**MajecSTIC 2009**
**Avignon, France, du 16 au 18 novembre 2009**

# Extraction automatisée de lignes et de fragments textuels dans les images de manuscrits d'auteur du 19ème siècle

Vincent MALLERON[1,2], Véronique EGLIN[1], Hubert EMPTOZ[1], Stéphanie DORD-CROUSLE[2], Philippe REGNIER[2]

1 : Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, F-69621, France
2 : Université de Lyon, CNRS, LIRE, UMR 5611 F-69007, France

Contact : vincent.malleron@liris.cnrs.fr

---

**Résumé**
Dans cet article nous proposons une nouvelle approche pour l'enrichissement des éditions électroniques de corpus littéraires grâce à l'estimation de la structure des documents manuscrits. Dans tout processus d'analyse de document manuscrit l'analyse de la structure est une étape importante : en effet, disposer de la position des lignes de texte, des paragraphes et des fragments permet d'envisager de nouveaux moyens d'exploiter les corpus littéraires. L'extraction de structure d'un document manuscrit est rendue difficile par les variations d'orientation de la ligne de base et des espaces interligne mais également par les chevauchements entre lignes et les occlusions. On propose un algorithme d'extraction des lignes de texte et des fragments textuels basé sur une analyse en composantes connexes. Une fois l'extraction des composantes connexes réalisée on construit un graphe d'adjacences pondéré et orienté : chaque composante connexe correspond à un nœud. Chaque nœud est l'origine de 4 arcs le reliant à ses plus proches voisins dans 4 directions : gauche, droite, haut et bas. En parallèle et via une approche similaire chaque composante connexe est identifiée comme appartenant à une des classes suivantes : haut de page, bas de page, gauche, droite ou intérieur du texte. Cette identification permet d'initialiser l'algorithme d'extraction des lignes qui utilise une recherche du plus court chemin entre connexité gauche et droite sur le graphe orienté décrit précédemment. L'extraction des fragments est ensuite réalisée via une fusion des lignes extraites par rapport à des critères de distance interligne et de variation d'orientation de la ligne de base.

**Mots-clés :** manuscrit, structure, lignes, fragments, graphe

## 1. Introduction

### 1.1. Le projet BOUVARD

Notre travail se positionne au sein d'un projet de sciences humaines : le projet BOUVARD. Ce projet, soutenu par l'ANR a pour objectif d'éditer en ligne, sous une forme technologique innovante, un ensemble patrimonial cohérent, d'importance scientifique et culturelle reconnue : les Dossiers de *Bouvard et Pécuchet*, le dernier roman - posthume et inachevé - de Gustave Flaubert, soit 2 400 feuillets (pages manuscrites et coupures de presse), conservés à la Bibliothèque municipale de Rouen. Ce chantier documentaire a servi à rédiger le premier volume de l'œuvre et aurait dû être réutilisé pour la composition du second volume, jamais écrit en raison de la mort soudaine du romancier. Outre cette particularité rédactionnelle, les Dossiers sont porteurs d'une dimension épistémologique singulière : constitués pour rédiger une « encyclopédie critique en farce », ils proposent une configuration critique des savoirs au XIXe siècle, originale et révélatrice. Or, en raison de leur important volume, de leur organisation complexe et indéfiniment mouvante, ainsi que de leurs contenus scientifiques extrêmement variés, les Dossiers ne peuvent pas être édités de manière satisfaisante sous une forme imprimée. C'est particulièrement vrai pour les pages préparées en vue du « second volume » du roman : les annotations que l'écrivain y a portées, indiquant le lieu probable du classement, sont souvent plurielles et obligent à conserver une structure modulable aux pages : la fixité d'une édition imprimée n'est donc pas adéquate pour l'édition des Dossiers de *Bouvard et Pécuchet*. La dimension pluridisciplinaire du projet qui allie

sciences humaines et sciences et technologies de l'information et de la communication le place dans une posture novatrice et particulièrement adaptée pour la réussite de l'édition électronique des dossiers de *Bouvard et Pécuchet*.

### 1.2. Objectifs

Une image de document est souvent difficilement exploitable à l'état brut, que ce soit pour le chercheur en sciences humaines qui n'y trouve qu'un substitut au manuscrit réel, ou pour l'informaticien qui ne dispose d'aucune information structurelle sur l'image pour mettre en place une chaîne de traitements. L'extraction de la structure physique des documents permet d'envisager de disposer d'informations complémentaires sur les images. On présentera dans cet article notre méthode pour l'extraction de la structure de ces documents ainsi qu'un certain nombre de pistes pour offrir de nouvelles possibilités de navigation à l'intérieur du Corpus des dossiers de *Bouvard et Pécuchet*.

### 1.3. Un corpus particulier mais une approche généralisable

Le choix du corpus des dossiers de *Bouvard et Pécuchet* pour réaliser ces travaux s'explique par les deux raisons suivantes : il s'agit d'une part d'un corpus riche, présentant une diversité intéressante dans les mises en pages ainsi que dans les styles d'écriture, et, d'autre part d'un corpus reconnu permettant de bénéficier de l'expertise de chercheurs en sciences humaines. Cette collaboration pluridisciplinaire sur un corpus particulier permet de mettre en lumière des besoins et des solutions adaptables pour la grande majorité des corpus. Les techniques mises au point ici sont en effet basées sur des observations et des propositions conjointes entres spécialistes SHS et STIC. De par la diversité du corpus et des méthodes mises au point, la transposition du travail réalisé pour les dossiers de *Bouvard et Pécuchet* ne nécessitera que des ajustements mineurs afin d'être utilisable pour un nombre important de corpus en sciences humaines, mais également pour tout corpus présentant une structure fragmentaire (corpus de notes, carnets de fouilles...), ainsi que pour divers documents (lettres, fax).

## 2. Etat de l'art

L'extraction de la structure des documents manuscrits reste un domaine de recherche ouvert : en effet la majorité des travaux réalisés jusqu'à présent se limitent à l'extraction des lignes de texte, et ce souvent dans des conditions typographiques particulières. Ces travaux peuvent généralement être catégorisés entre les approches *top-down* et *bottum-up*. Pour les approches *top-down* une page de document est d'abord segmentée en zones, les zones sont ensuite séparées en lignes et ainsi de suite. Les méthodes utilisant les projections de profil sont parmi les méthodes *top-down* les plus populaires pour l'extraction de structure dans les documents imprimés mais ne peuvent être appliquées pour les documents manuscrits que lorsque ceux ci sont très réguliers et lorsque l'espace interligne est suffisamment grand pour apparaitre clairement lors de la projection [2]. Ces méthodes ne sont donc pas adaptées pour la majorité des documents manuscrits. Les méthodes basées sur l'extraction et le regroupement des composantes connexes dans un document constituent une bonne partie des méthodes *bottom-up*. On part d'un élément structurel bas : la composante connexe, qui peut représenter dans le cas d'un document manuscrit un caractère, un ensemble de caractères ou un mot et on regroupe ces éléments selon un certain nombre de règles de façon à constituer les lignes de texte. [3] et [7] proposent des algorithmes basés sur l'extraction des composantes connexes couplées à la transformée de Hough. La transformée de Hough permet de disposer d'une estimation de l'orientation locale, et ainsi d'extraire des lignes curvilinéaires. Une revue détaillée des méthodes d'extraction de lignes dans les documents manuscrits est proposée par Likforman-Sulem et al. dans [6]. Dans [5], Yi Li et al. proposent une méthode d'extraction des lignes de texte dans les documents manuscrits basée sur la méthode du *level-set*. Cet algorithme n'utilise aucune connaissance *a priori* sur le document, et réalise une extraction précise des lignes de texte.

Si l'on étudie maintenant les travaux réalisés pour l'extraction de la structure des documents, la plupart de ceux ci sont réalisés sur des textes imprimés. Dans [11] T.M.Breuel et al. proposent un ensemble d'algorithmes pour l'extraction de la structure dans les documents imprimés : ces algorithmes sont résistants au bruit et largement adaptables à différentes mise en pages et à différents

langages, mais nécessitent la structure forte de l'imprimé : ces méthodes sont donc difficilement transposables pour l'extraction de structure pour les documents manuscrits. Kise et al. [4] présentent un algorithme pour l'extraction de la structure des pages de documents imprimés basés sur les Diagrammes de Voronoï. Un certain nombre d'applications des diagrammes de Voronoï pour les documents manuscrits ont été proposées, notamment par Lemaitre et al. dans [1] mais ces algorithmes requièrent toujours une structure forte du document ce qui n'est généralement pas le cas dans le cadre de notre étude.
L'extension des algorithmes dédiés aux documents imprimés pour les documents manuscrits est loin d'être triviale et ce spécialement quand la structure présente une grande variabilité. Des méthodes consacrées aux documents manuscrits ont donc été développées et montrent des résultats intéressants : dans [10], L.O'Gorman propose le *Dosctrum* qui permet de modéliser et d'extraire la structure des documents manuscrits ayant une mise en page régulière. Nicolas et al. dans [9] et [8] présentent un algorithme basé sur les modèles de Markov cachés et permettant de segmenter une page de texte manuscrit en différentes zones : l'arrière plan, les zones de texte, les ratures, les espaces inter-lignes et inter-mots sont ainsi délimités.

## 3. Notre Approche

### 3.1. Notions de structure d'un document

Avant de décrire notre approche de façon plus précise il est important de bien définir la notion de structure d'un document manuscrit, et plus particulièrement la notion de structure physique. En effet, on désigne par le nom de structure un ensemble de descripteurs d'un document. Parmi eux on peut citer la structure syntaxique ou la structure physique. La structure syntaxique décrit de façon précise la syntaxe du document : mots, phrases, syllabes, etc. La structure physique divise un document en éléments structurels : page, fragments, paragraphes, lignes, mots, caractères. La structure physique ne repose en aucun cas sur le sens : elle peut donc être extraite de façon indépendante au contenu du texte. On présentera dans les sections suivantes notre algorithme d'extraction de la structure physique d'un document. Ces algorithmes pourront être adaptés par la suite afin d'extraire la structure syntaxique d'un document, en se basant cette fois sur des indicateurs de contenu. Dans la suite de cet article on détaillera particulièrement l'extraction des lignes de texte et des fragments structurels. L'extraction des mots et des caractères est rendue difficile par la résolution des images dont nous disposons, qui sont issues d'une campagne de numérisation des microfilms. La résolution des images est de 200dpi, ce qui s'avère néanmoins suffisant pour l'extraction d'élements de structure de haut niveau (fragments ou lignes).

### 3.2. En Préliminaire,classification de la mise en page

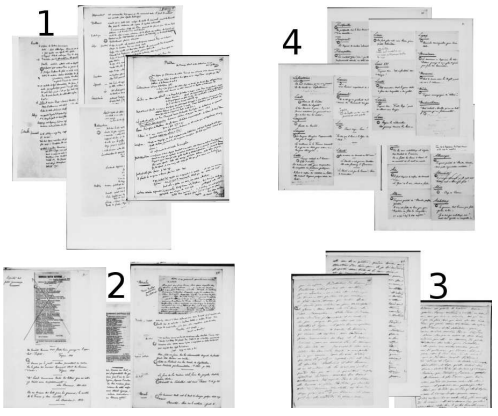


FIG. 1 – Différents types de mise en page

La figure 1 présente un assortiment des différents types d'images du corpus des dossiers de *Bouvard et Pécuchet*. Ces pages peuvent être classées de façon grossière dans les catégories suivantes : 1-Notes Bibliographiques, 2-Assemblage de Notes, 3-Correspondances et recopie de nouvelles, 4-Dictionnaire des idées reçues, 5-Extraits de journaux, imprimés. A l'exception des documents imprimés les images de ces catégories sont représentées sur la figure 1 . L'appartenance à l'une ou l'autre des catégories est définie en utilisant les trois descripteurs suivants : Nombre de tampons de bibliothèque, Densité de données dans la page, Dispersion du profil. Les tampons de bibliothèque ont été aposés par le conservateur pour identifier les différents fragments physiques rassemblés sur une même page. La localisation des tampons est réalisée via une simple mise en correspondance avec une image référence du tampon. La récupération de cette information permet de séparer les catégories 1,3 et 5 qui ne possèdent qu'un seul tampon par page des catégories 2 et 4 qui en possèdent 2 ou plus. La densité de données est le pourcentage de pixels de texte dans la page après binarisation. Cette densité est plus élevée pour les pages imprimées que pour les pages manuscrites, ce qui permet de discriminer les deux types de pages. Enfin la dispersion du profil caractérise la régularité de la mise en page, I correspondant au profil de projection vertical et X à la largeur de la page :

$$d = \frac{\sum(Max(I) - I(y))}{X}$$

Comme on peut le constater sur la figure 2, la dispersion du profil nous permet de distinguer de manière simple les pages de mise en page plus régulière comme les pages de catégorie 3. Cette dernière caractéristique nous permet donc d'obtenir une estimation efficace de l'appartenance typologique de chacune des pages du corpus.

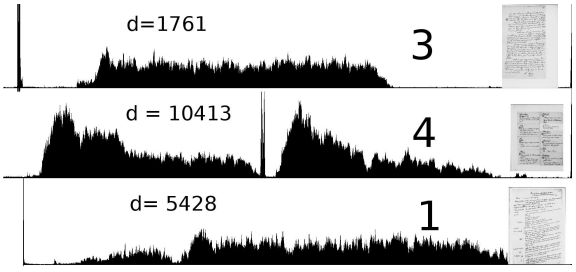


FIG. 2 – Dispersion du profil

### 3.3. Distance inter composantes connexes

Afin de s'affranchir des problèmes classiques engendrés par l'utilisation d'une distance euclidienne entre centre de gravité des composantes connexes on introduit une nouvelle mesure de distance inter composantes connexes. On considère deux formes manuscrites A et B qui peuvent représenter un mot, un fragment de mot ou encore un caractère. La distance entre A et B est donnée par la distance euclidienne la plus faible entre les contours de A et de B. Cette distance euclidienne est pondérée par le coefficient suivant :

$$\Delta\Theta = \alpha * (1 + |\frac{|\theta 1| - |\theta 2|}{|\theta 1| + |\theta 2|}|)$$

Ce coefficient se base sur le fait que l'orientation reste généralement constante à l'intérieur d'une même ligne. $\theta 1$ et $\theta 2$ sont obtenues via l'estimation d'une carte des orientations par la transformée de Hough à basse résolution. Cette distance est plus représentative des espaces inter-mots et inter-lignes qu'une simple distance euclidienne et permet d'obtenir une bonne représentation de la structure locale du voisinage d'une composante connexe. Un exemple de distance interconnexitée est représenté figure 3. La ligne de base obtenue grâce à la transformée de Hough est représentée en bleu, et la distance minimale entre contours en rouge. Dans le reste de cet arti-

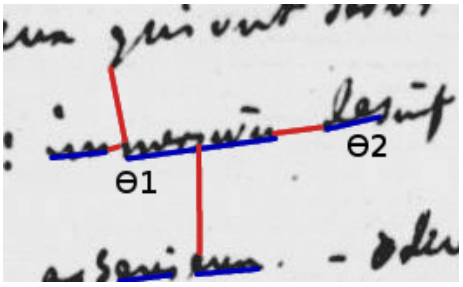


FIG. 3 – Distance inter composantes connexes

cle on utilisera cette distance dans différents intervalles d'orientation pour construire un graphe d'adjacences et extraire les lignes et les fragments.

### 3.4. Construction du graphe d'adjacence

Pour chaque composante connexe on recherche le plus proche voisin selon la distance proposée à la section 3 dans 4 directions de l'espace : haut, bas, gauche et droite. L'intervalle de recherche est donné par l'estimation de l'orientation décrite précédemment. La recherche est donc réalisée dans la direction de Hough, dans la direction orthogonale, la direction inverse et la direction inverse orthogonale. Une fois que l'on dispose des 4 voisins de chaque composante connexe on construit le digraphe pondéré $G = (V, A)$. $V = \{v_1, v_2, ..., v_n\}$, $A = \{e_1, e_2, ..., e_n\}$ ou $v_i$ représente une composante connexe et $e_i$ un arc. Le degré incident de G est de 4 : chaque nœud possède jusqu'à 4 arcs incidents ($e_i = (v_j, v_k)$) qui représentent le lien entre la composante connexe et ses voisins. Le poids des arcs est donné par la distance entre la composante connexe d'origine et la composante connexe pointée. Le graphe peut être reprojeté sur une page de manuscrit (figure 4). Les arcs orientés dans la direction de Hough directe sont colorés en cyan, les arcs en direction inverse en noir, les arcs orthogonaux en rouge ou en bleu selon leur poids et les arcs inverse orthogonaux sont colorés en vert. Quand deux arcs sont superposés seuls les arcs orthogonaux et directs sont représentés.

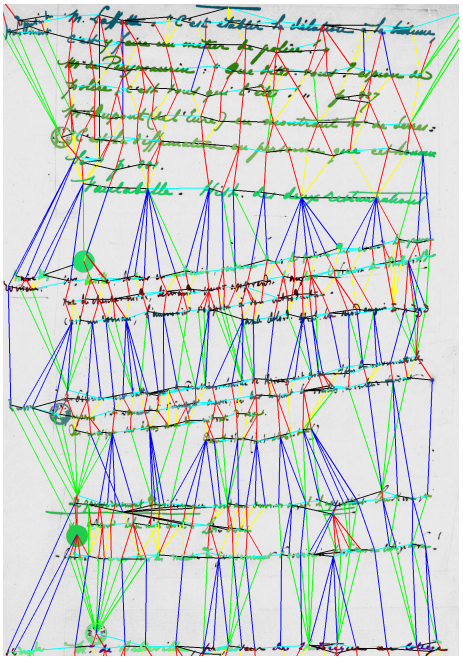

FIG. 4 – Reprojection du graphe, 1f188r

### 3.5. Extraction des bordures

Une fois le graphe d'adjacence construit, les bordures (composantes connexes situées en haut, en bas et sur les bords gauche et droit du document) peuvent aisément être extraites. En effet, de part la construction du graphe ces composantes sont nécessairement des composantes dont le degré incident est inférieur à 4. Une fois ces composantes identifiées on marque chacune des composantes connexes du graphe en les classant dans une des catégories. On peut observer ces bordures sur la figure 5. Les bordures gauches sont colorées en jaune, les droites en rouge, les bordures hautes en bleu et les basses en vert.

### 3.6. Extraction des lignes

La ligne de texte est un élément de structure essentiel : la connaissance de la position et de la forme de celles-ci permet d'envisager un grand nombre d'applications telles que la mise en relation entre l'image et la transcription. L'extraction des lignes utilise comme initialisation l'étape d'extraction des bordures : une ligne commence sur une composante étiquetée bord gauche et se termine sur une composante étiquetée bord droit. Le chemin le plus court dans notre graphe d'adjacence entre ces deux composantes permet d'extraire la ligne de texte. Une étape de traitement *a posteriori* est réalisée afin d'inclure les composantes connexes appartenant à la ligne mais non choisies dans le chemin.

### 3.7. Extraction des fragments

Le fragment est l'élément structurel le plus élevé de la page. Il peut s'agir d'un paragraphe de texte, d'un fragment collé ou encore d'un fragment de sens. Notre corpus est composé principalement de pages fragmentaires, ce qui justifie le développement d'algorithmes dédiés à l'extraction des fragments. L'extraction est réalisée en regroupant les lignes suivant des critères simples : la variation d'espace inter-lignes, d'orientation... Ces règles sont associées à la recherche du chemin le plus court entre les composantes étiquetées bordures gauches, droites, hautes ou basses. Le chemin ainsi calculé débute et s'achève sur le même noeud et décrit le contour du fragment. La fonction de coût utilisée pour le calcul du meilleur chemin est basée sur les valeurs de distances ainsi que les coups de transition entre composantes connexes hautes, gauches, basses et droites.

## 4. Résultats

Notre heuristique de classement a été testée sur une base de 280 images représentatives de la diversité du corpus. Le tableau 1 montre que les trois caractéristiques simples énoncées précédemment permettent un classement rapide de nos pages avec une fiabilité suffisante, ce qui nous permet d'adapter les traitements postérieurs au type de page.

| Classe | Nombre d'images | Rappel | Précision |
|---|---|---|---|
| Lettres&Nouvelles | 50 | 0.90 | 0.92 |
| Assemblages | 45 | 0.95 | 0.86 |
| Imprimés | 50 | 0.90 | 0.96 |
| Notes | 70 | 0.71 | 0.78 |
| Dictionnaire | 52 | 0.96 | 0.88 |
| Ensemble | 267 | 0.884 | 0.88 |

TAB. 1 – Rappel/Précision de notre classification

Notre algorithme pour l'extraction de lignes a également été testé sur un ensemble de pages représentatives de notre corpus. Aucune vérité terrain n'étant disponible, l'évaluation est basée sur un critère visuel : une ligne est considérée comme fausse lorsque moins de 80% de la ligne réelle est incluse dans la boite englobante de la ligne proposée. Le tableau 2 détaille les résultats de notre extraction de lignes sur différents types de pages.

La figure 6 présente les résultats de l'extraction de fragment sur une page du corpus. 4 des 5 fragments de cette page sont extraits correctement. Notre algorithme, basé sur des règles simples

| Page | Lignes Erronées | Lignes Correctes |
|---|---|---|
| 1 f 179 r | 1 | 15 |
| 1 f 007 v | 4 | 27 |
| 228 f 020 r | 2 | 21 |
| 4 f 234 r | 0 | 24 |
| 1 f 188 r | 1 | 24 |
| Pages Simples | 14 | 198 |
| Pages Complexes | 39 | 216 |

TAB. 2 – Résultats de l'extraction de lignes

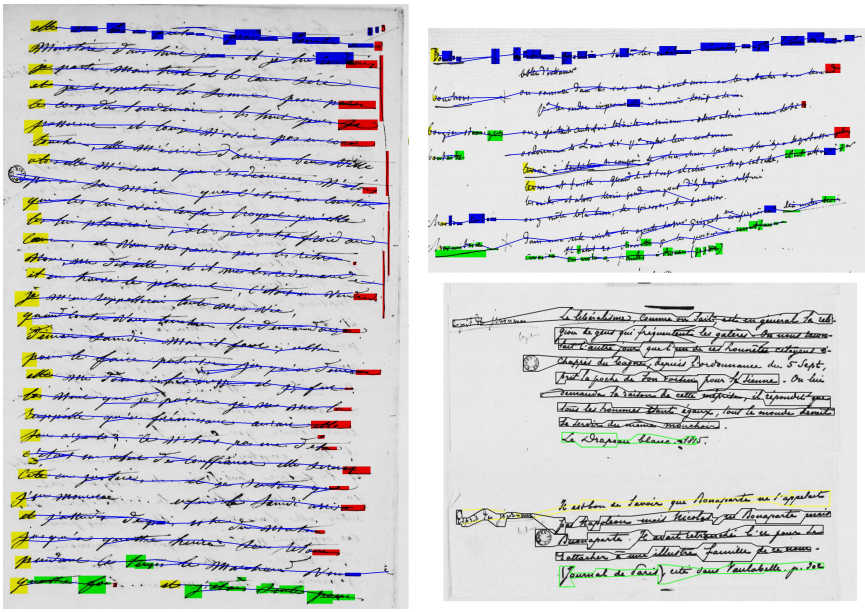

FIG. 5 – Résultat de l'extraction des lignes

fonctionne bien sur des pages de mise en page relativement simple. Des erreurs peuvent apparaître sur des pages plus complexes, notamment lorsque 2 fragments sont très proches, comme on peut le constater sur la figure 6.

## 5. Conclusions

Dans cet article on a proposé des méthodes pour extraire les éléments de structure physique d'un document manuscrit. En considérant que l'extraction des fragments est généralement une tache réalisée manuellement, et donc coûteuse, notre proposition offre des outils intéressants pour les chercheurs en sciences humaines. Les premiers tests réalisés montrent que notre approche est cohérente en regard de la complexité des pages du corpus. L'utilisation de la distance inter connexités définie section 3.3 nous permet de nous affranchir de la plupart des limitations dues à l'utilisation d'une approche par composante connexes. Ces algorithmes seront intégrés à la plateforme éditoriale des dossiers de *Bouvard et Pécuchet*

## Bibliographie

1. I.Leplumey A.Lemaitre, B. Coüasnon. Using a neighbourhood graph based on voronoï tessellation with dmos, a generic method for structured document recognition. *Graphics Recognition*, 3926 :267–278, 2006.

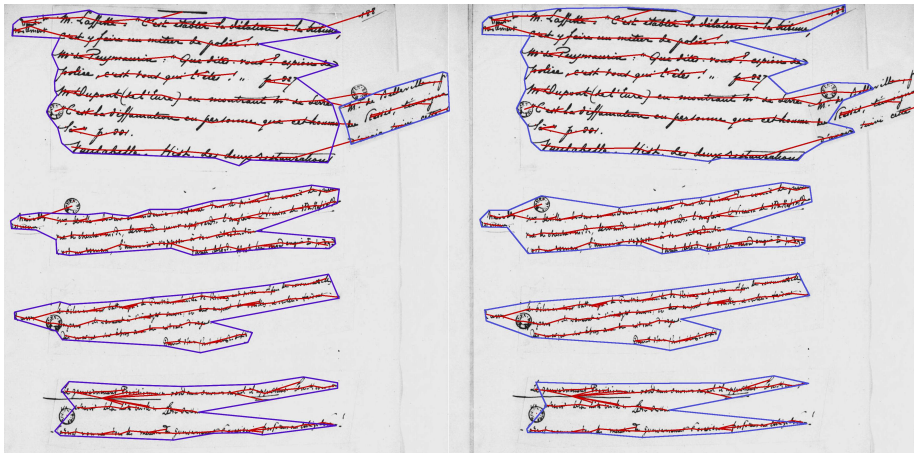

FIG. 6 – Fragments extraits : Vérité Terrain (Gauche), Simulation (Droite)

2. B.Yu et A.K. Jain. A robust and fast skew detection algorithm for generic documents. *PR*, 29(10) :1599–1629, 1996.
3. I.Pratikakis G.Louloudis, B.Gatos et C.Halatsis. Text line detection in handwritten documents. *PR*, 41(14) :3758–3772, 2008.
4. K. Kise, A. Sato, et M. Iwata. Segmentation of page images using the area voronoi diagram. *CVIU*.
5. Y. Li, Y.F. Zheng, D. Doermann, et S. Jaeger. Script-independent text line segmentation in freestyle handwritten documents. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
6. L. Likforman Sulem, A. Zahour, et B. Taconet. Text line segmentation of historical documents : a survey. *IJDAR*.
7. A.Hanimyan L.Likforman-Sulem et C.Faure. A hough based algorithm for extracting text lines in handwritten documents. In *ICDAR '95*, page 774, 1995.
8. S. Nicolas, T. Paquet, et L. Heutte. Complex handwritten page segmentation using contextual models. In *DIAL06*.
9. S. Nicolas, T. Paquet, et L. Heutte. A markovian approach for handwritten document segmentation. In *ICPR06*.
10. L. O'Gorman. The document spectrum for page layout analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(11) :1162–1173, 1993.
11. Joost Van Beusekom, Daniel Keysers, Faisal Shafait, et Thomas M. Breuel. Distance measures for layout-based document image retrieval. In *DIAL '06 : Proceedings of the Second International Conference on Document Image Analysis for Libraries*, pages 232–242, Washington, DC, USA, 2006. IEEE Computer Society.

*Ce travail a été réalisé grâce au financement du Cluster 13 de la région Rhône-Alpes : culture, patrimoine et création. Le projet BOUVARD est supporté par l'agence nationale pour la recherche ANR. Les images des dossiers de Bouvard et Pécuchet sont la propriété de la bibliothèque municipale de Rouen.*