L'ATR en pratique : lumière sur les techniques de transcription automatique à Genève

Pauline Jacsont, Elina Leblanc

L'ATR en pratique : lumière sur les techniques de transcription automatique à Genève

Pauline Jacsont, Elina Leblanc
Université de Genève
{prenom.nom}@unige.ch

Résumé

Afin de donner un aperçu des dernières avancées des technologies d'ATR à l'Université de Genève, cet article propose une synthèse recensant les outils et les pratiques dans ce domaine. Après avoir établi un état de l'art de la discipline, nous examinerons les solutions développées à l'Université de Genève. Enfin, à travers une analyse de cas concrets, nous mettrons en évidence les avancées scientifiques les plus prometteuses et les défis actuels dans le domaine.

1 Introduction

Que ce soit pour la recherche en sciences humaines ou pour la valorisation des documents patrimoniaux, les outils de reconnaissance automatique de texte (*Automatic Text Recognition*, ATR) jouent un rôle crucial. Ils englobent la reconnaissance automatique d'écriture manuscrite (*Handwritten Text Recognition*, HTR) et la reconnaissance optique de caractères (*Optical Character Recognition*, OCR) [1]. L'utilisation de l'ATR pour l'étude, la diffusion et la préservation du patrimoine textuel amplifie considérablement les possibilités de recherche et de compréhension de notre héritage textuel. Aussi, grâce à ces outils, il est possible de traiter et d'étudier de plus vastes ensembles de textes afin d'identifier des tendances, des motifs et des relations à grande échelle, offrant de nouvelles perspectives pour l'analyse des textes transmis jusqu'à notre époque (Gabay, 2023).

La réalisation de transcriptions automatiques passe par la construction de modèles [2] efficaces adaptés au corpus à transcrire. Pour cela, il est nécessaire de constituer un jeu de données d'entraînement de qualité, suffisamment volumineux [3]. Dans certains cas, il est possible de trouver des données d'entraînement en libre accès prêtes à être utilisées [4] qui présentent des caractéristiques similaires au corpus à transcrire ; cependant, il sera toujours préférable d'ajouter quelques-unes de ses propres données afin d'obtenir de meilleurs résultats [5].

La constitution du jeu de données est une étape cruciale qui nécessite des données représentatives du corpus à traiter *in fine*. Ces « données » prennent une forme double :

— Les numérisations des documents sélectionnés pour constituer la vérité de terrain ;
— des fichiers (le plus souvent en XML-ALTO ou XML-PAGE) contenant des informations précises sur la localisation des différentes zones présentent sur la page (on parle de segmentation) et le texte transcrit.

La préparation de la vérité de terrain comprend ainsi une étape de segmentation pour identifier les éléments de mise en page et les lignes de texte du document, ainsi qu'une étape de transcription. Ce travail peut être manuel (si l'on démarre un nouveau jeu de données) ou semi-automatisé (en utilisant un modèle imparfait pour accélérer le travail). Une fois que la vérité de terrain est prête à être exploitée, un entraînement est réalisé au cours duquel le moteur ATR, une fois paramétré correctement [6], « apprend » sa tâche de segmentation ou de transcription en réalisant plusieurs itérations sur

1. Les trois termes peuvent être utilisés pour désigner les technologies de transcription automatique. Or, aujourd'hui, on tend à privilégier le terme ATR afin de réduire la distinction entre la transcription des documents imprimés et celle des documents manuscrits, qui n'est pas toujours pertinente : en effet, certains imprimés nécessitent un traitement tout aussi complexe que celui des manuscrits.

2. Un modèle est un fichier entraîné à partir d'un jeu de données d'apprentissage (vérité de terrain) pour automatiser l'exécution d'une tâche telle que l'analyse de mise en page et la transcription des caractères.

3. Il est difficile d'estimer la quantité de données à fournir, car cela dépend de plusieurs facteurs tels que la langue, le système d'écriture, le type d'écriture, la qualité des numérisations, la régularité des documents, etc.

4. Le catalogue *HTR-United* (Chagué et al., 2021; Chagué et Clérice, 2023) en propose un grand nombre. Cf. https://htr-united.github.io.

5. On parle notamment de *fine-tuning*.

6. L'apprentissage machine requiert la définition d'un nombre conséquent de paramètres pour optimiser la tâche

l'ensemble des données présentes dans la vérité de terrain. Quand des modèles de segmentation et de transcription efficaces on été créés, ils peuvent alors être appliqués à l'ensemble du corpus au moyen du moteur de transcription.

Dans ce texte, nous étudierons et présenterons les pratiques de l'ATR à l'Université de Genève, afin de mieux appréhender la manière dont cette technologie joue un rôle crucial dans la préservation et l'accessibilité du patrimoine textuel, en facilitant la recherche, l'analyse et la gestion de l'information.
Tout d'abord, nous ferons un examen des différents moteurs et logiciels d'ATR disponibles afin de contextualiser les choix qui ont été faits par l'Université de Genève. Puis, nous nous pencherons sur l'infrastructure ATR mise en place par le projet FoNDUE à Genève. Enfin, nous exposerons quelques exemples concrets d'utilisation des données produites par les outils d'ATR au sein de projets de recherche de l'Université.

2 Quel outil choisir pour la reconnaissance automatique de texte ?

Lorsque l'on se lance dans un projet avec de la transcription automatique pour la première fois, il est inévitable de faire face à une décision cruciale : le choix de l'outil approprié. Parmi une multitude de monstres marins, d'appellations ésotériques et d'acronymes, il s'agit de déterminer l'outil qui correspond aux exigences de chaque projet. Les outils les plus connus, qui permettent aussi bien de transcrire des documents imprimés que des manuscrits, et les plus couramment utilisés sont *Transkribus*/HTR+, *Tesseract* et e-Scriptorium/*Kraken*. Afin de mieux comprendre les choix de l'Université de Genève, nous allons esquisser un bref comparatif de ces trois outils.

Transkribus. Développée par le projet READ (*Recognition and Enrichment of Archival Documents*), la première version de *Transkribus* est lancée en 2015, pour ensuite continuer à se développer et à s'améliorer au fil du temps. Elle serait aujourd'hui la plateforme la plus utilisée pour produire des transcriptions dans les instituts patrimoniaux (Nockels et al., 2022). La plateforme utilise le moteur de transcription HTR+ développé par l'Université de Rostock. L'un des principaux avantages de *Transkribus* est de proposer aux utilisateurs·rices une interface facile d'utilisation qui permet de charger des images de documents, de les transcrire et de les éditer (cf. fig. 1). Aussi, l'entraînement des modèles est complètement pris en charge par l'outil ce qui facilite le travail des chercheur·euses.

La plateforme contient également une librairie avec des modèles de transcription pour différents types de document et d'écriture ce qui facilite la mise en œuvre rapide de la reconnaissance de texte. Toutefois, l'outil n'est pas complètement *open source* et certaines fonctionnalités de *Transkribus*, parmi lesquelles la transcription du texte, sont payantes [7].

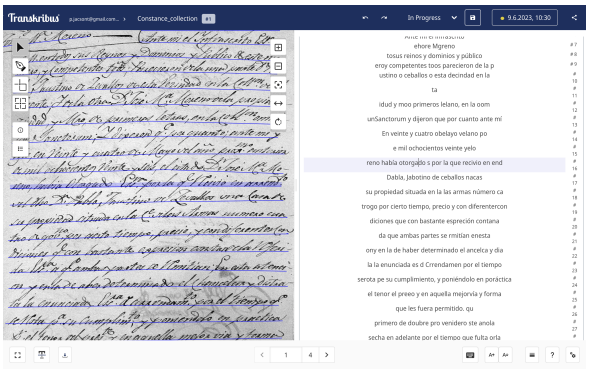

FIGURE 1 – Interface d'édition de la plateforme *Transkribus*. À gauche, *Distrato y cancelación*, *Archivo Histórico Provincial de Sevilla*, 2947, f. 721r.

Tesseract. Contrairement à *Transkribus*, *Tesseract* est un logiciel *open source* (Kay, 2007) initialement conçu pour les documents imprimés non historiques (par exemple des documents administratifs, factures, etc.), soit des documents avec peu de variation. Aujourd'hui, le développement de *Tesseract* est réalisé par Google et l'outil utilise des techniques d'apprentissage automatique pour améliorer sa précision de reconnaissance des caractères, ce qui le rend performant aussi bien sur des documents imprimés que manuscrits. Bien que l'outil utilise des modèles d'apprentissage automatique, tels que des réseaux de neurones, ses performances pour les documents manuscrits ne sont pas toujours satisfaisantes (Gatos et al., 2015). Enfin, il n'existe actuellement aucune interface graphique pour faciliter l'utilisation de ce moteur de transcription, qui se fait directement en ligne

d'apprentissage : *learning rate*, architecture du réseau de neurone…

7. Voir https://readcoop.eu/Transkribus/credits.

de commande ce qui peut être un obstacle pour certain·e·s utilisateurs·rices.

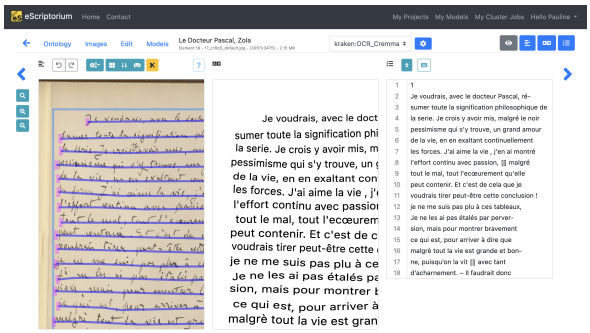

FIGURE 2 – Interface d'édition *eScriptorium*. À gauche, Manuscrit autographe du *Docteur Pascal* d'Émile Zola (MS. Z-6.3*), Bibliothèque Martin Bodmer, f° 1. Numérisation réalisée par le Bodmer Lab.

eScriptorium. L'interface *eScriptorium* qui intègre le moteur de transcription *Kraken* a le bénéfice de combiner aussi bien les qualités de *Transkribus* — grande performance sur les documents historiques, aussi bien imprimés que manuscrits (Kiessling, 2019) et une interface graphique facile d'utilisation (Stokes et al., 2021) — que celle de *Tesseract* — logiciel *open source*, soit un logiciel libre d'accès et complètement gratuit. L'Université de Genève a opté pour l'interface *eScriptorium*/*Kraken* non seulement pour les avantages précédemment cités, mais également en raison du succès d'expériences similaires menées à l'EPHE, l'INIRA et l'École des chartes. Cette décision s'inscrit également dans une démarche de recherche et d'expérimentation : contrairement à *Transkribus*, qui présente une certaine rigidité, *eScriptorium* est un outil plus modulaire grâce à sa nature *open source*, ce qui permet de l'adapter aux besoins spécifiques des différents projets. Afin de circonvenir le problème commun à tous les services précédemment cités, le projet FoNDUE a fait le choix de déployer une instance interfacée avec les services de *High Performance Computing* (HPC) de l'université, permettant ainsi d'exécuter efficacement les tâches nécessitant des calculs complexes.

3 FoNDUE, une infrastructure genevoise pour l'ATR

Le projet FoNDUE (FOrmes Numérisées et Détection Unifiée des Écritures) a pour objectif de mettre en place une infrastructure ATR qui facilite l'accès et l'exploitation des données textuelles tout en répondant aux besoins spécifiques des utilisateurs·rices. Au cœur de cette architecture, se trouvent les outils *open source eScriptorium* et *Kraken*, qui offrent une solution à la fois performante et simple d'utilisation pour la réalisation et l'utilisation de modèles de segmentation et de transcription avec le moteur *Kraken*.

L'un des points forts du projet est la garantie du traitement des données à Genève, ce qui permet de conserver le contrôle des données et de garantir (le cas échéant) leur confidentialité. Cela revêt une importance toute particulière pour la protection de la vie privée et la gestion de données sensibles (documents sous droit, documents privés…). FoNDUE vise également à garantir la reproductibilité des résultats, indépendamment du type de données partagées ou du moteur de reconnaissance de texte utilisé : il s'agit là d'un aspect crucial, notamment dans le contexte de la recherche scientifique. Des modèles de transcription automatique, produits par la communauté internationale des utilisateurs de *Kraken*, sont mis à disposition des utilisateurs·rices afin d'améliorer l'efficience des modèles de transcriptions : certaines de ces données ont été produites par les membres du projet FoNDUE, mettant ainsi à disposition des modèles spécifiques aux écritures gothiques françaises, cursives françaises et gothiques modernes allemandes.

Une autre caractéristique essentielle du projet est la mise à disposition d'une grande puissance de calcul et un accès direct aux GPU. Cela permet aux utilisateurs·rices d'exploiter pleinement les performances du moteur ATR pour des tâches de calcul intensives, telles que la reconnaissance de caractères complexes. Les utilisateurs·rices peuvent aussi tirer pleinement parti des performances du moteur ATR pour des tâches de calcul intensives sur de vastes ensembles de documents. De plus, les utilisateurs·rices expérimenté·e·s ont la possibilité de lancer leurs entraînements en ligne de commande via SLURM, ce qui leur permet d'avoir accès à tous les paramètres d'optimisation de l'architecture du moteur ATR pour un entraînement de modèle plus adapté au corpus à transcrire, et donc plus efficace, afin d'obtenir les meilleurs résultats de transcription.

Une autre caractéristique essentielle du projet est la mise à disposition d'une grande puissance de calcul gratuite grâce aux GPU de l'université. De plus, les utilisateurs·rices expérimenté·e·s ont la

possibilité de lancer leurs calculs sur le GPU en ligne de commande, ce qui leur permet de personnaliser l'architecture du moteur ATR afin d'obtenir un modèle plus adapté au modèle à transcrire et donc plus performant en termes de résultats.
Enfin, le projet FoNDUE soutient activement les chercheurs en fournissant une documentation régulièrement mise à jour et en proposant des formations [8]. Cela permet aux utilisateurs·rices de se familiariser avec les outils, de comprendre les meilleures pratiques et de maximiser leur efficacité dans l'exploitation des fonctionnalités offertes par l'ATR.

4 Cas concrets d'exploitation des données

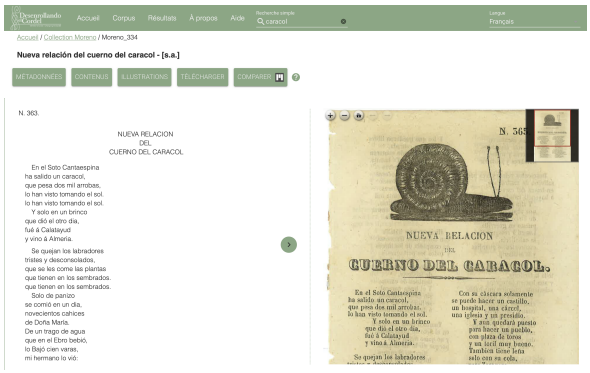

FIGURE 3 – Exemple de visualisation que propose la bibliothèque numérique *Desenrollando el Cordel* avec mise en regard de la transcription avec le facsimilé numérique.

L'utilisation la plus courante des données produites dans le cadre du projet FoNDUE est la numérisation et la préservation du patrimoine culturel. La transcription automatique joue un rôle essentiel dans la conservation à long terme de ces ressources et dans leur diffusion auprès d'un public plus large. Un exemple concret de cette utilisation se trouve dans le projet *Desenrollando el Cordel* [9], qui vise à promouvoir et à préserver des imprimés de colportage espagnols qui étaient initialement destinés à une durée de vie éphémère (Carta et Leblanc, 2021). Les prédictions ATR occupent une place centrale au sein de la bibliothèque numérique *Desenrollando el Cordel* (cf. fig. 3), qui utilise l'application TEI-Publisher et la base de données eXist-DB pour la gestion des fichiers XML TEI produits à partir de ces prédictions (Leblanc, 2023).

La réalisation de prédictions ATR sur des textes inédits offre aussi l'occasion de travaux linguistiques ou d'analyses textuelles avancés tels que l'analyse des structures linguistiques, des variations dialectales, l'identification de motifs, la recherche d'entités nommées ou la création de corpus linguistiques – des études que s'apprêtent à mener les membres d'un autre projet genevois : SETAF (Solfrini et al., 2023).

FIGURE 4 – Édition numérique de la *Toponomasia* avec visualisation sur une carte des toponymes sélectionnés par Gasparo Sardi.

De plus, les prédictions ATR produites peuvent être utiles pour la recherche historique, en permettant l'exploration et l'analyse de documents anciens tels que des manuscrits, des archives ou des textes imprimés. Cela peut contribuer à la compréhension de périodes historiques spécifiques, de mouvements intellectuels ou de contextes culturels. Dans le cadre d'un projet en partenariat avec l'Université de Genève et la Bibliothèque de la bourgeoisie de Berne, une transcription complète d'un manuscrit du XVI^{e} s a été réalisée. Ce manuscrit contient l'œuvre de Gasparo Sardi intitulée *Toponomasia* [10], un ouvrage toponymique qui établit un lien entre la nomenclature des toponymes antiques et celle de l'époque de l'auteur. La transcription de ce manuscrit a permis la création d'une édition numérique qui, pour mettre en valeur la dimension géographique de cette œuvre, propose aux utilisateurs·rices une carte géolocalisant les toponymes collectés par Gasparo Sardi (cf. fig. 4 et Jacsont 2022). Cette visualisation offre un autre moyen

8. Pour obtenir davantage d'informations, vous pouvez consulter la documentation de Fondue (Gabay et al., 2022) et les cours disponibles sur Mediaserveur : https://mediaserver.unige.ch/play/VN3-4929-2022-2023.

9. https://desenrollandoelcordel.unige.ch.

10. Il s'agit du codex 174 un manuscrit latin du XVI^{e} s écrit en cursive humaniste.

d'interroger le travail de l'auteur et sa relation avec la géographie antique, en analysant les choix qu'il a opérés pour la sélection de ces toponymes.

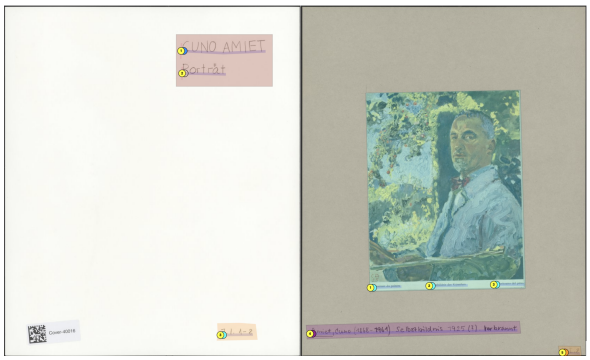

FIGURE 5 – Exemple de documents, venant des archives d'Heinrich Wölfflin, segmentés et transcrits.

L'application des outils d'ATR dans le domaine des archives présente de nombreux avantages en matière de traitement et d'accès aux documents. Grâce à l'ATR, l'indexation, la création de métadonnées et la recherche d'informations sont considérablement simplifiées. Ainsi, dans le cadre d'un projet collaboratif entre l'Université de Zurich et l'Université de Genève, la transcription des archives d'Heinrich Wölfflin, un historien de l'art ayant rassemblé une collection de plus de 50 000 photographies d'art (cf. fig. 5), a permis d'extraire un ensemble précieux d'informations, comprenant des notes manuscrites de Heinrich Wölfflin et des annotations de conservateurs : ces transcriptions serviront à la création d'une base de données (Jacsont et al., 2023).

5 Conclusion

Les technologies d'ATR jouent un rôle crucial dans la préservation et l'accessibilité du patrimoine textuel. L'Université de Genève, grâce à l'infrastructure et aux outils mis en place par le projet FoNDUE, contribue activement à l'avancement de cette discipline. Il reste encore beaucoup à explorer et à développer : les technologies d'ATR sont en constante évolution et de nombreuses expérimentations sont à réaliser pour améliorer les algorithmes d'apprentissage automatique et créer des modèles de reconnaissance plus efficaces, capables d'obtenir des résultats encore plus précis et fiables.
Pour progresser dans cette voie, il sera indispensable que, dans les années à venir, une collaboration s'établisse entre les institutions universitaires, les bibliothèques, les archives et tous les autres acteurs impliqués dans la préservation et la diffusion du patrimoine textuel. Ces échanges permettront de partager connaissances, expériences et ressources, contribuant ainsi à l'amélioration continue des outils et des pratiques de l'ATR.

Bibliographie

Jean-Baptiste Camps et Nicolas Perreaux. 2021. Reconnaissance optique des caractères et des écritures manuscrites – projet e-ndp. In *Séminaire du projet e-NDP - Notre-Dame de Paris et son cloître*, Paris.

Constance Carta et Elina Leblanc. 2021. Le projet "démêler le cordel" : une bibliothèque numérique pour l'étude de la littérature éphémère espagnole du xixe siècle. In *Humanistica 2021*, Rennes, France.

Alix Chagué. 2022. eScriptorium : une application libre pour la transcription automatique des manuscrits. *Arabesques*, 107 :25.

Alix Chagué et Thibault Clérice. 2023. "I'm here to fight for ground truth": HTR-United, a solution towards a common for HTR training data. In *Digital Humanities 2023 : Collaboration as Opportunity*, Graz, Austria. Alliance of Digital Humanities Organizations and University of Graz.

Alix Chagué, Thibault Clérice, et Laurent Romary. 2021. HTR-United : Mutualisons la vérité de terrain ! In *DHNord2021 - Publier, partager, réutiliser les données de la recherche : les data papers et leurs enjeux*, Lille, France. MESHS.

Thibault Clérice, Malamatenia Vlachou-Efstathiou, et Alix Chagué. 2023. CREMMA Medii Aevi: Literary manuscript text recognition in Latin. *Journal of Open Humanities Data*, 9 :4.

Béatrice Couture, Verret Farah, Gohier Maxime, et Deslandres Dominique. 2022. The challenges of HTR model training: Feedbacks from the project donner le goût de l'archive à l'ère numérique.

Simon Gabay. 2023. Fondue : Un outil pour construire les grands corpus de demain. In *MAGMAH*, Genève, Suisse.

Simon Gabay, Jean-Baptiste Camps, Ariane Pinche, et Nicola Carboni. 2021. SegmOnto - a controlled vocabulary to describe the layout of pages.

Simon Gabay, Pierre Kuenzli, Jean-Luc Flacone, et Christophe Charpilloz. 2022. FoNDUE: Documentation.

Basilis Gatos, Nikolaos Stamatopoulos, Georgios Louloudis, Giorgos Sfikas, George Retsinas, Vassilis Papavassiliou, Fotini Sunistira, et Vassilis Katsouros. 2015. Grpoly-db: An old greek polytonic document image database. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 646–650.

Pauline Jacsont. 2022. Mise en valeur du patrimoine textuel grâce aux éditions numériques "le cas du codex 174 de la bibliothèque de la bourgeoisie de berne". Mémoire de master, Université de Genève.

Pauline Jacsont, Simon Gabay, et Tristan Weddigen. 2023. Numériser les archives d'histoire de l'art: la collection de photographies d'heinrich wölfflin. In *Humanistica 2023*, Genève, Suisse. Association francophone des humanités numériques.

Philip Kahle, Sebastian Colutto, Günter Hackl, et Günter Mühlberger. 2017. Transkribus - a service platform for transcription, recognition and retrieval of historical documents. In *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 4, pages 19–24, Kyoto, Japon. International Association for Pattern Recognition.

Anthony Kay. 2007. Tesseract : An open-source optical character recognition engine. *Linux Jounal*, 2007(159) :2.

Benjamin Kiessling. 2019. Kraken - a universal text recognizer for the humanities. In *Digital Humanities Conference 2019 - Book of abstracts*, Utrecht, the Netherlands.

Benjamin Kiessling, Robin Tissot, Peter Stokes, et Daniel Stökl Ben Ezra. 2019. eScriptorium: An open source platform for historical document analysis. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 19–19, Sydney, Australia.

Elina Leblanc. 2023. Une bibliothèque numérique face à son public : l'exemple de l'enquête utilisateurs du projet Démêler le cordel. In *Humanistica 2023*, Bonnes pratiques, Genève, Suisse. Association francophone des humanités numériques.

Joe Nockels, Paul Gooding, Sarah Ames, et Melissa Terras. 2022. Understanding the application of handwritten text recognition technology in heritage contexts: a systematic review of Transkribus in published research. *Archival Science*, 22(3) :367–392.

Ariane Pinche. 2021. Compte-rendu de la séance n°1 du séminaire : "création de modèle(s) htr pour les documents médiévaux en ancien français et moyen français entre le Xe-XIVe siècle". In *Séminaire CREMMALAB - Création de modèle(s) HTR*, École nationale des chartes, Paris.

Ray Smith. 2007. An overview of the tesseract ocr engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 629–633, Curitiba, Brésil. International Association for Pattern Recognition.

Sonia Solfrini, Geneviève Gross, Brigitte Roux, Nathalie Szczech, Pierre-Olivier Beaulnes, Aurélia M Oliveira, et Daniela Solfaroli Camillocci. 2023. Étudier le " groupe de Neuchâtel ". In *Humanistica 2023*, Poster, Genève, Switzerland. Association francophone des humanités numériques.

Uwe Springmann, Florian Fink, et Klaus U. Schulz. 2016. Automatic quality evaluation and (semi-) automatic improvement of OCR models for historical printings. *ArXiv e-prints*.

Peter A. Stokes, Benjamin Kiessling, Daniel Stökl, Ben Ezra, Robin Tissot, et El Hassane Gargem. 2021. The e-scriptorium VRE for manuscript cultures. *Classics@ Journal*, 18(1).

Phillip Ströbel, Simon Clematide, Martin Volk, Raphael Schwitter, Tobias Hodel, et David Schoch. 2022. Evaluation of HTR models without ground truth material. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association.