Transcrire les fiches de lecture de Michel Foucault avec le logiciel Transkribus : compte rendu des tests

Marie-Laure Massot – CAPHÉS [UMS 3610 / ANR FFL]
Arianna Sforzini – Triangle [UMR 5206 / ANR FFL]
Vincent Ventresque – Triangle [UMR 5206 / ANR FFL]

1er juin 2018

Table des matières

1 Le projet ANR Foucault Fiches de Lecture 2

2 Un outil pour transcrire automatiquement des manuscrits 2

3 Créer des données d'apprentissage pour l'écriture de Foucault 3

4 Principales difficultés rencontrées 5

5 Bilan de l'expérimentation et perspectives 7

1 Le projet ANR Foucault Fiches de Lecture

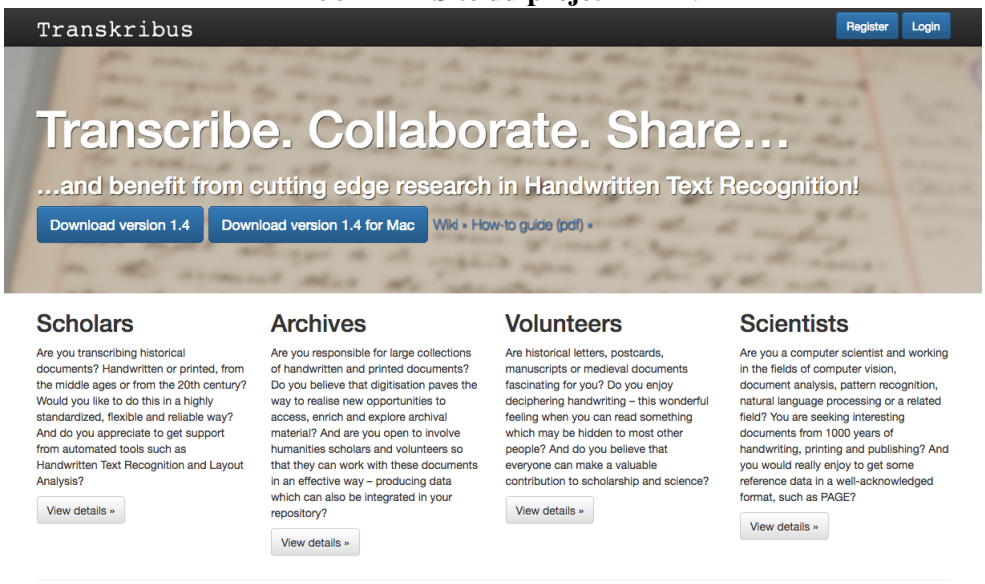
Le projet Foucault Fiches de lecture (FFL) a pour but d'explorer et de mettre à disposition en ligne un large ensemble de fiches de lecture (citations et références organisées et commentées) de Michel Foucault conservées à la BnF depuis 2013.

Il s'agit ainsi de numériser, décrire et enrichir les fiches de lecture de Foucault se rapportant à la préparation de ses livres, cours, conférences, mais aussi de permettre une nouvelle approche de son œuvre, fondée sur l'analyse des pratiques de lecture du philosophe et des cheminements de pensée qu'elles aident à retracer. Pour la mise en ligne des fiches de lecture et la production collective des données, l'équipe travaille au développement d'une plate-forme collaborative, basée sur les technologies RDF, permettant de rapprocher les contenus archivistiques et les données bibliographiques sur les sources consultées par Foucault.

Ce projet financé par l'ANR (2017-2020) et coordonné par Michel Senellart, professeur de philosophie à l'ENS Lyon, bénéficie des partenariats de l'ENS/PSL et de la BnF.

2 Un outil pour transcrire automatiquement des manuscrits

FIGURE 1 – **Site du projet READ**.



Une collaboration internationale avec le projet européen READ/Transkribus [1] a été mise en œuvre pour la transcription automatique des manuscrits de fiches de lecture. Transkribus est un logiciel de reconnaissance automatique de l'écriture manuscrite, accompagné d'une plateforme de transcription d'images numérisées de manuscrits et d'un OCR classique.

Principaux usages de Transkribus :

- Entraîner le moteur de reconnaissance d'écriture manuscrite (HTR : *Handwritten text recognition* [2]) puis l'utiliser pour transcrire automatiquement les images fournies ;
- Transcrire des documents pour une édition scientifique (notamment : interface de saisie *Wysiwyg* avec encodage TEI et création de balises personnalisées) ;
- Faire des recherches de termes dans les documents manuscrits, y compris sur des termes proches (recherche "floue", *keyword spotting* [3]).

1. https://read.transkribus.eu/transkribus.
2. Reconnaissance automatique de texte manuscrit.
3. Voir la documentation : https://transkribus.eu/wiki/images/1/1b/HowToTranscribe_Keyword_Search.pdf.

Transkribus est en effet un système expert, basé sur une technologie d'intelligence artificielle, i.e. un système capable « d'apprendre » à déchiffrer l'écriture d'un scripteur donné [4]. Cet apprentissage est réalisé à partir d'un jeu de données d'entraînement (*training dataset*), et aboutit à la production d'un modèle mathématique spécifique au scripteur.

Il est donc nécessaire de fournir des images numérisées accompagnées de leur transcription (texte numérique balisé), en assurant la correspondance ligne à ligne entre l'image et le texte. Une fois réalisé le modèle mathématique par l'équipe de Transkribus, il devient possible de transcrire automatiquement un lot d'images avec le moteur HTR. En outre, le modèle mathématique ainsi produit peut être amélioré par la suite, en complétant les données initiales du *training dataset*. Une démarche itérative peut donc être menée : à l'issue du premier entraînement, on utilise le moteur HTR pour produire un premier lot de transcriptions automatiques, que l'on corrige manuellement, et qui permettent de réaliser un deuxième entraînement, et ainsi de suite. Il faut cependant noter que le processus d'apprentissage nécessite au moins 200 images pour fonctionner efficacement : l'utilisation de Transkribus est donc pertinente à partir d'une quantité importante d'images à transcrire [5].

Les services proposés par la plateforme sont gratuits. Il suffit de s'inscrire sur le site web, de contacter l'équipe du projet READ, et de suivre les instructions du document d'introduction « Comment utiliser Transkribus en 10 étapes » [6] pour installer le logiciel et travailler avec ses propres documents après les avoir uploadés sur le serveur de Transkribus. La prise en main du logiciel requiert un peu de temps, mais le guide Transkribus et l'équipe participent activement à cet apprentissage [7].

3 Créer des données d'apprentissage pour l'écriture de Foucault

Dans un premier temps, un test a été réalisé par Vincent Ventresque avec 200 images tirées du manuscrit de *Théories et institutions pénales* : l'équipe disposait en effet d'une transcription déjà réalisée par Elisabetta Basso, ce qui a permis d'accélérer considérablement la production du premier *training dataset*.

Il restait à aligner cette transcription avec les images, et à la rendre aussi conforme que possible à l'original : Transkribus n'utilise pas de dictionnaire et ne cherche pas à reconnaître des mots, mais analyse les lignes de texte *caractère par caractère*. Pour s'assurer d'un apprentissage optimal, il est donc recommandé de respecter scrupuleusement la lettre du manuscrit : il a fallu rétablir les abréviations lorsqu'elles avaient été développées dans la transcription initiale, mais aussi les fautes d'orthographe et les accents manquants, et ajouter des balises pour repérer les passages peu clairs ou illisibles [8].

Cet alignement entre les images et le texte nécessite une phase de segmentation, à savoir de repérage des zones de texte (blocs et lignes) dans l'image. Au moment de notre test, il fallait découper manuellement les grandes zones de texte (*text regions*) sur chaque image [9], lancer le processus de reconnaissance automatique des lignes, corriger manuellement les lignes mal reconnues, et enfin, ajouter le texte numérique sur chaque ligne [10].

4. Cela peut également fonctionner pour plusieurs scripteurs, si leur nombre n'est pas très élevé pour un même corpus.

5. La transcription automatique permet de gagner du temps même si elle doit être corrigée, et fournit un texte numérique de départ dans lequel il est immédiatement possible de chercher des termes. Toutefois, le temps consacré à la production des données d'apprentissage représente un investissement non négligeable. Dans le cas des fiches de lecture de Foucault et en l'état actuel de l'outil, nous estimons que cet investissement devient justifié au-delà de 500 images à transcrire, parmi lesquelles 200 images peuvent servir pour un premier test.

6. Ce document est une traduction de « How to use Transkribus – in 10 steps (or less) » réalisée par Régis Schlagdenhauffen (EHESS – IRIS) : http://regis-schlagdenhauffen.eu/2018/01/comment-utiliser-transkribus-en-10-etapes/. Voir la version anglaise originale ici : https://transkribus.eu/wiki/images/7/77/How_to_use_TRANSKRIBUS_-_10_steps.pdf.

7. Par ailleurs, certaines institutions commencent à proposer des formations, comme par exemple l'EHESS le 19 janvier 2018.

8. Avec un encodage TEI et l'utilisation de mise en forme, cf. section suivante.

9. Depuis, la reconnaissance des lignes a été grandement améliorée, et il n'est plus nécessaire de découper manuellement les zones de texte, ce qui représente un gain de temps considérable – reste la correction des lignes mal reconnues.

10. Voir les captures d'écran des figures n°2 et 3. Transkribus nous demande de faire un lien direct entre les images du document et le texte transcrit correspondant. Pour cela, il fallait auparavant (ce n'est plus le cas) définir manuellement des régions de texte sur plusieurs images puis détecter automatiquement les lignes à l'aide du bouton « *Find lines in text regions* » dans l'onglet « Outils ». Il faut ensuite procéder à l'alignement entre lignes de base de l'image et lignes de la fenêtre de transcription (pour chaque ligne de base (*baseline*) de l'image, il y a une ligne correspondante dans l'éditeur de texte). Le texte doit donc être transcrit ligne par ligne, exactement comme il apparaît dans l'image.

À l'issue de cette première phase de production des données d'entraînement, qui a duré un peu plus de deux semaines, l'équipe du projet READ a produit (en quelques heures) un premier modèle de l'écriture de Foucault : le taux moyen de reconnaissance était de 85% par caractère. Ce résultat a été jugé très encourageant par l'équipe READ, et nous avons décidé de poursuivre avec des manuscrits tirés du corpus FFL, pour lesquels nous n'avions pas de transcription préexistante.

En effet, G. Mühlberger, responsable de l'équipe READ, nous a assuré qu'avec environ 400 images supplémentaires, le taux moyen de reconnaissance pourrait passer au-dessus de 90% par caractère, et que nous pourrions produire automatiquement une transcription automatique sur l'intégralité de notre corpus.

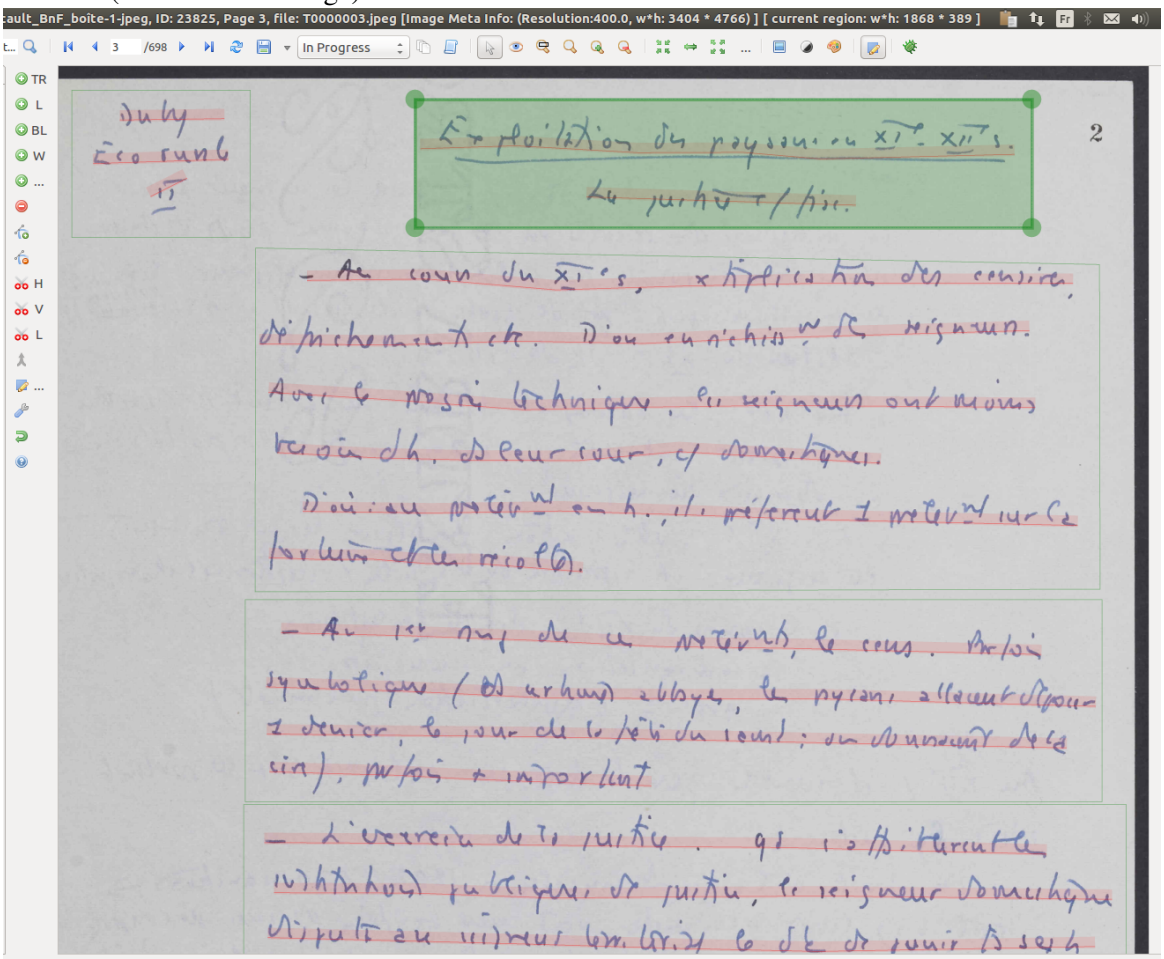
La deuxième phase a donc consisté à transcrire et aligner un peu plus de 400 images, choisies cette fois parmi le corpus dont le projet prévoyait la transcription.

Ce travail de transcription, réalisé par Marie-Laure Massot et Arianna Sforzini, portait sur des dossiers sélectionnés dans les boîtes suivantes :

- boîte n°1 (préparation des cours de 1971-1975 et *Surveiller et punir*),
- boîte n°51 (préparation de *La volonté de savoir*) [11].

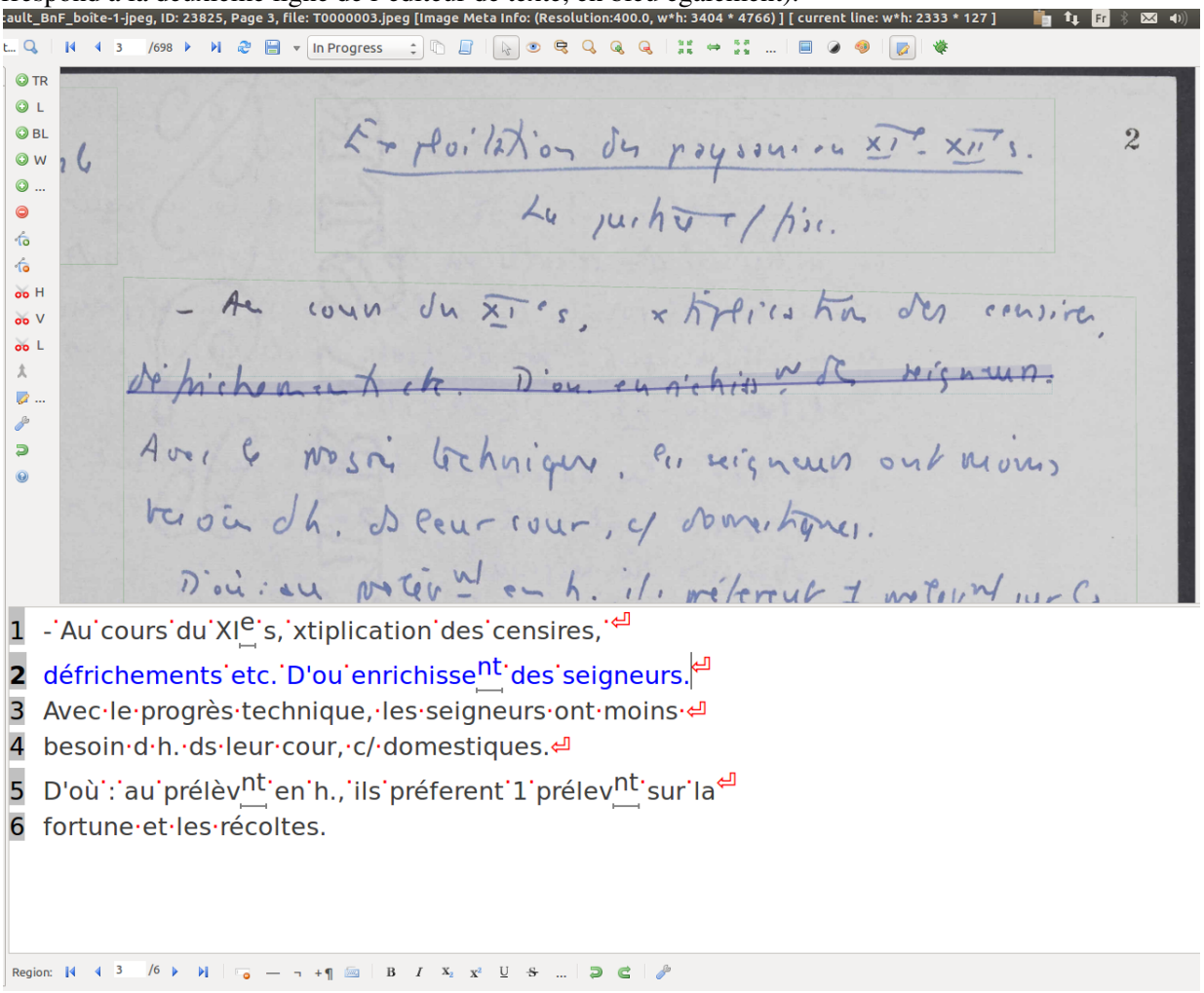
L'objectif était de disposer d'un modèle mathématique amélioré, et donc d'obtenir des transcriptions automatiques plus fidèles : cela permettrait à la fois de gagner du temps pour les transcriptions à venir, et de produire un texte numérique d'ores et déjà exploitable pour la recherche « plein texte », avant même la relecture-correction des transcriptions automatiques [12].

FIGURE 2 – **Segmentation du document** : régions de texte (*textregions* : cadres rectangulaires verts), lignes et lignes de base (*baselines* : en rouge).



11. Au total, 431 images supplémentaires : boîte 1 : 1-300 ; 300-324 ; 335-352 ; 369-380 ; boîte 51 : 33-62, 81-126.
12. Le projet FFL envisageait initialement de produire des transcriptions pour une partie limitée du corpus de fiches de lecture. Grâce à Transkribus, il devient possible de transcrire plus de fiches (gain de temps pour la lecture et la saisie) et de disposer immédiatement d'un corpus transcrit imparfaitement, certes, mais automatiquement.

FIGURE 3 – **Transcription ligne à ligne du manuscrit** (1er § du folio ; la ligne surlignée en bleu dans l'image correspond à la deuxième ligne de l'éditeur de texte, en bleu également).



4 Principales difficultés rencontrées

– L'écriture de Foucault

Il s'agit de fiches de lecture, avec beaucoup d'abréviations. L'écriture de Foucault et ses abréviations sont souvent difficiles à déchiffrer. Plusieurs abréviations sont d'ailleurs équivoques, pouvant signifier des mots différents selon les contextes. Foucault écrit aussi plusieurs lettres de la même manière, rendant encore une fois le déchiffrement de son écriture extrêmement difficile sans la prise en compte du contexte général (cela est particulièrement vrai par exemple pour les noms d'auteurs et les titres d'ouvrages des fiches de lecture). Pour les besoins du test Transkribus, nous avons repéré les mots illisibles, avec un encodage TEI (élément `<gap/>`), et une police de caractères rayés (`<strikethrough/>`) pour les mots douteux, mal formés ou raturés.

Il serait intéressant par la suite de soumettre certaines images à la communauté des spécialistes de Foucault, en ayant recensé les mots illisibles ou difficiles à déchiffrer. De même, un dictionnaire des mots et abréviations difficiles pourrait être envisagé, avec des captures d'écran des mots et abréviations, ce qui faciliterait les transcriptions futures, et plus généralement, l'accès aux manuscrits de Foucault pour les chercheurs. En outre, le projet s'attache à inventorier les ouvrages et auteurs cités, ainsi que les personnages historiques, en alignant ces "entités" avec les listes d'autorités et notices bibliographiques de la BnF et d'autres institutions [13]. Ce travail permettra aux chercheurs de connaître les sources utilisées par Foucault, mais aussi de déchiffrer plus facilement les fiches de lecture.

13. En particulier, la plate-forme développée par l'équipe FFL permet déjà d'enrichir les données de description des fiches de lecture avec les données RDF de `data.bnf.fr`.

Dans tous les cas, une relecture des transcriptions par un ou plusieurs spécialistes de Foucault devrait être envisagée à court ou moyen terme, sur la plateforme Transkribus ou sur une autre, plus facile d'accès.

– **Prise en main du logiciel et temps moyen par fiche**

Il faut quelques heures pour prendre en main et utiliser Transkribus efficacement dans le cadre d'un projet : c'est assez rapide pour les fonctions de bases nécessaires à la transcription (segmentation et saisie), mais certains automatismes s'acquièrent peu à peu et permettent de gagner du temps sur la saisie.

Ainsi, après s'être familiarisé avec l'écriture de Foucault et avoir transcrit une trentaine d'images, le temps moyen de transcription (segmentation, saisie) puis de relecture d'une fiche (2 images) était de 30 à 40 min en fonction de la longueur de la fiche (1 page et demie ou 2 pleines pages) et de la présence de noms propres difficiles à lire, pour lesquels il était nécessaire d'effectuer une recherche documentaire. Remarquons cependant que ces difficultés ne sont pas spécifiques à l'utilisation de Transkribus, et sont communes à tout travail de transcription.

– **Ergonomie du logiciel**

Il faut rappeler que Transkribus est encore en cours de développement, et que l'équipe READ se concentre sur l'amélioration des fonctionnalités de reconnaissance des zones de texte et des caractères. De ce fait, l'interface reste à améliorer pour faciliter la lecture des images et la saisie du texte numérique.

Par exemple, sur un ordinateur portable avec un écran de 13 pouces, il n'est pas toujours aisé de lire le manuscrit et de saisir la transcription en même temps. L'espace de saisie est très limité et la ligne en cours de saisie est surlignée en bleu sur l'image, ce qui peut nuire à la lisibilité. Il serait peut-être plus pratique de pouvoir choisir la disposition des fenêtres (fenêtres "flottantes"), pour mettre en regard l'image et le texte numérique plutôt que l'un au-dessus de l'autre.

Une limitation importante également : il n'est pas facile de naviguer d'une fiche à l'autre pour vérifier un mot, ni de comparer plusieurs images, ou même d'avoir une vision globale de la fiche pour rechercher un mot en particulier. Passer à l'image suivante ou précédente fait disparaître l'image en cours de consultation, et le temps de chargement de l'image et des données (zones de texte, transcription) peut être assez long, surtout si l'on ne dispose pas d'une bonne connexion internet. La comparaison de plusieurs images est particulièrement problématique : il existe une fonction d'affichage par vignettes (*thumbnails*), mais là aussi le chargement est long et peut parfois bloquer le logiciel. Actuellement, il n'est pas possible de choisir le nombre d'images à charger – toutes les vignettes d'une collection sont chargées en une seule fois –, or nous avons uploadé nos images par lots correspondant aux boîtes d'archives (entre 700 et 1140 images par boîte [14]).

– **Saisie à plusieurs mains**

Dans l'éventualité d'une saisie par plusieurs personnes, certains problèmes se posent pour uniformiser les transcriptions, car les interprétations peuvent être multiples.

Par exemple, Foucault écrit souvent presque en continu, mot après mot sans espace visible. Certains transcripteurs auront tendance à séparer les mots, d'autres à rester plus fidèles à l'image et à les attacher. Ou encore, il abrège certains mots (par ex. : « x » pour « point »), mais pas systématiquement. Certains transcripteurs auront tendance à développer l'abréviation, d'autres non. Enfin, les accents et les apostrophes sont rarement marqués distinctement par Foucault, ils sont souvent ligaturés aux lettres précédentes ou suivantes, on les devine plus qu'on ne les voit. Là non plus Foucault n'est pas systématique.

Dans tous ces cas, l'interprétation du transcripteur occupe une part importante dans la saisie de ce type de document. Or il paraît impossible d'envisager la saisie d'un corpus aussi étendu et complexe sinon collaborativement, et l'un des objectifs du projet FFL est précisément d'ouvrir à la communauté des spécialistes de Foucault un espace de transcription et d'annotation collaboratives. Il sera donc indispensable de mettre à disposition un outil collaboratif plus simple d'accès que Transkribus [15] et d'établir un guide de saisie et de transcription définissant les principes et règles communs à tous les contributeurs.

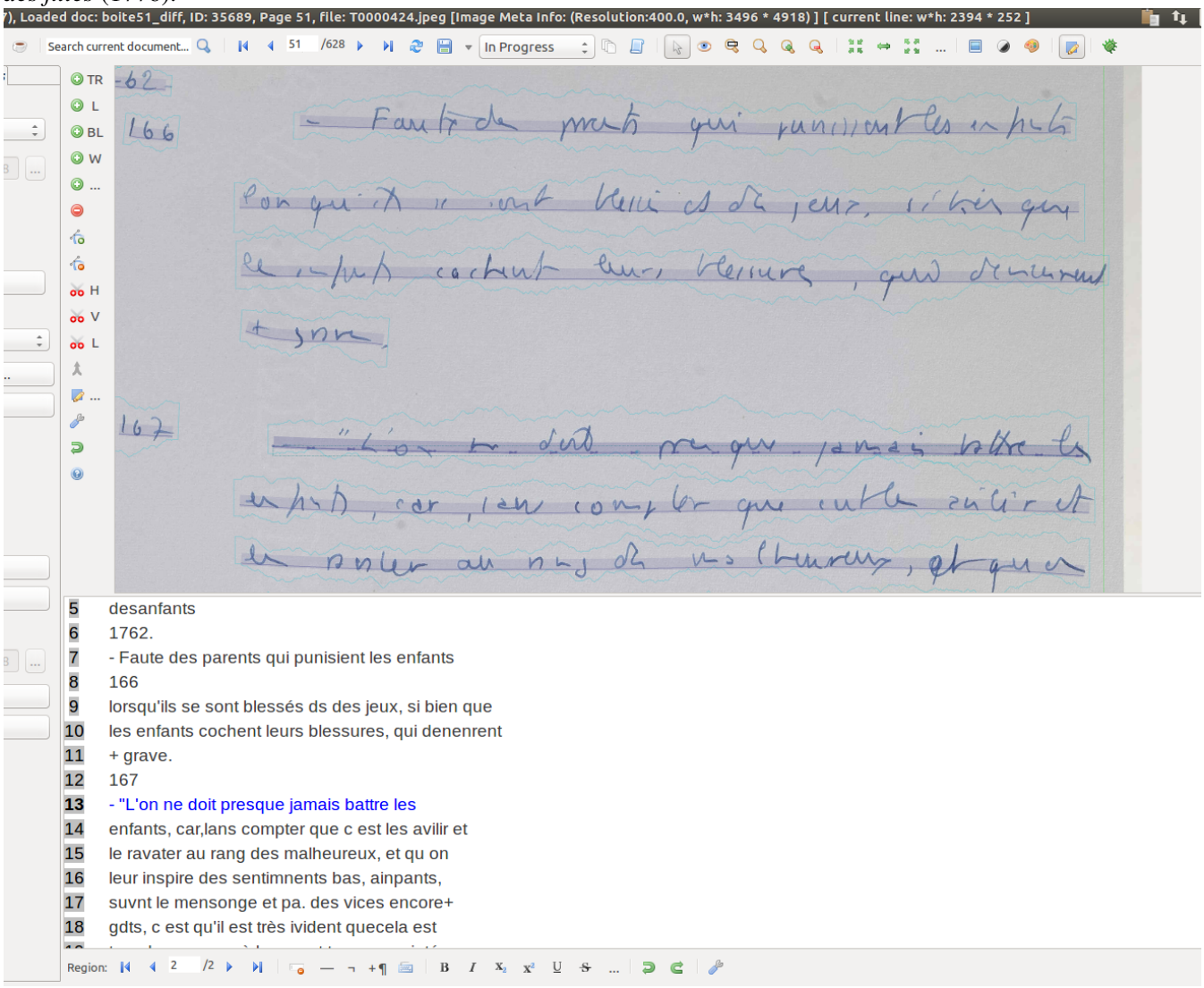
14. Pour de prochaines phases d'entraînement, il pourrait être judicieux de créer des "sous-collections", en limitant le nombre d'images par collection. D'un autre côté, cela entraînerait une manipulation supplémentaire pour circuler dans le corpus. Une autre solution pourrait être de travailler avec un dossier local d'images, si la synchronisation avec le serveur est possible.

15. Il serait peu réaliste, notamment, de demander à chaque contributeur potentiel d'installer le logiciel, et de jongler entre la plate-forme du projet FFL et l'interface de Transkribus. Nous envisageons donc l'utilisation de la plate-forme eman (cf. *infra*).

5 Bilan de l'expérimentation et perspectives

Après cet « entraînement » sur environ 600 images transcrites manuellement (transcriptions réalisées par Marie-Laure Massot et Arianna Sforzini, et réutilisation d'une transcription d'Elisabetta Basso pour le manuscrit de *Théories et institutions pénales*), les résultats des tests du logiciel Transkribus sont très encourageants : nous obtenons un taux moyen de réussite de 92% sur les caractères [16]. En outre, les futures transcriptions corrigées pourront être réinjectées dans Transkribus pour compléter l'entraînement et accroître le taux de reconnaissance.

FIGURE 4 – **Transcription automatique** : boîte 51, fiche sur Venel, *Essai sur la santé et l'éducation médicinale des filles* (1776).

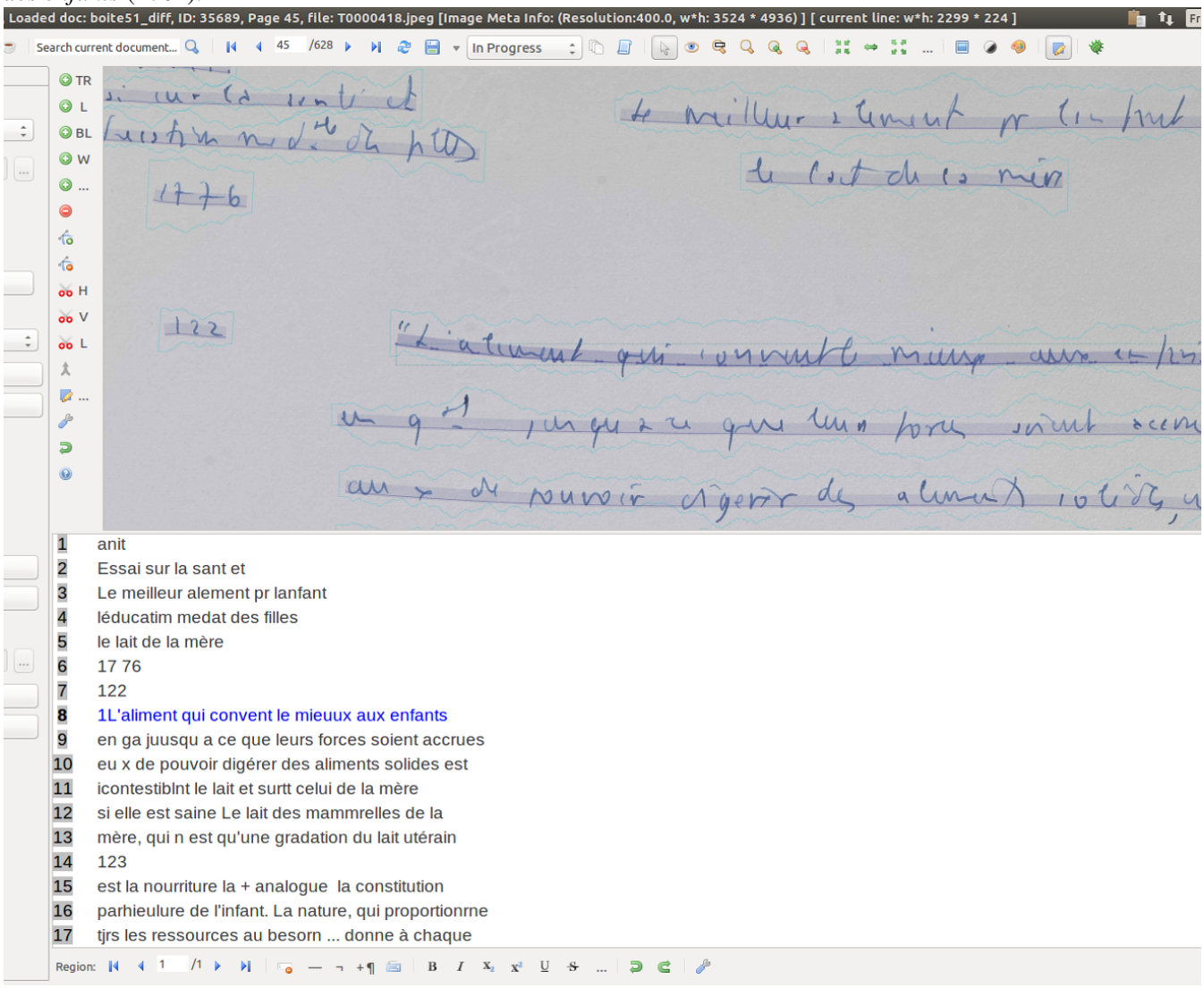


La transcription automatique des fiches doit évidemment être reprise manuellement par des correcteurs. On remarque aussi une certaine différence dans le degré de reconnaissance automatique des caractères selon les différentes boîtes exploitées : en particulier, le logiciel est sensible à la transparence du papier des fiches numérisées et lit parfois aussi les caractères au verso. Les développements ultérieurs du logiciel pourraient peut-être corriger cette difficulté [17].

Les résultats obtenus demeurent néanmoins très positifs ; la transcription automatique permet une transcription manuelle plus rapide et aide dans la reconnaissance de certains mots foucaldiens difficiles à lire. Nous

16. Voir les captures d'écran des figures n°4 et 5.
17. Par exemple, par la prise en compte des différences de contraste entre les caractères du recto et ceux du verso qui apparaissent en transparence.

FIGURE 5 – **Transcription automatique** : boîte 51, fiche sur Ballexserd, *Dissertation sur l'éducation physique des enfants* (1762).



envisageons donc de produire des transcriptions automatiques pour l'ensemble des images déjà numérisées ou à numériser. Il ne sera évidemment pas possible de corriger l'ensemble de ces transcriptions dans le seul cadre du projet ANR, c'est pourquoi nous réfléchissons également à l'utilisation de la plate-forme Omeka/eman [18], pour mettre en place un système de correction collaborative de ces transcriptions automatiques. À terme, nous pourrons créer des comptes pour les personnes intéressées par ce travail. Il est en effet plus intéressant de créer un espace de partage autour des manuscrits de Foucault, ouvrant la possibilité de compléter progressivement les transcriptions et annotations du corpus, que de proposer une série de transcriptions en "lecture seule" et en nombre limité.

Enfin, malgré leur imperfection, les transcriptions automatiques sont déjà utilisables pour la recherche « plein texte ». Une des fonctionnalités les plus pertinentes de Transkribus pour les chercheurs est sans doute la recherche « plein texte » dans les images ou les textes transcrits de Foucault. Il est possible, même en présence d'une transcription erronée, de lancer une recherche de mots-clés (recherche "floue", *keyword spotting*). Le logiciel ne reconnaît pas correctement le mot manuscrit de premier abord lors du processus de transcription automatique (HTR), mais il est capable de reconnaître la similitude entre ce mot et le terme qu'on souhaite rechercher. Évidemment, cela pourrait être décisif, sur un corpus de plus de 10 000 fiches de lecture, de pouvoir effectuer des recherches transversales par concept, auteur, ouvrage, y compris sur des termes reconnus

18. Développée par Richard Walter et l'ITEM : http://eman-archives.org/.

imparfaitement.

Cet outil de recherche n'est pourtant accessible que pour les utilisateurs de Transkribus, qui ont téléchargé et installé le logiciel. À moyen terme, de plus, l'accès à Transkribus deviendra payant. Nous envisageons donc d'implémenter des fonctionnalités de recherche floue [19] au sein de la plate-forme Omeka/eman, et du prototype d'annotation basé sur RDF. En outre, il faut rappeler que Transkribus n'utilise pas nativement de dictionnaire pour la phase de transcription automatique [20], 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.

19. Voir par exemple la fonctionnalité *FuzzyQuery* de Lucene : https://www.tutorialspoint.com/lucene/lucene_fuzzyquery.htm.

20. Il est possible d'ajouter un dictionnaire, mais la procédure paraît complexe et ne semble pas améliorer significativement les résultats du module HTR.