



HAL
open science

Corpus, méthodes et ressources pour la transcription automatique des documents manuscrits patrimoniaux francophones contemporains

Alix Chagué

► To cite this version:

Alix Chagué. Corpus, méthodes et ressources pour la transcription automatique des documents manuscrits patrimoniaux francophones contemporains. 89e Congrès de l'Acfas, Section 310 - Le numérique dans les sciences humaines: édition et visualisation, ACFAS; Micahel Eberle Sinatra; Marcello Vitali-Rosati, May 2022, Montréal, Canada. hal-03664788

HAL Id: hal-03664788

<https://inria.hal.science/hal-03664788v1>

Submitted on 3 Jun 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Alix Chagué

Corpus, méthodes et ressources pour la transcription automatique des documents manuscrits patrimoniaux francophones contemporains

Sous la direction de Laurent Romary, Emmanuel Chateau-Dutier et de Michael Sinatra

Université de Montréal et École Pratique des Hautes Etudes (Paris) ; CRIHN (UdeM) et ALMAnACH (Inria-Paris)

Contexte

TiTranskribus®



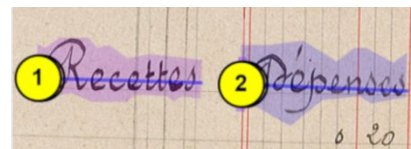
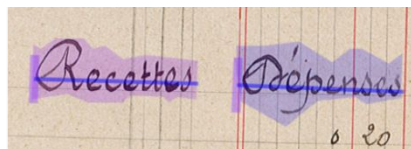
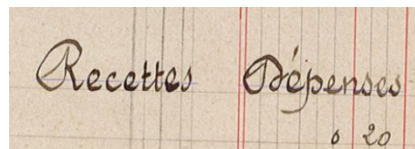
- HTR = Handwritten Text Recognition
- Des logiciels grand public pour l'HTR : Trankribus, eScriptrium

Usages :

- Faciliter l'accès à des collections manuscrites (moteur de recherche)
- Rendre compatible avec la fouille de texte
- Repenser les premiers maillons des chaînes de traitement pour l'édition numérique



Contexte : Les grandes étapes de l'HTR



Recettes Dépenses

Segmentation
détection de l'emplacement des lignes
(baseline et masque)

Analyse de la mise en page
ex : calcul de l'ordre de lecture
des segments

Transcription
chaque portion d'image est associée à
une chaîne de caractères

Machine à lire les lettres
illisibles



Présentation du problème

- appropriation empirique par les GLAMS et les HN
- = manque un cadre méthodologique et conceptuel pour appliquer l'HTR
- à partir de l'exemple de documents contemporains français

Méthodologie en chantier

- commencer par une analyse critique des étapes du workflow



Focus : universalisation des modèles de transcription

Difficultés :

- très grande variation dans la forme des lettres
- graphèmes et d'abréviations propres à chacun-e
- accidents et lisibilité naturelle des mots

2 situations :

- imprécision de la transcription compensée par recherche floue
- entraînement de modèles spécifiques à une écriture ou une famille d'écritures



Focus : Les données d'entraînement

- mutualiser les données d'entraînement
- homogénéiser les pratiques de transcription
- décrire ces jeux de données
- construire des jeux équilibrés



HTR
United

- initiative pour le partage libre de métadonnées sur les jeux de données
- laboratoire d'expérimentation (schéma de description, seuil de qualité, créer des réflexes)

What's next?

- continuer à prendre connaissance de l'état de l'art
- sélectionner les questions de recherche les plus pertinentes
- commencer à analyser les écritures françaises des XIXe et XXe siècles notamment à partir des données d'HTR-United

Merci !



Références des illustrations

- Whistler (1858). Reading by lamplight (Detail) [Etching]. Rijksmuseum.
- Pellan (1960). Machine à lire les lettres illisibles [Crayon à bille sur papier quadrillé]. Musée national des beaux-arts du Québec.
- Rizzuto (1957). [View of construction workers on steel beams with cranes in background] [Gelatin silver print]. Library of Congress.
- Van den Valckert (1624). Four Regents and the Housemaster of the Leper-House, Amsterdam (Detail) [Oil on panel]. Rijksmuseum.
- Gijsbrechts (1668). Trompe l'oeil. Board Partition with Letter Rack and Music Book [Oil on canvas]. Statens Museum for Kunst.
- Julien (1894). « Des lettres? Oui, oui, oui! Criait le bon diable tout essoufflés ». Illustration pour Chouinard, conte de Louis Fréchette (Detail) [Encre de Chine sur papier]. Musée national des beaux-arts du Québec.