-%C3%A9tapes-ou-moins.pdf

1) Il faut télécharger les manuscrits.

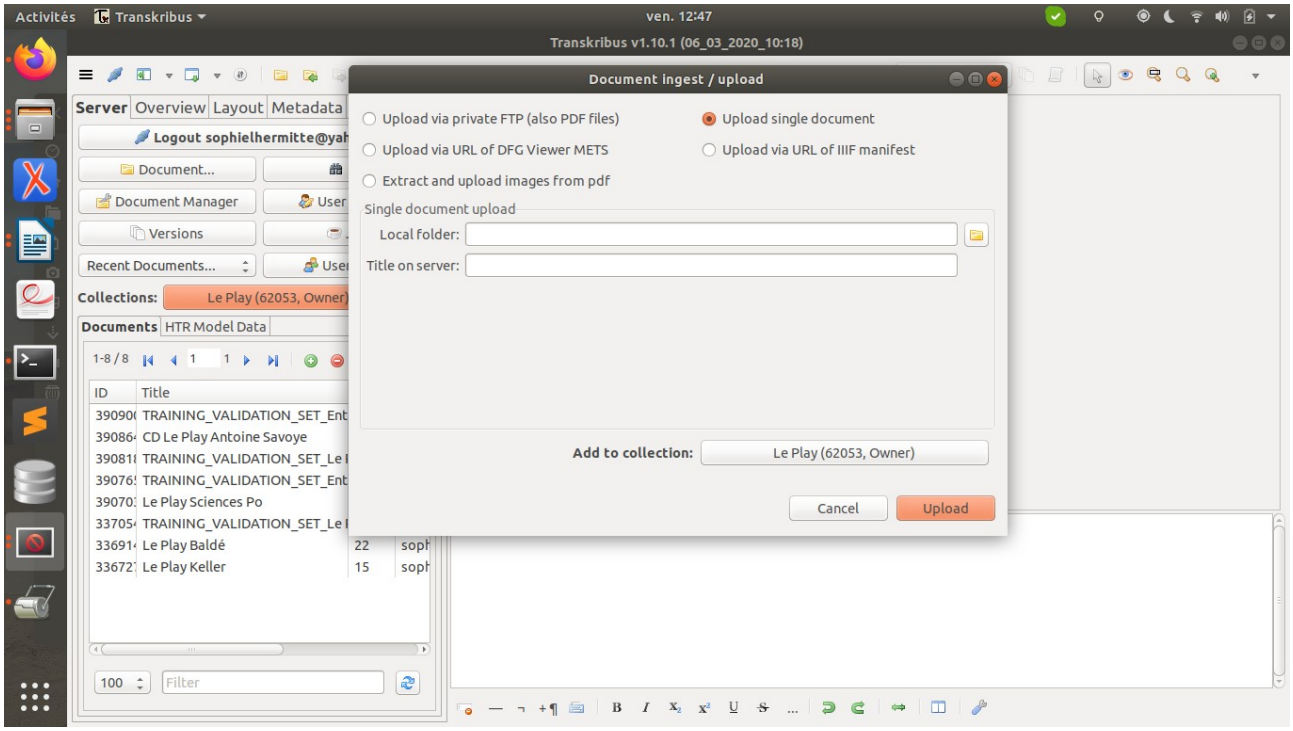
Vérifier qu'on est dans la bonne collection.
Ici, c'est la collection Le Play.



- Cliquer sur l'icône pour télécharger les données en haut à gauche (dossier avec une flèche verte vers la gauche)



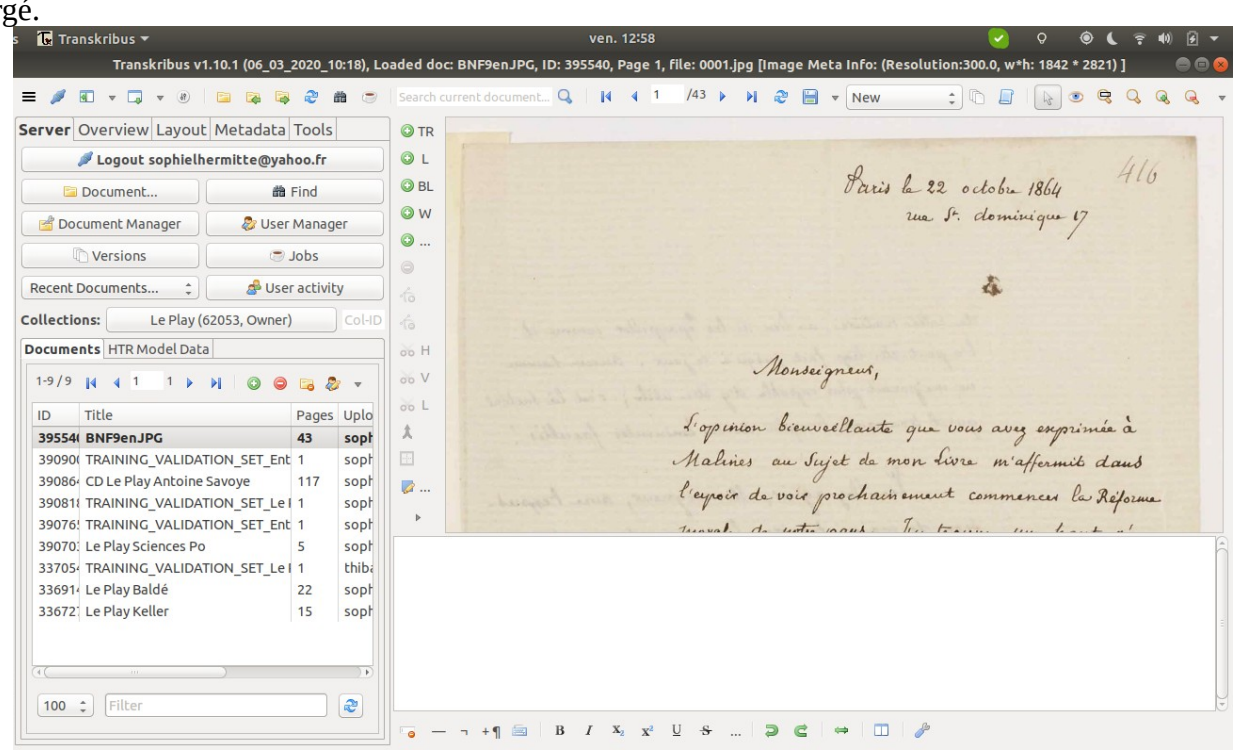
Cela nous ouvre une fenêtre. Cliquer sur le dossier pour télécharger un dossier/des fichiers depuis notre ordinateur. On peut télécharger plusieurs dossiers à la fois.



Valider.
Charger.

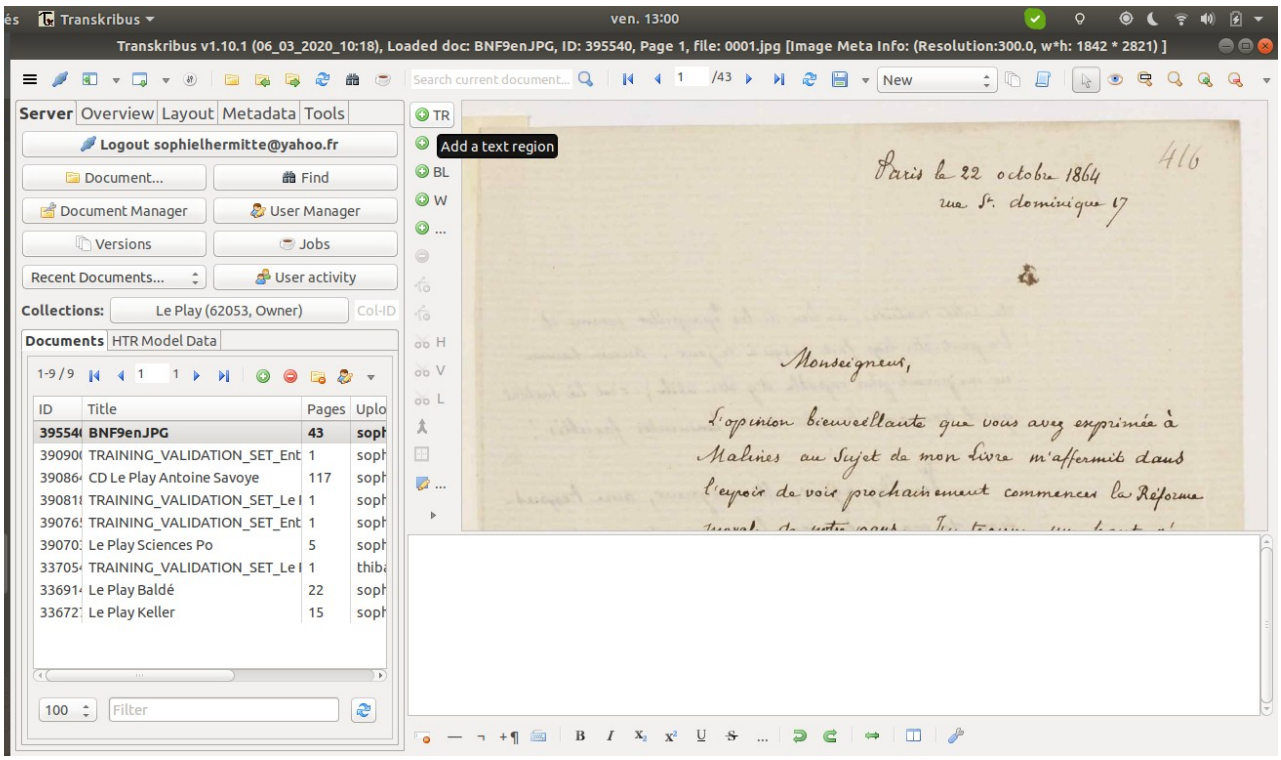
Rafraîchir la page une fois que c’est chargé pour que le dossier téléchargé apparaisse en double-cliquant sur la collection.

Une fois qu'il apparaît (cela peut prendre plus ou moins de temps), double-cliquer sur le dossier chargé.

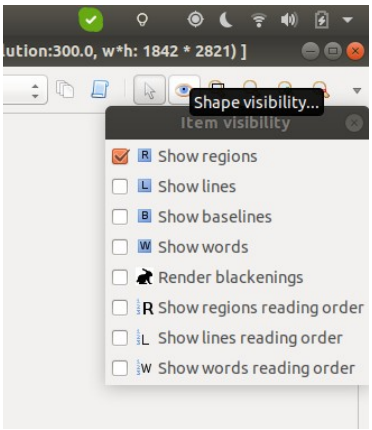


Le dossier chargé (BNF9enJPG) apparaît (il faut attendre un peu le temps que cela charge).

2) Il faut à présent diviser manuellement les régions de texte(TR). Ce n'est pas obligatoire.
Pour cela, cliquer au milieu, en haut, sur le + vert TR : add a text region

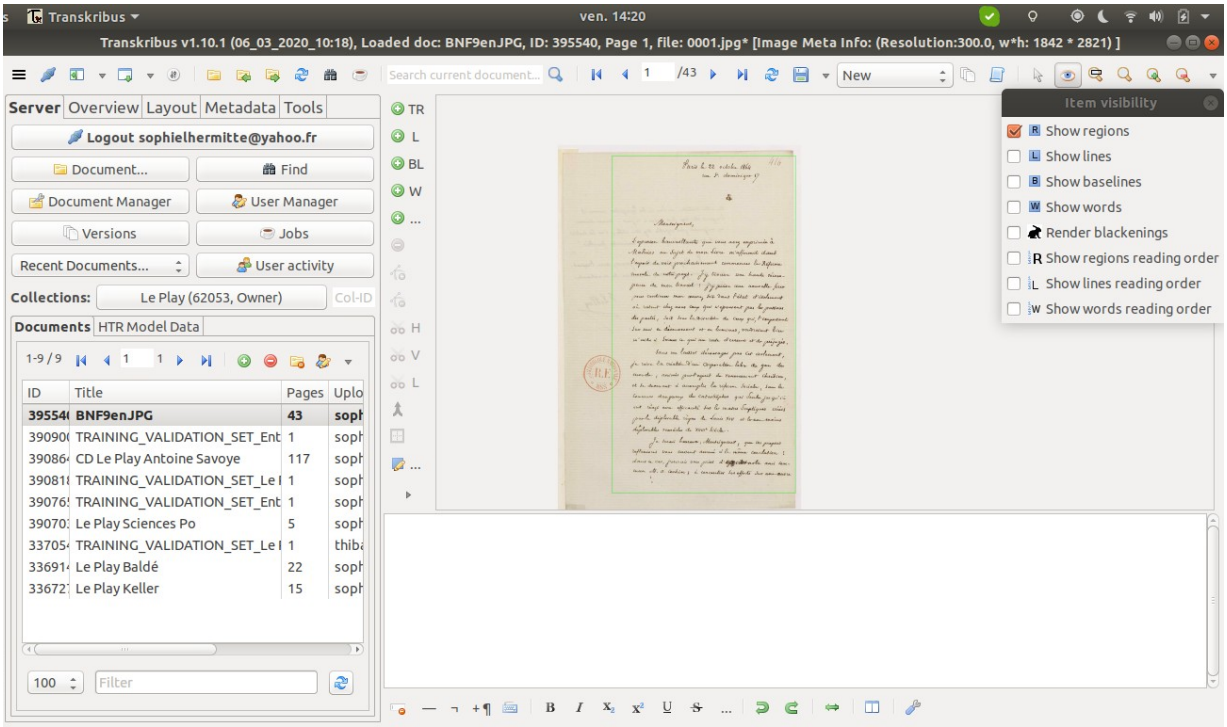


Auparavant, il faut bien vérifier en cliquant sur l'œil en haut à gauche que l'on voit bien les régions de texte.

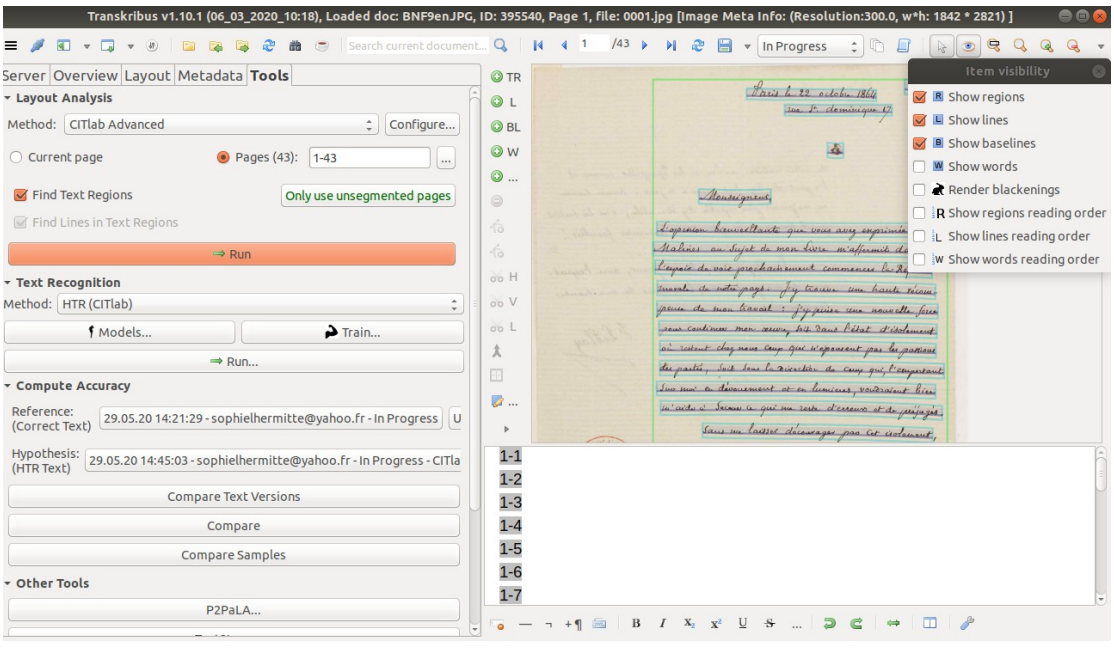


C'est parti !

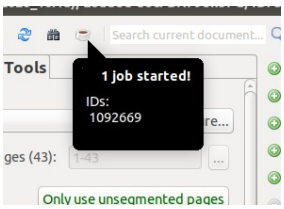
Les régions de textes apparaissent en vert.
Penser à sauvegarder à chaque page (ctrl S) c'est plus sûr.



Aller ensuite dans l'onglet tools, Layout Analysis, pour que la machine reconnaisse elle-même les lignes du texte.
Attention à bien sélectionner toutes les pages.
Cliquer sur Run.
Cela se fait assez rapidement.



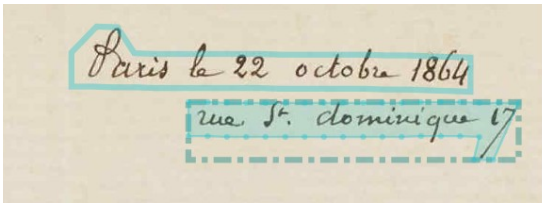
Cliquer sur l'icône café en haut à gauche, pour voir si cela tourne ou si c'est fini. Une fenêtre s'ouvre pour renseigner sur l'avancement du travail.



Type	State	Doc-Ic	Pages	Username	Description	Errors	Created	Started	Finished	ID
Layout analysi	RUNNING	39554(	43	sophielhermit	Running Layou	0	29.05.2020 14:36:	29.05.2020 14:36:		1092669
Create Docum	FINISHED	39554(		sophielhermit	Done, duratio	0	29.05.2020 12:50:	29.05.2020 12:50:	29.05.2020 12:51:	1092299
CITlab Handwi	FINISHED	39090(	1	sophielhermit	Done, duratio	0	22.05.2020 15:46:	22.05.2020 15:46:	22.05.2020 15:46:	1079908
Layout analysi	FINISHED	39086-	1-117	sophielhermit	Done, duratio	0	22.05.2020 15:27:	22.05.2020 15:27:	22.05.2020 15:33:	1079883
Create Docum	FINISHED	39086-		sophielhermit	Done, duratio	0	22.05.2020 14:44:	22.05.2020 14:44:	22.05.2020 14:46:	1079828
CITlab HTR+ Ti	FINISHED	-1		sophielhermit	Done, duratio	0	22.05.2020 14:09:	22.05.2020 14:09:	22.05.2020 15:46:	1079772
CITlab Handwi	FINISHED	39081(	1	sophielhermit	Done, duratio	0	22.05.2020 13:01:	22.05.2020 13:01:	22.05.2020 13:01:	1079688

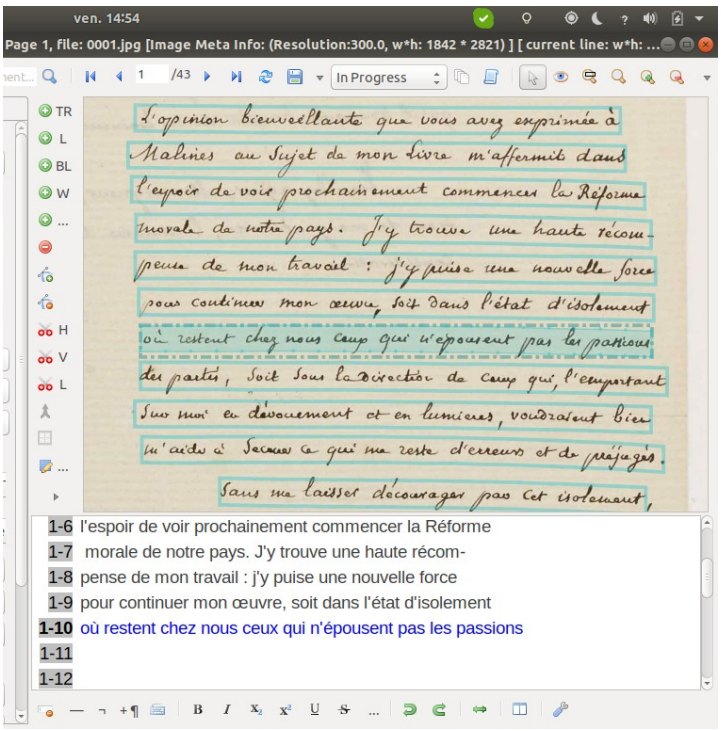
3) Transcrire les lettres

a) S'assurer que la machine a bien fait le travail en divisant le texte en lignes. Le plus important, ce sont les base lines (BL) en violet. (pour voir tous ces détails, bien appuyer sur l'œil en haut à droite).

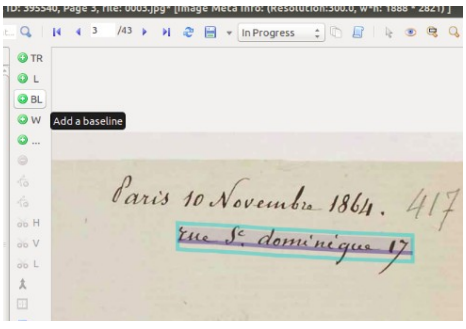


Puis il faut que les lettres soient bien prises en entier, faire attention surtout aux lettres qui montent (comme les « l ») ou qui descendent bas (comme les « p ») et qui souvent ne sont pas bien prises en compte. Ou encore les virgules en fin de ligne.
Il faut supprimer ce qui ne compte pas pour l'entraînement, c'est à dire ce qui n'est pas de la main de l'écrivain que l'on veut faire reconnaître à la machine (l'écriture dactylographiée par exemple, ou les tâches d'encre qui sont parfois prises en compte, ou encore la numérotation qui est souvent faite par le destinataire de la lettre et non son expéditeur). Sélectionner ce qu'on veut supprimer et appuyer sur la touche suppr. de l'ordinateur. Si c'est toute une région de texte, on peut la supprimer directement et toutes les lignes seront de ce fait supprimées.

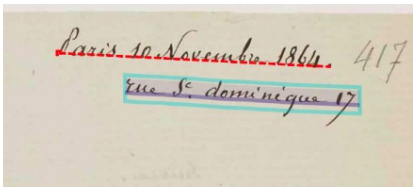
b) Transcrire
A chaque ligne sa transcription.
Copier coller les transcriptions déjà réalisées par les étudiants.
Attentions aux fautes qui se sont glissées.
Essayer de comprendre les mots illisibles.
Quand un mot n'est pas clair, il est possible de le signaler sur Transkribus dans l'onglet metadata : on peut tager « unclear » par exemple.
Il faut également signaler les mots barrés, ou ceux qui comportent des particularités, en utilisant par exemple l'icône en bas, dans les icônes de style.

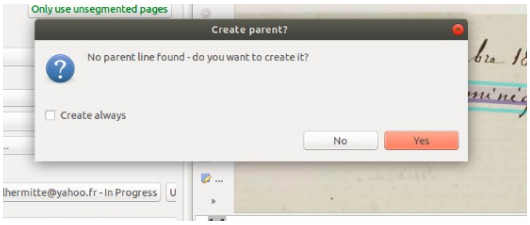


Pour ajouter une ligne si nécessaire, cliquer sur +BL



Cliquer avec la souris à la base de la ligne,
de gauche à droite (et non l'inverse !), et appuyer
sur la touche entrée de l'ordinateur.





Une fenêtre s'ouvre. Cliquer sur OK.
Et le tour est joué !

Ainsi de suite pour chaque dossier à entrer.
L'étape suivante sera la création du modèle avec ces données. Pour cela, voir le Tuto 2.