Transcription assistée par reconnaissance optique avec Transkribus

Régis Schlagdenhauffen

Transcription assistée par reconnaissance optique avec *Transkribus* : L'expérience du journal intime d'Eugène Wilhelm (1885-1951)

Régis Schlagdenhauffen[1]

EHESS, Paris, France

* regis.schlagdenhauffen@ehess.fr

Résumé

Cet article propose de restituer une « expérience utilisateur » du logiciel Transkribus en contexte francophone. Il s'appuie sur le projet de transcription semi-automatisée du journal intime du juriste Eugène Wilhelm (1866-1951). Ce journal comporte deux défis principaux : le premier est lié à la durée de la rédaction, 66 années, qui engendre des variations dans la forme de l'écriture, cette dernière devenant de plus en plus « illisible » le temps passant. Le second défi est lié à l'emploi concomitant de deux alphabets ; romain pour tout ce qui relève du quotidien et grec pour le for privé.

L'expérience utilisateur restituée dans cette contribution s'articule autour de deux aspects. Dans un premier temps, après avoir présenté le projet et les spécificités liées à l'usage de l'outil, les principaux obstacles rencontrés et les solutions apportées pour y remédier seront synthétisés. Puis, je reviendrai sur l'expérience collaborative de transcription conduite avec des étudiants en salle de cours en présentant les difficultés observées et les solutions trouvées pour y remédier. En conclusion, je proposerai un bilan relatif à l'utilisation de ce logiciel d'HTR (Human Text Recognition) en contexte francophone et en situation d'enseignement.

INTRODUCTION

Cet article propose un retour réflexif sur une expérience de transcription collaborative conduite à l'Ecole des hautes études en sciences sociales (EHESS) depuis 2017, notamment dans le cadre d'un séminaire conduit avec des étudiants de Master[1]. Il s'inscrit dans le champ des recherches en humanités numériques [Mounier, 2018] et se fonde sur la notion d'expérience utilisateur.

En contexte francophone, [Massot *et al*., 2018] sont les premiers à avoir proposé une réflexion critique de l'utilisation de cette application dans le contexte de la transcription assistée des fiches de lecture de Michel Foucault. Suite à cela, [Perrin, 2019] a proposé un tutoriel relatif à la transcription collaborative des archives archéologiques du site de Bibracte qui restitue point par point l'expérience utilisateur appliquée à cet outil[2].

L'expérience utilisateur (UX) désigne l'ensemble des perceptions, interactions et ressentis qu'un utilisateur éprouve vis-à-vis d'un produit ou d'un service avant, pendant et après son utilisation [Christine, Trognot, 2015]. Appliquée aux logiciels de reconnaissance d'écriture elle peut soit se référer aux imprimés soit aux manuscrits. Dans le premier cas, il s'agit de logiciels dits OCR (*Optical characters recognition*) dont l'usage est désormais développé dans de

[1] Article publié en anglais : Régis Schlagdenhauffen. Optical Recognition Assisted Transcription with Transkribus: The Experiment concerning Eugène Wilhelm's Personal Diary (1885-1951). *Journal of Data Mining and Digital Humanities*, Episciences.org, 2020, Atelier Digit_Hum. ⟨hal-02520508v3⟩

[2] https://f.hypotheses.org/wp-content/blogs.dir/7177/files/2019/11/Tutoriel_Transkribus_V2.pdf

nombreux domaines ; dans le second, il s'agit d'HTR (*Human text recognition*), champ en plein développement car devant relever des défis supplémentaires par rapport à l'OCR : singularité de chaque écriture humaine *vs*. standardisation des polices de caractères, variabilité des formes d'écriture, utilisation d'abréviation, présence de ratures, caviardage, etc. En outre, tandis que les logiciels d'OCR sont déjà relativement anciens [Schantz, 1982] ,les logiciels permettant d'automatiser tout ou partie du processus de transcription d'une écriture humaine restent rares. L'un d'eux, Transkribus, a été développé dans le cadre de deux programmes européens H2020 successifs par un consortium d'universitaires (Recognition and Enrichment of Archival Documents, READ project)[3]. Le logiciel, développé depuis 2016, permet au moyen d'un outil d'expert (*expert client*) de télécharger des documents et images au sein d'une collection privée mais partageable, de segmenter des images en blocs, puis lignes et mots à l'aide d'outils d'analyse de la mise en page ; de lier le texte à l'image, puis de transcrire le texte dans n'importe quelle langue avec n'importe quel jeu de caractères. Enfin, il permet d'exporter les documents transcris dans plusieurs formats (DOCX, TEI, RTF, PDF, XML).

La suite du propos sera organisée comme suit. Tout d'abord nous présenterons le projet TransDiary-TEI[4] et ses spécificités en regard de l'application Transkribus. Nous aborderons les questions liées à la spécificité du corpus exploité ainsi qu'à l'amélioration continue de l'utilisation de l'outil. Ensuite, nous reviendrons sur l'expérience utilisateur en problématisant les difficultés rencontrées. Suite à cela, nous dresserons un bilan de l'utilisation du logiciel Transkribus appliquée au contexte particulier du projet TransDiary-TEI.

I LE PROJET TRANSDIARY-TEI

1.1 Caractéristiques du journal d'Eugène Wilhelm

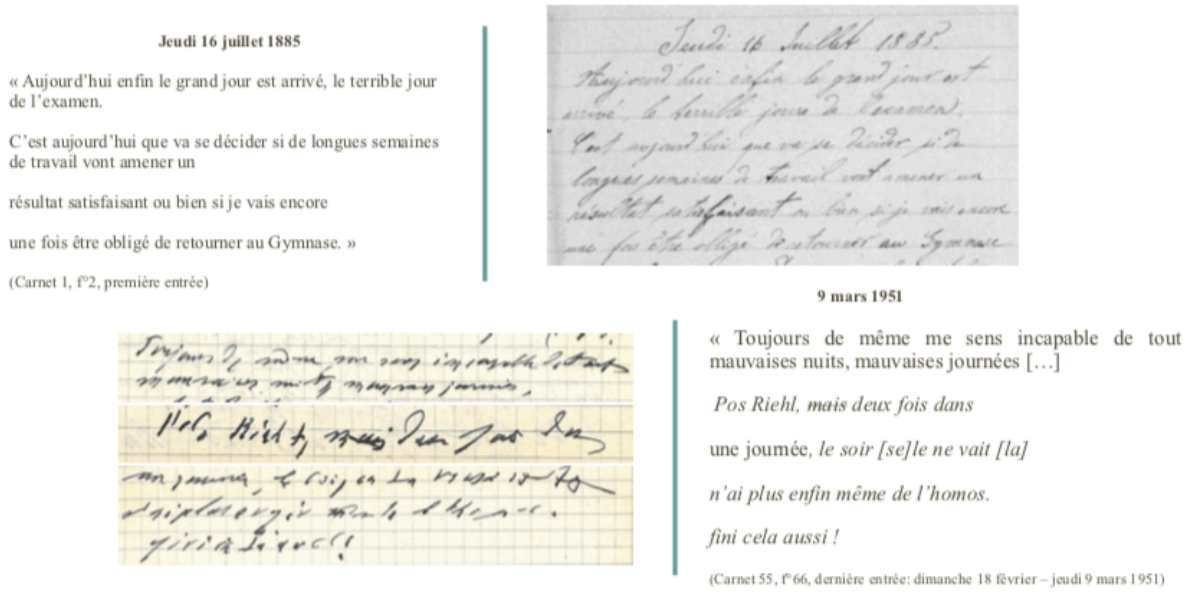
Le projet TransDiary-TEI a pour objectif la transcription et l'édition en TEI du journal intime du juriste Eugène Wilhelm (Strasbourg, 1866-1951). Ce dernier se présente sous la forme de 55 carnets numérotés, recouverts de moleskine pour la plupart, soit 8538 pages rédigées en français, mais émaillées de fragments en alphabet grec. Il a été tenu durant 66 années (des 19 aux 85 ans du diariste) avec une belle régularité. Il restitue des éléments contextuels sur sa carrière professionnelle, sa vie quotidienne, sa trajectoire intellectuelle et bien sûr la situation politique européenne entre 1885 et 1951 ainsi que le récit des relations sexuelles du diariste avec des hommes et des femmes de tous âges et de toutes conditions sociales. Les carnets sont de taille variable et recouvrent cinq grandes périodes historiques : le Kaiserreich, c'est-à-dire la période durant laquelle l'Alsace est administrée en tant que terre d'Empire allemande, soit les carnets 1 à 25 (3146 p.), la période correspondant à la Première Guerre mondiale, carnets 26 à 30 (906 p.), l'entre-deux-guerres, carnets 31 à 41 (2200 p.), la Seconde Guerre mondiale, carnets 42 à 51 (1416 p.) et l'après-guerre, carnets 52 à 55 (870 p.). Les spécificités du journal d'Eugène Wilhelm ont été présentées par [Dubout, 2016 et 2018], [Dubout et Schlagdenhauffen, 2014] et [Schlagdenhauffen, 2014 et 2015] : la présence de résumés ultérieurement rédigés par le diariste au début de chaque carnet, la présence de bilans annuels, ainsi que le choix d'employer deux alphabets distincts représentant respectivement les ¾ du journal pour l'alphabet romain et le ¼ pour l'alphabet grec tout en étant toujours tenu en langue française.

[3] https://read.transkribus.eu
[4] https://cahier.hypotheses.org/transdiary-tei

Afin de rendre ce journal intime d'une extraordinaire richesse accessible au plus grand nombre, l'idée du projet TransDiary-TEI consiste à en proposer une édition numérique. Un tel projet comporte divers avantages notamment en raison de la taille du journal, de la durée de sa tenue, et de la possibilité offerte d'améliorer continuellement la qualité de la transcription tout en l'enrichissant de métadonnées. Ces deux derniers volets du projet concernent d'une part les entités nommées (noms propres, de personnes et de lieux[5]) qui peuvent porter à confusion, homonymie ou nécessiter des compléments d'information avec l'évolution de la recherche ; d'autre part des métadonnées relatives à des renvois à l'intérieur et à l'extérieur du corpus, permettant le suivi du parcours du diariste à travers le temps et l'espace, la cartographie des réseaux qui sont les siens, et la mise en relation avec des bases de données internationales ou sites spécialisés pour créer un univers global et digital autour d'Eugène Wilhelm. Enfin, une édition numérique permettra également de donner un accès à la matérialité des documents stockés dans l'entrepôt des données de la recherche de l'EHESS, *Didomena*[6]. Dans le cadre du projet TransDiary-TEI, nous transcrivons en équipe à l'EHESS le journal au moyen de l'outil Transkribus et c'est cette expérience qui va être relatée dans les parties qui suivent.

1.2 Spécificités de l'outil Transkribus appliqué au projet

Les journaux intimes constituent un champ encore peu exploré par les logiciels de reconnaissance d'écriture humaine (HTR) tout comme par la TEI - à l'inverse de lettres, registres ou de documents d'archives publiques. Aussi, dans le cas du journal intime d'Eugène Wilhelm, deux phénomènes sont à prendre compte simultanément : la durée d'écriture et sa variabilité. En effet, entre les 19 et les 85 ans du diariste, l'écriture connaît une évolution qui rend difficile pour l'heure l'établissement d'un modèle d'entraînement unique. En effet, tandis que l'écriture est régulière et pour ainsi dire « scolaire » à des débuts, elle se complexifie avec l'âge pour devenir difficilement lisible sur la fin de la vie du diariste (cf. Figure 1).

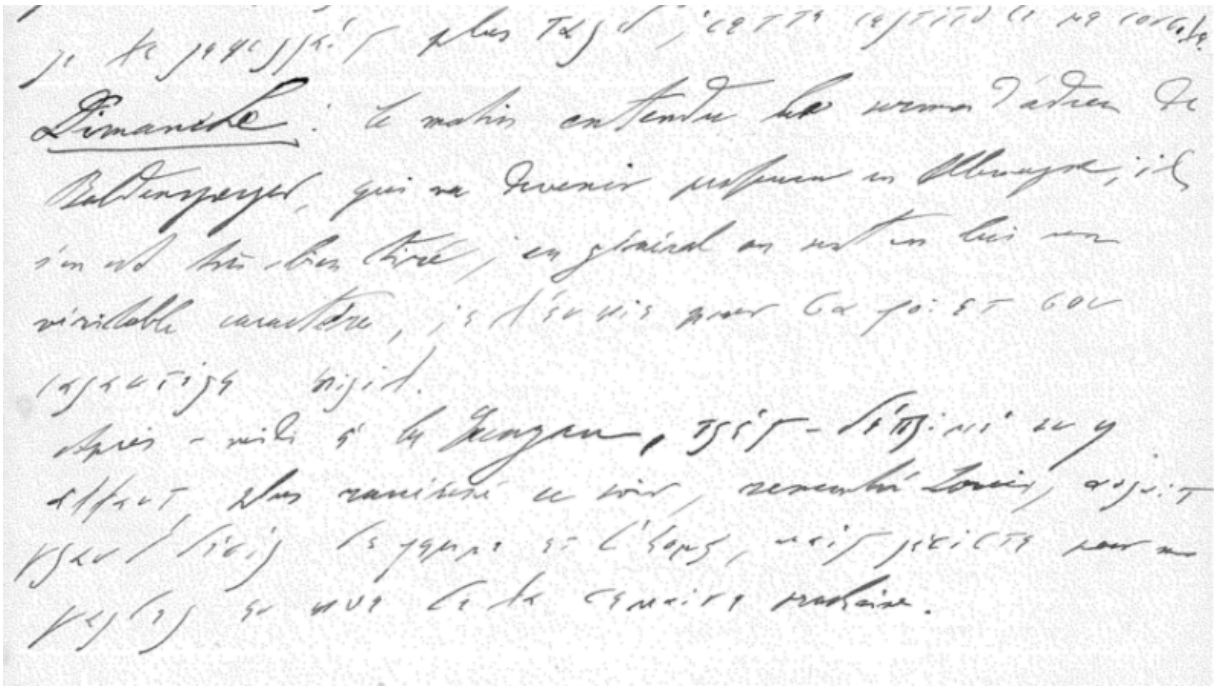


[Figure 1. *Variations de l'écriture d'Eugène Wilhelm, Comparaison carnet 1 (1885) et Carnet 55 (1951)*].

[5] *Cf.* par ex. Brando Carmen *et al.*, "Adaptation et évaluation de systèmes de reconnaissance et de résolution des entités nommées pour le cas de textes littéraires français du 19ème siècle". http://psig.huma-num.fr/HumaNS/wp-content/uploads/2018/10/8.Article-SAGeo_adaptation-et-évaluation-de-systèmes-de-REN-et-NEL.pdf/

[6] https://didomena.ehess.fr

À cela, s'ajoute l'usage quasi-alterné des alphabets latins et grecs (cf. Figure 2).



[Figure 2. *Alternance des alphabets grecs et latins, Carnet 10, 1890, f°6/68*]

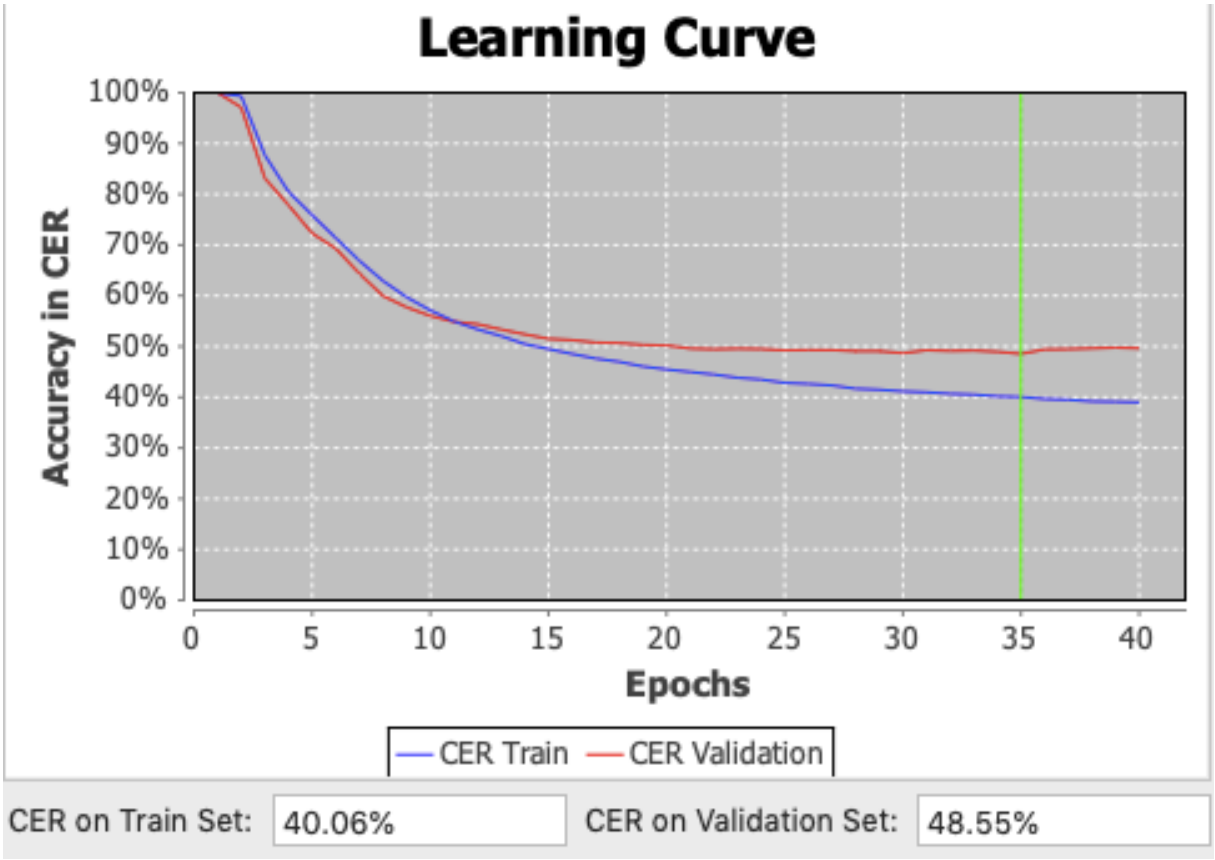
Sachant cela, les premiers essais réalisés en 2017 indiquaient un taux d'erreur élevé (soit 40 % en mode entraînement et 49 % en mode validation) que nous nous employons depuis à réduire, notamment au moyen de l'usage d'un dictionnaire idoine, mais aussi d'un entraînement plus efficace et attentif aux variations de l'écriture sur la longue durée.

Concernant un autre aspect du projet, relatif à l'encodage en TEI, quelques modèles existent déjà. Mais ils restent insatisfaisants comme le souligne [Nelson, 2017] ou le rappellent, pour l'espace francophone, [Soudan, 2012]. Cependant, en dehors de ces considérations – et comme dans le cas d'autres projets de transcription automatisée via un logiciel d'HTR –, il s'est d'abord agi de produire des transcriptions afin de pouvoir entraîner un modèle ainsi que le recommande le guide d'utilisation de Transkribus[7].

1.3 Utilisation et entraînement de l'HTR

Dans le but d'entraîner un modèle permettant la transcription automatisée nous avons tout d'abord livré une centaine de pages à l'équipe du consortium READ-Transkribus. Force est de constater qu'à ce stade, les résultats n'étaient pas probants puisque le taux d'erreur restait supérieur à 40 % à ce stade (cf. Figure 3).

[7] https://transkribus.eu/wiki/images/1/1d/Comment_utiliser_Transkribus_-_en_10_étapes_ou_moins_with_Screenshots.pdf



[Figure 3. *Modèle « Eugène Wilhelm t2i_M1 », réalisé le 02.02.2018*]

Selon le 1er modèle développé par l'équipe Transkribus, dit modèle « Eugene_Wilhelm-t2i_M1 » (cf. Figure 3), la reconnaissance automatisée restait laborieuse et le taux d'erreur trop élevé pour être satisfaisant. Toutefois dès la création de ce premier modèle, nous avons observé que l'HTR reconnaissait plus facilement les caractères grecs que romains. Cela s'explique notamment en raison du fait que les lettres écrites en grec ne sont pas liées mais détachées.

Une autre explication réside dans la qualité des numérisations. Celles ayant été réalisées en basse résolution et pour la plupart du temps en noir et blanc rendent la reconnaissance automatique des lignes plus laborieuse. À ce stade, la reconnaissance automatique des lignes de base a été abandonnée et nous lui avons préféré un traçage manuel des zones de texte et lignes, qui reste une solution efficace mais chronophage.

Non adossé à un dictionnaire, le premier modèle a montré les limites de la reconnaissance automatique d'une écriture humaine en français. En effet, comme l'ont souligné [Massot *et al.*, 2018] : « Transkribus n'utilise pas de dictionnaire et ne cherche pas à reconnaître des mots, mais analyse les lignes de texte *caractère par caractère* ». Aussi, les premières transcriptions automatiques, réalisées sans aucun dictionnaire transcrivaient une langue méconnue digne du manuscrit de Voynich !

Par conséquent, les effets attendus de la transcription automatisée ne se sont pas faits sentir durant cette première phase. L'intervention de Carmen Brando, docteure en informatique et responsable de la plateforme géomatique de l'EHESS associée aux conseils de Joachim Dornbusch, du Pôle numérique recherche de l'EHESS, fut à ce moment déterminante. Nous avons alors convenu, en partant des transcriptions réalisées précédemment de créer un dictionnaire d'Eugène Wilhelm permettant de faciliter la transcription automatisée en partant

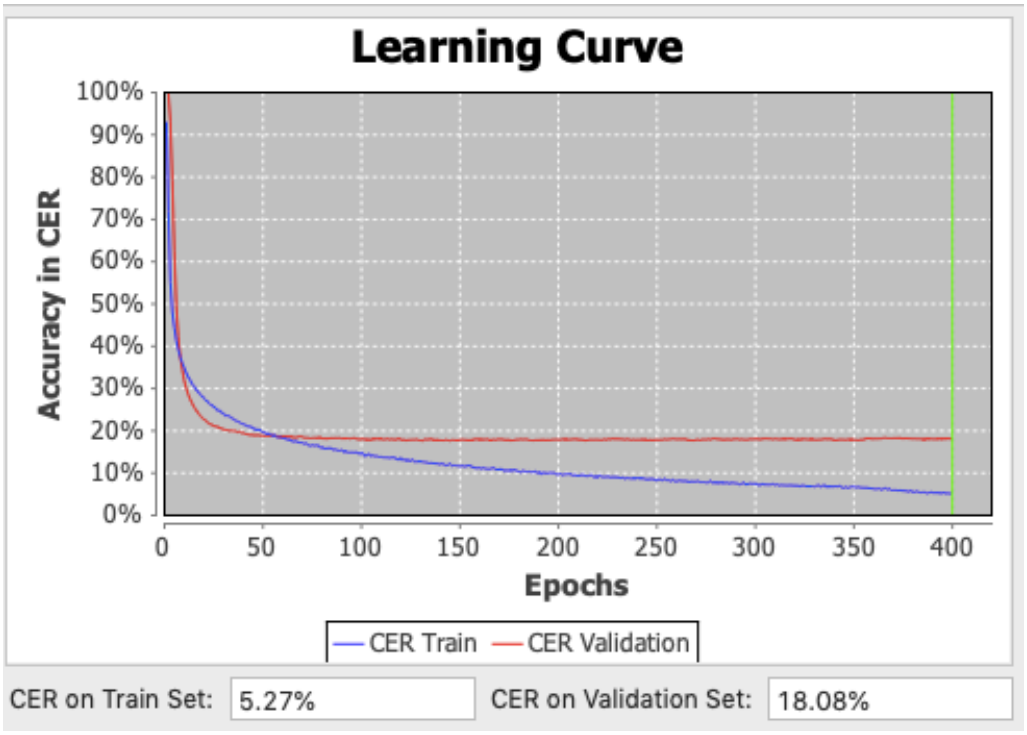
de l'hypothèse que l'utilisation d'un dictionnaire spécifique supposait que le vocabulaire de la partie transcrite soit plus ou moins homogène avec celui de la partie à transcrire. Ce choix fut par ailleurs recommandé par l'équipe de Transkribus qui considérait que « *there is one thing to take notice of when using a (general) dictionary: The bigger (in number of words) the dictionary is, the slower the recognition process will be. A general dictionary is therefore often not really usable*. »

La création du dictionnaire a consisté en un étiquetage morphosyntaxique au moyen d'un traitement sur TXM (lexique-mots.txt) permettant la production d'une liste de mots (lexique-occurrences.csv). En ce qui concerne la réalisation du processus à partir du vocabulaire employé par Eugène Wilhelm, nous sommes partis des textes déjà transcrits (environ 2000 pages sur plus de 8000). Pour ce faire, il a fallu effectuer trois traitements consécutifs. Premièrement, il s'est agi d'extraire une liste de mots des textes par un processus de segmentation et de tokenisation des textes ; deuxièmement, nous avons procédé à un processus de lemmatisation, de sorte que chaque mot soit associé à sa catégorie grammaticale ; troisièmement, nous avons fléchi chaque mot (hormis les mots vides et les noms propres) à partir des règles de la grammaire française.

Au final, le dictionnaire comportant plus de 34000 formes fléchies a été d'un puissant recours.

En faisant le choix d'un dictionnaire spécifique à ce corpus nous avions l'intuition qui s'est révélée exacte que l'outil gagnerait en efficacité et rapidité.

Et, en effet, c'est à partir de ce moment qu'il a été possible d'aboutir à la réalisation d'un nouveau modèle de transcription automatisée plus efficace dit modèle « EW_2+ ». Ce dernier possède un taux d'erreur de 5,27 % en mode entraînement et de 18 % en mode validation (cf. Figure 4). Ce dernier modèle, actuellement utilisé, a permis de considérablement améliorer la transcription à partir de là, montrant l'efficacité de l'outil et ouvrant la voie à une expérience collaborative de transcription satisfaisante.



[Figure 4. Modèle EW_2+, réalisé le 15.12.2018]

II LA TRANSCRIPTION COLLABORATIVE AVEC LES ETUDIANTS

Dans cette partie nous aimerions revenir sur l'expérience de transcription collaborative conduite avec des étudiants de Master de l'EHESS. Ces derniers n'ont pas forcément de connaissance préalable dans l'usage des logiciels de transcription au début du séminaire. Le séminaire de 24 h/semestre, soit 2 h hebdomadaires, se tient désormais depuis 3 années à l'EHESS. Son objectif est double : d'une part initier les étudiants à l'utilisation de l'outil, d'autre part, leur permettre de maîtriser toutes les étapes jusqu'à l'entraînement d'un modèle. Dans ce cadre, nous avons utilisé le journal d'Eugène Wilhelm comme support pour la réalisation des différentes étapes.

2.1 Les séquences pédagogiques

Afin de faciliter l'exercice nous avons débuté avec un des premiers carnets du corpus, le carnet n°3 (du 7 août 1886 au 5 juin 1887), qui a pour particularité, comme les 10 autres premiers d'être « lisible » pour ce qui est des parties rédigées en alphabet romain. S'agissant des parties rédigées en grec, elles restaient difficilement accessibles aux étudiants du fait de l'absence désormais obligatoire d'un enseignement de grec ancien dans les filières scolaires françaises. Aussi, l'enseignement se décline-t-il selon les séquences pédagogiques suivantes.

Première séquence pédagogique : après une présentation générale de l'interface, les étudiants sont invités à prendre part à une « collection » créée spécialement dans le cadre du séminaire et regroupant plusieurs carnets dédiés à l'entraînement manuel. En effet, l'ensemble du journal a déjà été téléchargé sur l'application dans le cadre du projet TransDiary-TEI. Cette phase consiste en une première familiarisation avec l'outil et le corpus.

La seconde étape de l'enseignement est dédiée à l'initiation manuelle au traçage de zones de textes puis de lignes de base. Une telle démarche permet d'apprendre ces opérations qui peuvent par la suite être automatisées mais néanmoins nécessiter des ajustements le cas échéant. La capacité à tracer manuellement les zones de texte et les lignes est d'une grande importance dans la mesure où elle permet de palier les manquements éventuels de l'outil.

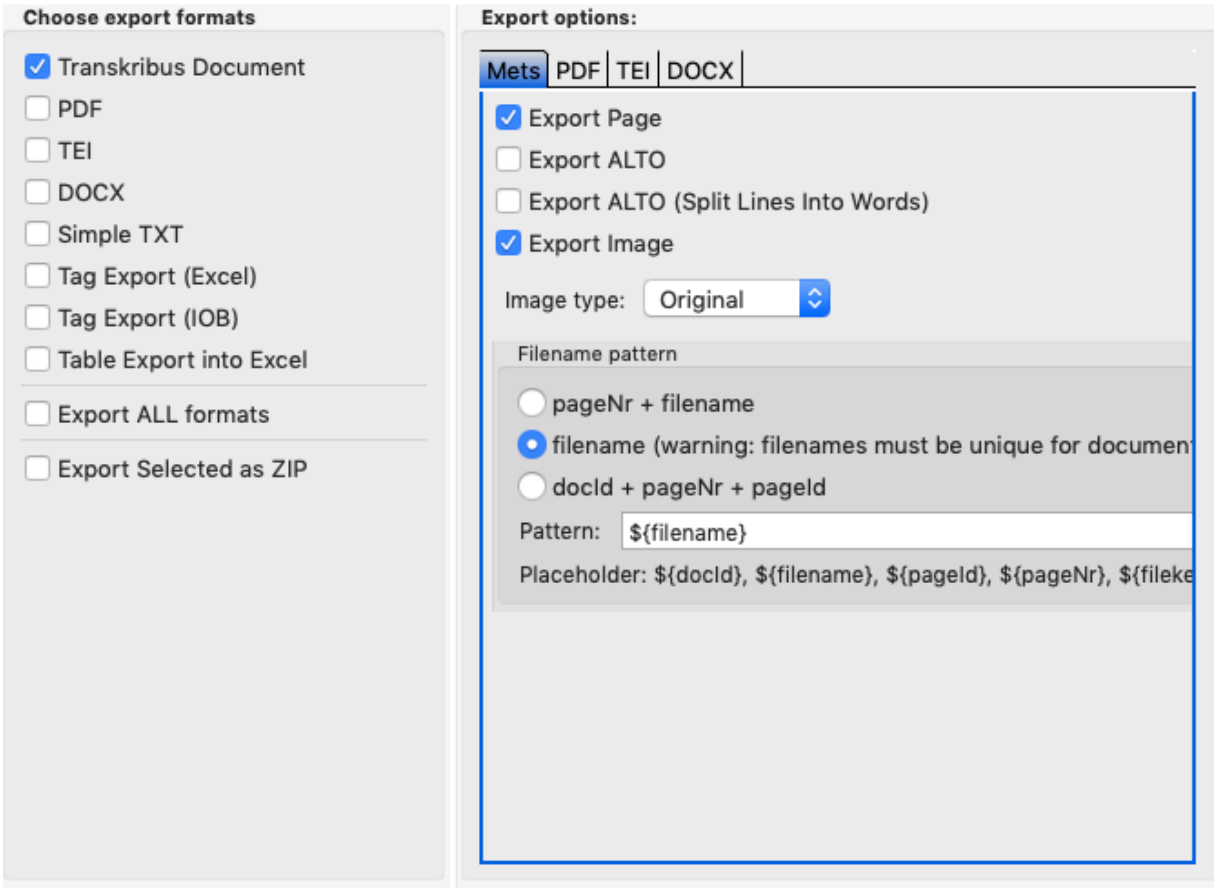
Une troisième étape consiste ensuite à l'automatisation des deux processus décrits précédemment puis à la réalisation des ajustements nécessaires (retraçage de certains lignes, ajout ou suppression de lignes non reconnues ou reconnues à tort par l'outil). Cette opération peut être facilitée par l'affichage dans Transkribus de la numérotation des lignes et/ou d'autres options permettant un ajustement optimal de l'interface en fonction des attentes du moment.

Une quatrième étape concerne l'approfondissement de l'habituation à l'écriture diaristique. Dans ce cadre, une initiation ou un rappel de l'alphabet grec vient en aide aux étudiants désireux de tout comprendre. Parallèlement, la lecture des pages du journal se fait collectivement dans le but de reconnaître l'écriture et de discuter de la meilleure interprétation de mots pouvant porter à confusion. Il s'agit à cet endroit principalement de noms propres (de lieux, de personnes) mais uniquement. Cette étape, qui peut sembler à la fois la plus simple car ne nécessitant pas de connaissances techniques est néanmoins aussi complexe car obligeant à une habituation. Elle s'est étalée sur plusieurs séances.

Une cinquième étape a ensuite consisté à employer un modèle HTR existant afin d'expérimenter la puissance de l'outil. Cette étape a permis d'accéder véritablement à la dimension collaborative de l'outil. Il s'agit aussi sans doute de l'une des étapes les plus stimulantes collectivement. Lors de chaque séance un étudiant avait en charge de diriger les

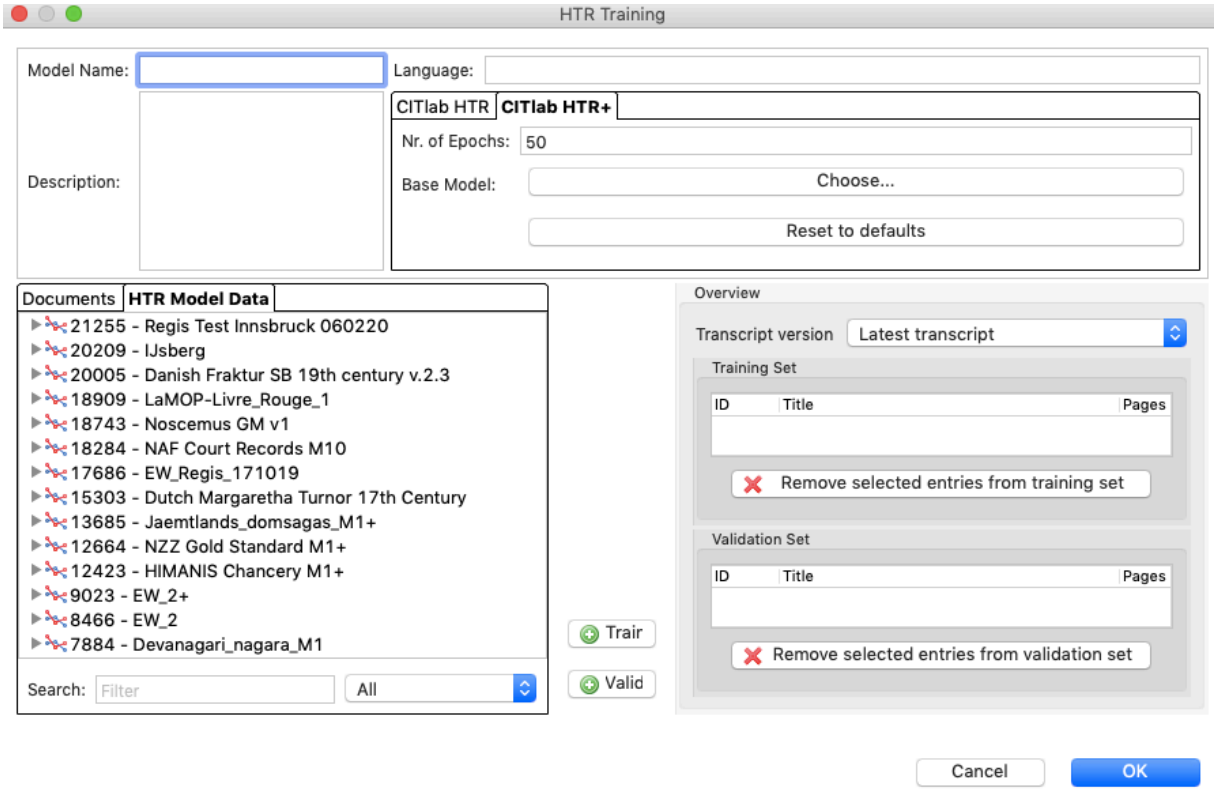
opérations réalisées sur la page en cours d'étude (en effet, seule une personne à la fois peut réaliser des modifications puis les enregistrer. Ceci constitue peut-être une des limites actuelles de l'outil dès lors que l'on travaille collectivement sur une même page). Après réalisation de la transcription automatisée, les opérations étaient de trois ordres. Premièrement, ajuster les lignes (rallonger ou raccourcir certaines lignes et/ou rajouter ou supprimer des lignes mal reconnues). Deuxièmement, il s'agissait de corriger les mots mal reconnus. Ces derniers, pouvaient relever de toutes catégories : verbes dont la conjugaison était défectueuse, adjectifs mal accordés, ou encore mots non-reconnus. C'est durant cette phase de reconnaissance et d'interprétation que nous discutons collectivement de la meilleure interprétation d'un mot difficilement lisible. Lorsqu'une hésitation persiste nous adjoignons une étiquette (tag) « unclear ». La question de l'étiquetage nous amène à prendre en considération la troisième opération qui est relative à l'enrichissement du texte en métadonnées. Par défaut, Transkribus propose un certain nombre de tags tels que : abréviation, adresse, date, lieu, organisation, personne, etc.

La sixième étape du séminaire consiste en l'utilisation des fonctions d'exportation de la transcription. Cette dernière est particulièrement intéressante dans la mesure où elle permet d'une part de sélectionner différents formats (docx, pdf, tei) d'autre part de sélectionner différentes options tel que les standards ALTO/METS, IOB, TEI, ou pour les exports de conserver les sauts de ligne et de page, marquer les mots « unclear » et d'une manière plus générale de conserver les tags sélectionnés, qui sont autant d'options destinées à faciliter l'édition numérique d'une source (cf. Figure 5).



[Figure 5. Formats d'export]

Enfin, une septième étape consiste en l'entraînement d'un modèle (cf. Figure 6)[8]. Pour l'heure et dans le cadre du séminaire, seul des entraînements de modèles dérivés des modèles déjà crées ont été réalisés. Ils peuvent être entraînés grâce à la collection de modèles de base disponibles permettant d'améliorer l'entraînement. Cette dernière étape a pour but d'initier les étudiants à la modélisation d'un entraînement mais n'a pas encore été réalisée avec des corpus autonomes.



[Figure 6. Interface d'entraînement de modèles]

Après cette présentation des différentes séquences pédagogiques d'un séminaire d'initiation à l'utlisation de l'outil Transkribus nous allons voir dans la partie qui suit les difficultés rencontrées et les solutions déployées pour y remédier.

2.2 Difficultés rencontrées et solutions déployées

Plusieurs difficultés ont été rencontrées jusqu'à présent. La première d'entre-elles est liée à la lisibilité générale de l'écriture du diariste. Tout comme dans le cas du projet de transcription des notes de Michel Foucault [Massot *et al.*, 2018], le diariste tronque certains mots, les abrège, ou plus généralement ne s'applique pas dans la tenue de son journal. Dans le contexte d'un enseignement de méthodologie appliqué au logiciel Transkribus cela pose un double défi. En effet, soit, il convient de travailler sur des parties du journal immédiatement lisibles (journal dit de jeunesse) auquel cas la puissance de l'outil n'est pas particulièrement perceptible ; soit il convient de travailler sur des entrées du journal plus difficile à décrypter par l'œil humain, auquel cas la puissance de l'outil peut être vérifiée, mais ceci rend plus complexe la vérification par un œil humain, qui plus est peu habitué à la singularité de l'écriture du diariste. A cela s'ajoute une autre singularité, l'usage de l'alphabet grec précédemment mentionné pour lequel

[8] Plus d'infos sur l'entraînement des modèles : http://regis-schlagdenhauffen.eu/wp-content/uploads/2019/10/Comment-entraîner-un-Modèle-dans-Transkribus.pdf

il a été convenu de transcrire les passages concernés en alphabet romain, en raison des limites actuelles des connaissances de l'alphabet grec.

Une autre difficulté est liée à la méconnaissance du contexte général, politique et historique, du diariste par les étudiants : par exemple l'époque du Reichsland pour l'Alsace (1871-1914) qui constitue la toile de fond du carnet n°3 nécessite un travail de contextualisation. Cette difficulté peut cependant être aussi considérée comme un atout car elle permet de passionnantes discussions en classe et l'acquisition par les étudiants de connaissances nouvelles sur un sujet peu étudié, tout du moins dans le contexte français. Il me semble d'ailleurs rétrospectivement que les discussions comptent parmi les moments les plus passionnants, qu'il s'agisse des événements relatés par le diariste que nous commentons, de la possibilité de le suivre, pas à pas, grâce à une cartographie des lieux qu'il a visités ou traversés, suivre ses déplacements, restituer leur logique.

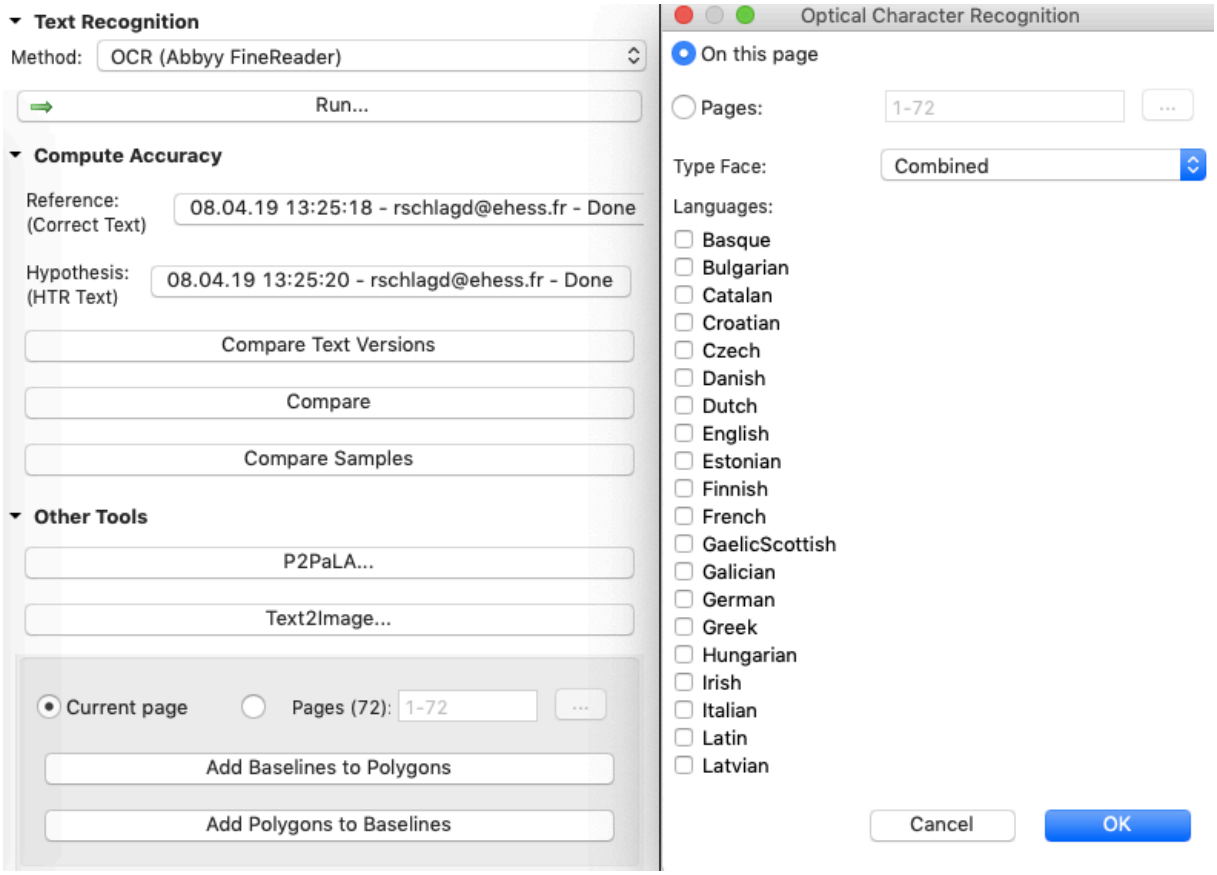
Au niveau de l'outil, la difficulté liée au monolinguisme de l'interface (en anglais) ne semble pas avoir joué de rôle déterminant. Dès lors que les différentes fonctionnalités ont été comprises et apprises, il est possible pour les étudiants de les tester, de les reproduire, et de comprendre *in fine* la logique générale qui prévaut à ce niveau d'utilisation de l'application.

A cela il convient encore d'ajouter les limites rencontrées par l'absence de dictionnaire intégré à Transkribus. Dans le cas du journal d'Eugène Wilhelm, comme nous l'avons souligné précédemment, les premières transcriptions automatisées étaient de mauvaise qualité notamment en raison de l'absence d'un dictionnaire. Pour y remédier, nous avons opté pour le choix suivant : créer un dictionnaire des mots employés par Eugène Wilhelm plutôt que d'injecter un dictionnaire de la langue française (*cf. supra*, partie 1.3)[9]. En effet, dans la mesure ou le vocabulaire de chaque individu reste limité – ou circonscrit – à certaines sphères, il n'y avait pas de sens d'adosser les transcriptions à un dictionnaire général comportant par exemple du vocabulaire issu des champs des sciences naturelles, de la médecine ou de l'artisanat qui sont autant de mots inutilisés dans le contexte d'écriture du journal intime d'un juriste.

Enfin, se pose bien entendu la question du rapport coût/bénéfice selon la taille du corpus. Dans notre cas, une partie du corpus avait fait l'objet d'une première transcription réalisée par Régis Schlagdenhauffen et Kevin Dubout principalement[10]. En effet, sur les plus de 8000 pages que comporte le corpus, environ 2000, soit ¼ du corpus, avaient fait l'objet d'une transcription préalable. Par conséquent, l'injection de données d'apprentissage dans l'application fut aisée. Et, dans un premier temps, un test a été réalisé par l'équipe Transkribus avec 200 images tirées du plusieurs carnets du journal. L'idée qui prévalait à ce moment était de restituer immédiatement les variations d'écriture survenues dans la tenue des carnets. L'usage d'une telle stratégie a permis d'accélérer considérablement la production d'un premier *training dataset*. Cependant, dans le cas d'étudiants qui souvent travaillent sur des corpus de moindre ampleur dans le cadre de la réalisation de leur mémoire, la question de l'investissement en temps et énergie reste entière. Aussi, paradoxalement, est-ce sans doute l'usage de l'outil OCR qui semble être dans l'immédiat l'outil le plus accessible pour des étudiants dès lors qu'ils travaillent aussi sur des corpus imprimés (cf. Figure 7). Ce dernier outil est en effet d'une grande facilité d'utilisation.

[9] L'auteur remercie Carmen Brando (EHESS) pour la réalisation de dictionnaire d'Eugène Wilhelm.

[10] L'auteur remercie notamment Kevin Dubout, Nicolas Eybalin et Sara Maïka pour leur collaboration durant les premières phases de transcription, tout comme Günter Hackl, Günter Mühlberger ainsi que Marie-Laurence Bonhomme et Carmen Brando pour les phases d'entraînement.



[Figure 7. Capture d'écran module OCR]

Comparé à d'autres outils, la fonction d'OCRisation de Transkribus possède tout d'abord comme avantage de pouvoir adosser les canevas et les transcriptions à une collection enregistrée sur le serveur. Ceci peut particulièrement être utile lorsque l'on travaille sur un périodique par exemple ou sur diverses publications imprimées d'un auteur. En outre, l'outil OCR propose les mêmes fonctionnalités d'exportation (variabilité des formats, prise en charge des étiquettes, balises, etc.). Par ailleurs, l'OCRisation étant une fonctionnalité désormais bien maîtrisée, les taux de reconnaissance sont élevés – et cela même sans l'utilisation de données d'entraînement préalabales ! Enfin, cet outil a le mérite, tout comme l'outil HTR de pouvoir facilement être réexploité dans le cadre d'une traduction semi-automatisée par exemple. Aussi, est-ce peut-être la possibilité de pouvoir transcrire conjointement des textes manuscrits et imprimés qui confère à Transkribus sa singulière puissance.

CONCLUSION

En guise de bilan, nous ne pouvons que confirmer certaines des limites observées par [Massot *et al.,* 2018] : « la transcription automatisée doit être reprise manuellement par des correcteurs, mais les résultats obtenus restent positifs car la transcription automatique permet une transcription manuelle plus rapide et aide dans la reconnaissance de certains mots ». En outre, malgré leur imperfection, les transcriptions automatiques sont déjà utilisables pour la recherche « plein texte » qui est considéré, à bien des égards, comme l'une des fonctionnalités les plus pertinentes de Transkribus. Enfin, « Transkribus n'utilise pas de dictionnaire pour la phase de transcription automatique, mais analyse les manuscrits lettre par lettre : les résultats pourraient donc être améliorés en utilisant des algorithmes de correction automatique par recherche de

similarités pour "nettoyer" les données produites automatiquement » [Massot *et al.*, 2018]. Il convient donc à cet endroit de retenir que seules des numérisations en haute définition et l'usage d'un modèle au taux d'erreur bas (inférieur ou égal à environ 5%) rendent l'expérience utilisateur agréable. Dans notre cas, la difficulté liée à la fiabilité de la reconnaissance par le modèle a été partiellement contournée grâce à l'usage d'un dictionnaire idoine mais reste quasi insurmontable pour tout autre corpus en français à l'heure actuelle.

Au niveau de l'expérience utilisateur, dans la mesure où l'interface est simple, nous n'avons pas relevé de difficulté particulière au niveau des étudiants. En fait, ce n'est pas tant l'outil qui semblait poser des difficultés aux étudiants mais bien plus la singularité de l'écriture du diariste. Enfin, concernant sa réutilisation à titre individuel dans le cadre de projets d'ampleur moindre, une des limites persistantes réside dans l'existence du peu de modèles publics réutilisables nécessitant la réalisation d'un nouveau modèle. Cependant, comme nous l'avons souligné, l'usage de l'outil d'OCRisation reste immédiatement accessible est c'est sans doute à ce niveau que des étudiants peuvent trouver des satisfactions immédiates. Par conséquent, le bilan est en demi-teinte concernant l'expérience utilisateur appliquée à des étudiants en sciences humaines et sociales (donc non archivistes). En effet, Transkribus n'est efficace que pour des corpus manuscrits de moyenne ou grande taille. Son utilisation aura sans doute un coût d'entrée trop élevé pour quiconque tentera de l'exploiter pour un corpus de maigre ampleur ou comportant une grande variété de scripteurs. Cependant, plus le logiciel sera utilisé et plus de modèles seront mis à disposition, plus la reconnaissance semi-automatisée d'une grande variété d'écritures manuscrites s'en verra facilitée.

BIBLIOGRAPHIE

Brando C., Soudani A., Meherzi Y., Bouhafs A., Frontini F., Dupont Y. and Melanie-Becquet F., Adaptation et évaluation de systèmes de reconnaissance et de résolution des entités nommées pour le cas de textes littéraires français du 19ème siècle. *Atelier Humanités Numériques Spatialisées (HumaNS'2018)*. 2018. https://hal.archives-ouvertes.fr/hal-01925816/file/8.Article-SAGeo_adaptation-et-évaluation-de-systèmes-de-REN-et-NEL.pdf

Brent N., Curating Object-Oriented Collections Using the TEI. *Journal of the Text Encoding Initiative*. 2017;9. http://journals.openedition.org/jtei/1680/

Cummings W., The William Godwin's Diaries Project. *Jahrbuch für Computerphilologie*. 2008;10. http://computerphilologie.de/jg08/cummings.pdf/

Dubout K., Durch Rezensionen zur Emanzipation? Die „Bibliographie der Homosexualität" (1900-1922) im Jahrbuch für sexuelle Zwischenstufen. *LIBREAS. Library Ideas*. 2016;29. https://edoc.hu-berlin.de/handle/18452/9746?show=full

Dubout K., *Der Richter und sein Tagebuch. Eugen Wilhelm als Elsässer und homosexueller Aktivist im deutschen Kaiserreich*, Campus Verlag (Francfort), 2018.

Dubout K. and Schlagdenhauffen R., Une archive inédite : le Journal intime d'Eugène Wilhelm (1866-1951). *Le Magasin du XIXe Siècle*. 2014 ;4:274-276.

Massot M.-L., Sforzini A. and Ventresque V., Transcrire les fiches de lecture de Michel Foucault avec le logiciel Transkribus: compte rendu des tests. 2018. 2018. hal-01794139v2.

Michel C. and Trognot G., L'expérience utilisateur au cœur de la stratégie. *I2D – Information, données & documents*. 2015;53(4) :40-41. 10.3917/i2d.154.0040

Mounier P., *Les humanités numériques*. FMSH eds. (Paris), 2018.

Perrin E., Bulliot, « Bibracte et moi. Transcription collaborative des archives archéologiques du site de Bibracte. 2019. https://f.hypotheses.org/wp-content/blogs.dir/7177/files/2019/11/Tutoriel_Transkribus_V2.pdf

Schantz H., The history of OCR, optical character recognition. *Manchester Center, Recognition Technologies Users Association*, Manchester, United Kingdom, 1982.

Schlagdenhauffen R., Une écriture du désir bisexuel est-elle possible ?. *Langage et Société*. 2014;148 :53-73.

Schlagdenhauffen R., Retour sur une controverse franco-allemande : l'Affaire Paris-Berlin (1904-1914). *In* González Bernaldo P. and Hilaire-Peréz L. (eds.), *Les savoirs-mondes. Mobilités et circulation des savoirs depuis le Moyen Âge.* Presses Universitaires de Rennes (Rennes), 2015:109-117.

Soudan C., Introduction. *Les Dossiers du Grihl, Faire une édition numérique savante et critique en TEI de manuscrits du XVIIe siècle*. 2012. http://journals.openedition.org/dossiersgrihl/5411