

TRANSKRIBUS : l'intelligence artificielle au service du patrimoine documentaire

Lancée en 2015, Transkribus est la première plateforme de reconnaissance automatique des écritures manuscrites. C'est aussi une interface de traitement du patrimoine documentaire.



Transkribus est la première plateforme de reconnaissance automatique des écritures manuscrites (RÉM ou HTR pour *Handwritten Text Recognition*) développée pour mettre en valeur le patrimoine documentaire. Elle a été lancée en 2015, dans le cadre du projet READ (*Research and Enrichment of Archival Documents*) mené par l'université d'Innsbruck (en collaboration avec un consortium de 13 autres universités et centres de recherches européens) et financé par la Commission européenne dans le cadre de l'initiative Horizon 2020 (2016-2019). Aujourd'hui prise en charge par la coopérative READ-COOP SCE, dont les membres sont essentiellement des centres d'archives, des bibliothèques, des universités ou des laboratoires de recherche, Transkribus compte plus de 80 000 utilisateurs partout dans le monde¹.

Tablant sur les avancées de la recherche en intelligence artificielle, Transkribus permet la reconnaissance d'écritures de tous les types (manuscrits ou imprimés), de toutes les époques et dans toutes les langues. Pour ce faire, Transkribus exploite une approche d'apprentissage machine basée sur des réseaux de neurones profonds (*deep neural network*) pour localiser avec précision les lignes de texte dans une image numérique² et pour reconnaître chaque caractère de ces lignes en les comparant statistiquement avec les données d'entraînement fournies par l'utilisateur. Avec une centaine de pages transcrites, les utilisateurs peuvent ainsi créer un modèle de reconnaissance spécifiquement adapté à la graphie et à la langue des textes qu'ils souhaitent travailler.

Jusqu'ici, quelque 12 000 modèles de reconnaissance ont été entraînés par les usagers, qui ont permis de transcrire plus de 31 millions de pages³, incluant des imprimés (ouvrages et journaux) anciens

et modernes et des documents manuscrits ou hybrides. Parmi ces modèles, une centaine sont publiquement accessibles à tous les utilisateurs et ce, dans 24 langues différentes, du XI^e au XXI^e siècles⁴.

DES TAUX D'ERREUR TRÈS FAIBLES

La « performance » des modèles varie évidemment selon la nature des données qui ont servi à les entraîner et la nature des documents à transcrire. Ainsi, il est possible, à partir de quelques dizaines de pages de transcriptions fiables (vérifiées attentivement), d'entraîner pour les imprimés anciens des modèles dont les taux d'erreur se situent sous la barre des 1 %. Avec les manuscrits, la variété des graphies et la qualité des images demeurent les principaux enjeux à relever mais pour des documents dont la graphie est assez constante (tels que les greffes de notaires), 150 pages de transcriptions suffisent pour atteindre des taux d'erreurs sous les 5 %. Les modèles accessibles à tous, entraînés à partir de données fournies par plusieurs utilisateurs, sont basés sur des corpus de plusieurs milliers, voire de dizaines de milliers de pages. Ces modèles, qui prennent en charge une grande variété de graphies dans une langue donnée, atteignent facilement des taux d'erreurs de 5 % à 10 %.

UN MODÈLE D'AFFAIRE COOPÉRATIF

Transkribus n'est pas simplement un logiciel de transcription automatisée, c'est aussi une interface de traitement du patrimoine documentaire permettant l'enrichissement des transcriptions par le balisage et le partage des données, et offrant de puissants outils de recherche. Entre autres, la plateforme comporte un module de balisage de métadonnées calqué sensiblement sur les principes de la TEI, qui facilite l'enrichissement des transcriptions avec des données

concernant les individus, les institutions, les lieux, les dates ou toute autre information jugée utile pour la recherche. Les métadonnées balisées peuvent même être ajoutées aux données d'entraînement des modèles de reconnaissance, si bien qu'une partie du balisage peut déjà être effectuée automatiquement par la machine.

Contrairement à certains outils de RÉM (dont eScriptorium), l'utilisation des algorithmes de reconnaissance de texte (manuscrit comme imprimé) a un coût dans Transkribus, qui varie en fonction du statut de l'utilisateur (membre ou non de la READ-Coop, étudiants aux cycles supérieurs), du volume d'achat et de sa récurrence (achat unique, abonnement annuel ou mensuel). Néanmoins, le modèle d'affaire coopératif assure que les revenus sont réinvestis dans l'entretien et l'amélioration des serveurs (extrêmement puissants) et le développement continu de la plateforme et de ses algorithmes, ainsi que dans l'ajout d'outils facilitant la diffusion des contenus transcrits (*read@search*) ou la transcription participative (*citizens&science*).

MAXIME GOHIER

Professeur d'histoire, université du Québec à Rimouski et directeur de Nouvelle-France numérique
Maxime_gohier@uqar.ca

[1] <https://readcoop.eu/transkribus>

[2] Max Weidemann, et al., *HTR Engine Based on NNs P2: Building Deep Architectures with TensorFlow*, READ-H2020 Project, 2017.

[3] Données à jour en août 2022
(<https://readcoop.eu/transkribus>).

[4] <https://readcoop.eu/transkribus/public-models>

[5] <https://nouvellefrancenumerique.info>