READ
Recognition and Enrichment of Archival Documents

Entraînement d'un modèle dans Transkribus

Version v 1.8.0
Dernière mise à jour de ce guide : 24.10.2019 [1]

Ce guide explique comment utiliser Transkribus pour réaliser un modèle de reconnaissance d'écriture manuscrite (modèle HTR+). Après avoir entraîné le modèle, celui-ci peut réaliser des transcriptions automatiques qui vous aideront à effectuer des recherches par mots dans vos documents.

Téléchargez le *Transkribus Expert Client* ou assurez-vous d'utiliser la dernière version :

- https://transkribus.eu/

Visitez le Wiki Transkribus pour plus d'informations et les Guides pratiques :

- https://transkribus.eu/wiki/

Transkribus et la technologie sous-jacente sont mis à disposition à travers les projets et plateformes suivants :

- https://read.transkribus.eu/
- https://transcriptorium.eu/
- https://github.com/transkribus/

Contact :

- L'équipe Transkribus : email@transkribus.eu

[1] Traduction française de *Modell Training in Transkribus* (v.1.8.0) établie par Régis Schlagdenhauffen (EHESS).

Sommaire

Le projet READ est financé par le programme de recherche et d'innovation Horizon 2020 dans le cadre du contrat de subvention n° 674943.

Remarques introductives

- La plate-forme Transkribus permet aux utilisateurs d'entraîner un modèle HTR+ de reconnaissance automatique de documents. Le modèle doit être entraîné pour reconnaître un style d'écriture particulier. Cela se fait en lui "montrant" les images et les transcriptions exactes correspondantes.
- La formation d'un modèle nécessite entre 5 000 et 15 000 mots (environ 25 à 75 pages) de matériel transcrit. Pour les textes imprimés, moins de données d'entraînement sont normalement requises que pour les textes manuscrits.
- La fonction d'entraînement n'est pas incluse dans la version standard de la plate-forme Transkribus. Si vous souhaitez former un modèle, veuillez contacter l'équipe de Transkribus (email@transkribus.eu). Vous aurez alors accès à cette fonction.

Préparation

- Nous vous recommandons de commencer le processus d'entraînement avec 5 000 à 15 000 mots de matériel transcrit, selon que vous travailliez avec des documents imprimés ou manuscrits.
- Le réseau neuronal en arrière-plan apprend rapidement et plus il y a de données d'entraînement, meilleur est le résultat.
- Vous pouvez créer des données d'entraînement pour votre modèle HTR+ en téléchargeant des images et en transcrivant le texte associé aux images. Un manuel complet peut être trouvé ici (en anglais, allemand) : Transcrire avec Transkribus.
- Si vous avez déjà des relevés de notes, vous pouvez les utiliser pour former votre modèle. Pour plus d'informations, consultez ce guide (en anglais, allemand) : Utiliser les transcriptions existantes pour former un modèle.

Entraînement

- Les réglages de base pour la formation d'un modèle se trouvent dans l'onglet **"Tool"** (*outils*) de la zone **"Text Recognition**" (Reconnaissance de texte).
- La "**Méthode HTR** (CITLab)" est actuellement la méthode d'entraînement la plus efficace.
- En cliquant sur le bouton "Modèles", vous pouvez voir quels modèles sont disponibles et avec quels documents ils ont été réalisés.
- Le bouton **"Train"** (*Entraînement*) vous permet d'accéder aux options de formation d'un modèle.

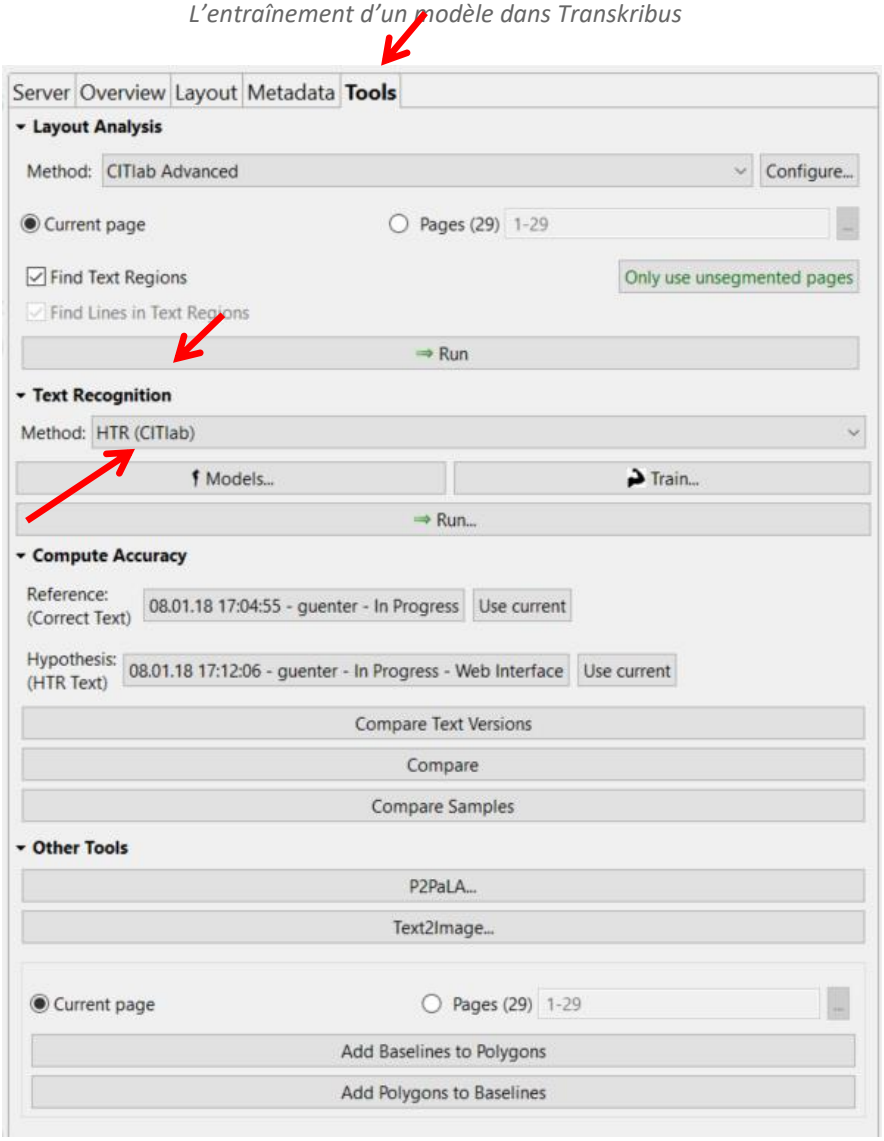

Figure 1 Domaine d'entraînement modèle

Mise en place de l'entraînement HTR+

- Pour accéder à la fenêtre "HTR+ Training", cliquez sur le bouton "Train" dans l'onglet "Tools".

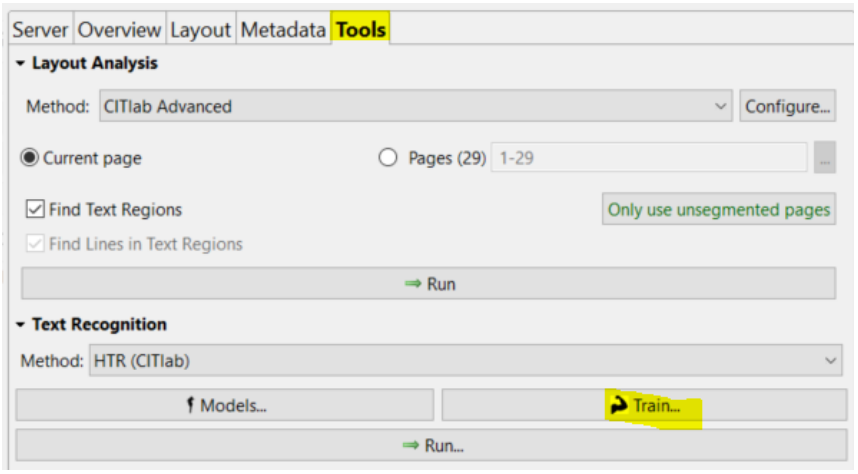

Figure 2 Ouvrir la fenêtre "HTR Training".

- La fenêtre suivante s'ouvre :

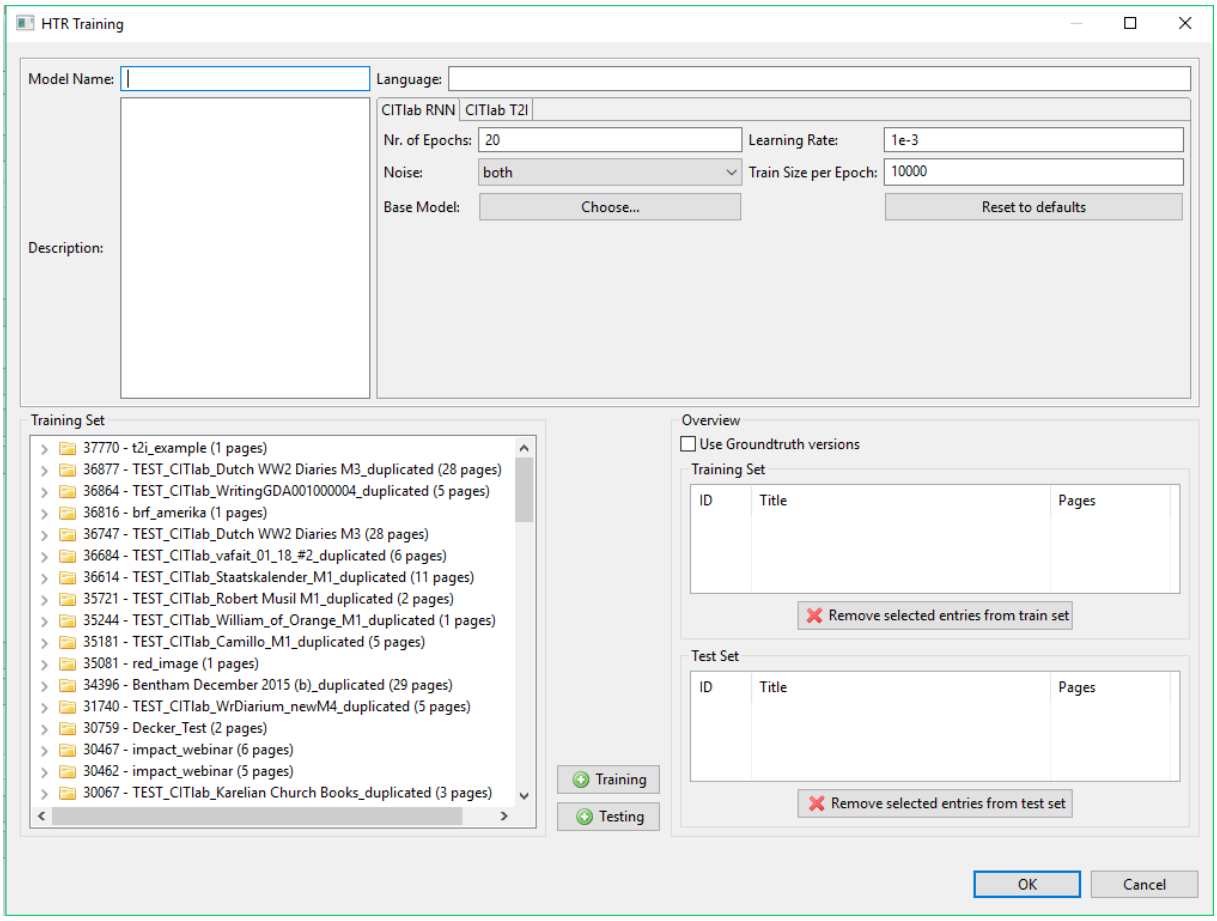

Figure 3 Fenêtre "HTR Training

- Dans la partie supérieure, saisissez les informations relatives à votre modèle.

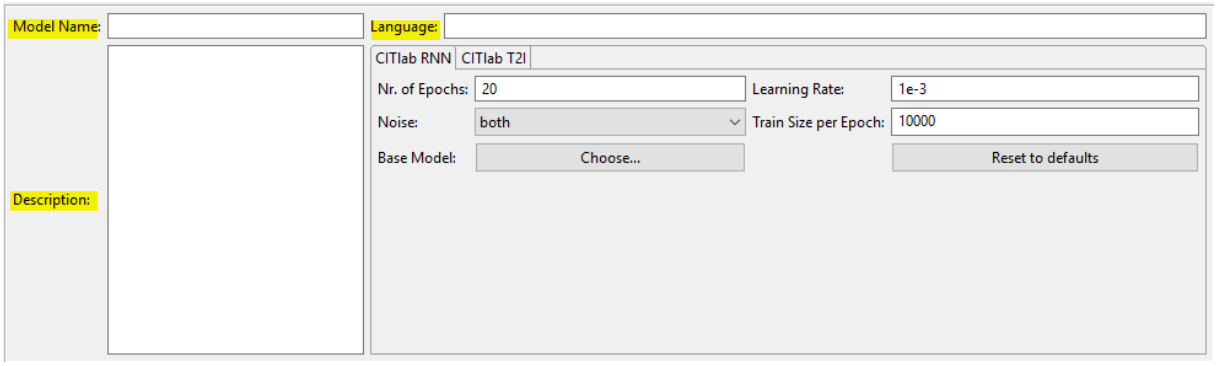

Figure 4 Insérer les détails du modèle

- Veuillez ajouter ce qui suit
    - Nom du modèle (au choix)
    - Langue (l'écriture dans le document)
    - Description (des documents et pages utilisés comme données d'entraînement et d'essai)

- Attention : le "Nombre d'époques" fait référence à la fréquence à laquelle les données de formation sont évaluées. Si vous augmentez le nombre d'époques, le processus de formation sera plus long ...

Modèle de base

- Si vous ne voulez pas partir de zéro et qu'il existe déjà un modèle sur lequel vous pouvez construire votre entraînement, Transkribus vous permet d'ajouter un modèle de base à votre entraînement. Les données contenues dans le modèle de base sont ainsi formées pour devenir le nouveau modèle.
- Pour bénéficier de cet avantage, les données du modèle de base doivent être similaires à celles du nouveau modèle.
- A l'aide d'un modèle de base, le processus d'entraînement peut être accéléré. Une amélioration doit être testée sur une base individuelle et ne peut pas toujours être garantie.
- Un grand avantage de l'utilisation d'un modèle de base réside dans le fait que vous pourrez commencer avec une plus petite quantité de données de formation et réduire ainsi le temps nécessaire à la transcription manuelle.
- Pour utiliser un modèle de base, sélectionnez le modèle souhaité avec "Choose..." à côté de "Base Model".

Set d'entraînement

- Ensuite, sélectionnez les pages que vous voulez utiliser comme ensemble de données d'entraînement.
- Pour ajouter toutes les pages de votre document au jeu d'entraînement, cliquez sur le dossier puis sur "**+Testing**".
- Pour ajouter une série de pages de votre document au jeu d'entraînement, double-cliquez sur le dossier, puis cliquez sur la première page que vous voulez ajouter, maintenez la touche Maj enfoncée et cliquez sur la dernière page que vous voulez ajouter. Cliquez ensuite sur "+Formation".
- Pour ajouter des pages individuelles de votre document au jeu d'entraînement, double-cliquez sur le dossier, puis maintenez la touche CTRL enfoncée et cliquez sur les pages que vous voulez utiliser comme données de formation. Lorsque vous les avez toutes sélectionnées, cliquez sur "+Formation".
- Les pages sélectionnées apparaîtront alors dans la zone "Training Set".

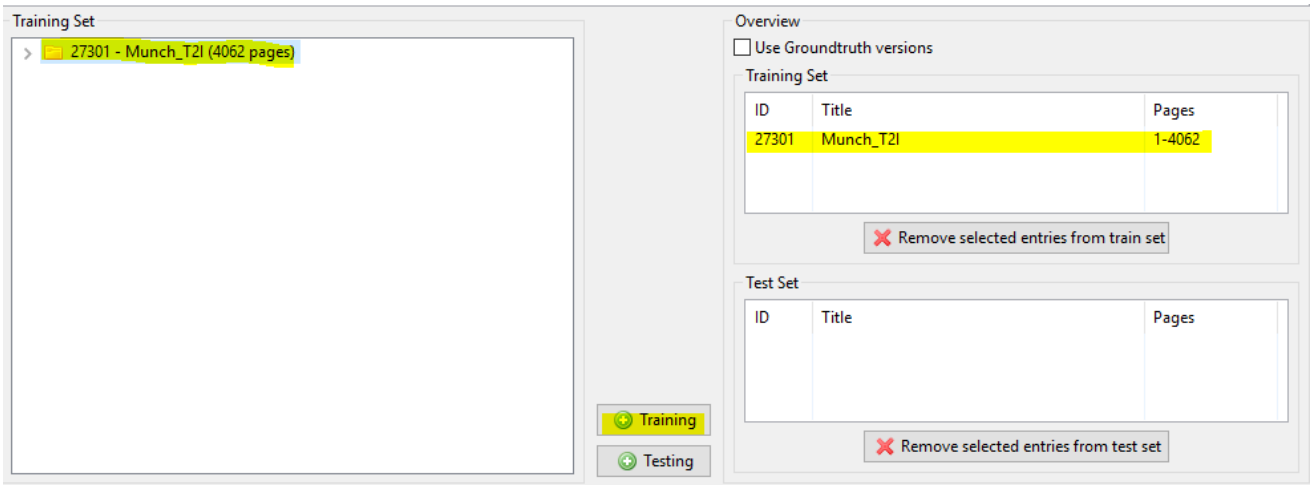

Figure 5 Ajout de toutes les pages au jeu d'apprentissage

Test

- Pendant le processus de formation, quelques pages sont mises de côté à titre de test. Elles ne sont pas utilisées pour l'entraînement du modèle HTR+. Elles sont plutôt utilisées pour tester les performances de votre modèle.
- Nous recommandons au moins une page d'ensemble de test pour chaque sous-ensemble de 50-100 pages de l'ensemble de formation.
- Les pages de votre ensemble de tests doivent refléter le style des pages de l'ensemble de formation.
- Plus votre jeu de tests contient de pages, plus l'entraînement sera long.
- Pour ajouter des pages à l'ensemble de test, suivez la même procédure que pour l'ensemble d'apprentissage, mais cliquez sur le bouton +Test.

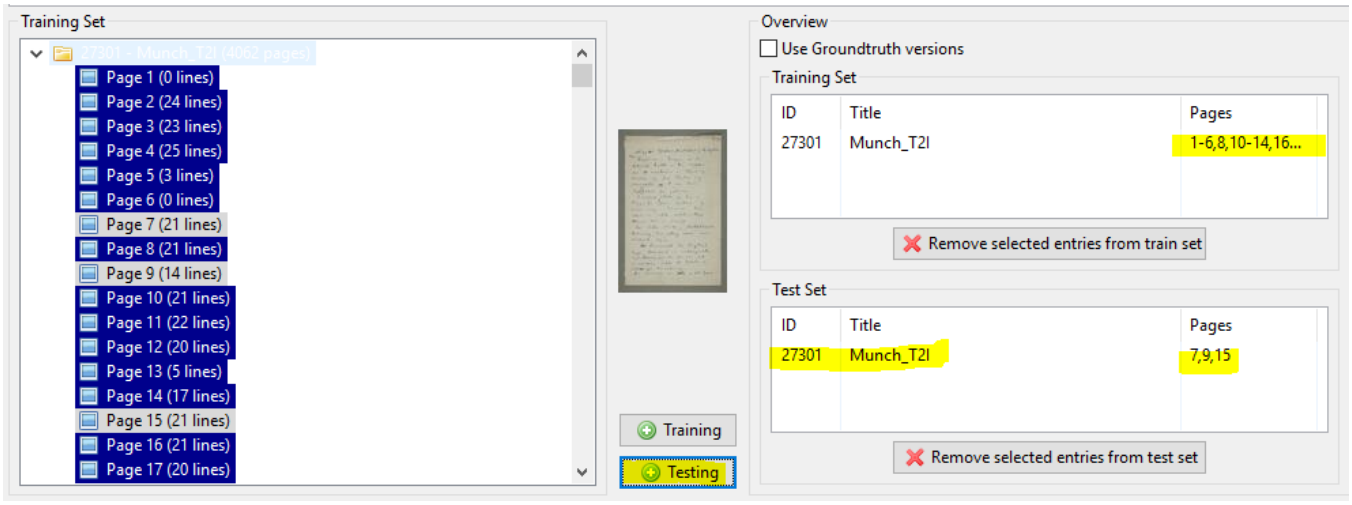

Figure 6 Ajouter des pages au jeu de test

- Pour retirer des pages du "Training Set" ou du "Test Set", cliquez sur la page et ensuite sur le "X" rouge.

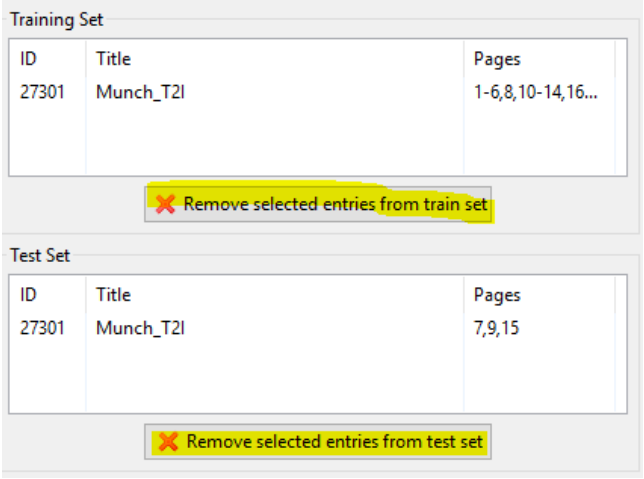

Figure 7 Supprimer des pages

- Vous pouvez lister les pages utilisées dans le jeu de test dans la description du modèle.
- Vous pouvez commencer la formation en cliquant sur "OK".

Suivre les progrès

- Vous pouvez suivre le déroulement de l‘entraînement en cliquant sur le bouton "Jobs" dans l'onglet "Server".

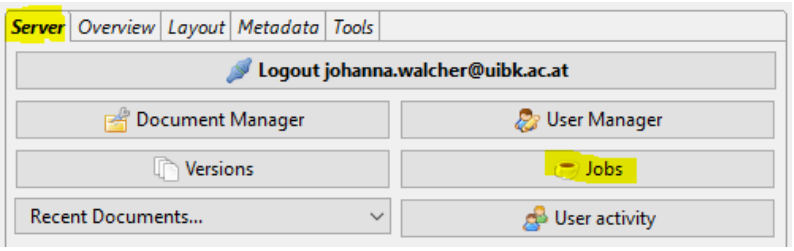

Figure 8 Contrôler le déroulement de l'entraînement à l'aide de la touche "Jobs“

- La fin de chaque époque ainsi que la fin de l‘entraînement est affichée dans la fenêtre "jobs".
- La formation d'un modèle HTR+ prend un certain temps. Vous pouvez faire d'autres travaux dans Transkribus ou fermer la plateforme sans que cela n‘interrompe l‘entraînement.

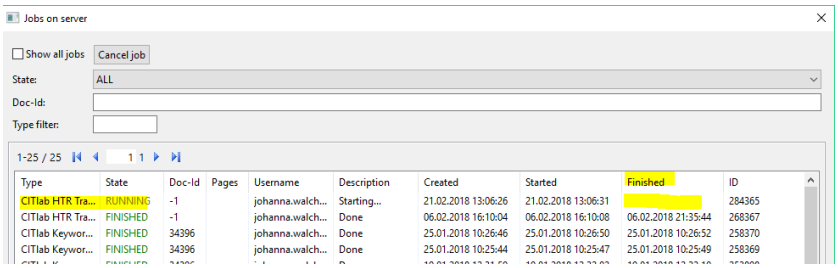

Figure 9 "Emplois sur le serveur" Vue d'ensemble

Après l’entraînement

- Une fois l‘entraînement terminé, le modèle est disponible dans votre collection.

- Pour appeler le modèle, cliquez sur "Models" dans l'onglet "Tools".

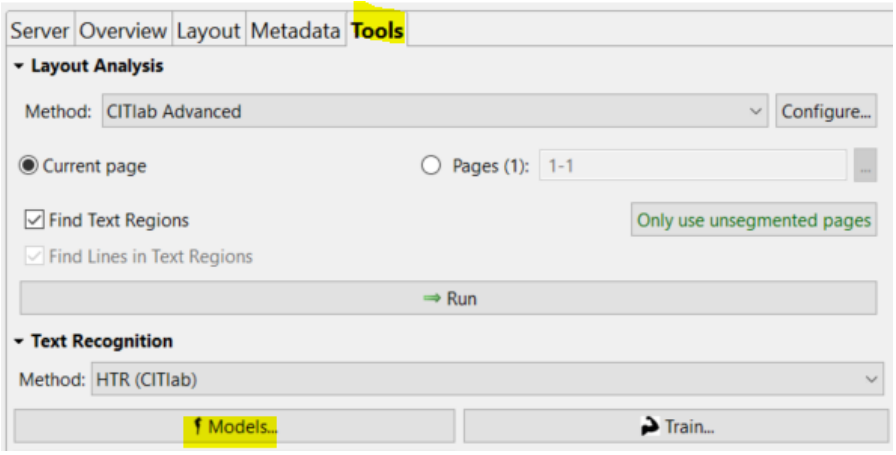

Figure 10 Ouvrir la fenêtre "Choisir un modèle

- La fenêtre suivante s'ouvre :

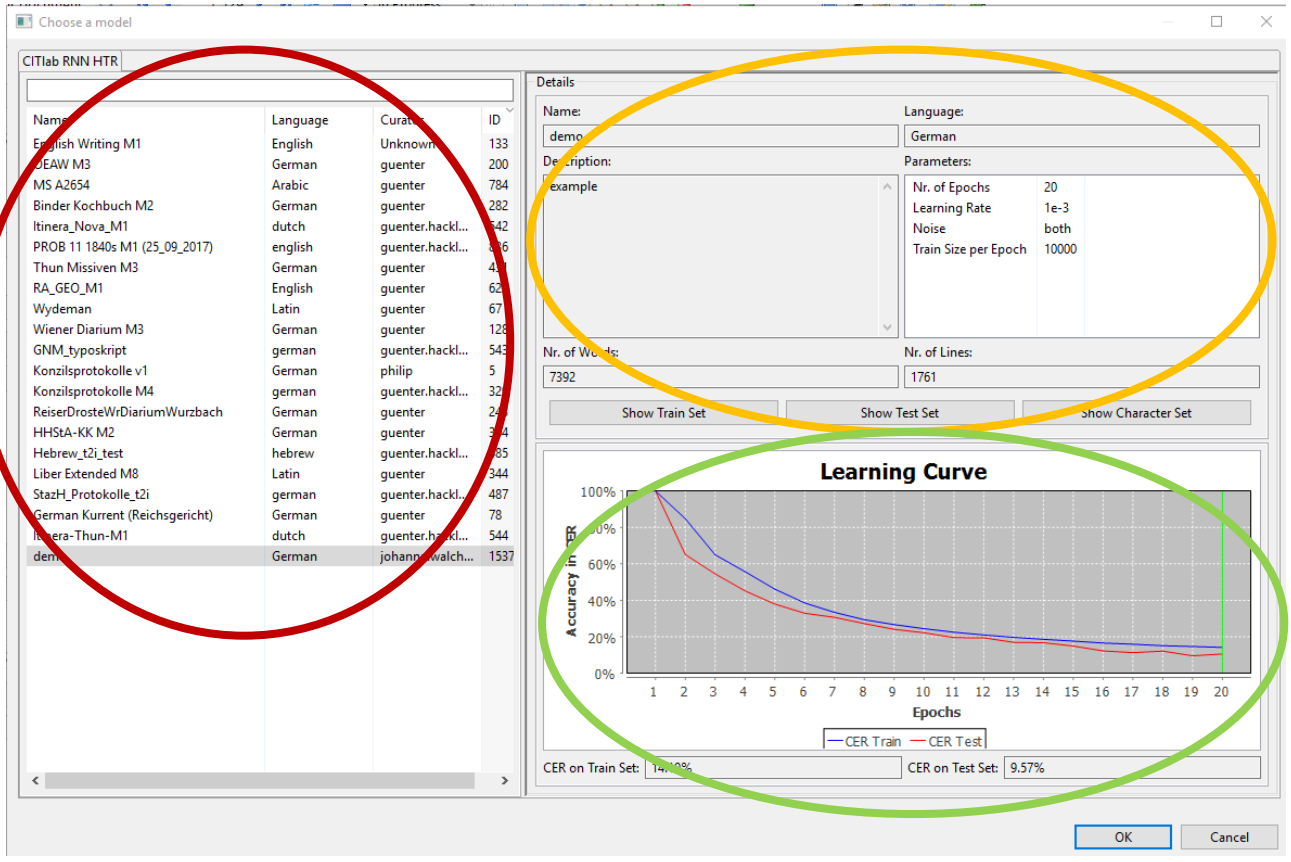

Figure 11 Fenêtre "Choisir un modèle

- Sur le côté gauche, vous voyez un aperçu de tous les modèles disponibles.
- En haut à droite vous pouvez voir les détails du modèle.
- En bas à droite vous pouvez voir la courbe d'apprentissage du modèle. Vous trouverez plus d'informations sur ces statistiques ci-dessous.

Statistique

- Le graphique de la courbe d'apprentissage illustre la précision de votre modèle.

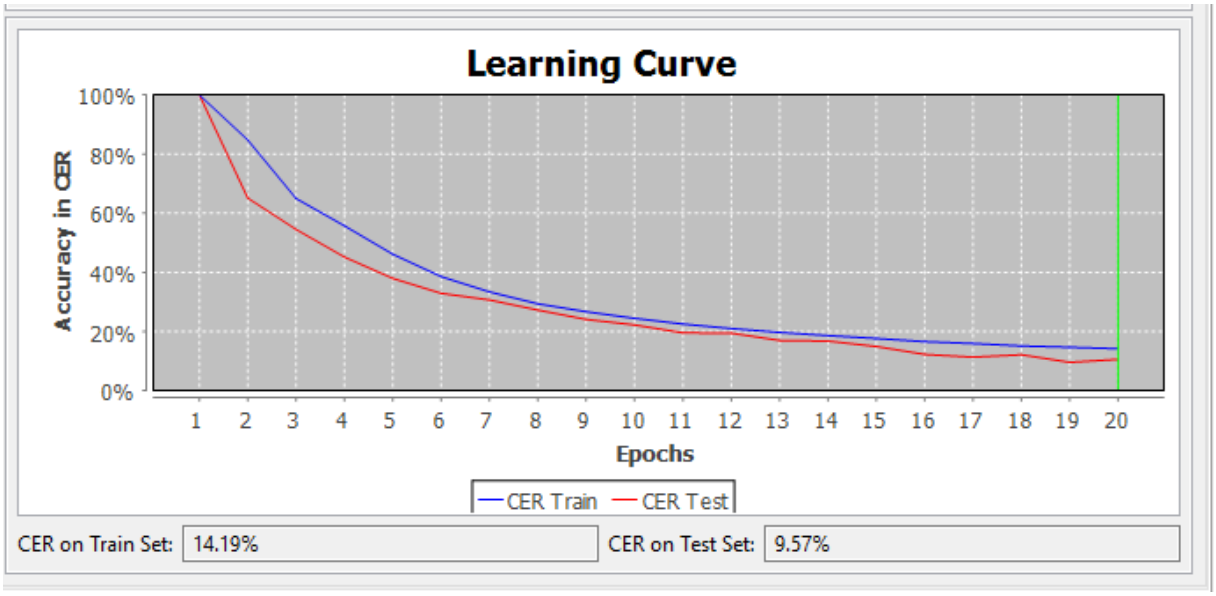

Figure 12 "Courbe d'apprentissage" de votre modèle

- L'axe Y définit la "Précision en CER" (voir figure 12).
- "CER" signifie **Character Error Rate**, c'est-à-dire le taux (en %) de caractères qui n'ont pas été correctement transcrits par l'HTR.
- **La"Précision en CER**" est affichée en pourcentage sur l'axe des ordonnées. La courbe commence toujours à 100 % et descend au fur et à mesure que l'entraînement progresse et que le modèle s'améliore.

- L'axe X est défini comme "**époques**".
- Pendant le processus de formation, Transkribus effectue une évaluation après chaque époque. Dans la figure 12, le "Training Set" était divisé en 20 époques.
- Lorsque vous entraînez un modèle, vous pouvez spécifier le nombre d'époques dans lesquelles le Training Set doit être divisé. Plus il y a d'époques, plus la formation dure longtemps.

- Le **graphique** montre une ligne rouge et une ligne bleue.
- La **ligne bleue** indique la progression de l'entraînement.
- La **ligne rouge** indique l'état d'avancement des évaluations dans le jeu de tests.
- Le programme s'entraîne d'abord dans le **Training Set**, puis se teste à l'aide des pages du **Test Set**.

- Au-dessous du graphique se trouvent deux pourcentages qui se réfèrent aux taux d'erreurs de l'ensemble d'apprentissage et de l'ensemble de test.
- Sur la figure 12, le modèle a un taux d'erreur (CER) de 14,19 % pour le Training Set et de 9,57 % pour le Test Set.
- La valeur de l'ensemble de test est significative car elle montre comment le modèle se comporte sur les pages où il n'a pas été formé.
- Les résultats avec un taux d'erreur (CER) de 10 % ou moins peuvent être utilisés pour les transcriptions automatisées.
- Les résultats avec un taux d'erreur de 20 à 30 % sont suffisants pour travailler avec la recherche de mots-clés. Pour plus d'informations, cliquez ici : Recherche dans les documents avec repérage de mots-clés (en anglais, allemand).

Créer des transcriptions HTR

- Avec votre modèle, vous pouvez maintenant générer automatiquement des transcriptions de documents de votre collection.
- Téléchargez d'abord votre document sur Transkribus.
- Segmentez ensuite votre document en zones de texte, lignes et lignes de base.
- Pour plus d'informations sur le téléchargement et la segmentation, veuillez consulter le guide Transcription avec Transkribus.
- Pour accéder à votre modèle, cliquez sur l'onglet Outils et allez dans la section Reconnaissance de texte.
- Cliquez sur "Exécuter" puis sur "Configurer". Sélectionnez votre modèle HTR dans la liste à gauche de l'écran et cliquez sur OK.
- Indiquez si vous voulez créer une transcription HTR à partir d'une ou plusieurs pages.
- Cliquez sur "Run" (Exécuter) pour lancer le processus de reconnaissance de texte.
- Dès que la reconnaissance de texte est terminée, les pages doivent être rechargées et la transcription automatique apparaît dans la zone de texte.

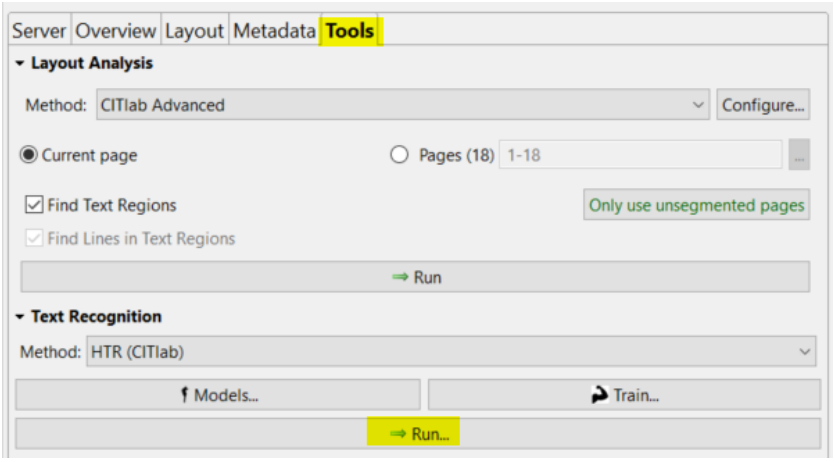

Figure 13 Démarrage du modèle

Partager un modèle

- Vous pouvez partager votre modèle HTR avec d'autres collections de Transkribus, mais aussi avec les vôtres ainsi qu'avec d'autres utilisateurs.
- Si vous souhaitez partager votre modèle avec une autre collection, vous devez avoir accès à cette collection.
- Cliquez avec le bouton droit de la souris sur le nom de votre modèle, à gauche de la fenêtre "Choose a model" ("Choisir un modèle").

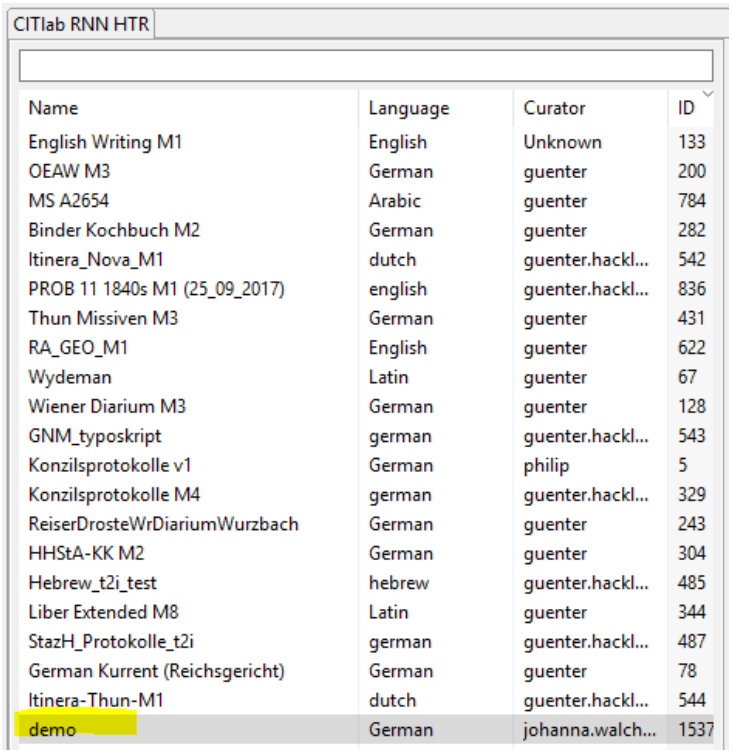

Figure 16 Fractionner un modèle en cliquant avec le bouton droit de la souris sur le modèle

- Puis sélectionnez "Share model…" ("*Partager le modèle*")
- La fenêtre "Choose a collection via double click" (*Choisir une collection par double clic*) s'ouvre.
- Dans la fenêtre suivante, cliquez sur la collection avec laquelle vous souhaitez partager le modèle, puis cliquez sur OK.
- Dans cette fenêtre, vous pouvez également créer une nouvelle collection pour le modèle en utilisant le bouton "Create".
- Cliquez sur "OK" pour confirmer.

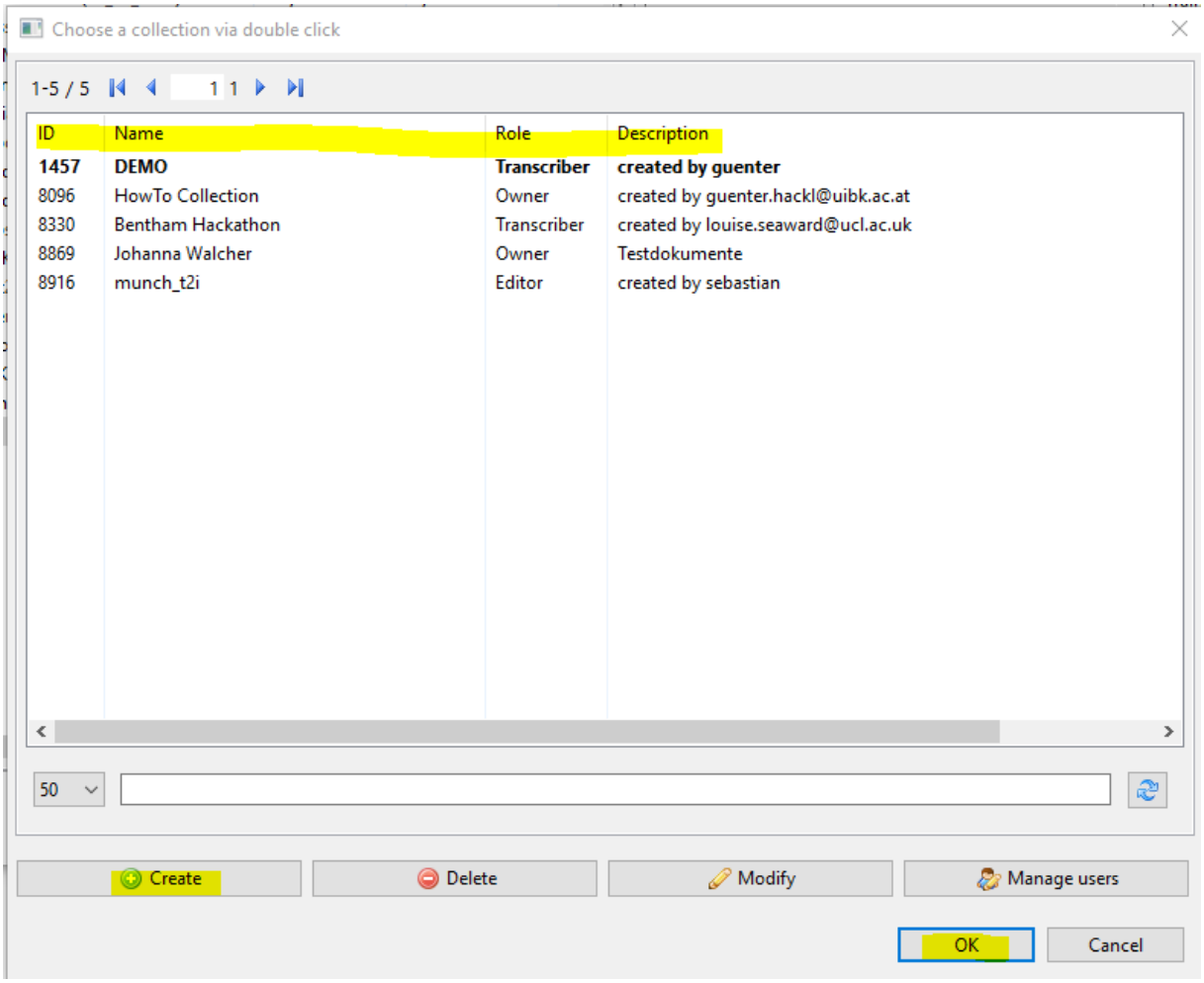

Figure 17 Modèle de division

- Une fois que vous avez sélectionné la collection, cliquez à nouveau sur "OK" et le modèle sera partagé.

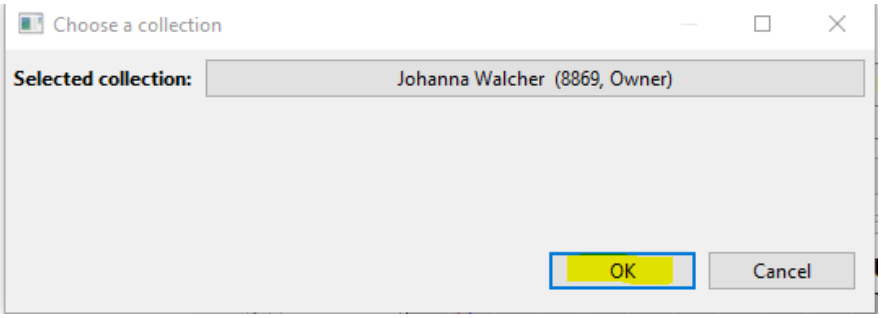

Figure 18 Confirmation du partage de modèle

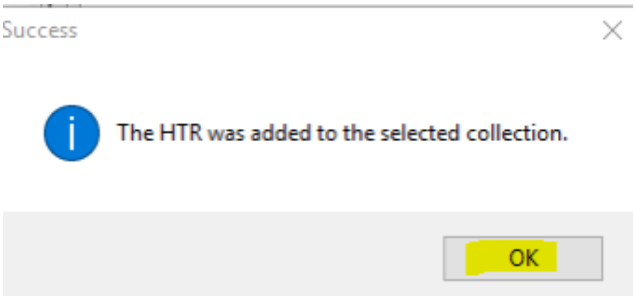

Figure 19 Le modèle a été divisé

Avantages de l’entraînement d’un modèle

- Après avoir terminé la formation, vous pouvez essayer votre modèle sur n'importe quel autre document historique avec une écriture similaire.
- Vous pouvez partager votre document avec d'autres personnes qui peuvent en bénéficier.
- Vous pouvez répéter l'entraînement avec plus de données pour obtenir des résultats encore meilleurs.
- Vous pouvez mesurer la précision de votre modèle à l'aide de la fonction (“Compute Accuracy”). Les résultats de l’HTR dépendent de la similitude et de la clarté de l’écriture dans le document historique.
- L'équipe de Transkribus travaille actuellement sur un algorithme qui transcrira automatiquement n'importe quel type de document sans avoir à préparer de données de formation. La technologie s'appuie principalement sur les données de formation traitées dans Transkribus.
- Plus il y a de données disponibles, plus la technologie sera efficace. Entraînez votre propre modèle et faites-en partie ! 😊

Crédits

Nous tenons à remercier les nombreux utilisateurs et utilisatrices qui ont contribué à l'amélioration du logiciel Transkribus par leurs commentaires.

Transkribus sera mis à la disposition du public dans le cadre du projet d'infrastructure H2020e READ (Recognition and Enrichment of Archival Documents), financé par la Commission européenne.