LECTAUREP : Lecture Automatique des Répertoires de Notaires Parisiens

Alix Chagué, Rostaing Aurélia

LECTAUREPParis Notary Record Books
Automated Reading

Alix Chagué (Inria, Université de Montréal)

Aurélia Rostaing (Archives nationales)

(diapo 1 – intro)

Le projet Lectaurep porte sur les répertoires de notaires parisiens de la période 1803-1940 conservés aux Archives nationales. Un échantillon d'images numériques de ce corpus a été traité par reconnaissance de caractères imprimés et manuscrits et, de manière exploratoire, par traitement automatique des langues, reconnaissance des entités nommées et éditorialisation

(diapo 2 – périmètre de LECTAUREP)

Un répertoire est un registre où le notaire consigne par ordre chronologique les actes notariés qu'il a établis. Les contenus de ses colonnes sont autant de métadonnées relatives aux actes décrits ; on peut les catégoriser en types d'actes, dates, noms d'agents, professions, noms géographiques, mots matière. Les instruments de recherche que sont les répertoires constituent de ce fait des corpus de recherche en eux-mêmes, où il est possible d'isoler des lots homogènes de données intéressant l'histoire économique et sociale.

(diapo 3 - Enjeux)

Notre but est de faciliter et de massifier l'**accès** à ces contenus en offrant à nos usagers un service de lecture enrichie et de fouille de données.

Nous voulons aussi partager le résultat de nos travaux et nos retours d'expérience afin de favoriser la mutualisation et l'interopérabilité des données et métadonnées produites par l'HTR. Il s'agit, pour cela, de convenir de bonnes pratiques communes, voire de standards, entre GLAM, chercheurs, généalogistes, prestataires de logiciels métier ou de services d'HTR.

Enfin, nous avons tenu à prendre en compte quinze ans d'un patrimoine numérique résultant aussi bien d'une numérisation rétrospective de microfilms, que d'une numérisation d'après originaux.

(diapo 4 – Échantillons)

Concrètement, nous avons échantillonné deux lots d'images en NB et en couleurs, puis nous avons ouvert deux chantiers plus homogènes sur le plan matériel, et par conséquent moins difficiles à traiter techniquement : un siècle de registres d'enregistrement de contrats de mariage de commerçants, et les répertoires d'un notaire du XVIIIe siècle.

Sur ces 4 lots, représentant **250 mains** au moins, environ **2000 pages ont été transcrites (soit quelques dizaines de mains)**, dont un gros quart a été relu. Des modèles d'HTR satisfaisants ont pu être affinés sur la base d'une vérité terrain de qualité afin de réduire les taux d'erreur par caractère.

(diapo 5 - Environnement technologique)

Pour la tâche de transcription, l'environnement technologique dans lequel s'inscrit le projet est essentiel : il convient de distinguer ce qui relève du *software*, la partie la plus visible, et ce qui relève du *hardware*.

(Software)

Le projet LECTAUREP s'inscrit dans une démarche de science ouverte. Tout naturellement, cela se retrouve dans le choix des logiciels :

- Premièrement **Kraken**[1], qui est un moteur d'HTR développé par Benjamin Kiessling depuis 2015, désormais sous l'égide de SCRIPTA-PSL. Kraken est compatible avec de nombreux systèmes d'écriture, alphabétiques ou non alphabétiques, et plusieurs sens de lecture. Il permet d'entraîner différents types de modèles : pour la transcription, mais aussi pour la segmentation, c'est-à-dire la détection des lignes et/ou des zones de texte sur l'image.

- Deuxièmement **eScriptorium**[2], une application web développée par SCRIPTA PSL depuis 2018. C'est un plan de travail virtuel pour la conduite de projet de transcription. Il sert de coquille ou d'interface graphique à des moteurs d'HTR, en l'occurrence kraken.

(Hardware)

1 Kiessling, B. (2021). *Mittagessen/kraken* (3.0.6) [Python]. https://github.com/mittagessen/kraken (Original work published 2015)
2 Tissot, R. (2021). *Scripta/eScriptorium* (0.10.2a) [Python]. https://gitlab.com/scripta/escriptorium/-/tree/v0.10.2a

Le projet LECTAUREP s'appuie donc depuis 2019 sur une application eScriptorium, qui est déployée sur un serveur minimaliste avec peu de capacités, dont le but était de permettre à tous les membres du projet de travailler ensemble sur la même base de données. Puis, une première montée en charge en 2020, avec une migration sur le serveur « Traces6 », mieux équipé (notamment doté de cartes graphiques, indispensables pour un entraînement efficace des modèles). Enfin, LECTAUREP fait partie des projets qui pourront tirer profit de la nouvelle montée en charge grâce au serveur CREMMA, financé par le DIM MAP. En plus d'être mieux équipé (plus de GPU, plus de mémoire), il propose une architecture modulaire compatible avec de futures améliorations. Sans cet environnement, la tâche de transcription n'est pas possible.

(diapo 6 – livrables contractuels 1)

En 2021, les résultats de LECTAUREP sont nombreux. Nous avons établi une méthode de production de modèles de transcription qui fonctionne bien : elle s'appuie sur l'utilisation de modèles dits "génériques" dont les taux d'erreur par caractères sont inférieurs à 10% (càd. le modèle fait une erreur pour moins d'1 lettre sur 10). Ces modèles sont entraînés sur des lots de mains variés, généralement au moins une dizaine.

On en possède deux, entraînés sur différents ensembles de transcription qui sont plus ou moins parfaite. Ces modèles servent plusieurs objectifs : 1) produire une première passe de transcription permettant, soit la publication d'une transcription certes faussée mais compatible avec une exploration des corpus dans le cadre d'une recherche floue, soit une pré-annotation des documents ce qui fait gagner du temps lors de la transcription manuelle. Au lieu de déchiffrer, on n'a qu'à corriger. 2) De plus, ces modèles servent de base pour affiner des modèles dits spécialisés, qui sont ré-entraînés sur des petits lots de données uniformes. De cette manière, on parvient rapidement à des taux d'erreur égaux voire inférieur à 5%, soit une faute tous les 20 caractères.

Bien entendu, dans le cadre d'une démarche ouverte, les conventions et pratiques de transcription élaborées sont documentées, de même que les expérimentations avec les données, les modèles et l'infrastructure.

(diapo 7 – livrables contractuels 2)

Les données de transcription sont le nerf de la guerre ; elles sont la base pour entraîner les modèles. LECTAUREP en a produit beaucoup, soit en faisant de la transcription entièrement à la main, soit en faisant de la reprise de transcription automatique.

Une partie de ces données est considérée comme « *gold* » : c'est-à-dire qu'elles ont été contrôlées et corrigées. Elles sont rendues publiques par l'intermédiaire de l'organisation HTR-United[3] et pourront aussi faire l'objet d'une publication par le biais de data.culture.gouv.fr. Le reste des transcriptions nécessite encore des corrections de la part du DMC et intégrera le corpus *gold* progressivement.

(diapo 8 – livrables bonus)

D'autres livrables n'avaient pas été anticipés par le projet. On peut mentionner une contribution directe et continue au projet SCRIPTA-PSL sous la forme de cas d'usages et de retours utilisateurs, sous la forme de développement de fonctionnalités qui sont intégrées au code source de l'application (mentionnons ici le travail d'Yves Tadjo dont le contrat est financé par LECTAUREP) et globalement sous la forme d'une documentation qui est mise à disposition de tous les utilisateurs de l'application.

Plus largement, les membres du projet sont engagés dans une démarche de partage d'expertise avec des utilisateurs porteurs de projets impliquant de l'HTR, avec des groupes de travail comme CREMMALab et avec la communauté des GLAMs.

Ce partage d'expertise prend aussi la forme de publications scientifiques sur les questions qui intéressent le projet, dont une partie fait l'objet de billets de blog sur un carnet hypothèses[4], ouvert à la faveur du confinement et du stage de Lucas Terriel en 2020.

(diapo 9 – KaMI : les métriques du labo au terrain)

Témoin d'une appropriation de la question de la mesure des performances des modèles « au contact du terrain », l'outil KaMI[5] rend possible une meilleure évaluation de la réussite des

3 Chagué, A. & Clérice, T. (2021). *HTR United, a centralization effort of HTR and OCR ground-truth repositories for French languages*. https://github.com/HTR-United/htr-united (Original work published 2020)

4 Voir : https://lectaurep.hypotheses.org/

5 Terriel, L., & Chagué, A. (2021). KaMI Lib (0.1.1) [Computer software]. https://gitlab.inria.fr/dh-projects/kami/kami-lib

modèles.

Il donne davantage de métriques, en combinant par exemple le taux d'erreur par caractères (CER), le taux d'erreur par mots (WER), la distance de Levenshtein et les opérations d'édition comme les substitutions, suppressions et ajouts. KaMI est agnostique : on peut s'en servir pour comparer deux chaînes de caractères, quelque soit le logiciel d'HTR utilisé pour les générer. Et enfin, il permet surtout de jouer avec des filtres afin de négocier la sévérité de l'évaluation en fonction de critères considérés comme importants : on peut par exemple ignorer les erreurs portant sur la reconnaissance des nombres dans un cas où on s'occuperait surtout de savoir si les lettres et les mots sont bien reconnus. Une telle évaluation est très utile pour anticiper la difficulté de la tâche de correction après l'application d'un modèle de transcription.

(diapo 10 : un corpus pour les spécialistes de TAL)

Le corpus de textes produit à l'occasion du projet LECTAUREP fait émerger des défis qui peuvent intéresser les spécialistes du traitement automatique des langues (NLP) car la langue utilisée dans les pages des répertoires est loin d'être naturelle : elle contient de nombreuses abréviations et entités nommées et est faite de phrases non verbales.

(diapo 11 – vers XMLTEI et TEI Publisher)

Enfin, les documents qui intéressent le projet ont permis de s'interroger sur la manière de rendre accessibles des documents dont la mise en page est complexe. LECTAUREP alimente une partie des exemples étudiés par l'équipe ALMAnaCH pour intégrer une application comme TEI Publisher[6] dans une chaîne de traitement généraliste dédiée à l'HTR et reposant sur une utilisation plus systématique de la TEI.

(diapo 12 – Conclusion 1)

Le projet Lectaurep a montré, à l'échelle d'une partie de l'échantillon initial, que des modèles d'HTR à large spectre fonctionnent suffisamment bien pour permettre une recherche floue sur des pages aux lignes d'écritures aérées (le corpus en NB du XIXe s. et une partie du corpus en couleurs du XXe s.).

6 e-editiones. (2021). *Eeditiones/tei-publisher-app* (7.1.0) [XQuery]. e-editiones.org. https://github.com/eeditiones/tei-publisher-app (Original work published 2020)

Les modèles de segmentation se heurtent pour le moment à un seuil quand les lignes sont trop serrées (soit une part importante du corpus du XXe s.) ; il serait donc utile de disposer d'outils permettant d'évaluer la qualité des modèles de segmentation, dont celle de l'HTR dépend, afin de pouvoir affiner ces modèles en se fondant sur des métriques.

(diapo 13 – Conclusion 2)

Le corpus cible de Lectaurep (évalué à + 1 M images) est monumental. La mise en production de moins d'1 % de ce corpus (par exemple l'une des 122 études notariales, ou bien une tranche chronologique d'une année sur 140) requiert une logistique participative, des infrastructures et une ingénierie de projet, concernant notamment les flux des images et des données.

Pour affiner un échantillonnage, il peut être utile d'approfondir la connaissance diplomatique de nos sources physiques et numériques. Avoir une idée du nombre de mains par registre, pouvoir préciser la répartition quantitative entre NB et couleurs grâce à une meilleure maîtrise des métadonnées permettrait peut-être d'optimiser et d'économiser les ressources nécessaires à l'intelligence artificielle.