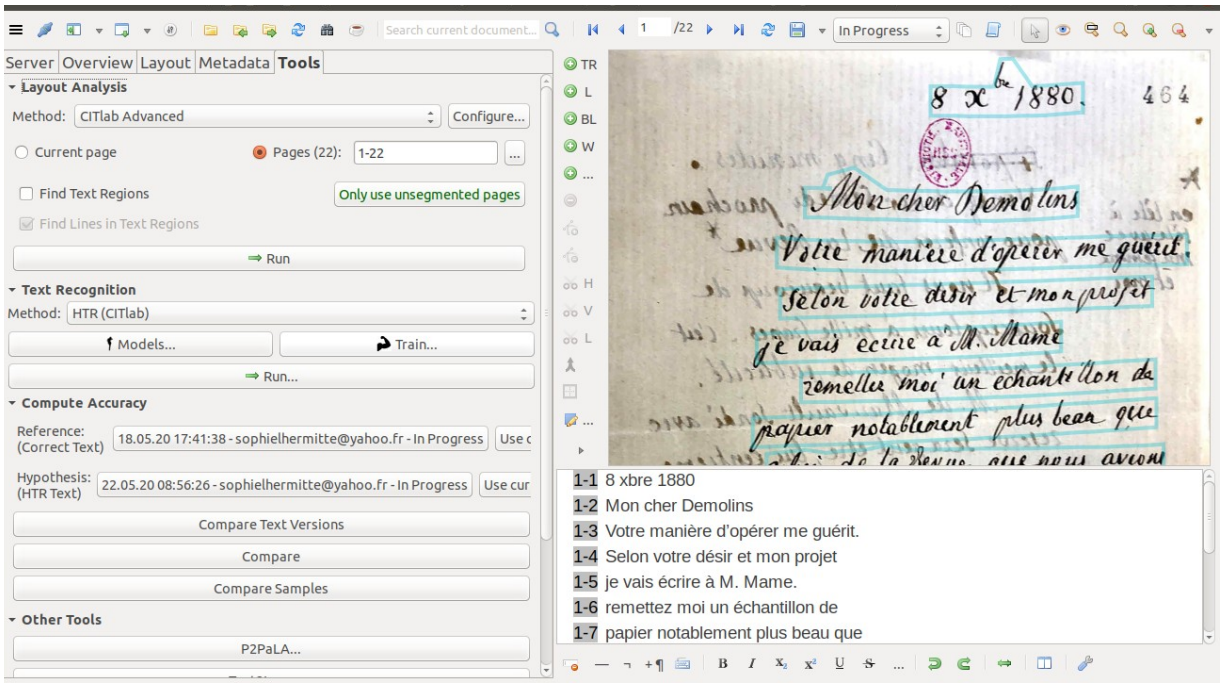
Tutoriel Transkribus
CRHXIX

2ème étape, entraînement d'un modèle

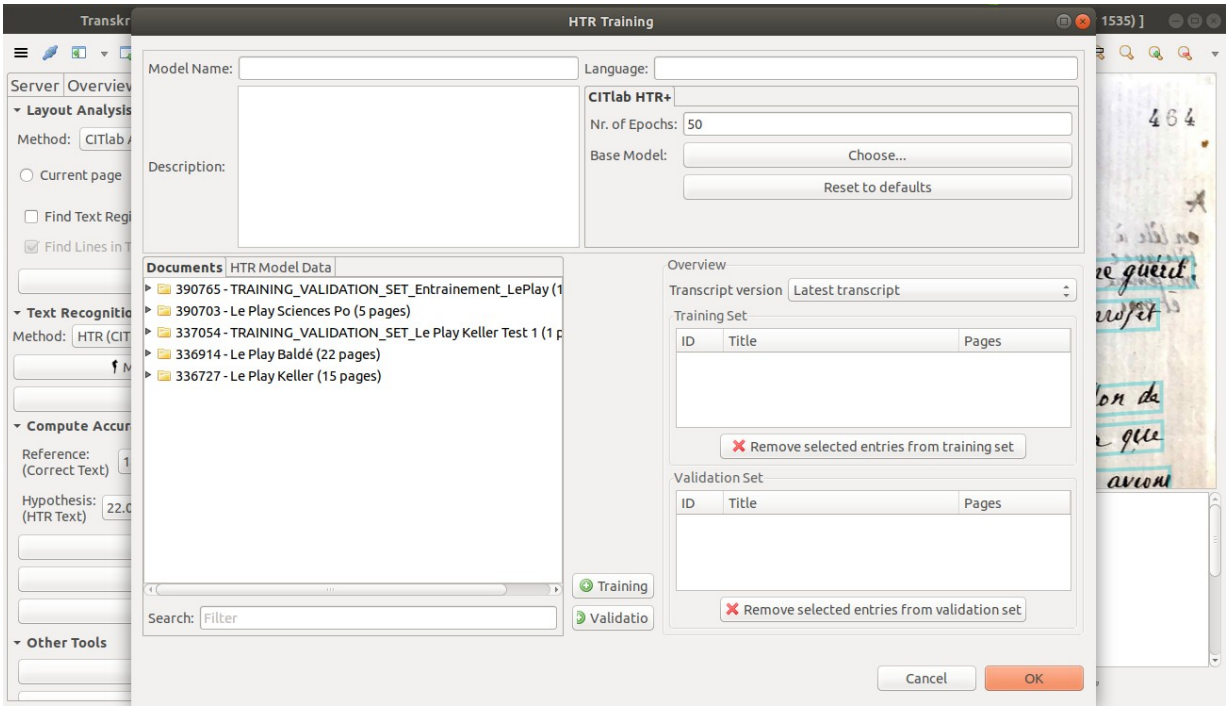
Par Lucie Slavik, stagiaire, en stage filé trois jours par semaine, 4 mai 2020 – 31 juillet 2020
Tutoriel réalisé le vendredi 22 mai 2020
Mis à jour le 8 juin 2020.

Exemple réalisé à partir du le matériel de formation utilisé pour la formation à Transkribus à l'Ecole nationale des Chartes, 6 mars 2020, avec Monsieur Thibault Clérice.
Utilisation du set Baldé.
Tuto de base qui a servi pour la présente adaptation au CRHXIX :
Comment_entraîner_un_Modèle_dans_Transkribus.pdf, disponible sur internet :
https://transkribus.eu/wiki/images/8/84/Comment_entra%C3%AEner_un_Mod%C3%A8le_dans_Transkribus.pdf

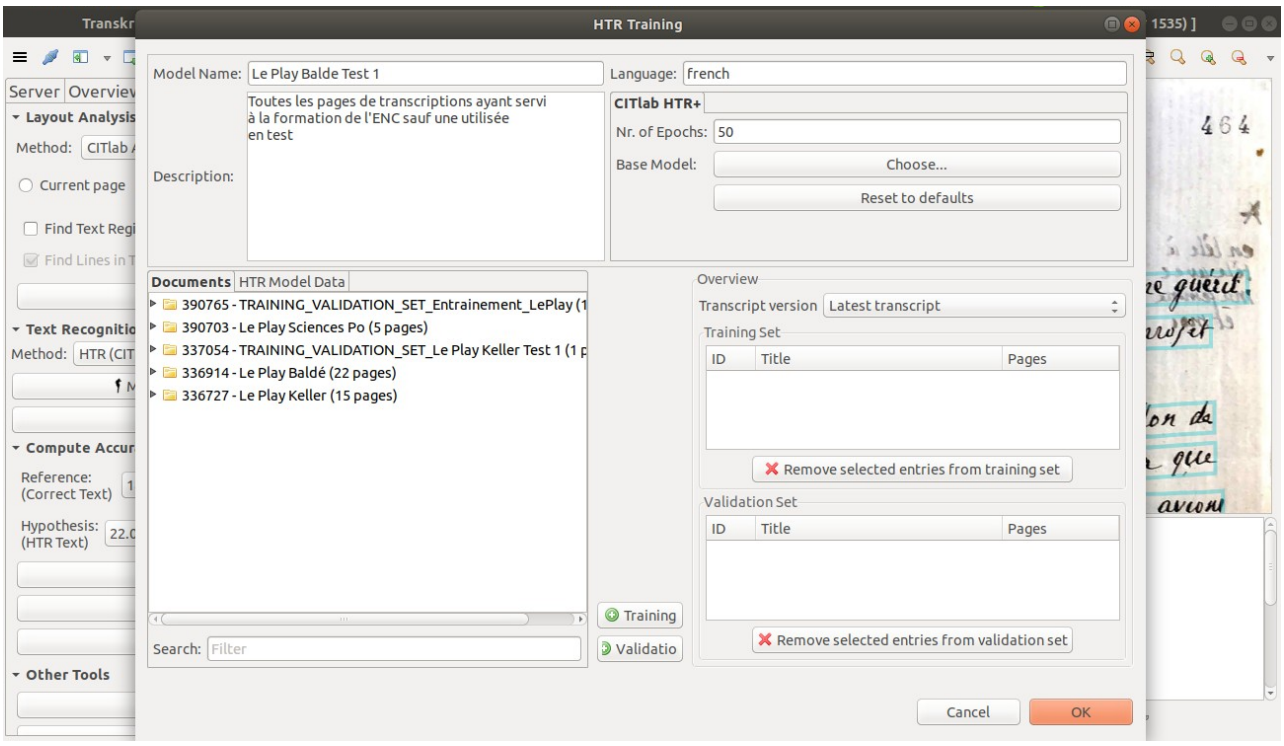
Une fois que les données pour l'entraînement sont prêtes (étape 1) aller dans l'onglet « tools » où se trouvent les outils pour l'entraînement d'un modèle, ici le modèle Le Play.
La sous-partie qui nous intéresse est « Text Recognition ».
Choisir la Méthode : HTR (CITLab)
Cliquer sur « Train ».



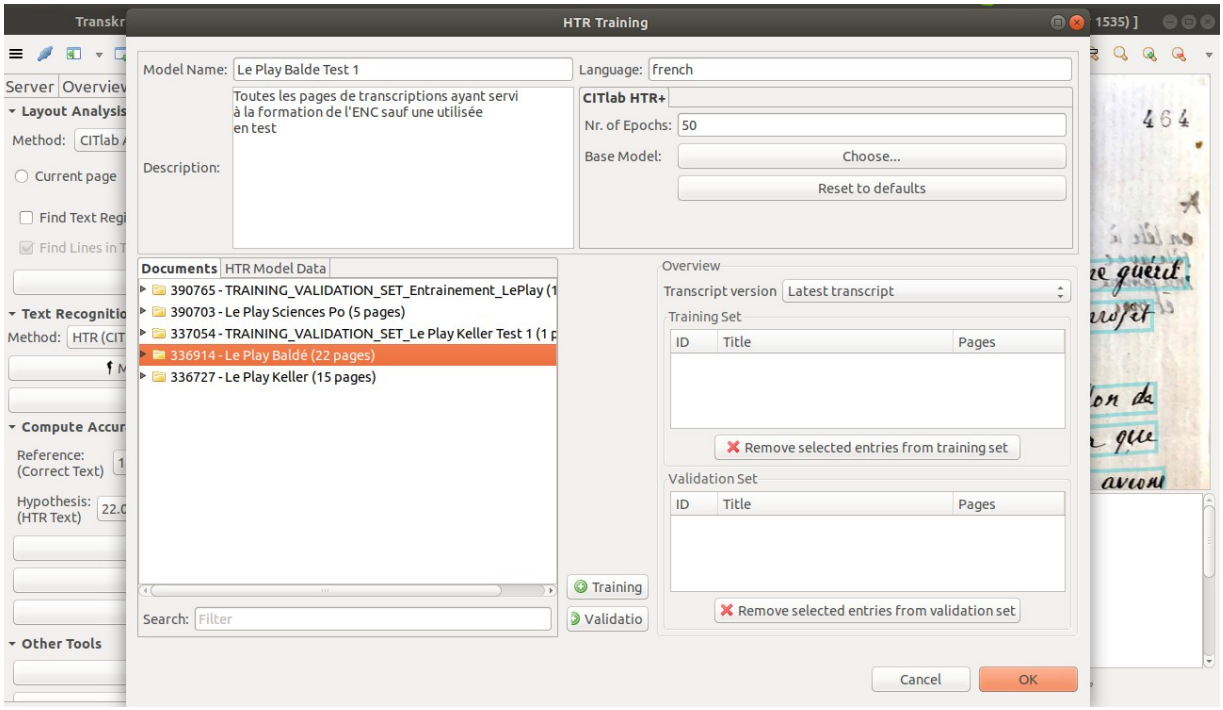
Une fenêtre s’ouvre alors, « HTR Training »



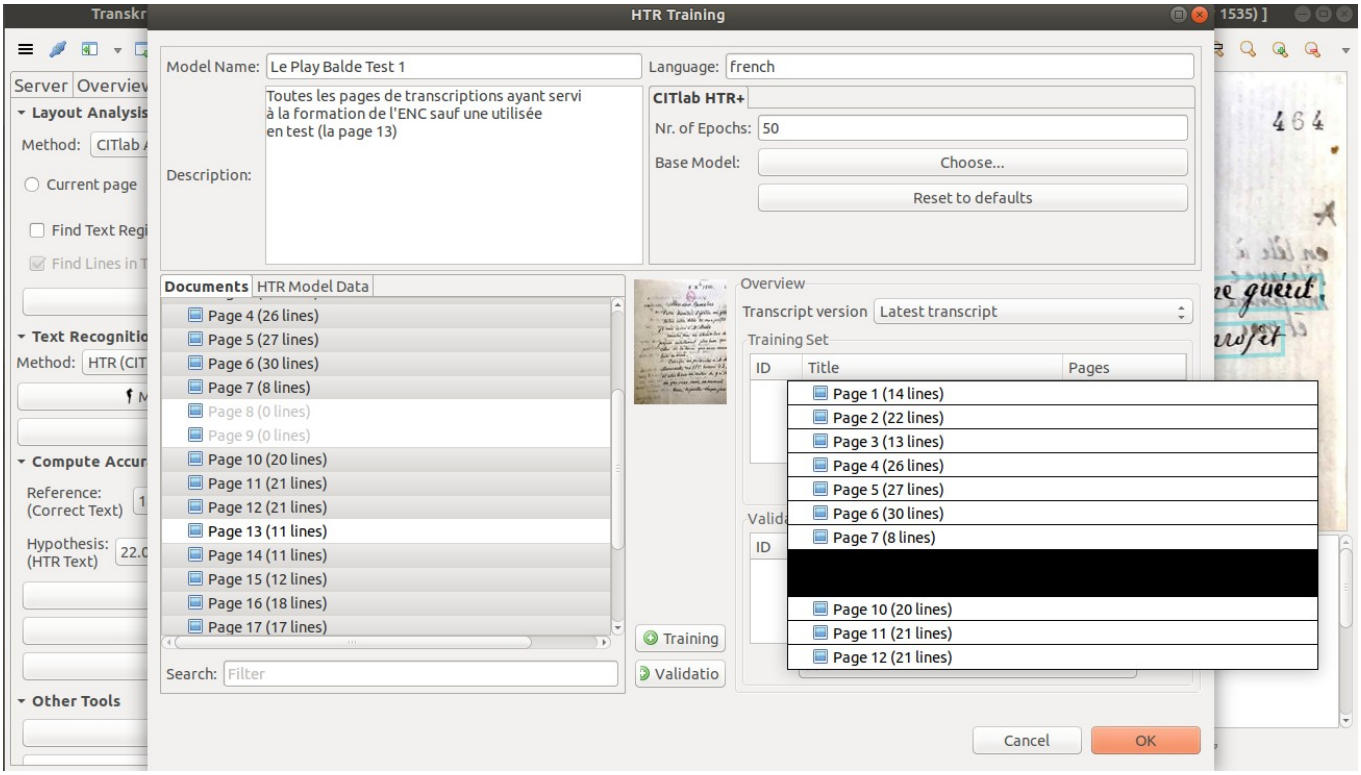
Il faut alors inscrire les modalités de l’entraînement.
Le dossier dont on veut se servir pour l’entraînement du modèle est en l’occurrence celui de 22 pages intitulé 336914 – Le Play Baldé (22 pages)
Dans le Model Name, à gauche, on indique le nom que l’on veut donner à notre entraînement, ici « Le Play Balde Test 1 »
Dans la description… on décrit !
On indique à droite la langue (Le Play écrit en français). (Vérifier à ce propos si c’est bien « french » qu’il faut indiquer, et non fr ou autre…).



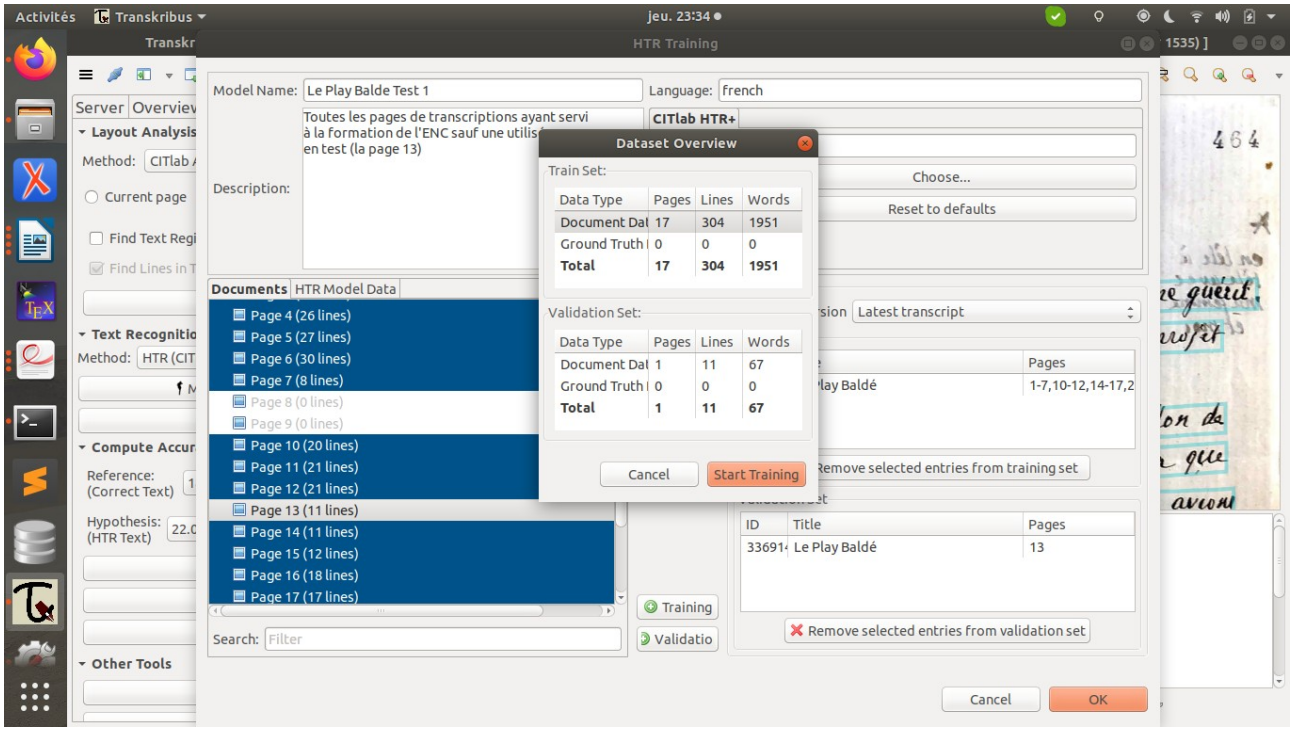
Puis on double-clique sur le dossier qui nous intéresse pour l'ouvrir :



Cela ouvre le document en question et tous les fichiers s'affichent.
On les sélectionne (ctrl + clic gauche) et on glisse l'intégralité des fichiers - 10% à droite (clic gauche en glissant la souris) pour les mettre dans Training set.
Ou on appuie sur le + vert Training si on a une souris qui ne permet pas de le glisser manuellement…



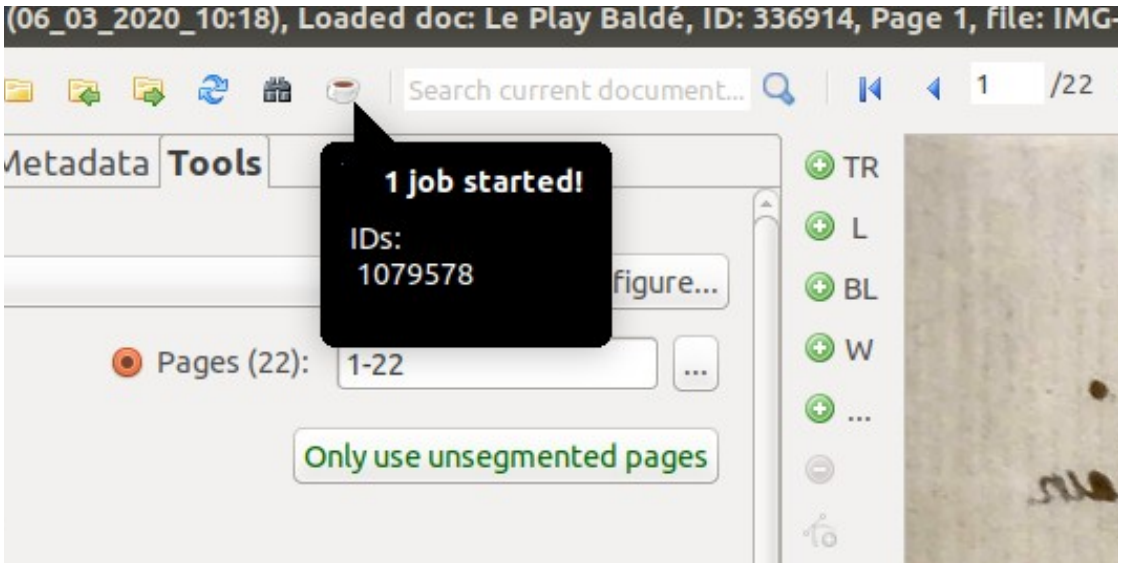
On fait la même chose avec les pages mises de côté (10% normalement, dans l'exemple nous n'en avons mis qu'une, il aurait été mieux d'en mettre deux) et qui se placent cette fois dans le set de validation : il permet de voir avec le set d'entraînement le taux de réussite ou non.
On la sélectionne et pareil, soit on la glisse en bas à droite avec la souris, soit on clique sur le +vert Validation.
On valide le tout en écrivant OK et la fenêtre Dataset Overview apparaît. On clique sur « Start Training »



Remarque : le nombre de mots est indiqué, tant pour le set d'entraînement que pour le set de validation.
Distinguer :

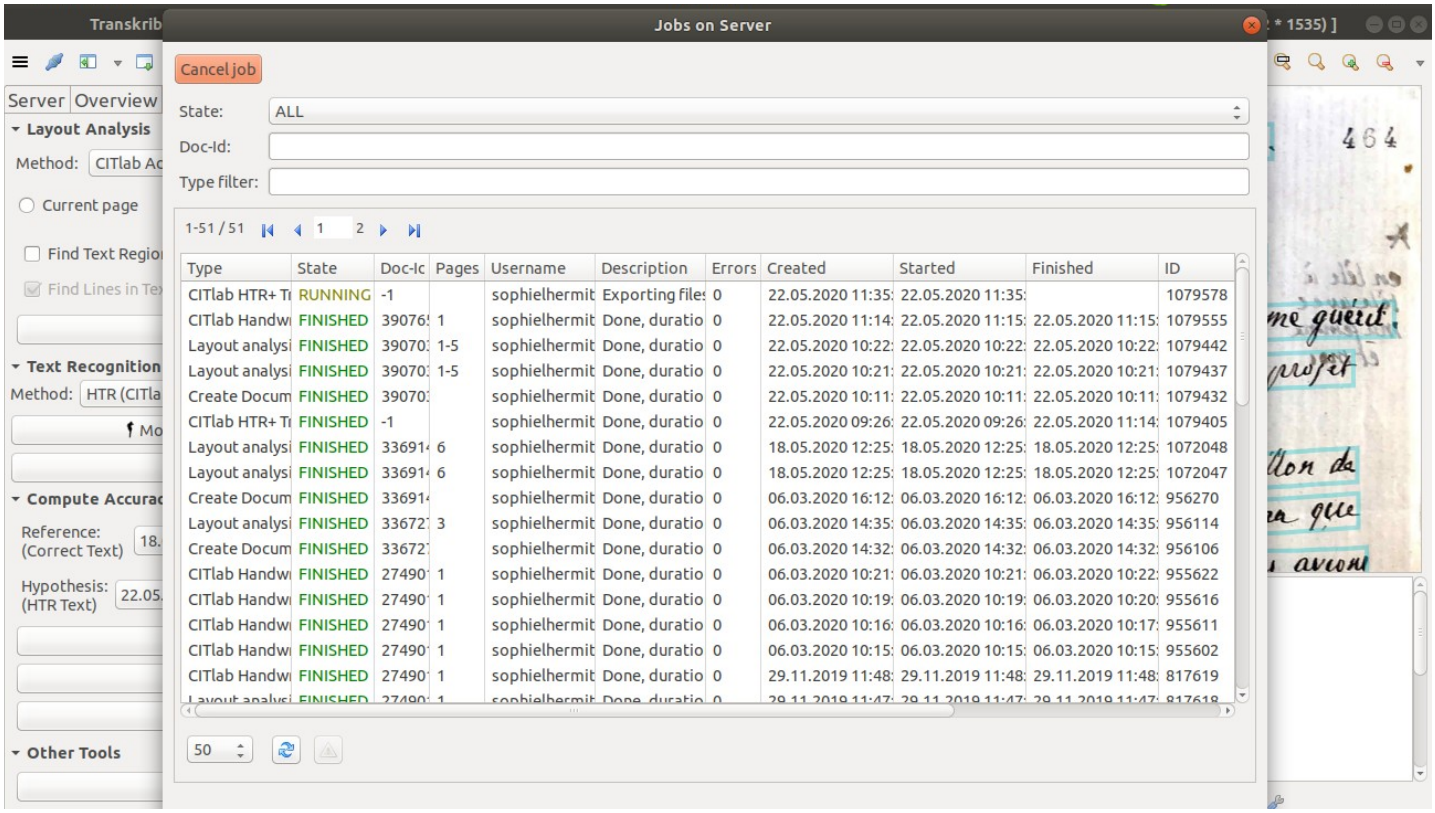
→ Le set d'entraînement ou Train Set, qui sont l'ensemble des données qui ont été chargées pour l'entrainement du modèle. On conseille de charger 20 000 mots pour un résultats plus satisfaisant.

→ Le set de validation ou Validation Set, qui représente 10 % du Train Set et qui compare ce que nous avons transcris avec ce que la machine reconnaît après avoir été entraînée. Il est conseillé de prendre à chaque fois le même set de validation pour se faire une meilleure idée des progrès réalisés par l'apprentissage de la machine.



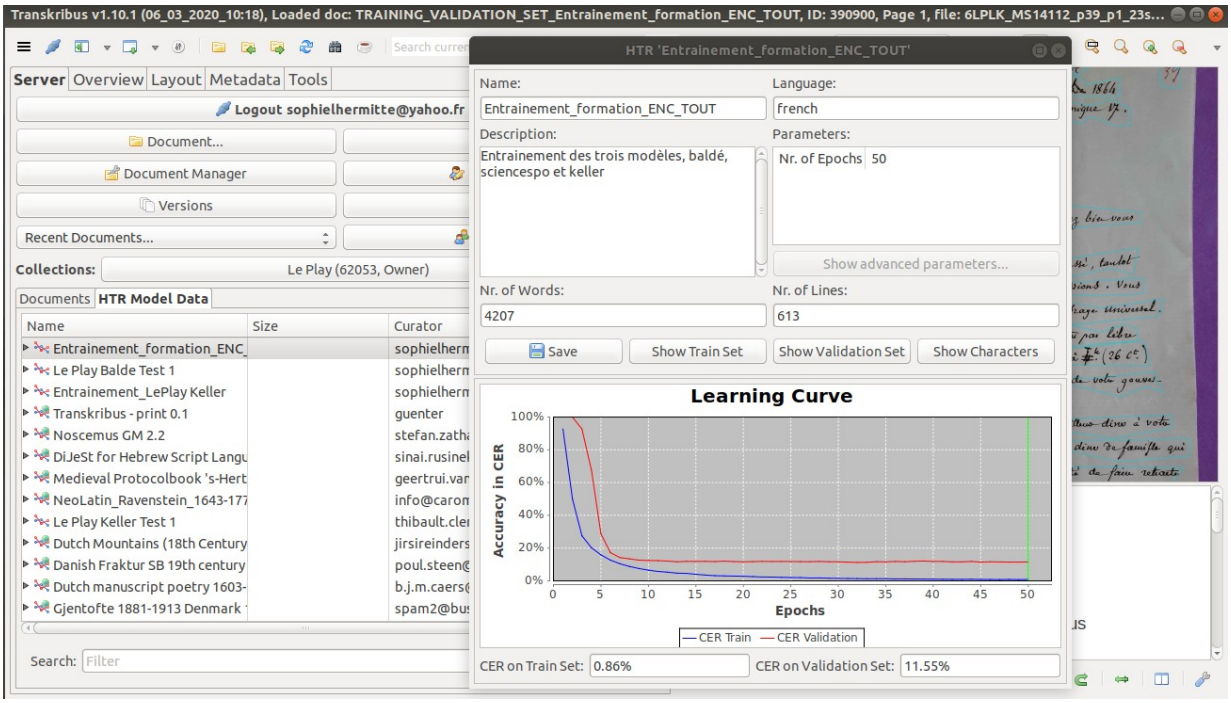
Et c’est parti ! Le set d’entraînement est en train de se créer !

On a le temps de faire une pause café… ou de continuer à travailler sur Transkribus, car tout se passe dans un serveur externe à notre ordinateur.
Si on clique sur l’icône café, une fenêtre s’affiche : on voit que c’est en train de tourner « running »



Pour ces 17 pages d’écriture de Le Play (22 lettres moins 5 : 1 pour le set de validation, 4 pages dactylo non utilisées) cela prend plus d’1h30. Heureusement, on peut continuer de travailler à côté. On peut même fermer Transkribus et il continue de « travailler ».

Après, on peut voir le taux de réussite de notre modèle en cliquant dans Server, HTR model data. Une fois que le modèle a été réalisé, on voit les résultats. (Ici, c’est la capture d’écran d’un modèle postérieur). Pour 4207 mots, on obtient donc environ 99 % de taux de réussite sur le Train Set (soit comme indiqué, 0,86 % d’erreur), et 88 % sur le set de validation (soit 11,55 % d’errreur), ce qui est un score plutôt bon.



Et qui encourage à continuer le chargement des données pour la transcription et l’entraînement d’un modèle plus fourni pour l’obtention d’un meilleur taux de réussite.