Théia

1 | 2024
Retours d'expérience en édition numérique de sources en histoire et histoire de l'art

Faciliter l'édition numérique avec les méthodes de reconnaissance automatique de texte

L'application de l'outil Transkribus dans le projet « Grand Tour digital »

Facilitating digital publishing with automatic text recognition methods. The application of the Transkribus tool in the 'Grand Tour digital' project.

Angela Göbel

http://publications-prairial.fr/theia/index.php?id=129

DOI : 10.35562/theia.129

Electronic reference
Angela Göbel, « Faciliter l'édition numérique avec les méthodes de reconnaissance automatique de texte », *Théia* [Online], 1 | 2024, Online since 17 avril 2025, connection on 25 octobre 2025. URL : http://publications-prairial.fr/theia/index.php?id=129

Faciliter l'édition numérique avec les méthodes de reconnaissance automatique de texte

L'application de l'outil Transkribus dans le projet « Grand Tour digital »

Facilitating digital publishing with automatic text recognition methods. The application of the Transkribus tool in the 'Grand Tour digital' project.

Angela Göbel

OUTLINE

TEXT

Introduction

1 L'utilisation des méthodes de reconnaissance automatique dans les projets d'éditions notamment numériques gagne de plus en plus en importance dans les sciences humaines et sociales. À partir du projet d'édition en cours « Grand Tour digital », cet article propose un retour d'expérience : il souhaite mettre en avant les méthodes mobilisées et discuter les possibilités et les limites qui sont progressivement apparues[1]. Réalisé à la Herzog August Bibliothek à Wolfenbüttel (HAB) en Allemagne, ce projet a pour but de numériser, d'exploiter et de visualiser des témoignages de voyage de formation ou du *Grand Tour* à l'aide des méthodes d'édition semi-automatiques.

2 Traitant un corpus d'une vingtaine de récits de voyage de l'époque moderne qui sont principalement conservés dans les fonds de la bibliothèque, le projet recourt à la plateforme de reconnaissance de texte et d'écriture manuscrite automatique Transkribus. Il apparaît que si cet outil est en mesure de faire face à un certain nombre de défis, la qualité de la numérisation dépend néanmoins fortement de l'état des manuscrits.

Le projet « Grand Tour digital »

Enjeux et buts du projet

3 À l'époque moderne, les voyages éducatifs tels que le *Grand Tour* étaient une étape cruciale dans le développement personnel des membres des élites nobles et bourgeoises. Leur signification particulière, jusqu'à présent peu étudiée dans la recherche sur les témoignages personnels (*Selbstzeugnisforschung*), réside dans le fait qu'ils capturaient et évaluaient moins les expériences en tant qu'altérité, mais mettaient plutôt en avant la réussite de la gestion des expériences dans des environnements étrangers. Ce type de témoignages personnels fait l'objet du projet d'édition « Grand Tour digital. *Digitalisierung, Erschließung und Visualisierung frühneuzeitlicher Selbstzeugnisse von Bildungsreisen unter Anwendung teilautomatisierter Editionsverfahren* [2] », soutenu financièrement par la *Deutsche Forschungsgemeinschaft* (DFG), de 2022 à 2025.

4 Employée dans le nom du projet, la notion de « Grand Tour » est utilisée dans un sens large. Par conséquent, le corpus de sources ne se limite pas uniquement aux récits du *Grand Tour* classique qui emmenait les voyageurs en France et surtout en Italie, mais inclut également d'autres voyages, tels les tournées à cheval des nobles-patriciens (*adlig-patriziche Kavarlierstour*), les pérégrinations académiques étudiantes (*studentische peregrinatio academica*) et les voyages savants professoraux, comme ceux distingués par Mathis Leibeteseder [3]. Ces distinctions permettent de classer ces témoignages personnels rendus disponibles à travers leurs numérisations. Les textes édités seront mis à disposition au cours du projet sur le portail des témoignages personnels (*Selbstzeugnisportal*) de la HAB [4]. Le travail est documenté dans un blog accompagnant

le projet [5], et une offre ultérieure pour la réutilisation du modèle HTR, des textes et des données de recherche sera également disponible via le serveur GitLab [6] de la HAB.

Corpus des sources

5 Au cœur du projet se trouvent 21 journaux de voyage, principalement rédigés en allemand entre les années 1550 et 1770 (pour un total d'environ 10 300 pages). Ils proposent des récits de voyage dans toute l'Europe et jusqu'à l'Empire ottoman et le Proche-Orient (Alep, Jérusalem). Ces journaux sont conservés principalement à la HAB à l'exception de deux manuscrits qui se trouvent aux Archives d'État de Basse-Saxe (Niedersächsisches Landesarchiv) à Wolfenbüttel.

6 Le projet s'intéresse particulièrement à cinq textes, dont l'un est accompagné de trois transcriptions, qui seront transcrits de manière partiellement automatisée grâce au logiciel de reconnaissance d'écriture manuscrite Transkribus, puis encodés en TEI-XML. L'utilisation de la reconnaissance d'entités nommées (*Named Entity Recognition*, REN) permet leur exploration, et leur visualisation est réalisée en combinant le texte avec l'itinéraire du voyage. Cette approche vise à expérimenter le développement d'un processus éditorial novateur.

7 Ces cinq récits ont été rédigés par les personnes suivantes :

- Barthold von Gadenstedt (HAB Cod. Guelf. 67.6 Extrav.)
- Le pharmacien Wagener (HAB Cod. Guelf. 267.1 Extrav.)
- Christian August de Schleswig-Holstein-Norburg (HAB Cod. Guelf. 221 Extrav.)
- Ernst Ferdinand et Heinrich Ferdinand de Brunswick-Wolfenbüttel-Bevern [7] (HAB Cod. Guelf. 149.14 Extrav.)
- Ludwig Rudolph de Brunswick-Wolfenbüttel (HAB Cod. Guelf. 89 Blank.)

8 S'ajoutent à cela 13 récits de voyage supplémentaires, qui sont numérisés. Leurs métadonnées sont explorées et seront mises à disposition via la base de données des manuscrits de la bibliothèque [8].

La reconnaissance automatique des récits de voyage avec l'outil Transkribus

9 La quantité de projets éditoriaux utilisant des logiciels de reconnaissance manuscrite, dont Transkribus, a augmenté de manière significative ces dernières années [9]. L'application de méthodes d'intelligence artificielle offre une nouvelle approche pour interagir avec et traiter des sources historiques, qu'elles soient imprimées ou manuscrites.

10 Transkribus est une plateforme dédiée à la transcription automatique, à l'analyse d'images et à la reconnaissance de structures de documents historiques grâce à l'intelligence artificielle. Le programme a été développé à la suite de deux projets consécutifs à l'Université d'Innsbruck : « tranScriptorium », de 2013 à 2015, et « READ » (*Recognition and Enrichment of Archival Documents*), de 2016 à 2019. Pour l'utilisateur, il est possible d'utiliser Transkribus soit via l'application téléchargeable sur le bureau, soit en ligne avec Transkribus Lite. L'inscription et l'utilisation sont gratuites, mais pour la transcription automatisée, l'utilisateur dispose d'un quota spécifique de « crédits ». Après leur épuisement, il peut en acheter de nouveaux dans différents packs.

11 Transkribus propose de nombreux modèles d'intelligence artificielle (IA) disponibles gratuitement. En janvier 2024, on en compte 138 créés pour des sources manuscrites et imprimées de différentes périodes, ainsi que pour de nombreuses langues européennes et extra-européennes. Pour ce projet, comment choisir le modèle qui convient à cette écriture manuscrite ? Quels critères sont importants, voire convaincants, pour la sélection du modèle de base ? Et pourquoi créer un nouveau modèle pour les écritures de la période moderne, alors que plusieurs modèles sont déjà disponibles [10] ?

12 Transkribus fait face à divers défis lors de la transcription automatisée de récits de voyage de l'époque moderne, tels que celui de Wagener, mais ces défis sont également applicables de manière générale aux témoignages personnels de cette époque :

Aspect de l'écriture et lisibilité : la qualité de la transcription dépend fortement de la forme de l'écriture, qui elle-même varie en fonction de la version originale ou d'une copie, ainsi que de l'état de conservation de la source.

- Écriture individuelle : chaque individu a une écriture unique. Les fautes d'orthographe, les ratures, les ajouts, ainsi que l'encre étalée, pressée ou décolorée compliquent la reconnaissance automatisée.
- Mise en page variée : les récits de voyage de l'époque moderne présentent souvent une mise en page variée, incluant le texte principal, des annotations marginales, des tableaux, des listes, des schémas ainsi que des esquisses de taille différente. S'ajoutent de manière répétée des compléments et corrections ultérieurs directement dans le corpus du texte, qui sont soit ajoutés en marge du texte, soit insérés entre les lignes.
- Polices et tailles de caractères variées : un texte peut contenir différentes polices et tailles de caractères, comprenant notamment la cursive allemande, l'écriture latine, ou d'autres encore.
- Utilisation de différentes langues : l'inclusion de différentes langues, comme des copies d'inscriptions ou des citations, représente un défi supplémentaire et peut demander au modèle d'être entraîné sur plusieurs langues.
- Exigences de contenu : les récits de voyage de l'époque moderne contiennent souvent un grand nombre de noms propres (personnes, lieux, œuvres d'art, titres de livres, etc.), des indications de date, des mesures et des caractères spéciaux qui doivent être précisément identifiés et transcrits.

13 Dans le cadre du projet sera développé un modèle entraîné à ces exigences qui sera partagé en accès libre à son achèvement.

À propos de l'application de Transkribus et le work-flow élaboré

Créer un premier modèle de reconnaissance de textes

14 L'objectif central du projet « Grand Tour digital » est la mise au point expérimentale d'un processus éditorial novateur en utilisant Transkribus. Le logiciel lui-même fournit de nombreuses instructions utiles pour l'édition de texte [11]. En outre, la recherche produit de plus en plus de rapports d'expérience, de discussions scientifiques et de conseils pratiques sur cet outil [12].

15 Lors d'une première étape, après avoir transcrit manuellement une vingtaine des pages du journal de voyage de Wagener dans Transkribus, on a pu partir des pages saisies et d'un modèle déjà existant pour permettre l'entraînement d'un nouveau modèle. Pour sa création, nous avons choisi le modèle existant « Transkribus German handwriting M1 » avec une faible erreur ou taux d'erreur de caractères (CER) [13] de 4,70 % comme modèle de base. Ce modèle a été réentraîné à partir de la transcription manuelle des premières pages (p. 7 à 26).

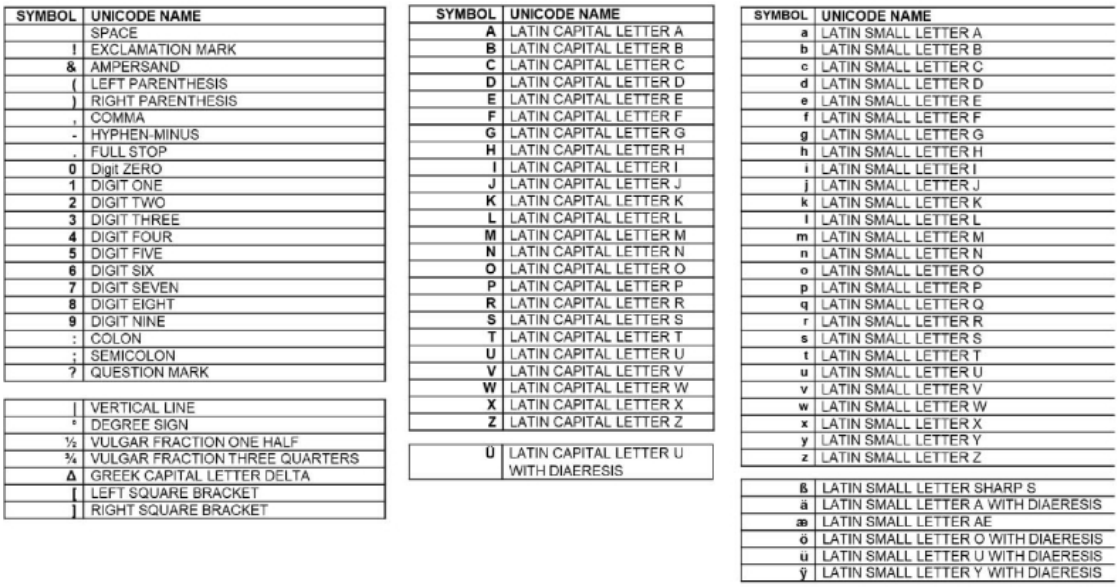
16 Une fois entraîné, ce modèle a été appliqué aux autres pages du manuscrit et ajusté dans des étapes alternées de transcription partiellement automatisée, de correction et de nouvel entraînement du modèle. L'idée centrale derrière ces étapes de travail était d'adapter le programme aux particularités de l'écriture de Wagener afin de réduire progressivement le taux d'erreur dans la reconnaissance manuscrite au cours de la transcription du texte. Dans cette perspective, il est important de comprendre les caractéristiques spécifiques du manuscrit qui a composé cette première base de départ. En ce qui concerne l'indexation du texte partiellement automatisée dans Transkribus, le manuscrit présente généralement une écriture claire avec peu de ratures ou d'ajouts. Parfois, l'encre a traversé le papier, ce qui rend la lecture du manuscrit difficile tant pour l'homme que pour la machine. À certains endroits de son récit de voyage, Wagener a également ajouté des notes sur de petits morceaux de papier collés à l'intérieur de son journal. La langue du texte est principalement l'allemand, mais on trouve à plusieurs reprises des citations transcrites d'inscriptions en français et en latin, surtout à la fin du récit [14].

17 Pour au mieux réussir la transcription automatisée dans Transkribus, le texte a été préparé de manière à minimiser autant que possible les erreurs de détection des champs de texte. À cette fin, le marquage des champs de texte peut soit être parcouru automatiquement et corrigé par la suite, soit être créé directement manuellement. Selon le manuscrit, l'une ou l'autre option est recommandée pour travailler de manière aussi efficace que possible. Dans le cas du manuscrit de Wagener, le marquage automatique des champs de texte a été effectué pour les pages restantes du manuscrit, puis, lors de l'examen ultérieur, de petites corrections ont été apportées en cas d'erreurs dans le marquage du texte principal et des notes en marge ainsi que des lignes manquantes ont été ajoutées. Ce travail se fait plus difficile lorsque les champs de texte ou même les mots se chevauchent et se superposent, mais il est également possible de corriger le tout manuellement dans ce cas.

18 Lors de l'entraînement d'un nouveau modèle, Transkribus distingue entre les pages d'entraînement et de validation. Les pages d'entraînement lors du premier passage (p. 7 à 21) étaient celles où le modèle était formé et les pages de validation (p. 22 à 26) étaient celles où le modèle était automatiquement vérifié et le taux d'erreur calculé [15]. Au cours de cette première itération, l'ensemble d'entraînement, associé au modèle de base avait atteint un taux d'erreur de 2,41 %. Les pages de validation, idéalement représentatives des particularités du manuscrit, avaient un taux d'erreur de 11 %. Afin d'optimiser les résultats sur les pages de validation lors des ajustements ultérieurs du modèle, des pages de manuscrit non consécutives ont été sélectionnées à intervalles réguliers (par exemple, par intervalles de cinq pages). Les 86 pages restant à traiter ont été lues automatiquement par Transkribus par tranches d'environ 20 pages chacune, puis corrigées manuellement et réentraînées avec les pages précédentes. En cas de pages fortement variables, il est également possible – à l'instar de la suggestion et de la documentation de Jacob Möhrke dans son rapport d'atelier – de trier les pages de manuscrit en fonction de leur qualité textuelle et de retirer du jeu d'entraînement les pages qualifiées de « non utilisables [16] ». Ce processus itératif visait à optimiser progressivement le modèle de transcription.

19 Le logiciel Transkribus propose lui-même une vue d'ensemble du jeu de caractères (*character set*) entraîné pour chaque modèle formé. Pour le manuscrit de Wagener, les caractères suivants ont fait sujet de l'entraînement :

Fig. 1. Vue d'ensemble des signes entraînés dans Transkribus à l'aide des premières pages du manuscrit Wagener.



SYMBOL	UNICODE NAME
	SPACE
!	EXCLAMATION MARK
&	AMPERSAND
(	LEFT PARENTHESIS
)	RIGHT PARENTHESIS
,	COMMA
-	HYPHEN-MINUS
.	FULL STOP
0	Digit ZERO
1	DIGIT ONE
2	DIGIT TWO
3	DIGIT THREE
4	DIGIT FOUR
5	DIGIT FIVE
6	DIGIT SIX
7	DIGIT SEVEN
8	DIGIT EIGHT
9	DIGIT NINE
:	COLON
;	SEMICOLON
?	QUESTION MARK

|	VERTICAL LINE
°	DEGREE SIGN
½	VULGAR FRACTION ONE HALF
¾	VULGAR FRACTION THREE QUARTERS
Δ	GREEK CAPITAL LETTER DELTA
[	LEFT SQUARE BRACKET
]	RIGHT SQUARE BRACKET

SYMBOL	UNICODE NAME
A	LATIN CAPITAL LETTER A
B	LATIN CAPITAL LETTER B
C	LATIN CAPITAL LETTER C
D	LATIN CAPITAL LETTER D
E	LATIN CAPITAL LETTER E
F	LATIN CAPITAL LETTER F
G	LATIN CAPITAL LETTER G
H	LATIN CAPITAL LETTER H
I	LATIN CAPITAL LETTER I
J	LATIN CAPITAL LETTER J
K	LATIN CAPITAL LETTER K
L	LATIN CAPITAL LETTER L
M	LATIN CAPITAL LETTER M
N	LATIN CAPITAL LETTER N
O	LATIN CAPITAL LETTER O
P	LATIN CAPITAL LETTER P
R	LATIN CAPITAL LETTER R
S	LATIN CAPITAL LETTER S
T	LATIN CAPITAL LETTER T
U	LATIN CAPITAL LETTER U
V	LATIN CAPITAL LETTER V
W	LATIN CAPITAL LETTER W
X	LATIN CAPITAL LETTER X
Z	LATIN CAPITAL LETTER Z

Ü	LATIN CAPITAL LETTER U WITH DIAERESIS

SYMBOL	UNICODE NAME
a	LATIN SMALL LETTER A
b	LATIN SMALL LETTER B
c	LATIN SMALL LETTER C
d	LATIN SMALL LETTER D
e	LATIN SMALL LETTER E
f	LATIN SMALL LETTER F
g	LATIN SMALL LETTER G
h	LATIN SMALL LETTER H
i	LATIN SMALL LETTER I
j	LATIN SMALL LETTER J
k	LATIN SMALL LETTER K
l	LATIN SMALL LETTER L
m	LATIN SMALL LETTER M
n	LATIN SMALL LETTER N
o	LATIN SMALL LETTER O
p	LATIN SMALL LETTER P
q	LATIN SMALL LETTER Q
r	LATIN SMALL LETTER R
s	LATIN SMALL LETTER S
t	LATIN SMALL LETTER T
u	LATIN SMALL LETTER U
v	LATIN SMALL LETTER V
w	LATIN SMALL LETTER W
x	LATIN SMALL LETTER X
y	LATIN SMALL LETTER Y
z	LATIN SMALL LETTER Z

ß	LATIN SMALL LETTER SHARP S
ä	LATIN SMALL LETTER A WITH DIAERESIS
æ	LATIN SMALL LETTER AE
ö	LATIN SMALL LETTER O WITH DIAERESIS
ü	LATIN SMALL LETTER U WITH DIAERESIS
ÿ	LATIN SMALL LETTER Y WITH DIAERESIS

20 Cette vue d'ensemble ne couvre pas tous les caractères possibles de l'écriture cursive latine, mais se limite aux seuls caractères utilisés dans le manuscrit. Cependant, il n'est pas tout à fait évident de savoir si cette liste inclut également des caractères du modèle de base sélectionné, dans notre cas « Transkribus German Handwriting M1 », ou si elle se limite exclusivement aux premières 26 pages du manuscrit de Wagener. Malheureusement, dans la présentation du modèle, il n'existe pas de référence explicite au modèle de base et des informations approfondies sur ce dernier font défaut, rendant une analyse plus approfondie et une meilleure compréhension de la procédure difficile.

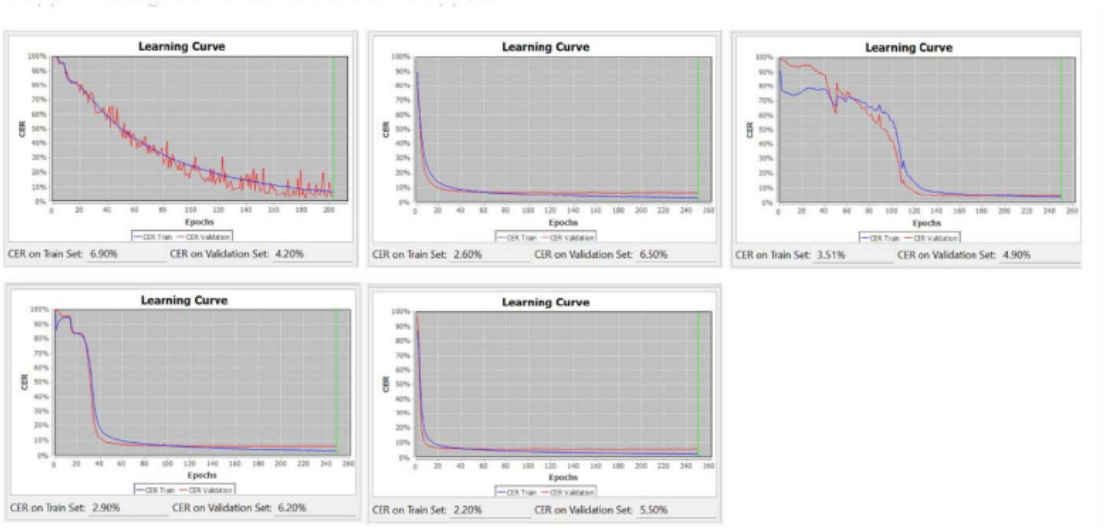
21 Le jeu de caractères généré lors de la première formation avec Transkribus a été élargi et complété avec des caractères manquants au cours des formations et ajustements ultérieurs :

Fig. 2. Vue d'ensemble des signes rajoutés lors des entrainements supplémentaires du modèle HTR.

SYMBOL	UNICODE NAME
‘	APOSTROPHE
+	PLUS SIGN
/	SOLIDUS
Q	LATIN CAPITAL LETTER Q
Y	LATIN CAPITAL LETTER Y
Æ	LATIN CAPITAL LETTER AE
Ö	LATIN CAPITAL LETTER O QITH DIAERESIS
à	LATIN SMALL LETTER A WITH GRAVE
è	LATIN SMALL LETTER E WITH GRAVE
ō	LATIN SMALL LETTER O WITH MACRON
œ	LATIN SMALL LIGATURE OE
ū	LATIN SMALL LETTER U WITH MACRON
Ɔ	LATIN CAPITAL LETTER OPEN O
ƚƀ	L B BAR SYMBOL
△	WHITE UP-POINTING TRIANGLE
⏉	ALCHEMICAL SYMBOL FOR VITRIOL

22 En résumé, lors de la formation du modèle sur l'exemple du manuscrit Wagener, les courbes d'apprentissage suivantes se sont développées :

Fig. 3. Vue des cinq courbes d'apprentissage réalisées lors des entraînements du modèle au manuscrit Wagener. De gauche en haut à droite en bas chaque courbe représente un entrainement.



23 Cette optimisation du modèle fonctionne-t-elle ? D'après les courbes d'apprentissage, une amélioration réelle n'a pas été totale, voire elle a échoué, mais pour quelles raisons ? Les courbes d'apprentissage démontrent une fois de plus que la qualité et le succès ou l'échec de la reconnaissance manuscrite dépendent fortement des pages et soulignent également qu'une sélection liée la qualité de certaines pages pour l'entraînement du modèle serait avantageuse. Une amélioration progressive du modèle n'est pas clairement perceptible dans le cas de l'écriture manuscrite de Wagener – le taux d'erreur, en particulier pour les pages de validation, fluctue entre 4 et 6 %. C'est encore nettement trop élevé dans tous les cas. Cependant, pour les pages d'entraînement, elle diminue de 7 % à environ 2 %. À la lumière de ces premiers résultats, deux questions se posent pour le traitement continu du texte : comment le résultat de l'entraînement du modèle diffère-t-il du résultat de l'application du modèle ? Y a-t-il un meilleur apprentissage en répétant l'entraînement de l'écriture manuscrite en éliminant les pages « inutilisables », ou altère-t-on progressivement le résultat global de cette manière ?

24 Au regard de ces questions, il est intéressant de savoir s'il peut y avoir une amélioration au cours du traitement ultérieur des autres manuscrits avec ce même modèle et si cette amélioration dépendra

de l'inclusion ou de l'exclusion de certaines pages de manuscrit. En réalité, le développement des courbes d'apprentissage pour le premier manuscrit n'a pas toujours été optimal et offre encore suffisamment d'espace pour une amélioration dans le cadre du travail continu de ce projet.

Évaluer la transcription automatique

25 Finalement, qu'est-ce qui a bien fonctionné, qu'est-ce qui n'a pas fonctionné ? Quels problèmes ont été identifiés et quelles solutions Transkribus propose-t-il ? Des évaluations comparables de transcriptions automatisées ont déjà été réalisées dans le cadre d'autres projets et les contributions scientifiques qui en ont résulté soulignent les opportunités de succès ainsi que les difficultés, telles que la dépendance à l'égard de la forme de l'écriture ou le risque de surapprentissage d'un modèle [17].

26 Les *insights* tirés de divers projets et manuscrits sur lesquels Transkribus a été testé montrent clairement que même avec des modèles bien entraînés, une précision de reconnaissance manuscrite à 100 % ne peut jamais être atteinte par une IA : « La reconnaissance automatique de l'écriture manuscrite conduira à un taux d'erreur même avec des modèles bien entraînés. Ainsi, aucun texte philologiquement irréprochable ne sera généré sans correction manuelle ultérieure (post-traitement) [18] ».

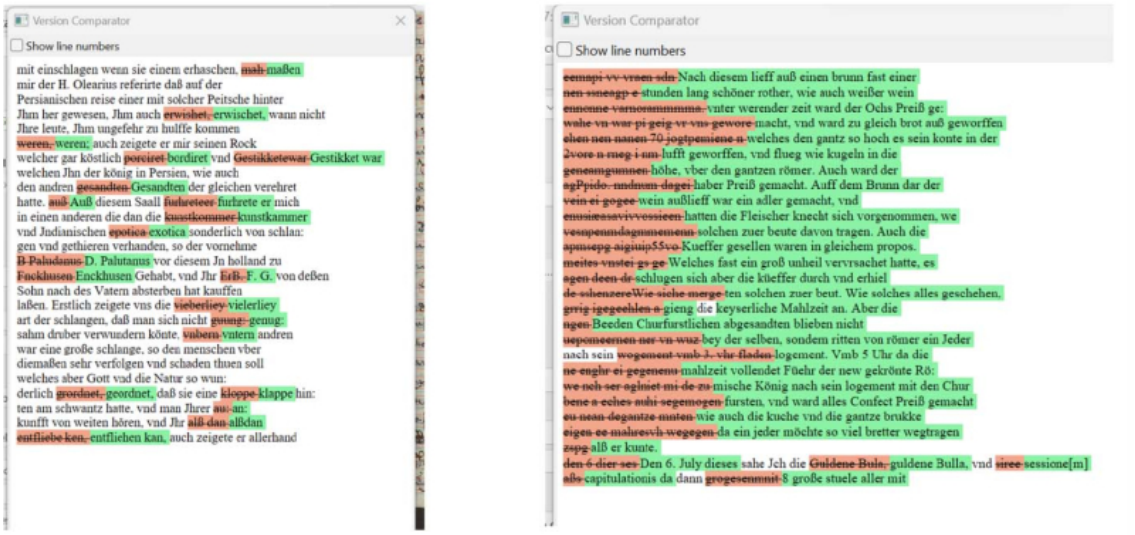
27 À partir du manuscrit de Wagener, nous avons pu examiner plus précisément l'évaluation dans Transkribus, présenter et remettre en question les possibilités d'évaluation automatisée des résultats, identifier les types d'erreurs et discuter de leurs causes possibles. Transkribus offre lui-même la possibilité d'évaluer la transcription effectuée directement dans le programme. Il convient de faire la distinction entre l'évaluation du taux d'erreur (CER) sur les pages de test et de validation et l'application du modèle entraîné sur les autres pages du manuscrit. La précision de la transcription peut être calculée à différents niveaux. Pour la comparaison, deux variantes de texte doivent être sélectionnées. Il faut d'abord une « Référence », un texte correct servant de référence pour l'identification des erreurs. Transkribus recommande « [e]n tant que Référence, choisissez une version de page qui a été correctement transcrite (*Ground Truth* :

transcription manuelle aussi proche que possible du texte original). Pour obtenir la valeur la plus significative, il serait préférable d'utiliser des pages d'un ensemble d'échantillons qui n'ont pas été utilisées dans l'entraînement et qui sont donc nouvelles pour le modèle [19] ». Pour Wagener, nous avons utilisé la version corrigée la plus récente, qui est la plus proche du manuscrit du point de vue orthographique.

28 En regard de ce texte de référence, il faut établir une « hypothèse », c'est-à-dire le texte « hypothétique » à comparer avec la variante correcte. Cela peut être la transcription automatisée par Transkribus, mais aussi des variantes textuelles ultérieures, afin d 'examiner ou de suivre certaines corrections entre collaborateurs. Transkribus recommande » en tant qu'Hypothèse, choisissez la version qui a été générée automatiquement avec un modèle de reconnaissance de texte manuscrit (HTR) et sur laquelle vous aimeriez voir à quel point le résultat est satisfaisant [20] ». Ainsi, nous avons choisi la reconnaissance de texte automatisée (décodage PyLaia). La possibilité offerte par Transkribus de comparer différentes versions de traitement de texte permet une confrontation directe entre la transcription partiellement automatisée et la correction manuelle effectuée. Cependant, il est difficile de comprendre exactement comment fonctionnent la comparaison et l'évaluation qui en découle. Contrairement à d'autres programmes de reconnaissance de texte manuscrit, cela n'est pas transparent dans le cas de Transkribus.

29 Les problèmes rencontrés lors de l'évaluation automatique concernent principalement la détection des lignes, surtout lorsqu'elle est corrigée ultérieurement ou lorsqu'il y a des améliorations dans l'ordre des lignes. Par exemple, les lignes peuvent être difficile à identifier si l'une d'entre elles est divisée en deux et que les deux éléments sont fusionnés, si une note est ajoutée au-dessus de la ligne, ou si une note en marge est séparée du texte principal après coup. Dans ces cas, l'ensemble du texte se déplace, ce qui entraîne également des parties considérées comme incorrectes ou affichées comme telles, bien qu'elles soient au moins partiellement correctes. En effet, Transkribus prétend attribuer une erreur à chaque caractère mal reconnu.

Fig. 4. Les deux captures d'écran montrent le résultat de la transcription automatique le plus réussi (à gauche) et le plus mauvais (à droite) [Des corrections ultérieures n'ont pas pu être prises en compte ici]



30 Dans cette perspective, un problème supplémentaire d'évaluation avec Transkribus est lié au fait que le système se base sur les mots et non sur les caractères individuels. Pour un observateur extérieur, les résultats semblent donc plutôt moyens, même s'ils sont en réalité nettement meilleurs. On le comprend lorsque l'on compare l'évaluation de l'écriture manuscrite de Wagener au niveau des mots et des caractères : au niveau des mots, le taux d'erreur est généralement plus élevé qu'au niveau des caractères.

31 Étant donné que Transkribus fournit l'évaluation automatique de la transcription au niveau des caractères et alors que pour les mots il le fait uniquement en pourcentages, nous avons effectué une évaluation manuelle de la transcription automatisée et nous avons compté les pages. Lors de l'évaluation manuelle de la transcription, nous avons veillé à compter toutes les reconnaissances divergentes. L'objectif était de découvrir et de comprendre la fiabilité de l'évaluation automatisée de Transkribus. Ce qui est étonnant, c'est qu'apparaissent des écarts significatifs dans les pourcentages. Ainsi, la somme des valeurs, que ce soit pour le taux d'erreur de mots (WER) et la précision des mots, ou pour le taux d'erreur de caractères (CER) et la précision des caractères, n'a pas atteint les 100 %. Il est

également frappant de noter que les valeurs lors de l'évaluation du taux d'erreur de mots (WER) dans l'évaluation automatisée correspondent approximativement à mes propres calculs ; cependant, au niveau des caractères, Transkribus présente un taux d'erreur de caractères de 41,66 %, alors qu'après le calcul manuel, seulement près de 17 % des caractères ont été mal lus. De plus, le résultat global est meilleur que ce que Transkribus avait prévu. Ainsi, pour l'écriture manuscrite de Wagener, nous avons atteint une précision d'à peine 58 % au niveau des mots mais à plus de 80 % au niveau des caractères. En tenant compte du faible nombre de pages de l'écriture manuscrite de Wagener sur lesquelles Transkribus a pu apprendre, le résultat obtenu ici est déjà bon, bien qu'il reste possible et souhaitable de l'améliorer davantage.

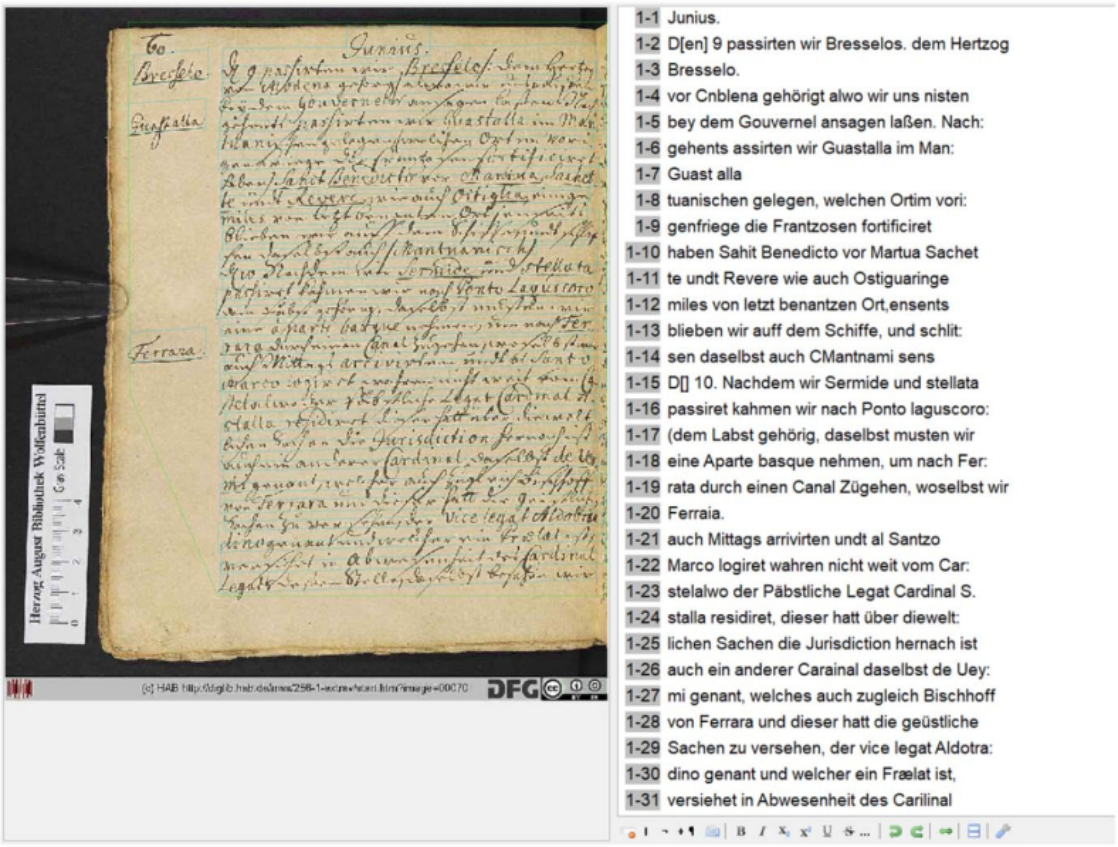
32 En plus de l'affichage de la fréquence des erreurs, une évaluation des pages manuscrites individuelles au niveau des mots et des caractères offre un aperçu plus précis du déroulement et des résultats des différentes itérations d'entraînement du modèle. Le nouveau cycle d'entraînement du modèle a-t-il été bénéfique ? Comment le résultat de la transcription se compare-t-il avec le taux d'erreur prédéfini dans l'ensemble d'entraînement et de validation ?

33 Lorsqu'on examine les résultats de la transcription partiellement automatisée, une analyse plus approfondie des erreurs survenues (types d'erreurs) est recommandée. Cela inclut notamment des lectures complètement erronées de mots et parfois même de lignes entières, qui déformaient le contenu ou le rendaient méconnaissable, rendant ainsi la lecture du manuscrit plus difficile. Comme mentionné précédemment, ces erreurs dépendaient fortement des pages et étaient causées par l'apparence de l'écriture, l'encre détrempée, des ajouts de mots isolés ou des ratures qui n'étaient généralement pas reconnus par le programme. Sur certaines pages, les erreurs étaient également causées par des petits morceaux de papier mal reliés, qui ne posaient généralement pas de problème à la reconnaissance des champs de texte par le logiciel, mais qui compliquaient la reconnaissance des caractères [21]. Dans l'ensemble, des lignes entières étaient souvent mal lues, transcrivant une séquence de lettres incompréhensible, tandis que des lignes entièrement lues sans erreur étaient plutôt l'exception.

34 En général, les erreurs concernaient souvent des lettres individuelles (notamment la confusion entre u/v, s/f, m/n, c/e, b/l, etc.), des signes de ponctuation (virgules manquantes ou mal lues, ponctuations, parenthèses) ou des espaces mal lus, trop peu ou trop nombreux. Certaines lettres ont été lues en double, d'autres ont été ignorées ; la quantité de caractères dans la version corrigée ne correspondait pas toujours à celle de la version lue automatiquement. Parfois, des signes de ponctuation ont été identifiés comme des lettres et vice versa, parfois l'ordre a été inversé ou mélangé, par exemple, « *wire* » au lieu de « *wier* ». Certaines erreurs étaient malheureusement dues uniquement à des espaces manquants, supplémentaires ou inutiles. Moins graves sont également les erreurs liées aux accents et aux trémas. Il est à noter que le logiciel n'a presque pas reconnu les soulignements, les mises en évidence, les ratures, les élévations et les indices de caractères individuels. Cela peut cependant être dû au fait que les zones de texte précédemment marquées n'ont pas nécessairement été incluses dans l'entraînement du modèle et la tâche de reconnaissance manuscrite partiellement automatisée en tant que telle. En ce qui concerne les caractères spéciaux, le programme a appris au fil de l'ajustement du modèle et s'est amélioré. Les caractères étaient généralement bien lus (par exemple, dans les indications de quantités) et seulement quelques erreurs sont survenues. Dans l'ensemble, la reconnaissance des dates et des chiffres généraux était défectueuse.

35 Le modèle a pu être optimisé au fil de son utilisation ultérieure sur des manuscrits subséquents, atteignant jusqu'à présent un taux d'erreur de 0,70 % dans l'ensemble d'entraînement et 2,60 % dans l'ensemble de validation.

Fig. 5. La capture d'écran montre comme exemple le résultat de la transcription automatique du manuscrit HAB Cod. Guelf. 256.1 Extrait à l'aide de l'application du modèle de Wagener.



36 Cependant, Transkribus atteint rapidement ses limites. Des ajustements échouent lorsque le programme ne peut plus améliorer la reconnaissance d'une écriture manuscrite. Les entraînements n'aboutissent alors pas : dans les ensembles d'entraînement et de validation, apparaissent des taux d'erreur inexplicables et élevés, dépassant parfois les 80 %, voire les 90 %. Lorsque de tels échecs surviennent, aucune amélioration supplémentaire n'est possible et il faut revenir au modèle précédemment entraîné ayant obtenu les meilleurs résultats.

À propos des possibilités et limites de Transkribus

37 Comme le montre l'exemple du manuscrit de Wagener, le succès d'une reconnaissance automatisée de l'écriture manuscrite dépend fortement de chaque page et de la ligne transcrite. De plus, la lisibilité et la qualité du scan du manuscrit, ainsi que la présence de ratures et d'ajouts (notamment entre deux lignes), sont d'une importance particulière pour le succès de la reconnaissance automatisée de l'écriture manuscrite et peuvent l'influencer négativement. Transkribus rencontre peu de problèmes de césure de mots, de distinction entre majuscules et minuscules, de ponctuation, ainsi que de reconnaissance de caractères spéciaux et d'abréviations. Cependant, des difficultés surviennent lors de la détection d'espaces et de l'apprentissage de différentes orthographes pour les mêmes mots (*vnd/und*, *aus/auß*…). Les noms propres (notamment les lieux et les personnes), les chiffres et les indications de date sont particulièrement sujets aux erreurs. La question des règles de transcription prédéfinies est alors cruciale. Pour la reconnaissance de l'écriture manuscrite, l'idéal est de rester aussi proche que possible de l'original. Dans le cas de Wagener, cela concerne, par exemple, la distinction entre les lettres I et J, la majuscule et la minuscule, ainsi que la ponctuation. Si des corrections doivent être apportées dans le cadre d'une édition critique, il est recommandé de les ajouter à une étape ultérieure pour éviter de risquer d'entraîner des erreurs dans Transkribus.

38 L'application antérieure du modèle Wagener sur d'autres manuscrits a permis une amélioration du modèle. Cependant, l'IA atteint rapidement ses limites, surentraînant le modèle et entraînant des taux d'erreur inexplicablement élevés dans les ensembles d'entraînement et de validation. Une amélioration supplémentaire du modèle précédent n'est donc plus possible. La question se pose alors de savoir dans quelle mesure ce surentraînement est spécifique au programme Transkribus ou s'il peut être attribué de manière générale à tous les systèmes de transcription manuscrite automatisée (HTR) pilotés par une IA.

39 Transkribus semble avoir joué un rôle de pionnier au cours de ces dernières années, mais il ne constitue pas la seule possibilité d'exploiter des documents historiques grâce à l'IA. Comme l'a déjà souligné Elpida Perdiki, il existe plusieurs alternatives au programme de reconnaissance manuscrite Transkribus. Outre l'application open-source eScriptorium, elle en mentionne d'autres : « python systems implémentés avec la bibliothèque TensorFlow, tels que Kraken, un système OCR pour les documents historiques, et Tesseract, le moteur OCR développé par Google et principalement utilisé dans de nombreux projets [22] ». Il serait autant plus intéressant dans cette perspective, de faire une étude comparative entre ces différents outils d'HTR à partir d'une source manuscrite exemplaire.

NOTES

1 Le présent article contient des observations déjà partagées partiellement en langue allemande sur le blog accompagnant le projet, https://grandtourdig.hypotheses.org qui ont été présentement rassemblées, traduites et complétées par des recherches supplémentaires entreprises depuis. Pour lire les rapports et témoignages périodiquement partagés sur le blog, voir notamment les deux billets suivants : Angela Göbel, « Erste Schritte in Transkribus (1) – Modelltraining und Transkription der Handschrift Wagener », in : *Grand Tour digital. Ein Forschungsblog zum Editionsprojekt frühneuzeitlicher Selbstzeugnisse an der Herzog August Bibliothek Wolfenbüttel*, mise en ligne le 6 juillet 2023 (https://grandtourdig.hypotheses.org/290), et Angela Göbel, « Erste Schritte in Transkribus (2) – Auswertung der Transkription der Handschrift Wagener », in : *Grand Tour digital. Ein Forschungsblog zum Editionsprojekt frühneuzeitlicher Selbstzeugnisse an der Herzog August Bibliothek Wolfenbüttel*, mise en ligne le 20 décembre 2023 (https://grandtourdig.hypotheses.org/1022). Il s'agit ici de deux billets de blog qui discutent de plus près de l'application et de l'évaluation de l'outil Transkribus à partir de l'exemple du premier manuscrit traité. Le présent article propose un aperçu de nos recherches courantes et vise à mettre à disposition du public francophone les premiers résultats de recherche.

2 Grand Tour digital. Numérisation, exploration et visualisation de témoignages personnels de voyages éducatifs de l'époque moderne en utilisant des procédés d'édition partiellement automatisés

3 Voir Mathis Leibetseder, « Kavalierstour – Bildungsreise – Grand Tour: Reisen, Bildung und Wissenserwerb in der Frühen Neuzeit », dans Leibniz-Institut für Europäische Geschichte (IEG) Hrsg., *Europäische Geschichte Online* (EGO), mis en ligne le 14 août 2013, URL : http://www.ieg-ego.eu/leibetsederm-2013-de (consulté le 15 janvier 2024).

4 Voir le site du portail intitulé « Selbstzeugnisse der Frühen Neuzeit in der Herzog August Bibliothek », en ligne : https://selbstzeugnisse.hab.de. Le projet « Grand Tour digital » fait suite à deux projets d'édition précédents, déjà publiés sur le portail des témoignages personnels et disponibles en libre réutilisation. Voir Jacqueline Krone, David Maus et Inga Hanna Ralle (dir.), « Herzog August der Jüngere von Braunschweig-Wolfenbüttel (1579–1666): Ephemerides. Sive Diarium (1594–1635) », dans *Selbstzeugnisse der Frühen Neuzeit in der Herzog August Bibliothek*, Wolfenbüttel, HAB, 2017, URL : http://selbstzeugnisse.hab.de/edition_august, et Andreas Herz, Jan-Hendrik Hütten, Alexander Zirr et al. (dir.), « Digitale Edition der Tagebücher von Herzog Ludwig Rudolf und Herzogin Christine Luise von Braunschweig-Wolfenbüttel », dans*Selbstzeugnisse der Frühen Neuzeit in der Herzog August Bibliothek*, Wolfenbüttel, HAB, 2019–2022, URL : http://selbstzeugnisse.hab.de/edition_sz2.

5 https://grandtourdig.hypotheses.org/

6 https://git.hab.de/forschungsdaten

7 À côté de l'original, trois autres copies de ce récit de voyage sont conservées dans les collections de la Bibliothèque Herzog August (HAB) et des Archives d'État de Basse-Saxe à Wolfenbüttel. Ces textes font également l'objet de travaux de recherche dans le cadre du projet, visant d'une part à regrouper les collections dispersées, et d'autre part, à disposer d'un corpus de sources identique ou très similaire en tant que « groupe de contrôle » pour l'indexation semi-automatisée des textes avec Transkribus. Ces copies sont conservées dans HAB Cod. Guelf. 256.1 Extrav. ; Nds. LA WF, 1 Alt 20 Nr. 94 ; et Nds. LA WF, 95 ALT Nr. 44.

8 Une grande partie est déjà accessible en format numérisé via la base de données des manuscrits de la HAB, les entrées respectives dans le portail des manuscrits et le portail des témoignages personnels sont actuellement en cours de traitement et seront préparées d'ici la fin du projet. Voir « Auswahl nach Projekt: Grand Tour digital Digitalisierung, Erschließung und Visualisierung frühneuzeitlicher Selbstzeugnisse von Bildungsreisen unter Anwendung teilautomatisierter Editionsverfahren (21 Hss.) », dans

Handschriftendatenbank, URL : https://diglib.hab.de/?db=mss. Pour y accéder cliquez « Auswahl nach … », puis « Projekt », puis sélectionnez le présent projet. Une présentation plus détaillée du corpus est à trouver sur le blog accompagnant le projet, voir https://grandtourdig.hypotheses.org/uber.

9 Dans sa thèse de doctorat, Joe Nockels, par exemple, a étudié à l'Université d'Édimbourg l'utilisation du logiciel de transcription manuscrite Transkribus dans la recherche entre 2015 et 2020. Voir Joe Nockels et al., « Understanding the application of handwritten text recognition technology in heritage contexts: a systematic review of Transkribus in published research », *Archival Science* 22 (2022), p. 367–392.

10 Si l'on souhaite transcrire un manuscrit complet avec un modèle existant sans effectuer un nouvel entraînement, et que l'on a du mal à choisir parmi la multitude de modèles possibles, Transkribus offre la possibilité de la comparaison de motifs.

11 Le « centre de ressources » de Transkribus propose des informations détaillées, fournissant plusieurs guides pour l'utilisation du logiciel, voir URL : https://readcoop.eu/transkribus/resources/.

12 Voir par exemple le billet de blog présentant l'outil par Jan Horstmann et la littérature complémentaire répertoriée dans celui-ci : Jan Horstmann, « Transkribus », in : forText. *Literatur digital erforschen*, mis en ligne en 2018, URL: https://fortext.net/tools/tools/transkribus (consulté le 15 janvier 2024) ; Marc Rothballer, « Transkribus. Der Einsatz von maschinellem Lernen und Handwritten Text Recognition in der Erschließung historischer Dokumente », in : *FORUM. Zeitschrift des Bundes der öffentlichbestellten Vermessungsingenieure e.V.* 4 (2020), p. 29–35, URL : https://www.bdvi.de/application/files/9516/1459/2541/Transkribus_FORUM_4-2020.pdf (consulté le 15 janvier 2024), ainsi que le rapport d'expérience et d'atelier de Jacob Möhrke « Training eines Transkribus Sprachmodells », in : Begleitblog zum Regensburger Projekt DEHisRe – Digitale Editionen Historischer Reiseberichte, mis en ligne le 20 juillet 2020, https://dehisre.ios-regensburg.de/training-eines-transkribus-sprachmodells/ (consulté le 15 janvier 2024).

13 Le taux d'erreur de caractères (CER) compare, selon le glossaire de Transkribus, « pour une page donnée, le nombre total de caractères (n), espaces compris, au nombre minimum d'insertions (i), substitutions (s) et suppressions (d) de caractères) » (« for a given page, the total number of characters (n), including spaces, to the minimum number of insertions (i),

substitutions (s) and deletions (d) of characters »), nécessaires pour obtenir une correspondance à cent pour cent entre la reconnaissance manuscrite et la transcription correcte du document en question. La meilleure façon de former un nouveau modèle dans Transkribus est expliquée ici étape par étape. La page est régulièrement mise à jour et offre une bonne introduction à la formation de modèles personnalisés, qui ne sera pas répétée à cet endroit. Voir URL : https://readcoop.eu/glossary/character-error-rate-cer/.

14 Pour d'autres informations sur ce travail : Angela Göbel, Maximilian Görmar, « Zu Besuch in der Gottorfischen Kunstkammer. Berichte eines Apothekergesellen auf Reisen », *HABlog*, mis en ligne le 28 mars 2023. URL : https://www.hab.de/zu-besuch-in-der-gottorfischen-kunstkammer/ (consulté le 15 janvier 2024), ainsi que dans la recherche plus ancienne les articles d'Otto Hahne, « Erlebnisse des Apothekers Wagener auf seinen Wanderjahren und in Begleitung des Herzogs Ferdinand Albrechts I. zu Braunschweig und Lüneburg (1652–1659) », in : Braunschweigisches Magazin 33 (1927), p. 81–94, et ibid., « Die Reisen des Apothekers Wagener aus Itzehoe 1652–1659 », in : Pharmazeutische Zeitung 72 (1927), p. 951–952.

15 Il est important de noter que cet entraînement peut prendre un certain temps. En général, vous vous retrouverez dans une file d'attente et selon l'heure de la journée, cela peut prendre un peu plus de temps. Il est donc recommandé de prévoir toujours une demi-journée à une journée entière.

16 Voir à ce sujet Jakob Möhrke, *op. cit.*

17 Voir par exemple Elpida Perdiki, « Transkribus: Reviewing HTR training on (Greek) manuscripts », in : RIDE – A review journal for digital editions and resources 15 (2022), URL : https://ride.i-d-e.de/issues/issue-15/transkribus/ (consulté le 15 janvier 2024), ainsi que Ivan N. Petrov, Achim Rabus, « Linguistic Analysis of Church Slavonic Documents: A Mixed-Methods Approach », in : Scando-Slavica 69 (2023), N° 1, p. 25-38, URL : https://doi.org/10.1080/00806765.2023.2189617.

18 Gabriel Viehhauser, « Digitalisierung von Handschriften und frühen Drucken, OCR (Bericht über Kurzvorstellungen und Diskussionen Sektion 1) », dans Joachim Hamm, Albrecht Hausmann, Elisabeth Lienert et Gabriel Viehhauser (dir.) : *Digitale Mediävistik. Perspektiven der Digital Humanities für die Altgermanistik*, Oldenbourg 2022 (= BmE Themenheft 12), URL : https://doi.org/10.25619/BmE20223192 : « *Automatische Handschriftenerkennung wird selbst bei besttrainierten Modellen zu einer gewissen Fehlerrate in der Erkennung führen. Ohne nachträgliche manuelle*

Korrektur (post-processing) wird also kein philologisch einwandfreier Text entstehen ».

19 La façon de calculer la précision des modèles de reconnaissance manuscrite (HTR) est expliquée par les développeurs de Transkribus sur leur site : « En tant que "Référence", choisissez une version de page qui a été correctement transcrite (vérité de terrain : transcription manuelle aussi fidèle que possible au texte original). Pour obtenir la valeur la plus significative, il serait préférable d'utiliser des pages d'un ensemble d'échantillons qui n'ont pas été utilisées dans l'entraînement et qui sont donc nouvelles pour le modèle. » (« As '*Reference*', *choose a page version, which was correctly transcribed (Ground Truth: manual transcription as close to the original text as possible). To get out the most significant value it would be best to use pages from a sample set which have not been used in the training and therefore are new to the model* »). URL : https://readcoop.eu/glossary/compute-accuracy/.

20 Voir *ibid.* : « En tant que "Hypothèse", choisissez la version qui a été générée automatiquement avec un modèle HTR et sur laquelle vous aimeriez voir à quel point le résultat est satisfaisant » (« As '*Hypothesis*', *choose the version, which was automatically generated with an HTR-model and on which you would like to see, how good the result is* »).

21 Étant donné que cela ne concernait qu'un petit nombre de pages et de lignes de texte, l'effort n'a pas été fait pour extraire les pages concernées du fichier PDF généré, corriger leur orientation, et les réintégrer dans Transkribus. Cela aurait pris plus de temps que la simple correction des pages en question. Malheureusement, à notre connaissance, Transkribus ne propose pas actuellement de solution directe à ce problème dans le programme. La possibilité de rotation partielle de la vue du manuscrit ne semble pas avoir d'impact sur la reconnaissance de l'écriture manuscrite en elle-même et ne concerne malheureusement pas seulement une page. En redressant les pages scannées de travers, toutes les pages sont automatiquement redressées.

22 Elpida Perdiki, *op cit.* : « *python sytems implented with TensorFlow library, [..] Kraken, an OCR system for historical documents, and [...] Tesseract, the OCR engine developed by Google and mostly used in many projects* ». On peut utiliser Kraken, mentionné par Perdiki, soit via eScriptorium, où il est intégré, soit individuellement.

ABSTRACTS

Français

Le projet « Grand Tour digital » vise à numériser, explorer et visualiser des témoignages personnels de voyages éducatifs de l'époque moderne, mettant l'accent sur cinq textes principaux. Le projet vise à développer expérimentalement un processus éditorial novateur en utilisant Transkribus. Financé par la Deutsche Forschungsgemeinschaft (DFG) de 2022 à 2025, le projet utilise la plateforme Transkribus pour la transcription automatique de manuscrits, combinée à la reconnaissance d'entités nommées (REN) pour l'exploration et la visualisation des témoignages. Le corpus comprend 21 journaux de voyage, rédigés entre 1550 et 1770, conservés à la Bibliothèque Herzog August à Wolfenbüttel (HAB), avec une variété de voyages couvrant l'Europe, l'Empire ottoman et le Proche-Orient. La reconnaissance automatique des récits de voyage réalisée avec Transkribus fait face à des défis tels que la variété de l'écriture, l'individualité de chaque écriture, la mise en page complexe, les polices variées, l'utilisation de différentes langues et les exigences de contenu. Au cours du projet sera développé un modèle adapté à ces exigences. Le logiciel fournit des instructions utiles pour l'édition de texte, et la recherche génère des rapports d'expérience, des discussions scientifiques et des conseils pratiques sur l'outil. La transcription manuelle des pages du journal de voyage de Wagener dans Transkribus a permis d'entraîner un nouveau modèle basé sur le modèle existant « Transkribus German handwriting M1 » de l'Université de Greifswald. Ce modèle a été ajusté progressivement en transcrivant partiellement automatiquement, en corrigeant, et en réentraînant. L'objectif était d'adapter le programme aux particularités de l'écriture de Wagener, réduisant ainsi le taux d'erreur dans la reconnaissance manuscrite. Des ajustements manuels ont été apportés au marquage des champs de texte pour minimiser les erreurs de détection. L'évaluation du modèle a montré des taux d'erreur fluctuants, atteignant 2,41 % dans l'ensemble d'entraînement et 11 % dans l'ensemble de validation lors de la première itération. L'optimisation du modèle a continué avec des ajustements itératifs, élargissant le jeu de caractères entraîné. Les résultats montrent des améliorations, bien que des questions subsistent sur la meilleure façon de traiter les pages « non utilisables ». L'évaluation automatique dans Transkribus s'est concentrée sur le taux d'erreur (CER) et la précision des mots et des caractères, révélant des écarts significatifs entre les évaluations automatiques et manuelles. Malgré des améliorations dans le modèle, Transkribus atteint ses limites, avec des échecs d'ajustement conduisant à des taux d'erreur élevés. L'étude soulève des questions sur la spécificité de ces limites à Transkribus par rapport à d'autres systèmes de transcription automatisée. L'auteure suggère également d'explorer d'autres outils d'HTR tels que Kraken et Tesseract

pour une comparaison approfondie. En conclusion, le succès de la reconnaissance automatisée dépend fortement de la qualité des pages, de la lisibilité du manuscrit, et de la présence de ratures. Bien que Transkribus ait joué un rôle pionnier, d'autres alternatives méritent une étude comparative pour évaluer les performances des différents outils d'HTR.

English

The "Grand Tour digital" project aims to digitize, explore, and visualize personal accounts of educational journeys from the modern era, focusing on five main texts. The project aims to experimentally develop an innovative editorial process using Transkribus. Funded by the Deutsche Forschungsgemeinschaft (DFG) from 2022 to 2025, the project utilizes the Transkribus platform for the automatic transcription of manuscripts, combined with Named Entity Recognition (NER) for the exploration and visualization of testimonies. The corpus consists of 21 travel journals written between 1550 and 1770, held at the Herzog August Library in Wolfenbüttel (HAB), covering various journeys across Europe, the Ottoman Empire, and the Middle East. Automatic recognition of travel narratives, especially with Transkribus, faces challenges such as writing variety, individuality of each script, complex layout, diverse fonts, use of different languages, and content requirements. A model adapted to these requirements will be developed during the project. The software provides useful instructions for text editing, and research generates experience reports, scholarly discussions, and practical advice on the tool. Manual transcription of Wagener's travel journal pages in Transkribus was used to train a new model based on the existing "Transkribus German handwriting M1" model from the University of Greifswald. This model was gradually adjusted by partially automatic transcription, correction, and retraining. The goal was to adapt the program to Wagener's writing characteristics, thus reducing the error rate in handwriting recognition. Manual adjustments were made to text field labeling to minimize detection errors. Model evaluation showed fluctuating error rates, reaching 2.41% in the training set and 11% in the validation set during the first iteration. Model optimization continued with iterative adjustments, expanding the trained character set. Results show improvements, although questions remain about the best way to handle "unusable" pages. Automatic evaluation in Transkribus focused on Character Error Rate (CER) and word and character accuracy, revealing significant discrepancies between automatic and manual evaluations. Despite improvements in the model, Transkribus reaches its limits, with adjustment failures leading to high error rates. The study raises questions about the specificity of these limits in Transkribus compared to other automated transcription systems. The author also suggests exploring other HTR tools such as Kraken and Tesseract for a comprehensive comparison. In conclusion, the success of automated recognition heavily depends on page quality, manuscript readability, and the presence of erasures. Although Transkribus has played a pioneering role, other alternatives deserve a comparative study to assess the performance of different HTR tools.

INDEX

Mots-clés
Transkribus, reconnaissance d'écriture automatique, HTR, Grand Tour, récits de voyage, édition en ligne, intelligence artificielle

Keywords
Transkribus, handwritten text recognition, HTR, Grand Tour, travelogues, online editions, artificial intelligence

AUTHOR

Angela Göbel
Université Jean Moulin Lyon 3, LARHRA UMR 5190 Herzog August Bibliothek Wolfenbüttel
IDREF : https://www.idref.fr/28172489X