CAMPUS CONDORCET | HUMATHÈQUE

La transcription texte à texte : enjeux, méthodes, outils

Humathèque Condorcet
Service Formations des usagers

Alyx Taounza-Jeminet

Octobre 2025

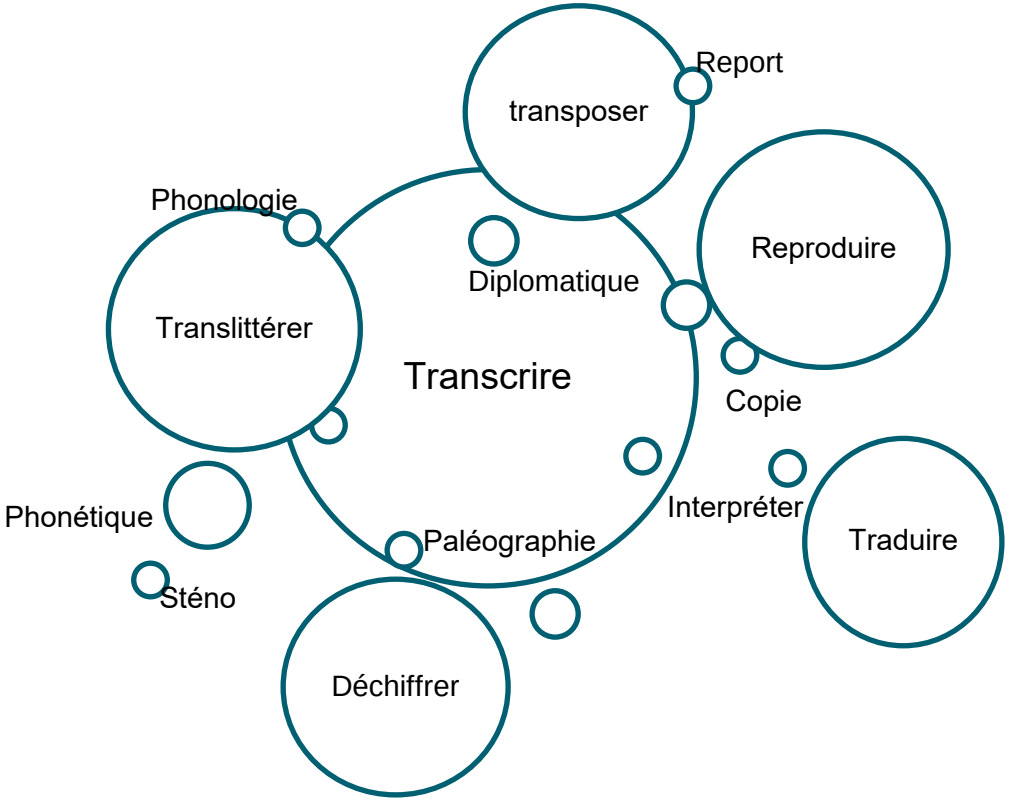
CAMPUS CONDORCET | HUMATHÈQUE

Les objectifs de la formation

1. Identifier les différentes formes de transcription et les recontextualiser
2. Découvrir l'étendue des outils de transcription et se familiariser avec un panel d'outils de transcription manuelle et automatique
3. Considérer la transcription comme étape du travail de recherche et formaliser une chaine de traitement

CAMPUS CONDORCET | HUMATHÈQUE

De quoi parlons-nous ?



CAMPUS CONDORCET | HUMATHÈQUE

De quoi parlons-nous ?

Terminologies

Transcription phonétique
Translittération
Romanisation
Speech to Text

Transcription diplomatique
Transcription allographétique
Transcription graphématique
Transcription imitative

Transcription manuelle
Transcription assistée par ordinateur (TAO)
Intelligent Character Recognition (ICR)
Optical Character Recognition (OCR)
Handwritten Character Recognition (HTR)
Automated Character Recognition (ATR)

CAMPUS CONDORCET | HUMATHÈQUE

De quoi ne parlons-nous pas ?

Aparté sur la translittération et la transcription phonétique

La translittération est un mode de conversion graphème à graphème d'un système d'écriture vers un autre. Une translittération, contrairement à une transcription phonétique ou phonologique, est réversible.

La romanisation est le fait de translittérer vers une écriture latine.

De nombreuses translittérations possèdent aujourd'hui des normes et systèmes internationaux.

« La transcription est l'opération visant à noter la prononciation d'une langue donnée au moyen du système de signes d'une langue de conversion. » En linguistique, sons et phonèmes sont convertis en signes selon des systèmes symbolisant l'oral, le plus connu étant l'alphabet phonétique international (API) et ses déclinaisons.

CAMPUS CONDORCET | HUMATHÈQUE

De quoi ne parlons-nous pas ?

Aparté sur la transcription de l'oral à l'écrit

Entretiens, enregistrements, les ressources orales sont d'une nature singulière mais ne sont pas si éloignés des documents écrits.

Les pratiques de transcription de l'oral ne font pas l'objet d'un formalisme unifié ou d'une norme mais de choix et interprétations qui ne peuvent qu'être encadrées par des bonnes pratiques. L'écart entre la source et sa transposition est d'autant plus important et intranscriptible de l'oral à l'écrit, bien que divers systèmes (métalangage, transcription phonétique) tentent de le combler.

Des logiciels de reconnaissance et transcription (ou *speech to text*) existent aujourd'hui qui facilitent et accélèrent ce travail : une solution pour les chercheurs en SHS est l'usage du logiciel Whisper via l'interface Sharedocs de Huma-Num.

CAMPUS CONDORCET | HUMATHÈQUE

De quoi parlons-nous ?

Ce que cette formation présente

- La transcription texte à texte
- Les outils de reconnaissance automatisée
- Des recommandations

Ce que cette formation ne couvre pas

- La transcription de l'oral
- La translittération
- L'encodage (XML-TEI)
- l'édition numérique

CAMPUS CONDORCET | HUMATHÈQUE

De quoi parlons-nous ?

Définitions de la transcription dans le contexte restreint de cette formation :

« […] l'effort de reporter précisément (dans les limites imposes par la typographie) ce qui constitue l'inscription textuelle d'un manuscrit. »

Vander Meulen, D. L. and Tanselle., G. T. (1999). A system of manuscript transcription. Studies in Bibliography,52: 201–12.

« […]aucune transcription de ces manuscrits dans un format lisible par un ordinateur ne pourra jamais être considérée "finale" ou "definitive". La transcription vers l'ordinateur est une pratique fondamentalement interprétative, composée d'une série d'actions de traduction d'un système de signes (celui du manuscrit) vers un autre (celui de l'ordinateur). »

Robinson, P., & Solopova, E. (1993). Guidelines for Transcription of the Manuscripts of the Wife of Bath's Prologue. In Canterbury Tales Project Occasional Papers (p. 19-52). Zenodo. https://doi.org/10.5281/zenodo.11954056

CAMPUS CONDORCET | HUMATHÈQUE

De quoi parlons-nous ?

Qui transcrit ?
Qu'est-ce qui est transcrit ?
Pour quels besoins ?
Selon quelles orientations ?
Avec quels outils ?

CAMPUS CONDORCET | HUMATHÈQUE

De quoi parlons-nous ?

Qui transcrit ?

- Les sténographes
- Les secrétaires et traducteurs de débats
- Les sous-titreurs
- Les généalogistes
- Les étudiants et chercheurs
  - En linguistique
  - En histoire
  - En sociologie
  - En littérature comparée
  - etc.
- L'intelligence artificielle ?

CAMPUS CONDORCET | HUMATHÈQUE

De quoi parlons-nous ?

Qu’est-ce qui est transcrit ?

- Documents administratifs
- Inscription épigraphique ou son estampage
- Discours, réunion, entretien, podcast…
- Fichier audio/vidéo
- Document imprimé
- Document manuscrit
- Une archive ou sa reproduction
- Des pensées ou sentiments… un paysage…

CAMPUS CONDORCET | HUMATHÈQUE

De quoi parlons-nous ?

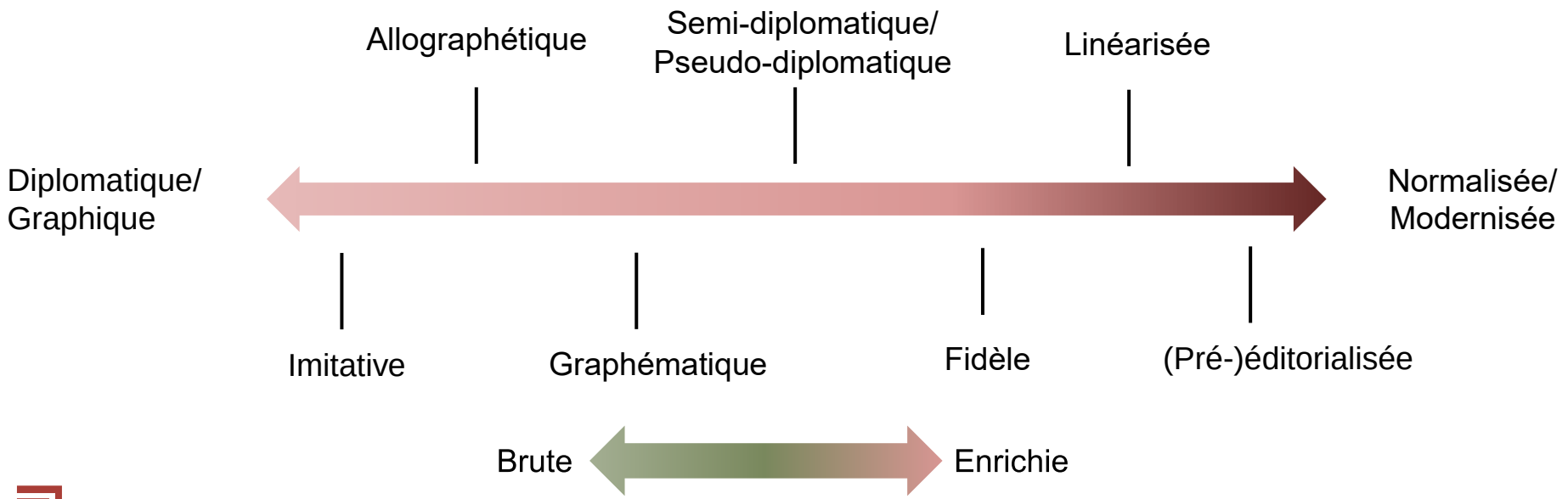
Pour quels besoins ?

- Citation
- Accessibilité et "cherchabilité" du texte
- Valorisation de collections
- Édition critique
- Étude génétique
- Enrichissement de texte et traitement du langage naturel : multimédia, reconnaissance d’entités nommées, etc.
- Analyse de données textuelles : lemmatisation, lexicométrie, textométrie, TAL, etc.

CAMPUS CONDORCET | HUMATHÈQUE

De quoi parlons-nous ?

Selon quelles orientations ? Termes courants



CAMPUS CONDORCET | HUMATHÈQUE

De quoi parlons-nous ?

Selon quelles normes ?

Les écrits anciens et les paléographes n'ont pas les mêmes besoins que les écrits modernes.

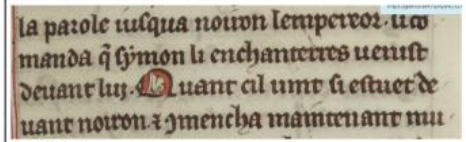
Une transcription pour traitement automatique des langues n'aura pas les mêmes contraintes qu'une transcription pour une édition critique.

Une édition papier n'aura pas les mêmes règles de transcription qu'une édition en ligne.

CAMPUS CONDORCET | HUMATHÈQUE

De quoi parlons-nous ?

Selon quelles normes ?


la parole iusqua noiron lempereor. li co-
manda q̄ symon li enchanterres uenist
deuant lui. Quant cil uint si estuet de-
uant noiron ⁊ ꝯmencha maintenant mu-

BnF, fr. 412

On transcrira :
« la parole iusqua noiron lempereor. li co-
manda q̄ symon li enchanterres uenist
deuant lui. Quant cil uint si estuet de-
uant noiron ⁊ ꝯmencha maintenant mu- »

TABLEAU 2 – Les variantes graphiques d'une même lettre

Exemple	Source	Transcrire	Unicode
Les	[MALINGRE] 1533a p. 5	Les	U+0073
Chreſtiente	[MALINGRE] 1533a p. 5	Chrestiente	U+0073
‹s› (Latin small letter s, U+0073) et ‹ſ› (Latin small letter long s, U+017F) sont transcrits par ‹s›.			

LA PONCTUATION :

Rétablir une ponctuation moderne avec des espaces après les ponctuation (*exemple,* ou *exemple ;* ou *exemple :*), avant les ponctuations (*exemple ?* ou *exemple !)* et des majuscules en début de phrase.

Utiliser les guillemets français dans le texte *« exemple ».*

LES NOMS PROPRES :

Rétablir la majuscule aux noms propres : *Le Bailli, Le Bon Génie, La Sorcière.*

Privilégier une graphie moderne : *Choisy-le-Roi* plutôt que *Choisy-le-Roy* et conserver ce choix de graphie dans tout le texte.

LES ACCENTS :

Rétablir l'accentuation selon les règles de grammaire et d'orthographe modernes, par exemple mettre l'accent sur le *à* majuscule : *À* et la cédille au ç majuscule : *Ç.*

CAMPUS CONDORCET | HUMATHÈQUE

De quoi parlons-nous ?

Selon quelles normes ?

Fondamental : garder une homogénéité de la transcription. Il faut se créer un guide avec des choix établis en fonction des besoins. Certains existent, notamment pour harmoniser les transcriptions automatisées ou transmettre les instructions formelles aux transcripteurs extérieurs.

CAMPUS CONDORCET | HUMATHÈQUE

Avec quels outils ?
Panorama

Pratiques individuelles : logiciels de traitement de texte

Pratiques participatives : plateformes en ligne

Pratiques automatisées : écosystèmes IA

Les outils ainsi classés ne sont pas forcément contraints à une seule catégorie. Les pratiques peuvent également être combinées.

Les logiciels de traitement et d'édition de texte

Microsoft Word

Libre Office

Google Docs

Sublime Text

Notepad++

Oxygen

Max

Tropy

Les logiciels de traitement et d'édition de texte

Pour transcrire des caractères « rares » ou anciens, il faut privilégier un encodage UTF-8 du texte. Une extension baptisée MUFI permet d'accéder à un jeu encore étendu pour les caractères médiévaux.

MUFI

Suivant la nature des documents et des projets, le travail de mise en page et l'encodage XML-TEI peuvent être concomitants de la transcription ou faire appel au même logiciel. Le pôle documents numériques de la MRSH a développé une gamme d'environnements spécifiques.

MRSH
NORMANDIE - CAEN
Maison de la Recherche en Sciences Humaines
CNRS - UNIVERSITÉ DE CAEN

L'annotation IIIF

Le protocole IIIF normalise le partage et la diffusion d'images en ligne. Il permet à différents outils de s'y greffer pour la visualisation, la manipulation, mais également l'annotation.

Liiive est une proposition expérimentale d'annotation collaborative basée sur le IIIF.
D'autres outils sont développés autour de ces questions.

Les outils de transcription manuelle participatifs

« Les sciences et recherches participatives sont des formes de production de connaissances scientifiques auxquelles participent des acteurs de la société civile, à titre individuel ou collectif, de façon active et délibérée. »

Charte des sciences participatives, France, 2017

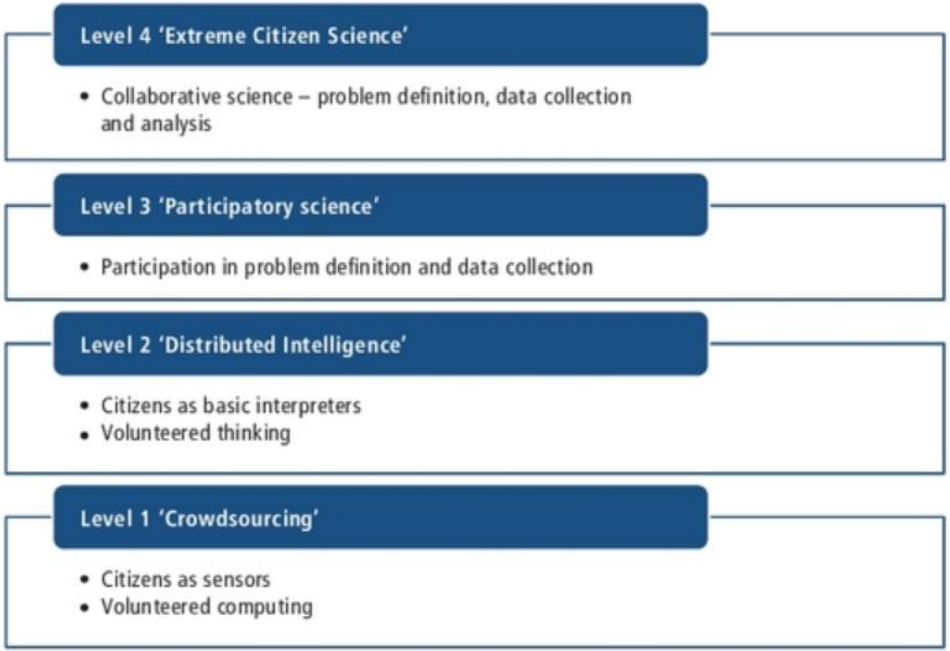
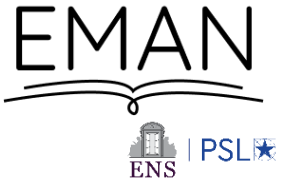


Fig. 4.1 Levels of participation in citizen science (Haklay 2013)

Les outils de transcription manuelle participatifs

Wikisource

Zooniverse



CAMPUS CONDORCET | HUMATHÈQUE

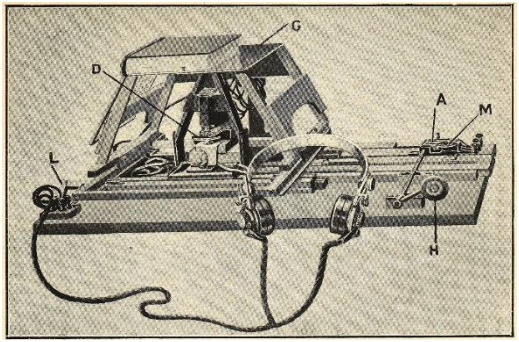
Les outils de transcription manuelle participatifs

Les intérêts de la transcription participative sont également ses contraintes :

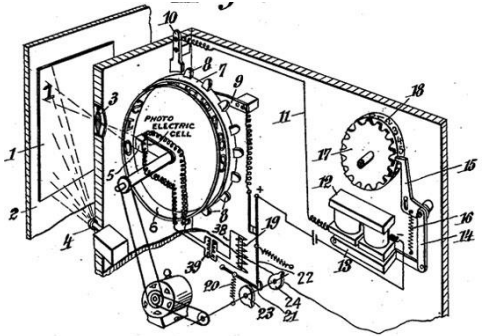
- Une plateforme dédiée est conçue : ce sont un site, une documentation, une communauté et une activité qui doivent être gérés.
- La valorisation et la visibilité du corpus débute dès sa mise en ligne pour transcription : les documents à transcrire doivent pouvoir être diffusés (tant au niveau de la qualité que juridiquement).
- La participation est ouverte (à tous ou à un public restreint) : le respect des consignes, la qualité et la rapidité de la transcription dépendent de la communauté.

Les outils de transcription automatisée

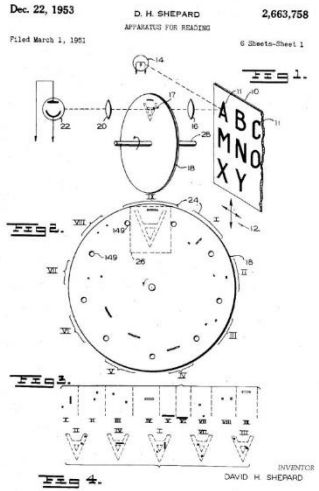
Historique de l'ATR



L'optophone d'Edmund Fournier d'Albe, 1913



La machine à lire de Gustav Tauschek, 1929



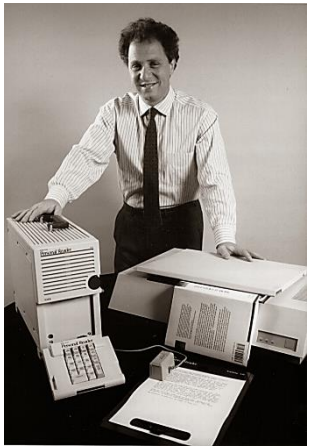
Le GISMO ou *Analyzing reader* de Shepard & Cook, 1953

Les outils de transcription automatisée

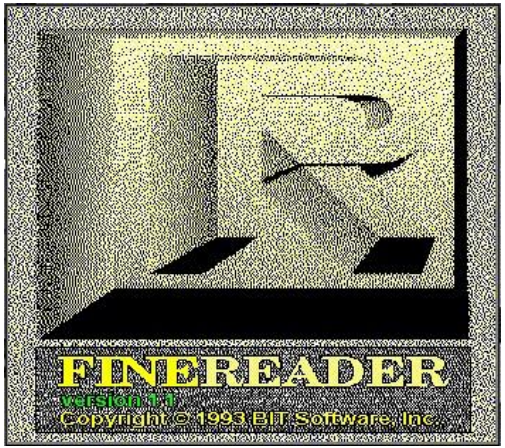
Historique de l'ATR

Figure 7-37. IBM 1287 Optical Reader Model 5

IBM 1287 Optical Reader utilisé en entreprise, 1964



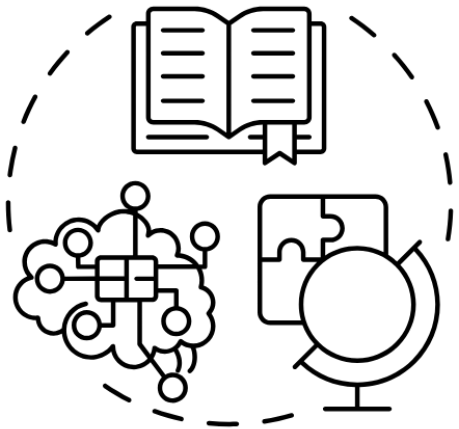
Ray Kurzweil, développeur de l'OCR omni-font et sa KRM, 1978



Splash Screen du logiciel FineReader 1.0 de ABBYY, 1993

Les outils de transcription automatisée

Historique de l'ATR



Deep Learning by bsd studio from Noun Project (CC BY 3.0)

L'introduction des réseaux neuronaux convolutifs et récurrents (CRNN) et l'apprentissage machine a permis de faire évoluer la reconnaissance automatique pour le manuscrit.

Aujourd'hui, les transformeurs et MLLM, plus rapides et pré-entraînés, prennent le pas sur ces réseaux.

La transcription automatisée reste aujourd'hui un processus qui demande une interaction entre l'humain et la machine, et la supervision humaine aux différentes étapes du traitement.

Les outils de transcription automatisée

Systèmes ATR contemporains :
Logiciels OCR : Abby - OCR4all – OCR-D – Tesseract
CRNN : Aletheia – Kraken – Monk – Pylaia
Transformeurs et MLLM : BERT – GPT – Claude- TrOCR

Transkribus® eS K4all CALFA TEKLIA MiKTEX P2T

Outils pour la recherche :
- Transkribus
- Kraken / eScriptorium
- OCR4all

Les prestataires privés :
- Teklia
- Calfa

Les outils pour typologies particulières :
- MixTeX
- Pix2Text
- mathpix

Les outils de transcription automatisée

Fonctionnement de l'ATR : le facteur humain

L'appel de l'IA peut sembler salvateur. Mais son utilisation demande des besoins humains particulier :

- Des compétences techniques : en code (python), en paléographie, etc.
- Du temps pour s'approprier l'outil
- Du temps pour entraîner la machine
- L'accès ou la création des ressources nécessaires (transcription existante, numérisations, etc.)

CAMPUS CONDORCET | HUMATHÈQUE

Les outils de transcription automatisée

Fonctionnement de l'ATR : les grandes étapes

1. Récupération des données
2. Traitement préparatoire (*preprocessing*)
3. Segmentation (*layout analysis*)
4. Transcription
5. Post-traitement (*postprocessing*)

Un glossaire de référence
https://harmoniseatr.hypotheses.org

CAMPUS CONDORCET | HUMATHÈQUE

Les outils de transcription automatisée

Fonctionnement de l'ATR : récupération des données (i.e. les images)

Scénario 1 : une numérisation existe

- Appel des images via protocole IIIF
- Chargement des fichiers en local

Scénario 2 : il n'y a pas de numérisation

- Faire une demande auprès de l'institution détentrice ou d'un prestataire
- Réaliser la numérisation soi-même (ex: avec une scantent)

Les outils de transcription automatisée

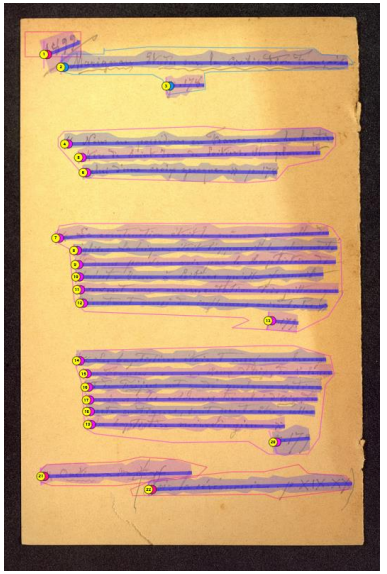
Fonctionnement de l'ATR : le preprocessing (facultatif)

Vérifier la qualité des images et traiter si nécessaire :

- Résolution (300dpi)
- Contraste (binariser l'image)
- Luminosité
- Rognage (pour éviter le bruit)
- Imagerie multispectrale
- Correction de la distorsion

Les outils de transcription automatisée

Fonctionnement de l'ATR : la segmentation ou zonage (*layout analysis*)



Détecter les zones, les segmenter, les ordonner :

1. Les lignes d'écriture (*baselines, centerline ou topline*)
2. Les masques ou polygones

La détection de lignes est essentielle pour l'extraction du texte. C'est uniquement le contenu visuel compris dans ce masque et adossé à la ligne qui va être transcrit.
Suivant le système d'écriture, il faut privilégier *baseline* ou *topline*).

CAMPUS CONDORCET | HUMATHÈQUE

Les outils de transcription automatisée

Fonctionnement de l'ATR : la segmentation ou zonage (*layout analysis*)

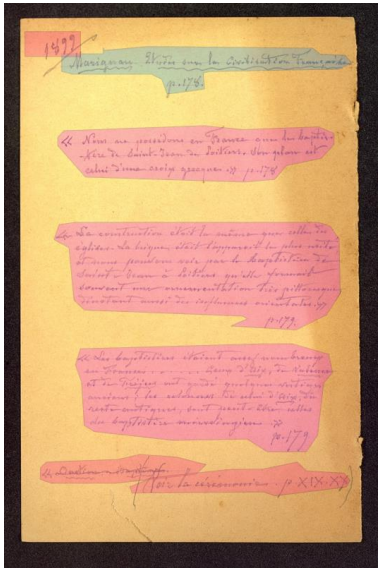
Détecter les zones et les segmenter :

3. Les régions

La détection de régions permet d'identifier et pré-encoder celles-ci (titres, marges, etc.) pour retenir la mise en page.

L'outil Segmonto permet d'intégrer un vocabulaire normalisé.

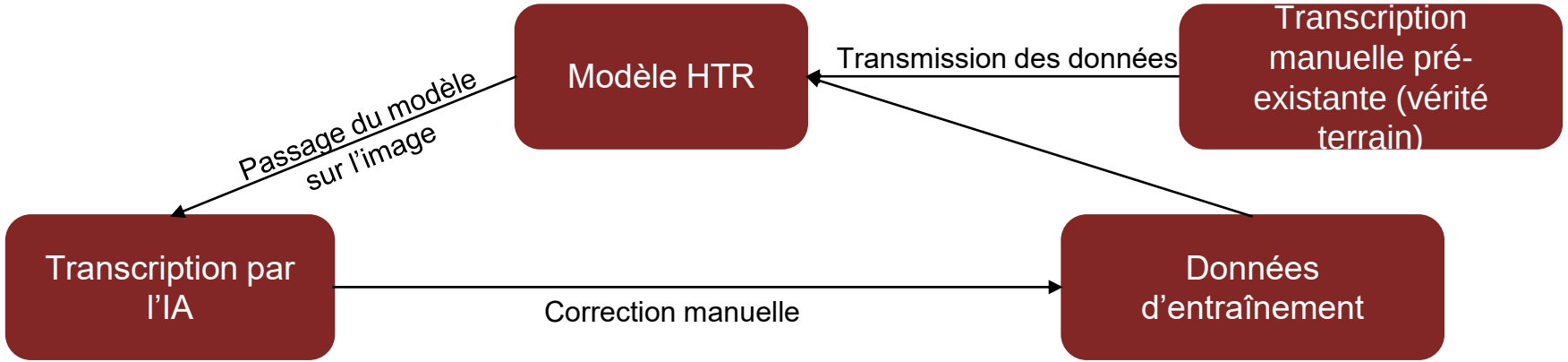
De nouveaux modèles de vision par ordinateur (*computer vision*) comme YOLOv5 apportent aujourd'hui de nouvelles capacités de détection.



Les outils de transcription automatisée

Fonctionnement de l'ATR : transcrire

La transcription n'est pas une étape unique et définitive. Transcrire demande de faire appel à un modèle de transcription et de l'entraîner jusqu'à arriver à un résultat satisfaisant ou à un plateau.



Les outils de transcription automatisée

Fonctionnement de l'ATR : les modèles

Il existe des modèles de segmentation comme des modèles de transcription.

Les modèles sont en général spécialisés : il fonctionneront sur des textes d'une écriture, d'une période, d'une mise en page, d'une graphie similaires.
Des modèles génériques sont en développement : ils possèdent des capacités de prédiction largement étendues.

La communauté de l'ATR met à disposition des modèles à la suite de projets sur Zenodo par exemple, en plus de ceux proposés par certains outils.

Les outils de transcription automatisée

Fonctionnement de l'ATR : le taux d'erreur

La transcription produite contiendra des erreurs : suppression d'un caractère, remplacement par un autre, ajout d'un caractère absent (hallucination).

Pour calculer le taux d'erreur, on compare une transcription faite par nous avec une transcription des mêmes pages faite par la machine :

- CER ou taux d'erreur par caractère
- WER ou taux d'erreur par mot

Une transcription peut être exploitée à différents niveaux de qualité :

- 80% permet déjà une recherche de texte
- 90% permet un post-traitement efficace
- 95% restreint les erreurs à des mots rares ou inconnus

Les outils de transcription automatisée

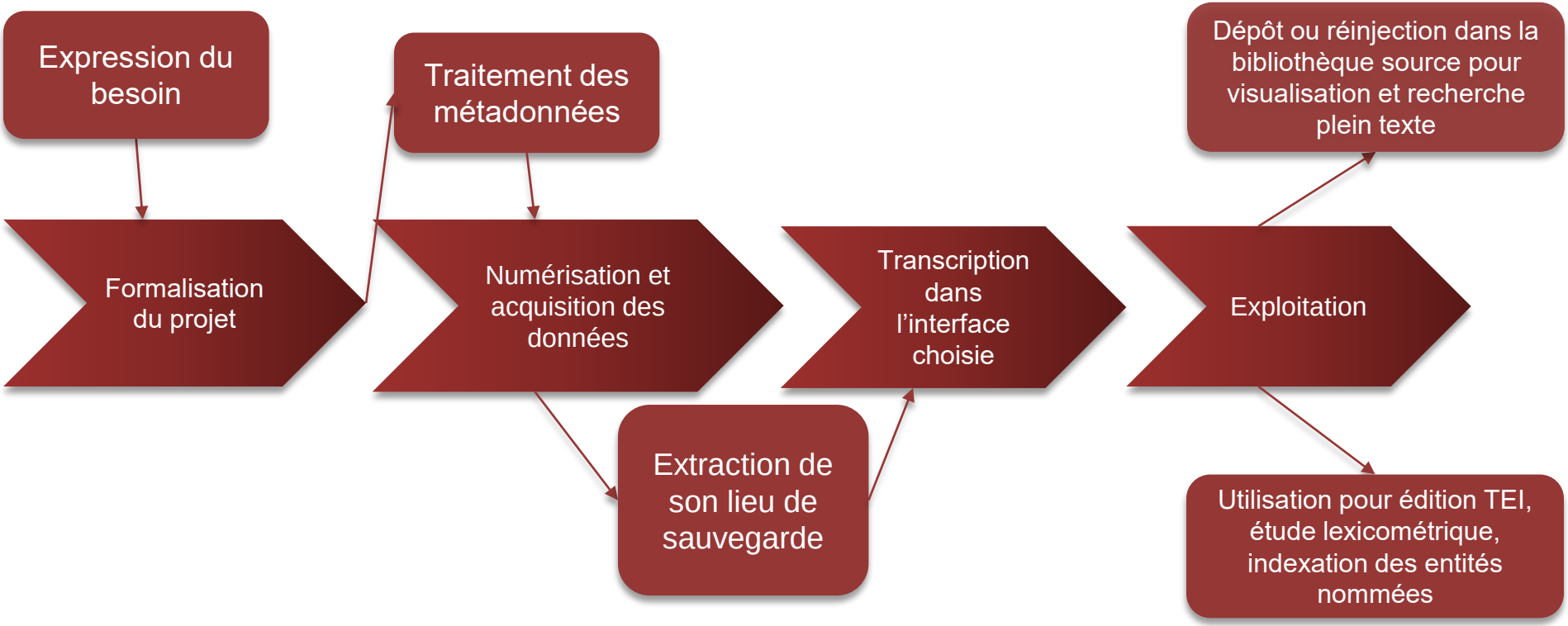
Fonctionnement de l'ATR : post-traitement et export des jeux de données

Une fois une transcription satisfaisante obtenue, les données sont exportables sous des formats XML (Alto ou Page) ou simples TXT. Ils peuvent à leur tour être traités par un modèle de langue pour relever des erreurs de transcriptions, des développements d'abbréviations, convertis pour un encodage TEI, enrichis, etc.

Ces exports peuvent avoir une utilité supplémentaire : entraîner d'autres modèles pour de futurs projets.
Dans le contexte d'une science ouverte et de données réutilisables, l'initiative HTR United propose de récolter et documenter des sets de transcriptions.

CAMPUS CONDORCET | HUMATHÈQUE

La transcription au cœur d'un processus technique



CAMPUS CONDORCET | HUMATHÈQUE

La transcription comme étape du travail de recherche

- Dans quel contexte souhaitez-vous transcrire ?

Un projet de recherche ? Un projet de valorisation patrimoniale ?

- Que souhaitez-vous proposer à partir de la transcription ?

Simple accès au texte ? Edition numérique enrichie ? Textométrie ?

- Quelle est la nature du corpus ?

Des documents imprimés ? Manuscrits ? Combien de pages ? Combien de mains ? Quelle est la qualité des numérisations ? Quelle complexité ?

- De quels moyens bénéficiez-vous ?

Financements, temps humain, équipe et expertise ?

CAMPUS CONDORCET | HUMATHÈQUE

Pour tout renseignement :

formations.humatheque@campus-condorcet.fr

Les prochaines permanences RePOD :

21 octobre 2025

18 novembre 2025

16 décembre 2025

20 janvier 2026

Le goût de l'archive passe par ce geste artisan, lent et peu rentable, où l'on recopie les textes, morceaux après morceaux, sans en transformer ni la forme, ni l'orthographe, ni la ponctuation. Sans trop même y penser. En y pensant continûment. Comme si la main, ce faisant, permettait à l'esprit d'être simultanément complice et étranger au temps et à ces femmes et à ces hommes en train de se dire.

[...]

L'archive recopiée à la main est un morceau de temps apprivoisé ; plus tard on découpera les thèmes, on formulera des interprétations. Cela prend beaucoup de temps et parfois fait mal à l'épaule en tiraillant le cou ; mais avec lui, du sens se découvre. »

A. Farge, *Le goût de l'archive*, Paris, Seuil, 1989, p. 24-25

Webographie

http://theleme.enc.sorbonne.fr/cours/edition_epoque_moderne/edition_des_textes

http://www.item.ens.fr/articles-en-ligne/structuration-des-manuscrits-du-corpus-a-la-region/

https://gout-numerique.net

https://harmoniseatr.hypotheses.org/

https://distam.hypotheses.org/

https://www.irht.cnrs.fr/fr/formation/les-stages

https://wiki-arhn.larhra.fr/lib/exe/fetch.php?media=seminaire:20240208_ariane_pinche.pdf

https://cahier.hypotheses.org/guides/guide-correspondance

https://cahier.hypotheses.org/guides/publication-editions-textes

https://www.unige.ch/lettres/humanites-numeriques/recherche/projets/projets-de-la-chaire/fondue

Bibliographie

Alix Chagué, Floriane Chiffoleau, Matthias Gille Levenson, Hugo Scheithauer, Ariane Pinche. *Chaînes d'acquisition, de traitement et de publication du texte*. Consortium Ariane - Axe 1. 2024. ⟨hal-04734959⟩

Alix Chagué, Thibault Clérice, Ariane Pinche, Benjamin Kiessling, Peter Stokes, et al.. *Apprendre à Lire aux Machines*. 2025. ⟨hal-05163931⟩

Nougaret, Christine et al. *L'édition critique des textes contemporains, XIXe-XXIe siècle*. Paris: École nationale des Chartes, 2015. Print.

Crasson Aurèle, et Crasson Aurèle. *L'édition du manuscrit : de l'archive de création au scriptorium électronique* : [séminaire général de l'Institut des textes et manuscrits modernes, ITEM, organisé en 2003-2004]. Louvain-la-Neuve: Academia-Bruylant, 2008. Print.

Lucas, Noëmie. *OCR/HTR et graphie arabe : Les manuscrits arabes à l'heure de la reconnaissance automatique des écritures*. N.p., 2022. Print.

Ariane Pinche, « Des plumes aux pixels : actualité de l'édition scientifique numérique », Théia [En ligne], 1 | 2024, mis en ligne le 26 novembre 2024, consulté le 07 janvier 2025. URL : https://publications-prairial.fr/theia/index.php?id=268

Caers, Bram. "Teaching Handwritten Text Recognition: Can New Technologies Save Old skills?". Quaerendo 54.2-3 (2024): 198-209. https://doi.org/10.1163/15700690-bja10024 Web.