

Du fac-similé à la donnée structurée

Pipeline de traitement d'un dictionnaire ancien

15 janvier 2026
Séminaire VLC, STIH

Anaïs Chambat

ATER de linguistique et d'informatique pour les SHS
Sorbonne Université, STIH (UR 4509)

Doctorante en Sciences du langage
CY Cergy Paris Université, LT2D (UR 7518)

CONTEXTE ET OBJECTIFS DE LA RECHERCHE

Le dictionnaire comme espace de construction du savoir

- Réceptacle d'une **représentation socioculturelle et lexiculturelle de la langue**, il est le reflet des réalités qu'elle désigne et de leur mise en mots (Galisson, 1988 et 1999)
- Quel que soit son programme linguistique, **sa temporalité n'est pas celle de la science**, mais bien « l'ensemble des connaissances antécédantes » (Auroux, 2006), celles d'aujourd'hui et peut-être de demain.

Question centrale

→ Comment appréhender cet « horizon de rétrospection » lexicographique ? (Auroux, 1986)

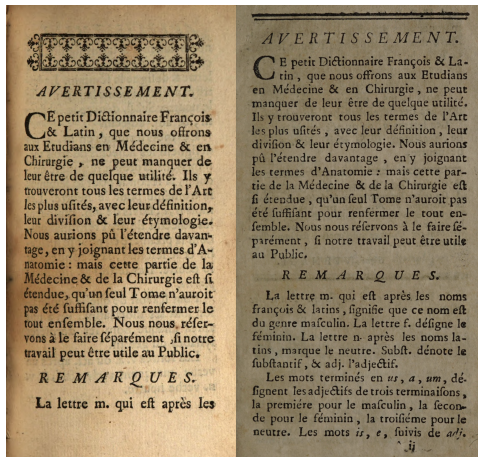
Premiers dictionnaires de médecine rédigés en français

- **1736** : 340 articles, 285 pages (échantillon-test)
- **1741-1760** : 5 éditions - 1450 articles, 490 pages
- **1764-1793** : 4 éditions - 4950 articles, 855 pages

Quelques mots sur leurs auteurs

- Col De Vilars, E. Professeur d'Anatomie et de Chirurgie, Doyen de la Faculté de médecine de Paris (1740-1743)
- Lavoisien, J.-F. Chef des hôpitaux des armées du Roi
- Praticiens avant d'être lexicographes

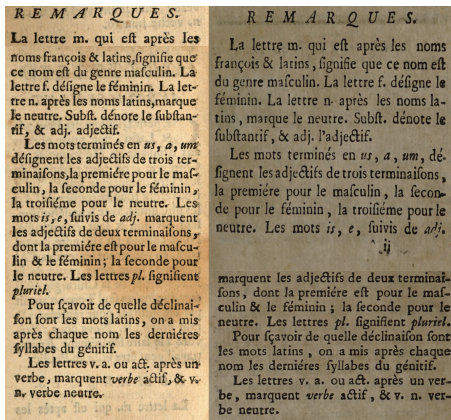
La dictionnairique en action



- Mise en page différente :
18 lignes à droite /
16 à gauche
- Aucune différence... Les deux contenus sont strictement identiques !

Extrait de la préface des éditions de 1741 et de 1753.

La dictionnaire en action



- Mise en page différente :
26 lignes à droite /
22 à gauche
- Un seul mot d'écart :
l'adjectif → adjectif

Extrait de la préface des éditions de 1741 et de 1753.

Enjeux des dictionnaires réédités

Les dictionnaires réédités conservent une structure relativement stable, mais les modifications peuvent être de différentes natures :

- **Discrètes** : reformulations, actualisation d'une référence, ajouts de synonymes ou de renvois, etc.
- **Massives** : évolutions sémantiques (changement de définition, d'exemple, d'auteur cité), glissements épistémologiques (passage d'un cadre théorique à un autre), modifications structurelles (ajout ou réécriture complète d'un article, déplacement de définitions), etc.

Enjeux méthodologiques

- Retracer l'historicité des révisions successives
- Analyser les transformations à différentes échelles
- Quantifier les révisions sémantiques, structurelles et formelles dans une perspective diachronique
- Restituer la dynamique de renouvellement du lexique
- Construire un protocole reproductible, fondé sur des outils *open source*, qui pourrait être appliqué à d'autres corpus lexicographiques anciens

COLLECTE ET ÉVALUATION DES SOURCES PRIMAIRES

→ Un OCR à 99% équivaut à 1 erreur tous les 100 mots.

Adopter une approche computationnelle du corpus

- **Collecter les différentes versions** disponibles
- **Évaluer la netteté** et la lisibilité des images
- **Croiser plusieurs métriques** afin d'évaluer la qualité des images de façon unitaire et comparative.
- Production d'un **score agrégé de qualité** pour chaque réservoir d'images considéré.

Sources et formats d'images disponibles

■ Standards ouverts (IIIF)

- **Gallica** : Bibliothèque nationale de France
- **Wellcome Library** : Ressources médicales, historiques
- **Numerabilis** : Bibliothèques numériques locales
- **Internet Archive** : Archives numériques

■ Plateformes commerciales

- **Google Books** :
 - **Formats** : EPUB/PDF (extraction d'images possible)
 - **Limites** : Copyright, OCR imparfait

■ Webscraping

- **Collecte** ciblée (ex : HathiTrust)
- **Risques** : Légalité, blocages techniques

Agrégation des mesures

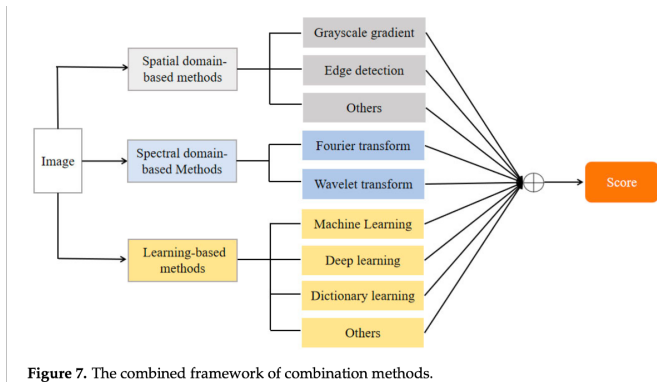


Figure 7. The combined framework of combination methods.

Zhu, Yu, Wang, Ke, Zhi (2023, 9).

Exemples de résultats obtenus

RÉPERTOIRE: CV1736_wellcome-collection_b3300884x_0001

CARACTÉRISTIQUES TECHNIQUES:

- └─ Taille moyenne (px) ■■■■■■■■■■ 0.2 (Très faible)
- └─ Taille fichier (ko) ■■■■■■■■■■ 0.6 (Moyenne)
- └─ Mémoire utilisée (ko) ■■■■■■■■■■ 0.6 (Moyenne)
- └─ Score technique agrégé: 0.467

ANALYSES UNITAIRES CLASSIQUES:

- └─ Netteté basée sur les gradients | Score: 0.601
- └─ Détection des bords et des contours | Score: 0.367
- └─ Analyses spectrales | Score: 0.383
- └─ Richesse informationnelle | Score: 0.631

SCORE AGRÉGÉ DE QUALITÉ D'IMAGE:

- └─ ★★ 0.495

RÉPERTOIRE: CV1736_wellcome-collection_b33027006_0003

CARACTÉRISTIQUES TECHNIQUES:

- └─ Taille moyenne (px) ■■■■■■■■■■ 0.2 (Très faible)
- └─ Taille fichier (ko) ■■■■■■■■■■ 0.6 (Moyenne)
- └─ Mémoire utilisée (ko) ■■■■■■■■■■ 0.6 (Moyenne)
- └─ Score technique agrégé: 0.467

ANALYSES UNITAIRES CLASSIQUES:

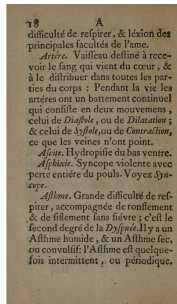
- └─ Netteté basée sur les gradients | Score: 0.430
- └─ Détection des bords et des contours | Score: 0.403
- └─ Analyses spectrales | Score: 0.376
- └─ Richesse informationnelle | Score: 0.581

SCORE AGRÉGÉ DE QUALITÉ D'IMAGE:

- └─ ★★ 0.448

La netteté d'une image dépend

- Format d'enregistrement utilisé (JPG, PNG, etc.).
- Modalités de mises à disposition (serveur IIIF, archive, etc.).



ABBY FineReader Pro 15 (propriétaire)	Transkribus (freemium)	eScriptorium (open-source)
Artère, Vaisseau destiné à recevoir le sang qui vient du cœur, & à le distribuer dans toutes les parties du corps : Pendant la vie les artères ont un battement continu qui consiste en deux mouvements, celui de Diafole, ou de Dilatation; & celui de Sy/lole, ou de Contraction, ce que les veines n'ont point. Afcite. Hydropisie du bas ventre. Aphixie. Syncope violente avec perte entière du pouls. Voyez Syncope.	Artere, Vaisseau destiné à recevoir le sang qui vient du coeur, & à le distribuer dans toutes les parties du corps : Pendant la vie les artères ont un battement continu qui consiste en deux mouvements celui de Diastole, ou de Dilatation; & celui de Systole, ou de Contraction, ce que les veines n'ont point. Ascite. Hydropisie du bas ventre. Asphixie. Syncope violente avec perte entière du pouls. Voyez Syncope.	Artère, Vaisseau destiné à recevoir le sang qui vient du cœur, & à le distribuer dans toutes les parties du corps : Pendant la vie les artères ont un battement continu qui consiste en deux mouvements, celui de Diastole, ou de Dilatation; & celui de Systole, ou de Contraction, ce que les veines n'ont point. Ascite. Hydropisie du bas ventre. Asphixie. Syncope violente avec perte entière du pouls. Voyez Syncope.

Exemple de la page 18 de l'ouvrage de 1736.

Quelle place pour les agents conversationnels ?

A

difficulté de respirer, & lésion des principales facultés de l'âme.

Artère. Vaisseau destiné à recevoir le sang qui vient du cœur, & à le distribuer dans toutes les parties du corps : Pendant la vie les artères ont un battement continu qui consiste en deux mouvemens, celui de *Diastole*, ou de *Dilatation* ; & celui de *Systole*, ou de *Contraction*, ce que les veines n'ont point.

Ascite. Hydropisie du bas ventre.

Asphixie. Syncope violente avec perte entière du pouls. Voyez *Syncope*.

Asthme. Grande difficulté de respirer, accompagnée de ronflement & de sifflement sans fièvre ; c'est le second degré de la *Dyspnée*. Il y a un Asthme humide, & un Asthme sec, ou convulsif : l'Asthme est quelquefois intermittent, ou périodique.

■ Un résultat bluffant :

- Ajout pertinent de l'italique
- Segmentation correcte des entrées
- Prise en charge des « s » longs

■ OUI, MAIS

- Aucune transparence sur les choix de transcription opérés (modernisation ?)
- Modèle computationnel coûteux
- Résultats variables suivant les modèles utilisés et les prompts

INFORMATISATION DU CORPUS

■ Partie 1 : Dictionnaire entier

0. Sortie OCR brute

- EScriptorium (Kraken + CATMuS PRINT)

1. Nettoyage automatique des césures

2. Nettoyage automatique du bruit structurel

3. Étape manuelle

- Séparation des lettres de la nomenclature

4. Segmentation automatique des articles

- Découpage automatique des entrées en autant de fichiers TXT
- Corrections OCR de la nomenclature

■ Partie 2 : À l'échelle des articles

5. Correction automatique des abréviations

6. Détection automatique des caractères grecs

7. Relecture manuelle

- Application Flask avec l'image en vis-à-vis
- Constitution de la vérité de terrain
- 1400 articles (~1 min/article)

8. Alignement automatique des articles

- Identification des différences inter-éditions
- Évaluation de la similarité (sémantique, textuelle et lexicale)
- Classification des modifications textuelles

Différents états du corpus

Texte issu de l'OCR	Texte post-traité et partitionné
Abdomen. La partie antérieure du bas ventre. Abscès. Tumeur qui renferme du pus, ordinairement accompagnée de douleur, & quelquefois de fièvre. Les abscess sont formés par un amas de sang, ou d'humeurs qui étant hors de la voie de la circulation dégénèrent en pus: il y en a d'internes, & d'externes.	ABDOMEN.txt Abdomen. La partie antérieure du bas ventre. ABSCES.txt Abscès. Tumeur qui renferme du pus, ordinairement accompagnée de douleur, & quelquefois de fièvre. Les abscess sont formés par un amas de sang, ou d'humeurs qui étant hors de la voie de la circulation dégénèrent en pus: il y en a d'internes, & d'externes.

Exemple de la page 18 de l'ouvrage de 1736.

Arborescence des fichiers générés

- Un dossier contenant le texte brut des articles.
- Un dossier de métadonnées contenant la liste des articles, la liste des renvois, des synonymies et un tableau de synthèse.

```

sortie/
|— articles/
|   |— ABDOMEN.txt
|   |— ABSCES.txt
|   |— ...
|— meta/
|   |— liste_articles.txt
|   |— renvois.txt
|   |— synonymies.txt
|   |— articles.csv

```

Index	Entrée	Synonymies	Renvois	Extrait Renvoi
ABDOMEN	Abdomen	Non	Non	
ABSCES	Abscès	Non	Non	
AIGUE	Aiguë	Non	Non	
ANASARQUE	Anasarque	Non	Oui	Leucophlégmatie
ANEVRISME	Anévrisme	Non	Non	
ANGINE	Angine	Non	Oui	Squinancie
ANTHRAX	Anthrax	Non	Oui	Charbon

Application de post-correction

ACOUSTIQUES. adj. pl. & subst. *Acoustica*, n. pl. Remèdes propres aux maladies de l'ouïe. Ce mot est Grec [ακουστικῶν], qui appartient à l'ouïe, du verbe [ἔξω], audio, j'entends; par cette raison les Anatomistes donnent aussi le nom d'*acoustique* au nerf auditif.

ACRE. adj. *Acer*, *acris*, *acre*, adj. ce qui est piquant & mordicant, ou composé de particules salines, rudes, inégales, anguleuses, pointues, & qui, à raison de sa surface âpre & raboteuse, fait une impression désagréable sur la langue, & même l'écorche, suivant son degré d'acrimonie. Tels sont les fruits sauvages, le Poivre, le Tythimale, la Moutarde, les Cantharides, l'Arsenic & tous les corrosifs. C'est aussi à cause de cette tiffure que les Chymistes donnent le nom de sel âcre au sel alkali; parce qu'ils se le représentent sous la figure de petites molécules hérissées de pointes: mais l'acrimonie est commune à toute sorte de sels.

ACROCHORDON. f. m. Espèce de verrue, ainsi appelée en grec ἀκροχορδόν, de ἀκρον, sommet, extrémité, & de χορδή, corde, parce qu'elle ressemble par son extrémité à une corde coupée, ou parce qu'étant attachée à la peau par un pédicule fort mince, elle pend comme une corde, ce qui fait

ACOUSTIQUES.txt 10/128

ACOUSTIQUES. adj. pl. & subst. *Acoustica*, n. pl. Remèdes propres aux maladies de l'ouïe. Ce mot est Grec [ακουστικῶν], qui appartient à l'ouïe, du verbe [ἔξω], audio, j'entends; par cette raison les Anatomistes donnent aussi le nom d'*acoustique* au nerf auditif.

Save Previous Next All files

ACOUSTIQUES.txt

RÉSULTATS

Quantifier les modifications textuelles

filename	classification	layout	difflib	cosine	jaccard	sbert
0-FRONTISPICE.txt	 MAJOR CONTENT REDUCTION	False	0.6416	0.8188	0.6732	0.8901
1-AVERTISSEMENT.txt	 MULTIPLE SEMANTIC ADJUSTMENTS	False	0.9726	0.9881	0.9764	0.9947
2-REMARQUES.txt	 ORTHOGRAPHIC-FOCUSED NORMALIZATION	False	0.9944	0.9956	0.9913	0.9978
ABSCES.txt	 ORTHOGRAPHIC-FOCUSED NORMALIZATION	True	0.9697	0.9870	0.9744	0.9884
ABSORBANS.txt	 MODERATE CONTENT EXPANSION	False	0.8806	0.9402	0.8844	0.9508
ACCES.txt	 MINOR GRAMMATICAL NORMALIZATION	True	0.9888	0.9889	0.9779	0.9922
ACCIDENT.txt	 COMPREHENSIVE SURFACE NORMALIZATION	False	0.9861	0.9861	0.9725	0.9838
AGLUTINATIF.txt	 COMPLETE IDENTITY	True	1.0000	1.0000	1.0000	1.0000
ANATOMIE.txt	 MODERATE TYPOGRAPHICAL REVISION	False	0.9239	0.9877	0.9755	0.9941
ANGIOLOGIE.txt	 CONTENT IDENTITY	False	1.0000	1.0000	1.0000	1.0000

Catégoriser les modifications textuelles

ADDITIONS (total: 1)

ABSORBANS.txt: 'Ils ne se dissolvent point comme eux dans les liqueurs aqueuses.'

DELETIONS (total: 1)

0-FRONTISPICE.txt: 'HERISSANT, ruë S. Jacques, à S. Paul & à S. Hilaire.'

REPLACEMENTS (total: 2)

0-FRONTISPICE.txt: 'COIGNARD, ruë S. Jacques, à la Bible d'or, LE MERCIER & BOUDET, ruë S. Jacques, au Livre d'or. ROLLIN, Quai des Augustins, à S. Athanase. DELESPINE , Imp. Lib. ord.' → 'Chez DELESPINE, Imprimeur-Libraire ordinaire'
0-FRONTISPICE.txt: 'XLI.' → 'LIII.'

Catégoriser les modifications textuelles

CAPITALIZATION (total: 2)

ABSORBANS.txt: 'yeux' → 'Yeux'

ABSORBANS.txt: 'os' → 'Os'

ACCENTS (total: 11)

0-FRONTISPICE.txt: 'DEFINITION,' → 'DÉFINITION,'

0-FRONTISPICE.txt: 'ruë' → 'rue'

0-FRONTISPICE.txt: 'Privilege' → 'Privilège'

1-AVERTISSEMENT.txt: 'étenduë,' → 'étendue,'

2-REMARQUES.txt: 'premiere' → 'première'

2-REMARQUES.txt: 'dernieres' → 'dernières'

ABSCES.txt: 'd'Apostème' → 'd'Apostême,'

ABSORBANS.txt: 'Medicamens' → 'Médicamens'

ABSORBANS.txt: 'veritables' → 'véritables'

ACCIDENT.txt: 'aussi-tôt' → 'aussi-tot'

PUNCTUATION (total: 8)

2-REMARQUES.txt: 'n,' → 'n.'

2-REMARQUES.txt: 'actif' → 'actif,'

ABSCES.txt: 'ûs,' → 'ûs.'

ABSORBANS.txt: 'Absorbentia.' → 'Absorbentia,'

ACCIDENT.txt: 'délire,' → 'délire'

ANATOMIE.txt: 'apprend' → 'apprend,'

ANATOMIE.txt: 'dissection' → 'dissection,'

ANATOMIE.txt: 'grec' → 'grec,'

Catégoriser les modifications textuelles

SPELLING_VARIANTS (total: 5)

0-FRONTISPICE.txt: 'Roy' → 'Roi'
 0-FRONTISPICE.txt: 'Roy.' → 'Roi.'
 ABSCES.txt: 'contre nature' → 'contre-nature'
 ABSORBANS.txt: 'calcinez' → 'calcinés'
 ACCIDENT.txt: 'l'érérysipéle,' → 'l'érésipéle,'

ABBREVIATIONS (total: 9)

ABSCES.txt: 's.'
 ABSORBANS.txt: 'plur.' → 'pl.'
 ABSORBANS.txt: 's.'
 ABSORBANS.txt: 'pl.'
 ACCES.txt: 's.'
 ACCES.txt: 's.'
 ACCIDENT.txt: 's.'
 ACCIDENT.txt: 's.'
 ANATOMIE.txt: 's.' → 'subst.'

SEMANTIC_CHANGES (total: 2)

1-AVERTISSEMENT.txt: 'avantageux' → 'utile'
 1-AVERTISSEMENT.txt: 'bien public.' → 'Public.'

Proposition de classification

Niveau	Description
	Identité (aucun ou très peu de changements)
	Normalisation de surface (grammaire, typographie, orthographe - sans impact sémantique)
	Ajustements mineurs de contenu (faible impact)
	Révisions modérées de contenu (impact moyen)
	Réécritures majeures (impact élevé)
	Relations complexes ou non classifiées

Note : Cette échelle permet de quantifier et catégoriser les modifications entre versions successives du texte, depuis les corrections superficielles jusqu'aux transformations profondes.

- **Col De Vilars, E.** (1736). *Recueil alphabétique de Prognostics dangereux & mortels sur les différentes maladies de l'Homme. Précédé d'une explication des maladies, & de quelques termes de Médecine*, Paris, 290 p.
- **Col De Vilars, E.** (1741, 1747, 1753, 1759, 1760). *Dictionnaire françois-latin des termes de médecine et de chirurgie, avec leur définition, leur division et leur étymologie*, Paris, 490 p.
- **Lavoisien, J.-F.** (1764, 1771, 1781, 1793). *Dictionnaire portatif de médecine, d'anatomie, de chirurgie, de pharmacie, de chymie, d'histoire naturelle, de botanique et de physique*, Paris, 716 p.

Bibliographie indicative

- **Auroux, S.** (1986). « Histoire des sciences et entropie des systèmes scientifiques : les horizons de rétrospection ». *Archives et documents de la SHESL, première série*, 7, 1-26.
- **Auroux, S.** (2006). « Les modes d'historicisation ». *Histoire, Épistémologie, Langage*, 28, 105-116.
- **Galisson, R.** (1988). « Cultures et lexicultures. Pour une approche dictionnaire de la culture partagée ». *Annexes des Cahiers de linguistique hispanique médiévale*, 7, 325-341.
- **Galisson, R.** (1999). « La pragmatique lexiculturelle pour accéder autrement à une autre culture, par un autre lexique ». *ÉLA. Études de linguistique appliquée*, 116, 477-496.
- **Zhu, M., Yu, L., Wang, Z., Ke, Z., & Zhi, C.** (2023). « Review : A Survey on Objective Evaluation of Image Sharpness ». *Applied Sciences*, 13(4), 2652. <https://doi.org/10.3390/app13042652>

Anaïs Chambat

- `anais.chambat@sorbonne-universite.fr`
- `https://achambat.github.io/`

Merci de votre attention !