Données ouvertes, données propres, et autres vies : Testaments de Poilus et CREMMA

Alix Chagué, Thibault Clérice

Données ouvertes, données propres, et autres vies : Testaments de Poilus et CREMMA

Alix Chagué[1, 2, 3] and Thibault Clérice[1]

[1]ALMAnaCH, Inria, Paris, France
[2]Université de Montréal, Montréal, Canada
[3]École Pratique des Hautes Études, Paris, France

Décembre 2023

1 Introduction

La reconnaissance automatique de textes manuscrits (Handwritten Text Recognition ou HTR) est un domaine de l'intelligence artificielle dont l'objectif est de prédire une ligne de texte en partant de l'image d'un document manuscrit [1]. Des modèles de transcription sont chargés de produire une prédiction et sont entraînés à partir de transcriptions exemples, aussi appelées « vérité de terrain ». Dans ce chapitre, nous nous intéressons plus particulièrement à la réutilisation des données du projet Testaments de Poilus [2] pour créer des données d'entraînement pour l'HTR, réalisée dans le cadre du projet CREMMA (Consortium pour la Reconnaissance des Écritures Manuscrites des Matériaux Anciens).

CREMMA a pour ambition de mettre à la disposition de la communauté scientifique d'Île-de-France l'application eScriptorium, conçue pour la conduite de projets de transcription automatique [3], par l'intermédiaire d'un serveur adossé à une configuration garantissant des capacités de calcul adéquates. En plus de l'infrastructure proposée par CREMMA, nous souhaitions doter les utilisateur·rices de modèles de transcription dits génériques. Tandis qu'un modèle spécialisé est entraîné sur une ou très peu d'écritures (on parle de « mains »), un modèle générique doit être entraîné à partir d'une vérité de terrain qui rassemble un très grand nombre d'exemples d'écri-

1. Victoria Ruiz-Parrado, Ruben Heradio, Ernesto Aranda-Escolastico, Ángel Sánchez et José F. Vélez, « A Bibliometric Analysis of Off-Line Handwritten Document Analysis Literature (1990–2020) », *Pattern Recognition*, 125 (mai 2022), p. 108513, DOI : 10.1016/j.patcog.2021.108513 ; Joe Nockels, Paul Gooding, Sarah Ames et Melissa Terras, « Understanding the Application of Handwritten Text Recognition Technology in Heritage Contexts : A Systematic Review of Transkribus in Published Research », *Archival Science* (, juin 2022), DOI : 10.1007/s10502-022-09397-0.

2. Emmanuelle de Champs, « Des bénévoles au service du patrimoine écrit. De l'Oxford English Dictionary aux Testaments de Poilus » (, 2021), p. 47, p. 47.

3. Benjamin Kiessling, Robin Tissot, Peter Anthony Stokes et Daniel Stökl Ben Ezra, « eScriptorium : An Open Source Platform for Historical Document Analysis », dans *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, 2019 (2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)), t. 2, p. 19-19, DOI : 10.1109/ICDARW.2019.10032.

tures [4]. Proposer des modèles de transcription génériques aux utilisateur·rices de la plateforme présentent plusieurs avantages : comme nous l'évoquerons plus bas, ils permettent de pré-annoter les documents en vue d'une correction manuelle postérieure ; ils sont parfois suffisamment performants pour être appliqués tels quels sur des documents et créer ainsi très rapidement un corpus textuel mobilisable dans le cadre d'un travail de fouille de texte, par exemple ; et enfin, ils sont essentiels pour accélérer la production de modèles de transcription spécialisés puisqu'ils peuvent être utilisés comme des modèles « fondations [5] » et être finetunés [6] avec relativement peu de données. Une partie du budget du projet CREMMA [7] a donc été dédiée à la création d'une vérité de terrain pour ces modèles.

Cette démarche s'inscrit dans le cadre d'une prise en compte de la nécessité de redéfinir la notion d'importance matérielle pour les jeux de données d'entraînement. La génération et la disponibilité de jeux de données massifs n'est en effet pas à concevoir uniquement en termes de nombre d'images et de nombre de caractères mais aussi en nombre de mains. Cette diversification des mains est un défi majeur pour l'apprentissage automatique. Il est relativement « facile » désormais d'entraîner un modèle à reconnaître une main unique, sur un document en particulier : dès 2015, avec des outils OCR qui supportaient l'apprentissage tels qu'Ocropy, Ariane Pinche et Jean-Baptiste Camps obtenaient déjà sur des manuscrits uniques des scores dépassant les 85% de réussite [8]. En revanche, pour les modèles génériques, qui doivent pouvoir, par définition, généraliser leur connaissances, la question de la disponibilité de données variées est centrale. On peut alors considérer les trois publics de l'HTR, à savoir : le grand public pour la reconnaissance de leurs notes manuscrites, la recherche et les institutions patrimoniales pour la mise à disposition ou la fouille de corpus, et enfin le privé pour la mise à disposition d'outils pour les deux premiers publics. CREMMA s'adresse principalement au public de la recherche.

La production des transcriptions d'entraînement peut être rapidement coûteuse car elle est réalisée manuellement. Les projets CREMMA, puis désormais HTRomance, ont permis d'établir une estimation du temps d'annotation nécessaire. On a ainsi pu observer que la transcription de documents contenant des écritures cursives, telles qu'elle a été réalisée dans le cadre du projet HTRomance, s'effectue en moyenne à un rythme avec une médiane à 3100 caractères par heure (Cf. Figure 1). Notons, que désormais, grâce à l'existence de modèles génériques tels que Manu McFrench [9] ou

4. Tobias Hodel, David Schoch, Christa Schneider et Jake Purcell, « General Models for Handwritten Text Recognition : Feasibility and State-of-the Art. German Kurrent as an Example », *Journal of Open Humanities Data*, 7–0 (juill. 2021), p. 13, DOI : 10.5334/johd.46, p. 12.

5. Pour cette notion, voir Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, *et al.*, « On the Opportunities and Risks of Foundation Models » (, juill. 2022), DOI : 10.48550/arXiv.2108.07258.

6. Le *finetuning* consiste à partir d'un modèle déjà existant et à l'entraîner une deuxième fois en utilisant des nouvelles données. De cette manière, le modèle bénéficie de connaissances de bases acquises durant le premier entraînement et se spécialise avec les données du deuxième entraînement. Cette méthode permet d'entraîner plus vite et avec moins de données, pour un résultat souvent plus robuste.

7. Soit environ 10 000€ sur les 50 000€ du budget total.

8. Jean-Baptiste Camps, *Homemade manuscript OCR (1) : OCRopy*, Billet, févr. 2017.

9. Alix Chagué, Thibault Clérice, Jade Norindr, Maxime Humeau, Baudoin Davoury, Elsa Van Kote, Anaïs Mazoue, Margaux Faure et Soline Doat, « Manu McFrench, from Zero to Hero : Impact of Using a Generic Handwriting Recognition Model for Smaller Datasets », dans *Digital Humanities 2023 : Collaboration as Opportunity*, 2023.

CREMMA Medieval [10], il est possible de dédier moins de temps à la transcription initiale des documents. En effet, ces modèles peuvent être utilisés pour pré-annoter le corpus et la correction de la pré-annotation prend alors moins de temps qu'une transcription partant de zéro.

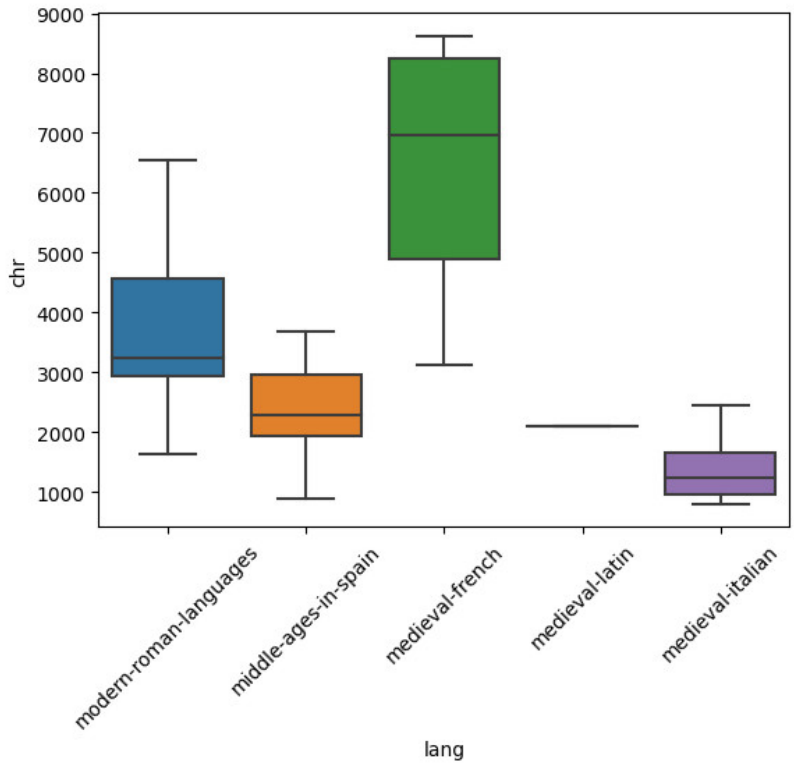

FIGURE 1 – Estimation du nombre de caractères transcrits en une heure pour chaque corpus du projet HTRomance, en fonction des documents (Auteur : Thibault Clérice).

Dans le cadre du projet CREMMA, sept étudiant·es du cursus « Technologies numériques appliquées à l'histoire » de l'École nationale des chartes ont été recruté·es pour des contrats de vacation : Baudoin Davoury, Soline Doat, Margaux Faure, Maxime Humeau, Anaïs Mazoue, Jade Norindr et Elsa Van Kote [11].

La réutilisation par CREMMA des données du projet Testaments de Poilus s'inscrit dans le contexte d'une optimisation des efforts de production de données d'entraînement pour la création d'un modèle générique. Nous commencerons par expliciter l'intérêt de s'appuyer sur des données préexistantes dans le cadre de cette démarche ainsi que les ressources sur lesquelles il est possible de s'appuyer. Ceci étant posé, nous aborderons les nombreux intérêts présentés par Testaments de Poilus en particulier, que nous comparerons à d'autres jeux de données similaires. Pour finir, nous détaillerons les moyens mis en œuvre pour produire un nouveau jeu de données à partir de celui proposé par le projet Testaments de Poilus.

10. T. Clérice, Ariane Pinche et Malamatenia Vlachou-Efstathiou, *Generic CREMMA Model for Medieval Manuscripts (Latin and Old French), 8-15th Century*, 2023, DOI : 10.5281/zenodo.7631619.

11. Nous donnons leurs noms car il nous semble essentiel de rendre visible le travail d'annotation qu'ils et elles ont fourni. Il s'agirait de ne pas reproduire un effacement du travail effectué par ce qu'on appelle parfois les « petites mains », tel qu'ont pu le mettre en lumière des chercheuses comme Melissa Terras dans M. Terras, *For Ada Lovelace Day – Father Busa's Female Punch Card Operatives*, oct. 2013.

2 Réexploitation des données

Proposer des vacations de transcription à des étudiant·es en cours de formation pose un défi qui est celui de la maîtrise de la lecture des textes à transcrire. Autrement dit, la compétence en paléographie est un critère important pour anticiper la rapidité avec laquelle la tâche de transcription est effectuée. Afin d'augmenter la rapidité de transcription, et ainsi pouvoir baisser le coût de sa production sans réduire le salaire octroyé aux vacataires, il est crucial de pouvoir s'appuyer sur des données préexistantes.

Plusieurs types de projets ont conduit à la production de données qui peuvent être réemployées pour créer des données d'entraînement pour l'HTR.

Les éditions papiers de textes variés, comme des mémoires, des thèses, des éditions critiques, etc, sont un premier moyen d'accéder à la transcription d'un document manuscrit. Ils présentent la particularité de fournir un texte édité, qui dépasse donc la transcription. Suivant que ces éditions portent sur des traditions textuelles pour les documents à plusieurs témoins (comme une partie des manuscrits médiévaux littéraires) ou sur des unica [12], ils comportent des corrections, sont le résultat de choix éditoriaux, ou encore ont fait l'objet d'une normalisation. Par exemple, pour un document à plusieurs témoins, le texte proposé peut être celui d'un témoin en particulier, sélectionné par les éditeurs, ou bien celui d'un témoin virtuel agrégeant les variations des différents témoins existants. S'agissant de la normalisation, celle-ci concerne en particulier le traitement des abréviations qui sont, dans les éditions de textes, souvent résolues. Numérisés ou non, ces éditions peuvent servir de support à la lecture pour les transcripteur·rices, mais ne peuvent, dans une perspective graphématique conservatrice [13], proposer autre chose qu'une base à adapter. Si, par chance, l'édition est accessible facilement pour les vacataires, il demeure le problème de trouver des éditions dont les textes sont disponibles sous formes de manuscrits numérisés. Le problème inverse existe aussi : trouver un manuscrit dont l'édition – sous forme d'un mémoire par exemple – est indisponible. Quelques éditions ou recherches en histoire peuvent parfois fournir des transcriptions de plusieurs manuscrits, de plusieurs mains (comptes médiévaux, échanges épistolaires, etc.) mais pour ces cas-là, la difficulté pour les institutions du patrimoine de numériser l'ensemble de leurs collections complique souvent la recherche des manuscrits numérisés. Malgré ses défauts et difficultés, ce type de ressources ne doit cependant pas être négligé : elles ont bel et bien permis de produire une partie des données créées dans le cadre des vacations CREMMA, notamment en ce qui concerne les textes médicaux latins.

D'autres supports, numériques cette fois, permettent d'accéder à des transcriptions pré-existantes. C'est le cas des jeux de données produits dans le cadre de campagnes de crowdsourcing [14], et c'est aussi le cas des éditions numériques. Considérons-les séparément.

12. Document à témoin unique, comme une lettre, une charte, ou un texte littéraire dont on ne connaît qu'un seul exemplaire.

13. Ariane Pinche en propose la définition suivante : « transcription qui préserve la suite des lettres et réduit chaque forme à son sens dans un système alphabétique », dans A. Pinche, *Guide de Transcription Pour Les Manuscrits Du Xe Au XVe Siècle*, juin 2022, p. 3.

14. L'expression *crowdsourcing* fait référence à un mode de production de contenus ou de données dans lequel il est fait appel au grand public pour exécuter une tâche. Le projet Testaments de Poilu est un projet de *crowdsourcing* dans lequel le grand public a été sollicité pour transcrire et encoder une collection de documents manuscrits.

Premièrement, les jeux de données issus de campagnes de crowdsourcing présentent un niveau de qualité inégal, mais néanmoins utilisable [15]. Leur limitation vient surtout du fait que les règles de transcription peuvent être très variables d'une campagne à l'autre ou d'une plateforme à l'autre, avec différents degrés de corrections post-annotation. La plate-forme FromThePage [16], lieu d'organisation de différents transcribathons [17], regorge de pages transcrites, accompagnées des numérisations correspondantes. Dans un même transcribathon, il arrive qu'une équipe veuille aller plus loin dans la tâche proposée, et fournisse par exemple un encodage des abréviations avec leur résolution, quand d'autres proposent uniquement les transcriptions résolues.

Deuxièmement, à l'inverse, les éditions numériques de manuscrits sont très finement corrigées, mais elles sont rares, surtout pour des textes français. Cette rareté s'additionne à la probabilité de trouver un document numérisé. Pour des éditions de texte à tradition complète, à défaut du ou des manuscrits principalement utilisés par les éditeurs, il est imaginable de se satisfaire de documents plus récents (et donc potentiellement plus fautifs) : il faudra alors espérer que les variations récentes – souvent non enregistrées dans l'appareil de notes de l'éditeur – ne soient pas trop difficiles à lire. Cependant, contrairement aux éditions papiers, ces éditions ont l'avantage de fournir – quand le manuscrit est aussi accessible – des éditions faciles à aligner numériquement, qu'il ne reste plus qu'à adapter aux règles de transcriptions de nos projets.

3 L'intérêt de Testaments de Poilus

Le corpus constitué par le projet Testaments de Poilus présente plusieurs intérêts dans la perspective d'une réutilisation des données pour la production de vérité de terrain pour l'HTR.

En tout premier lieu, Testaments de Poilus propose une édition de très bonne qualité (les transcriptions ont été relues et sont homogènes) avec des conditions de réutilisation clairement explicitées : le corpus de transcription est sous licence Creative Commons CC BY, et les images sont libres de droit.

Deuxièmement, Testaments de Poilus rassemble un relativement grand nombre de documents manuscrits sur lesquels on trouve des exemples de cursives, accompagnés de métadonnées pour chacun de ces documents. Cela permet de produire une vérité de terrain qui prend en compte des critères comme l'identité des testamentaires pour assurer une certaine variété de mains dans le corpus. De ce point de vue, Testaments de Poilus est le seul corpus permettant d'exploiter facilement une masse de documents avec ces objectifs : la TEI permet de récupérer la transcription souhaitée, mais aussi de créer des filtres sur les métadonnées pour garantir une grande variété dans le corpus.

15. Voir en particulier Tim Causer, Kris Grint, Anna-Maria Sichani et M. Terras, « 'Making Such Bargain' : Transcribe Bentham and the Quality and Cost-Effectiveness of Crowdsourced Transcription », *Digital Scholarship in the Humanities*, 33–3 (sept. 2018), p. 467-487, DOI : 10.1093/llc/fqx064, p. 467-487 ; ou encore Solène Tarride, Tristan Faine, Mélodie Boillet, Harold Mouchère et Christopher Kermorvant, *Handwritten Text Recognition from Crowdsourced Annotations*, juin 2023, DOI : 10.48550/arXiv.2306.10878, arXiv : 2306.10878 [cs].

16. Accessible à https://fromthepage.com/.

17. De *transcribe* (transcrire) et -athon, suffixe pris de hackathon : une sorte de compétition, amicale ou non, visant à produire des données de transcription sur un sujet particulier.

En comparaison, l'édition Sentences Commentary Text Archive (SCTA), qui s'appuie elle aussi sur la TEI et qui a été réutilisée pour fournir une aide à la transcription avec les images des manuscrits disponibles, présente plusieurs manuscrits et œuvres, et donc plusieurs mains. Cependant, il s'agit à chaque fois du même genre de texte. Avec la SCTA, on a donc une variété de mains mais pas une variété thématique : le vocabulaire est assez constant d'un manuscrit à l'autre, ils font références aux mêmes noms, etc. En dehors de la SCTA, des éditions comme HyperDonat, qui propose l'édition d'un commentaire grammatical à travers plusieurs manuscrits, ou la thèse d'Ariane Pinche sur Wauchier de Denain [18], qui présente une compilation de neuf textes et leurs variantes sur une dizaine de manuscrits, représentent de facto une réduction encore plus importantes de mains et de diversité de vocabulaire : non seulement ils tiennent du même genre, mais aussi du même texte.

Au contraire, l'édition de Testaments de Poilus permet de collecter une grande variété de mains mais aussi, dans une certaine mesure, une variété thématique. Ce sont bien des testaments, donc le sujet et une partie du vocabulaire ne varient pas : on retrouve ainsi régulièrement la formule « Ceci est mon testament ». Cependant, avec autant d'auteurs que de documents, cela signifie que le vocabulaire varie d'un testament à l'autre. Cela concerne aussi bien les personnes, les biens et les situations que la tonalité.

Il nous faut certainement nous arrêter sur l'intérêt de disposer d'un corpus diversifié dans le cadre de la production de modèles d'HTR. Cette diversité (Cf. Figure 2) s'exprime sur plusieurs aspects :

— Géographique : les lieux auxquels il est fait référence ne sont pas les mêmes, on a également dans la langue des régionalismes qui s'expriment parfois à l'écrit, allant de la variété lexicale à la variation orthographique (plus rare).
— Matériel : les hommes qui ont rédigé ces testaments l'ont fait sur des supports et avec des outils d'écritures très divers. Si l'on considère le texte écrit comme un signal, on peut alors intégrer dans le bruit (ce à quoi le modèle doit être résistant) non seulement la différence de main, mais aussi la couleur du papier, le type d'encre, l'épaisseur du trait, etc. En entraînant un modèle avec des bruits différents, on s'attend logiquement à obtenir un modèle plus confortable dans sa capture du signal et donc mieux capable de généraliser quel que soit l'aspect de l'écriture et du document. Les documents eux-mêmes nous parviennent dans des états variés, ce qui, encore une fois, joue dans la capacité d'un modèle de transcription à être robuste aux accidents présents sur les documents auxquels il est ensuite appliqué.
— Social : le corpus de texte a été rédigé par des hommes d'âges, de conditions sociales et de niveaux d'enseignement différents. On a donc tantôt des écritures très fines et élaborées, tantôt des écritures plus maladroites. Entre les deux, une très grande variété de manière de tracer les lettres.

18. A. Pinche, *Edition Nativement Numérique Du Recueil Hagiographique "Li Seint Confessor" de Wauchier de Denain d'après Le Manuscrit Fr. 412 de La Bibliothèque Nationale de France*, These de Doctorat, Lyon, 2021.

4 Courte comparaison avec d'autres jeux de données

Il existe des tentatives précédentes concernant la mise à disposition de jeux de données d'entraînement pour les manuscrits. Par exemple, le jeu de données IAM [19] contient 1537 images représentant l'écriture de 637 auteurs différents en anglais. Les données sont produites sous la forme d'un formulaire, avec un texte imprimé à recopier par un·e scripteur·ice à la main. Distribué sous licence non commerciale et uniquement à des fins de recherche, ce jeu de données propose une diversité textuelle plus importante que Testaments de Poilus, avec des textes issus d'un corpus littéraire.

Inspiré de ce genre d'approche, CREMMA a produit des jeux de données similaires pour le français avec des données capturées aléatoirement de Wikipédia France, mais contenant parfois des extraits d'autres langues par effet de citation. Là, le nombre de mains est important, la diversité lexicale est forte, mais la diversité de matériel reste bien en deçà de la variété de supports et d'encres trouvée dans les testaments de poilus.

RIMES [20] est un jeu de données produit dans le cadre du projet éponyme, composé à partir de la collecte de 5600 lettres factices (environ 12000 lignes) rédigées par 1300 scripteur·rices à partir de prompts les invitant à simuler une situation où ils/elles rédigeaient un courrier à une entreprise ou à une administration français.

Le corpus des Données du recensement du Valais [21] contient quelques centaines de formulaires de recensement du XIX^{e} siècle dont seules les parties manuscrites ont été remplies. Elles sont majoritairement composées de données onomastiques (noms de famille, prénom, lieux), de dates et de nombres. Elles peuvent contenir des noms de métier. Les formulaires de recensement étaient remplis par les habitant·es du foyer, ou, à défaut, par une connaissance.

19. U.-V. Marti et H. Bunke, « The IAM-database : An English Sentence Database for Offline Handwriting Recognition », *International Journal on Document Analysis and Recognition*, 5–1 (nov. 2002), p. 39-46, DOI : 10.1007/s100320200071.

20. Emmanuèle Grosicki, Matthieu Carré, Jean-Marie Brodin et Edouard Geoffrois, « Results of the RIMES Evaluation Campaign for Handwritten Mail Processing », *2009 10th International Conference on Document Analysis and Recognition* (, 2009), p. 941-945, DOI : 10.1109/ICDAR.2009.224.

21. Dubois Alain, T. Clérice, Delphine Mamie, Schlaeppi Darius, Clémence Rudaz et Marie-Caroline Schmied, *Tables Du Recensement Du Valais*, oct. 2022.

22. Un *token* est un ensemble de signes de ponctuation ([^\w\s]+) ou un ensemble de caractères hors signe ponctuation et espaces ([\w]+).

Corpus	**Valais**	**CREMMA AN Testament de Poilus**	**CREMMA MSS 19**	**IAM**	**CREMMA Wiki**	**RIMES**
Langue	Fr / De	Fr	Fr / En	En	Fr/ Unk	Fr
Tokens*	106657	19896	14561	**118375**	18840	109807
- dont uniques	4561	3929	3711	**13208**	5363	6570
Fréquence moyenne des tokens	**23.38**	5.06	3.92	8.96	3.51	16.71
Part du corpus représenté par les 10 tokens les plus fréquents	**41.59 %**	23.29 %	26.19 %	28.09 %	27.38 %	30.39 %
- hors ponctuation	**28.90 %**	18.97 %	14.83 %	18.89 %	20.58 %	19.47 %
Nombre moyen de tokens [22] uniques par document	41.25	62.17	**123**	56.07	45	N/A
Nombre médian de tokens uniques par document	40.00	56.5	**111**	56	42	N/A
Nombre moyen de tokens par ligne	1.28	5.97	8.05	8.86	**11.01**	9.07
Nombre médian de tokens par document	79	72.00	**176.00**	77	57	N/A
Nombre moyen de tokens par document	83.07 +/- 48.70	88.03 +/- 72.5	211 +/- 136.49	76.92	**304**	N/A

TABLE 1 – Comparaison de l'importance matérielle et de la diversité de vocabulaire de jeux de données similaires à CREMMA-AN-TestamentsdePoilus ou créés dans le cadre de CREMMA.

5 Mise en œuvre

Pour créer de la vérité de terrain à partir du corpus Testaments de Poilus, il fallait d'une part aligner les transcriptions et les images, et d'autre part vérifier que le texte correspondait aux règles de transcriptions prévues pour les corpus CREMMA [23].

Nous avons procédé à la création de deux sous-ensembles (batch 1 et batch 2) correspondants à chaque fois à un échantillonnage des 432 testaments édités et publiés dans le répertoire Github ArchivesNationalesFR/editionTestamentsDePoilus. Le jeu de données original était filtré de telle manière qu'un sous-ensemble contient un testament par département, sélectionné au hasard. Le premier sous-ensemble contenait ainsi 84 testaments (pour un total de 130 images) et le second en contenait 62 (pour 96 images). Cette chute du nombre de testament d'un sous-ensemble à un autre est simplement le fruit des sélections des éditions de Testaments de Poilus : tous les départements ne sont pas également représentés, certains départements ne possédant qu'un testament, là où d'autres en ont plusieurs dizaines. Sachant par avance que nous ne pourrions adapter l'ensemble des testaments, ce filtrage permettait de conserver l'attrait principal de l'édition pour l'HTR, à savoir les diversités de main, de lieu, de vocabulaire, etc.

Pour transformer l'édition en un jeu de données d'entraînement pour l'HTR, nous avons procédé par plusieurs étapes. Pour chaque sous-ensemble, nous récupérions les images ainsi que les fichiers XML TEI. Ces derniers étaient lus par un script, qui en extrayait uniquement le contenu textuel, ligne par ligne, en ne gardant que les formes non résolues des abréviations. Ensuite, les images étaient chargées dans l'application eScriptorium avant qu'on y applique un modèle de segmentation. La segmentation désigne l'étape préalable à la transcription et qui consiste à détecter l'emplacement des lignes de texte sur l'image et à analyser sa mise en page afin de regrouper les lignes, en paragraphes par exemple. Comme Testaments de Poilus s'est limité à la transcription des textes allographes, nous n'avons conservé que les lignes correspondant à cette partie du texte. Nous avons ensuite appliqué aux lignes et aux régions détectées le vocabulaire contrôlé SegmOnto [24], ce qui permettait de distinguer, par exemple, le texte interlinéaire du texte normal ou encore les zones de tampon, les notes marginales de la zone de texte principale. Ensuite, la transcription extraite des fichiers XML était intégrée dans le résultat de la segmentation, avant d'être modifiée, si nécessaire, pour correspondre aux règles de transcription des corpus CREMMA. Par exemple, une portion de texte suscrit, généralement utilisée pour une abréviation, est signalée à l'aide du signe « ⁀ » : cela donne par exemple « M⁀me » pour l'abréviation de Madame, lorsqu'elle est écrite « M^{me} » [25]. L'ensemble a ensuite été exporté hors eScriptorium, sous la forme de fichiers XML ALTO [26], standard utilisé pour le partage

23. Les normes de transcriptions utilisées pour l'alignement de CREMMA-AN-TestamentsdePoilus suivent les premières approches utilisées lors du projet LECTAUREP, formalisées par Alix Chagué dans A. Chagué et T. Clérice, *Règles générales de transcription pour les corpus CREMMA*, sept. 2022. En 2023, les membres des différents projets utilisant des règles de transcription similaires (graphématiques conservatrices) ont décidé de normaliser leurs approches en diachronie et de préparer cette normalisation sous le nom de CATMuS (*Consistent Approach for Transcribing Manuscripts*).

24. Simon Gabay, A. Pinche, Kelly Christensen, J.B. Camps et Nicolas Carboni, *SegmOnto, A Controlled Vocabulary to Describe the Layout of Pages*, Geneva/Lyon/¨Paris, 2023.

25. À l'heure de l'écriture de ce chapitre, il s'agit d'un élément pouvant être amené à changer afin de normaliser les pratiques avec les normes proposées par Ariane Pinche dans le cadre du projet CATMuS.

26. Library of Congress, *Analyzed Layout and Text Object (ALTO)*, 2020.

des données de l'HTR.

Les fichiers XML ALTO et les images, accompagnés des scripts utilisés pour la composition des sous-ensembles et l'extraction des transcriptions, ont été publiés dans un nouveau répertoire Github : HTR-United/CREMMA-AN-TestamentsDePoilus. Ce répertoire suit les recommandations établies par l'organisation HTR-United [27] et permet d'appliquer des outils de validations des données tels que HTRVX [28], qui vérifient la qualité des fichiers XML générés et la conformité aux principes de SegmOnto. Le répertoire est synchronisé avec Zenodo, et dispose donc d'un identifiant unique, en plus d'être versionné : 10.5281/ZENODO.10177106 [29].

Finalement, le corpus CREMMA-AN-TestamentsDePoilus, comme l'ensemble des jeux de données produits dans le cadre des vacations CREMMA, a été signalé dans le catalogue HTR-United. Ce signalement permet de le rendre visible à l'ensemble de la communauté qui peut alors s'en servir pour entraîner un modèle de transcription ou pour composer un nouveau jeu de données d'entraînement.

L'ensemble des transcripteur·rices originaux, ainsi que les responsables du projet Testaments de Poilus, ont été repris comme co-auteur·rices de notre jeu de données. S'y sont ajoutés les porteurs du projet CREMMA, ainsi que les annotatrices chargées de l'alignement. Ces informations d'autorité sont difficiles à intégrer à l'intérieur des fichiers XML ALTO, mais elles sont systématiquement indiquées dans l'ensemble des moyens de citations du jeu de données : description pour le catalogage HTR-United, métadonnées Zenodo, fichier « CITATION.CFF ».

Grâce à sa publication, une vie ultérieure pour les données du projet Testaments de Poilus a ainsi été de contribuer à l'entraînement d'un modèle générique pour le français moderne et contemporain (XVII^{e}-XXI^{e} siècles) que nous avons évoqué plus haut : Manu McFrench [30].

Conclusion

Le projet Testaments de Poilus n'avait (peut-être) pas à l'esprit, lors de sa conception, la possible réutilisation de ses données par des collègues pour de la reconnaissance d'écriture. Cependant, guidé par une pratique des principes FAIR (Findable, Accessible, Interoperable, Reusable), le projet a rendu possible sa ré-exploitation par d'autres projets, en particulier par CREMMA.

Lors de l'établissement du nouveau jeu de données dérivé de Testaments de Poilus, la question de la citation est devenue majeure, et est un problème peut-être rapidement oublié quand il s'agit de données FAIR. En créant un nouvel objet, nous pouvions de facto invisibiliser l'objet original, et donc ses autorités. En l'occurrence, contrairement aux travaux effectués à l'aide des éditions numériques ou papier où le document était support de lecture, il s'agit bien de la transcription originale des auteurs, adaptée à nos normes, que nous avons exploitée. Dans ce cadre, si le jeu de données enfant

27. L'organisation GitHub HTR-United peut être trouvée à l'URL suivante : https://github.com/HTR-United.

28. T. Clérice et A. Pinche, *HTRVX, HTR Validation with XSD*, sept. 2021, DOI : 10.5281/zenodo.5359963.

29. A. Chagué, T. Clérice, A. Mazoue et Elsa Van Kote, *CREMMA-AN-TestamentDePoilus*, nov. 2023, DOI : 10.5281/ZENODO.10177106.

30. Id., « Manu McFrench, from Zero to Hero… ».

ne faisait que citer le jeu de données parent, il y avait un risque de perdre cette trace, surtout dans la mesure où, contrairement aux documents TEI, la métadonnée d'autorité est extérieure aux fichiers ALTO : elle n'est inscrite que dans des fichiers de catalogage. Cette question de la transmission ou de redondance de l'autorité, quand il s'agit de documents dérivés, est un enjeu important pour la reconnaissance des responsabilités mais surtout du travail. Similairement aux créateurs de logiciels de recherche, les créateurs de corpus vivent en effet cette drôle de situation où, sans leur travail, peu de modèles (en sciences de l'informatique) ou d'analyses (en sciences humaines) voient le jour, et où ils font cependant partie de la population la plus invisibilisé dans les articles. Des initiatives comme la partie « Linguistic Resource » de LREC Coling, comme les cartes de dataset dans HuggingFace, de même que les datapapers, pourraient être des réponses à cette situation, si les changements de pratiques suivent la même trajectoire (ce qui semble être le cas).

CREMMA, grâce à cette masse de données textuelles alignées avec des images, a réussi à mettre à disposition, avec un budget finalement réduit de quelques milliers d'euros, deux modèles génériques : l'un pour la littérature médiévale, l'autre pour le français en cursive. La disponibilité des données de Testaments de Poilus a ainsi permis de produire des modèles génériques réexploitables : le meilleur exemple concerne probablement le projet EpiSearch (JDMDH) qui, ayant eu accès au premier modèle Manu McFrench avant sa publication, a réussi à exploiter avec un très haut taux de réussite des documents manuscrits en latin et italien du XVIII^{e} siècle [31].

Après cette expérience les Testaments de Poilus montrent, à travers la réutilisation totale de leurs données (image, texte, métadonnées) la capacité pour des projets à être exploités autrement et à nourrir d'autres pans de recherche inattendus. Testaments de Poilus et CREMMA-AN-TestamentsDePoilus sont une preuve supplémentaire qu'une FAIRisation des données est un atout scientifique important. Qui sait, peut-être que les données d'entraînement HTR seront utilisées un jour pour de la linguistique de corpus ?

Références

Alain (Dubois), Clérice (Thibault), Mamie (Delphine), Darius (Schlaeppi), Rudaz (Clémence) et Schmied (Marie-Caroline), *Tables Du Recensement Du Valais*, oct. 2022.

Bommasani (Rishi), Hudson (Drew A.), Adeli (Ehsan), Altman (Russ), Arora (Simran), von Arx (Sydney), Bernstein (Michael S.), Bohg (Jeannette), Bosselut (Antoine), Brunskill (Emma), *et al.*, « On the Opportunities and Risks of Foundation Models » (, juill. 2022), doi : 10.48550/arXiv.2108.07258.

Calvelli (Lorenzo), Boschetti (Federico) et Tommasi (Tatiana), *EpiSearch. Identifying Ancient Inscriptions in Epigraphic Manuscripts*, déc. 2022.

Camps (Jean-Baptiste), *Homemade manuscript OCR (1) : OCRopy*, Billet, févr. 2017.

Causer (Tim), Grint (Kris), Sichani (Anna-Maria) et Terras (Melissa), « 'Making Such Bargain' : Transcribe Bentham and the Quality and Cost-Effectiveness of Crowdsourced Transcription », *Digital Scholarship in the Humanities*, 33–3 (sept. 2018), p. 467-487, doi : 10.1093/llc/fqx064.

31. Lorenzo Calvelli, Federico Boschetti et Tatiana Tommasi, *EpiSearch. Identifying Ancient Inscriptions in Epigraphic Manuscripts*, déc. 2022.

Chagué (Alix) et Clérice (Thibault), *Règles générales de transcription pour les corpus CREMMA*, sept. 2022.
Chagué (Alix), Clérice (Thibault), Mazoue (Anaïs) et Van Kote (Elsa), *CREMMA-AN-TestamentDePoilus*, nov. 2023, DOI : 10.5281/ZENODO.10177106.
– « Manu McFrench, from Zero to Hero : Impact of Using a Generic Handwriting Recognition Model for Smaller Datasets », dans *Digital Humanities 2023 : Collaboration as Opportunity*, 2023.
Clérice (Thibault) et Pinche (Ariane), *HTRVX, HTR Validation with XSD*, sept. 2021, DOI : 10.5281/zenodo.5359963.
Clérice (Thibault), Pinche (Ariane) et Vlachou-Efstathiou (Malamatenia), *Generic CREMMA Model for Medieval Manuscripts (Latin and Old French), 8-15th Century*, 2023, DOI : 10.5281/zenodo.7631619.
de Champs (Emmanuelle), « Des bénévoles au service du patrimoine écrit. De l'Oxford English Dictionary aux Testaments de Poilus » (, 2021), p. 47.
Gabay (Simon), Pinche (Ariane), Christensen (Kelly), Camps (Jean-Baptiste) et Carboni (Nicolas), *SegmOnto, A Controlled Vocabulary to Describe the Layout of Pages*, Geneva/Lyon/¨Paris, 2023.
Grosicki (Emmanuèle), Carré (Matthieu), Brodin (Jean-Marie) et Geoffrois (Edouard), « Results of the RIMES Evaluation Campaign for Handwritten Mail Processing », *2009 10th International Conference on Document Analysis and Recognition* (, 2009), p. 941-945, DOI : 10.1109/ICDAR.2009.224.
Hodel (Tobias), Schoch (David), Schneider (Christa) et Purcell (Jake), « General Models for Handwritten Text Recognition : Feasibility and State-of-the Art. German Kurrent as an Example », *Journal of Open Humanities Data*, 7–0 (juill. 2021), p. 13, DOI : 10.5334/johd.46.
Kiessling (Benjamin), Tissot (Robin), Stokes (Peter Anthony) et Stökl Ben Ezra (Daniel), « eScriptorium : An Open Source Platform for Historical Document Analysis », dans *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, 2019 (2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)), t. 2, p. 19-19, DOI : 10.1109/ICDARW.2019.10032.
Library of Congress, *Analyzed Layout and Text Object (ALTO)*, 2020.
Marti (U.-V.) et Bunke (H.), « The IAM-database : An English Sentence Database for Offline Handwriting Recognition », *International Journal on Document Analysis and Recognition*, 5–1 (nov. 2002), p. 39-46, DOI : 10.1007/s100320200071.
Nockels (Joe), Gooding (Paul), Ames (Sarah) et Terras (Melissa), « Understanding the Application of Handwritten Text Recognition Technology in Heritage Contexts : A Systematic Review of Transkribus in Published Research », *Archival Science* (, juin 2022), DOI : 10.1007/s10502-022-09397-0.
Pinche (Ariane), *Edition Nativement Numérique Du Recueil Hagiographique "Li Seint Confessor" de Wauchier de Denain d'après Le Manuscrit Fr. 412 de La Bibliothèque Nationale de France*, These de Doctorat, Lyon, 2021.
– *Guide de Transcription Pour Les Manuscrits Du Xe Au XVe Siècle*, juin 2022.
Ruiz-Parrado (Victoria), Heradio (Ruben), Aranda-Escolastico (Ernesto), Sánchez (Ángel) et Vélez (José F.), « A Bibliometric Analysis of Off-Line Handwritten Document Analysis Literature (1990–2020) », *Pattern Recognition*, 125 (mai 2022), p. 108513, DOI : 10.1016/j.patcog.2021.108513.
Tarride (Solène), Faine (Tristan), Boillet (Mélodie), Mouchère (Harold) et Kermorvant (Christopher), *Handwritten Text Recognition from Crowdsourced*

Annotations, juin 2023, DOI : 10.48550/arXiv.2306.10878, arXiv : 2306.10878 [cs].

TERRAS (Melissa), *For Ada Lovelace Day – Father Busa's Female Punch Card Operatives*, oct. 2013.

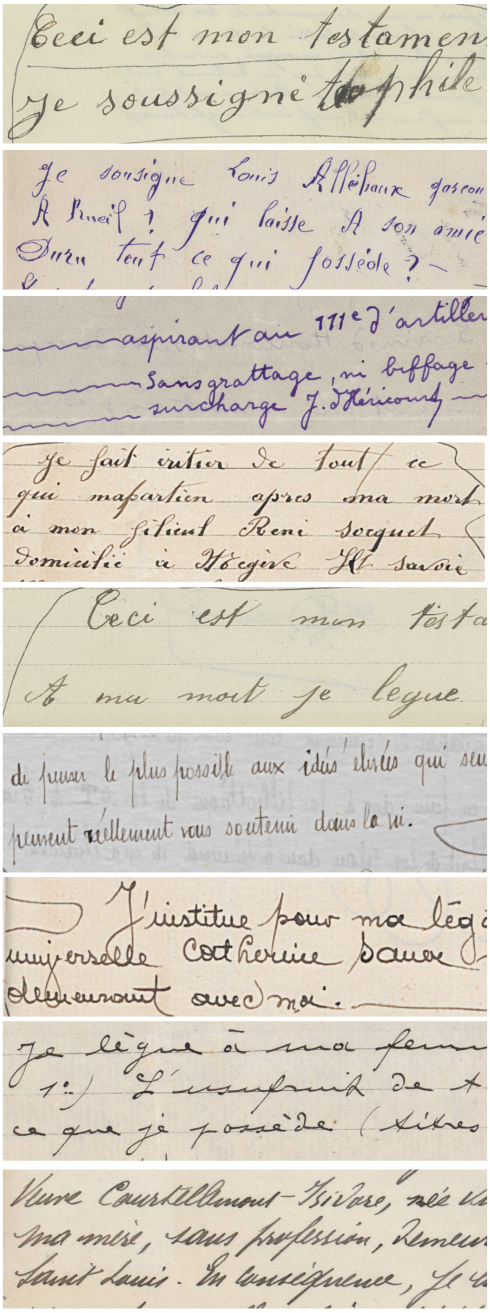

FIGURE 2 – Sélection de lignes extraites des numérisations des Testaments de Poilus, illustrant la diversité matérielle et d'aspect des testaments composant le corpus.