# Combinaison de connaissances physiques et textuelles pour la reconnaissance d'images de registres anciens

Solène Tarride

DOCTORAT BRETAGNE LOIRE / MATHSTIC

INSA RENNES

# THÈSE DE DOCTORAT DE

L'INSTITUT NATIONAL DES SCIENCES APPLIQUÉES RENNES

ÉCOLE DOCTORALE N° 601
*Mathématiques et Sciences et Technologies de l'Information et de la Communication*
Spécialité : *Informatique*

Par

**Solène TARRIDE**

## « Combinaison de connaissances physiques et textuelles pour la reconnaissance d'images de registres anciens »

**Thèse présentée et soutenue à Rennes, le 11 avril 2022**
**Unité de recherche : IRISA (UMR 6074)**
**Thèse N° : 22ISAR 07 / D22 - 07**

### Rapporteurs avant soutenance :

| | |
|---|---|
| Nicholas JOURNET | Maître de Conférences (HDR) à l'Université de Bordeaux |
| Thierry PAQUET | Professeur à l'Université de Rouen Normandie |

### Composition du Jury :

| | | |
|---|---|---|
| Président : | Harold MOUCHERE | Professeur à l'Université de Nantes |
| Examinateurs : | Véronique EGLIN | Professeur à l'INSA Lyon |
| | Josep LLADÓS | Professeur associé à l'Université Autonome de Barcelone |
| | Jean-Yves RAMEL | Professeur à l'Université de Tours |
| Dir. de thèse : | Bertrand COUASNON | Maître de conférence (HDR) à l'INSA de Rennes |
| Co-encadrante : | Aurélie LEMAITRE | Maître de conférence (HDR) à l'Université de Rennes 2 |

### Invité(s) :

| | |
|---|---|
| Sophie TARDIVEL | Présidente de Doptim |

# Remerciements

Tout d'abord, je tiens à remercier les membres du jury pour l'intérêt qu'ils ont porté à mon travail. Je remercie tout particulièrement Thierry Paquet et Nicholas Journet d'avoir accepté le rôle de rapporteur, et pour leur relecture attentive de mon manuscrit. Je remercie également Véronique Eglin, Jean-Yves Ramel, Josep Lladós et Harold Mouchère d'avoir accepté d'examiner mon travail de thèse.

Je souhaite remercier chaleureusement mes encadrants : Aurélie Lemaitre, Bertrand Coüasnon et Sophie Tardivel. J'ai beaucoup appris à vos côtés, et j'ai surtout pris beaucoup de plaisir à travailler avec vous. Merci de m'avoir laissé une grande autonomie pour explorer, expérimenter et aller au bout de mes idées, tout en restant toujours disponibles pour échanger, m'orienter vers de nouvelles pistes de recherche, et me rassurer dans les moments de doute. Cet équilibre m'a permis de prendre confiance en moi et a confirmé mon goût pour le travail de recherche.

J'adresse également toute ma reconnaissance aux nombreuses personnes qui ont contribué à l'avancement de ces travaux. Je pense en particulier à Aurélie Lemaitre, Ivan Leplumey, Ricarda Leroux, Philippe Hervagault, et plus largement tout le Cercle Généalogique de l'Est de l'Ille-Et-Vilaine, qui ont participé à l'effort de transcription des actes. L'intérêt que vous avez porté à mon sujet de thèse a renforcé ma motivation et mon intérêt pour la généalogie. Je souhaite aussi remercier Aurélie Lemaitre (encore !) et Jean Camillerapp pour l'aide précieuse apportée avec DMOS, et Killian Barrere pour notre collaboration sur la génération de documents synthétiques.

Je remercie également tous mes collègues d'Intuidoc pour la chaleureuse ambiance de travail et les nombreuses pauses-café, les pique-niques, les apéros, les escape games et les soirées jeux. Merci aussi à toute la Dopteam pour les nombreux moments de convivialité au bureau et à la mer.

Merci à mes parents, Isabelle et Bruno, de m'avoir toujours encouragée à être curieuse et ambitieuse. C'est grâce à vous si j'en suis là aujourd'hui.

Enfin, un grand merci à Alexandre d'avoir rendu cette période si agréable et facile à vivre, et de m'avoir soutenue (et supportée) pendant la rédaction de cette thèse.

# TABLE DES MATIÈRES

# Acronymes

**AD** Archives Départementales. 18, 20, 26, 27, 34–37, 44

**AP** Average Precision. 88

**BMS** Baptême Mariage Sépulture. 25, 28

**CE** Cross Entropy. 125, 126

**CGE35** Cercle Généalogique de l'Est de l'Ille-et-Vilaine. 23, 44

**CTC** Connectionnist Temporal Classification. 125, 126

**FN** False Negative. 89, 99, 100, 105

**FP** False Positive. 89, 99, 100, 105

**GAN** Generative Adversarial Networks. 192

**HMM** Hidden Markov Models. 72

**HTR** Handwritten Text Recognition. 70

**IE** Information Extraction. 70

**NER** Named Entity Recognition. 70

**NMD** Naissance Mariage Décès. 26, 28

**seq2seq** Séquence à séquence. 66, 121

**TN** True Negative. 99, 100, 105

**TP** True Positive. 89, 99, 100, 105

# TABLE DES FIGURES

# Liste des tableaux

# Résumé

Cette thèse CIFRE est une collaboration entre Doptim et l'IRISA, dans le cadre du contrat ANRT n°2018/0896. Nos travaux portent sur la reconnaissance de registres de population français datant du XVIè au XVIIIè siècle : les registres paroissiaux. Ces documents sont structurés en actes de baptême, mariage et sépulture et contiennent donc des informations importantes pour les généalogistes souhaitant retracer leur histoire familiale.

Les outils développés dans cette thèse sont destinés à enrichir Geneafinder, le site de généalogie développé par Doptim. Nos contributions pour la reconnaissance de registres paroissiaux s'articulent en trois axes.

Le premier axe porte sur la reconnaissance de la structure de ces registres. Nous proposons une méthode hybride capable de découper ces documents en actes, en s'appuyant sur l'apprentissage de motifs structurels (signatures, lignes de texte, bords de pages) et sur le groupement logique de ces motifs à l'aide de règles. Nous démontrons l'intérêt de cette approche dans un contexte où peu de documents sont disponibles pour l'apprentissage.

Le deuxième axe adresse la reconnaissance du contenu textuel de ces documents, dans le but d'identifier les mots importants présents dans les actes : la date, les noms, les prénoms, ou encore les métiers. Nous adaptons une architecture neuronale avec mécanisme d'attention pour la reconnaissance d'écriture manuscrite, et démontrons l'intérêt de ce mécanisme pour la reconnaissance conjointe d'écriture manuscrite et d'entités nommées. Nous proposons également différentes stratégies d'apprentissage basées sur cette architecture.

Le troisième axe s'articule autour de la génération de documents synthétiques. Nous proposons une méthode pour générer des actes synthétiques, en modélisant les structures de phrases récurrentes des actes et en implémentant des transformations permettant de déformer et dégrader les images générées grâce à des polices manuscrites. Nous montrons l'intérêt de cette approche pour réduire le besoin en transcriptions manuelles, qui est un enjeu majeur de ce travail.

# Introduction

En 1972, l'Organisation des Nations unies pour l'éducation, la science et la culture a initié une politique globale de préservation de notre héritage culturel [UNESCO 1972]. La numérisation des documents à forte valeur historique, culturelle ou patrimoniale est l'une des étapes clés de cet effort de préservation. En effet, celle-ci réduit considérablement les risques de dégradations des œuvres originales, qui peuvent être consultées sans être manipulées. En outre, elle facilite l'archivage et la diffusion de ces documents à grande échelle. Depuis les années 1990, les opérations de numérisation se sont multipliées dans les services d'archives, les musées, ou encore les bibliothèques. De nombreuses collections de documents historiques ont été numérisées, puis progressivement mises en ligne dans les années 2000. Aujourd'hui, des bibliothèques numériques telles que Gallica [1] ou Europeana [2] nous permettent d'explorer notre héritage culturel à travers des collections d'art, d'articles de presse, et d'ouvrages de géographie, de médecine ou de sport.

La numérisation a également impacté les registres de population. En France, de nombreuses collections de documents retraçant la vie de nos ancêtres ont été numérisées : état civil, registres matricules, tableaux de recensement, et archives notariales. La diffusion et l'analyse de ces documents intéressent particulièrement les généalogistes amateurs et professionnels [Gildas 1988], car ils permettent de retrouver des ancêtres et d'établir des liens de filiation entre eux. Mais l'exploitation macroscopique des informations contenues dans ces registres pourraient également intéresser les géographes, historiens, sociologues, et démographes [Gras 1939 ; Goubert 1954]. En effet, la mise en relation entre ces collections permettrait d'analyser et de comprendre les évolutions temporelles et géographiques de divers phénomènes : les déplacements, les relations sociales, la longévité et les causes de décès, la fécondité et les relations familiales, ou encore la propagation des épidémies.

Certains généalogistes amateurs et membres d'associations de généalogie ont entrepris d'extraire des informations à partir de ces documents, notamment grâce à des plateformes d'annotation collaborative. Cependant, face à la masse de documents numérisés, l'effort de lecture et de transcription des actes est considérable. Il semble aujourd'hui

1. Bibliothèque numérique de la Bibliothèque nationale de France : `https://gallica.bnf.fr`
2. Bibliothèque numérique de la Commission Européenne : `https://www.europeana.eu`

indispensable de développer des outils capables d'extraire le contenu textuel des documents de manière automatique. Ces dernières années, de nombreux chercheurs en vision par ordinateur et en traitement du langage naturel se sont penchés sur la reconnaissance automatique de documents numérisés. Les avancés scientifiques et technologiques dans ces domaines permettent désormais d'envisager, sous certaines conditions, une extraction automatique d'information à partir de documents numérisés. La mise au point d'un système de reconnaissance de documents d'archives faciliterait considérablement la recherche généalogique.

## Contexte applicatif et objectifs

C'est dans cet objectif que Doptim [3], entreprise Rennaise spécialisée en sciences des données et intelligence artificielle, a proposé ce sujet de thèse. Doptim développe Geneafinder [4] : un site de généalogie, dédié aux généalogistes amateurs qui souhaitent effectuer des recherches dans les archives pour compléter leur arbre généalogique. L'objectif pour Doptim est de proposer de nouveaux services innovants aux utilisateurs de Geneafinder, en particulier pour faciliter la lecture des documents anciens.

Dans ce contexte, nous nous intéressons à un type de registre de population français : les registres paroissiaux. Ces documents étaient rédigés par les prêtres depuis le milieu du XVI$^{e}$ siècle, et compilent les actes de baptême, de mariage religieux, et de sépulture des membres de la paroisse. Ces registres ont donc une forte valeur historique, car ils contiennent des informations localisées et datées concernant la population française sur plusieurs siècles. Tous ces documents ont été numérisés par les Archives Départementales (AD), et sont à présents consultables gratuitement en ligne dans la plupart des départements.

L'objectif des travaux réalisés pendant cette thèse est de développer un prototype de reconnaissance de registres paroissiaux, afin d'enrichir le site Geneafinder. Plusieurs outils sont envisagés pour faciliter le travail des utilisateurs. Dans un premier temps, Doptim souhaite proposer une interface permettant la localisation des actes et les lignes de texte à partir des images de registres. Les utilisateurs pourront ainsi assigner un label et saisir une transcription pour chaque zone de texte détectée. Dans un second temps, Doptim espère développer un outil capable de reconnaître automatiquement le texte manuscrit sur ces

3. `https://doptim.eu/`
4. `https://geneafinder.com/`

documents, ainsi que la localisation et l'extraction des informations les plus pertinentes : noms, prénoms, dates ou métiers. Le développement de tels outils permettrait d'envisager une indexation des images de registres. Les utilisateurs pourraient ainsi effectuer des recherches par mots clés dans ces images, et en particulier par nom. De plus, une extraction d'information fiable et massive de ces données ouvrirait également la voie à la construction automatique d'arbres généalogiques.

La réalisation de ces outils est un challenge scientifique qui s'inscrit dans le domaine de la reconnaissance automatique de documents, situé à la frontière entre la vision par ordinateur et le traitement du langage naturel. Plusieurs étapes sont nécessaires à la reconnaissance de ces documents. Celles-ci sont illustrées sur la figure 1.

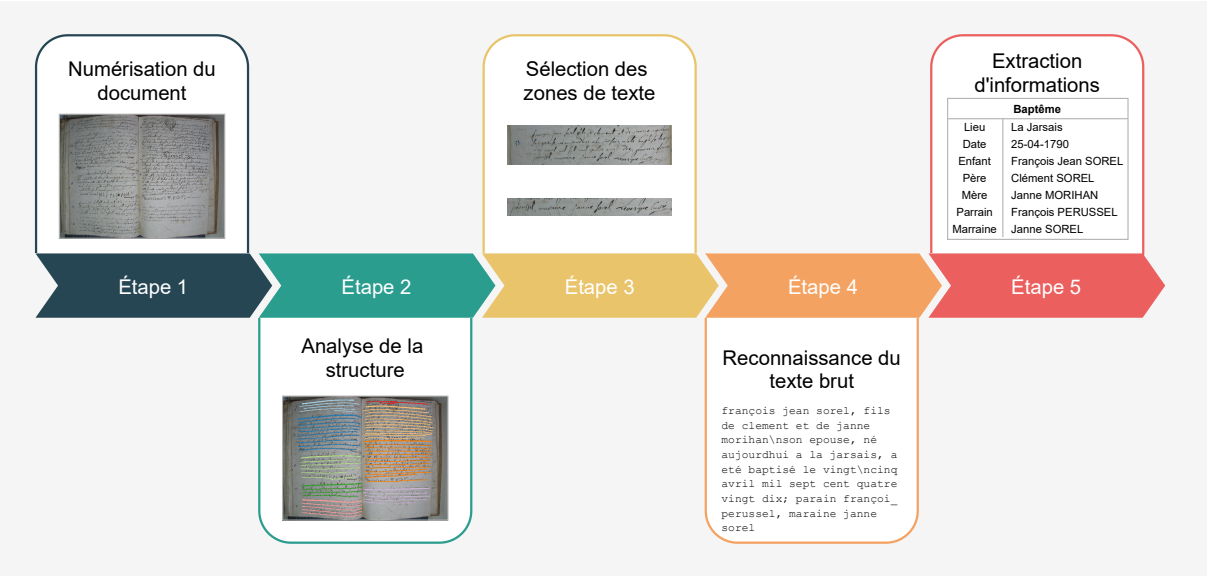



**Figure 1** – Chaîne de traitement d'un système de reconnaissance automatique de document. Une fois le document numérisé, une analyse de sa mise en page est effectuée. Les zones de texte homogènes (lignes, paragraphes, pages) sont extraites puis analysées par un modèle de reconnaissance d'écriture manuscrite. La transcription obtenue est à son tour analysée afin d'extraire les informations pertinentes.

La première phase d'analyse consiste **reconnaître la structure** de ces documents, en séparant les éléments de différentes natures : lignes de texte, illustrations, sceaux, tableaux, ou encore signatures. Les registres paroissiaux étant des documents complètement textuels, l'enjeu principal est de localiser les lignes de texte et les actes. Cette première étape permet, d'une part, de faciliter la lecture des registres lors de leur consultation en ligne, et d'autre part, de simplifier la phase de reconnaissance de texte et d'extraction d'information.

La deuxième étape consiste à **reconnaître le texte manuscrit** sur ces documents. Cette deuxième étape permettrait d'aider les généalogistes à déchiffrer l'écriture des prêtres. Si la tâche de reconnaissance de texte manuscrit historique est très complexe, certaines similarités entre les actes d'un même type peuvent tout de même être exploitées. En particulier, les prêtres employaient un vocabulaire relativement restreint, ainsi que des structures de phrase récurrentes.

Enfin, la dernière étape consiste à enrichir la transcription obtenue automatique, en **localisant les informations pertinentes** pour la reconstruction des histoires familiales : les noms, les prénoms, les dates, les lieux, ou encore les métiers. En effet, l'analyse du texte contenu dans les actes doit permette de remplir des bases de connaissances structurées, afin d'indexer ces documents.

## Verrous scientifiques

La difficulté principale de ce travail vient de la complexité et de l'hétérogénéité du corpus à traiter. Nous identifions quatre difficultés majeures dans le traitement automatique de ces registres.

Un premier verrou vient donc de la dégradation de ces documents, malgré des conditions de préservation optimales aux Archives Départementales. La qualité du papier et de l'encre sont deux sources de dégradations majeures, car l'encre peut être baveuse, brunie, ou partiellement effacée et le papier tâché, déchiré ou troué. Par conséquent, certains actes sont difficilement lisibles.

Le deuxième verrou, vient de l'hétérogénéité des conditions de numérisation des documents. En effet, dans certains départements, la numérisation des archives s'est effectuée avec différents modèles d'appareil photo. Ainsi, la résolution des images, ainsi que les conditions d'éclairage ou de prise de vue sont variables, ce qui complique l'application de systèmes de reconnaissance automatique.

Le troisième verrou vient de la variabilité des mises en page. Les registres sont des documents faiblement structurés : ils sont organisés en actes, mais souvent sans la présence de séparateurs physiques nets. Ainsi, il est parfois difficile d'identifier et de séparer visuellement les différents actes présents dans une même page.

Le dernier verrou vient de la complexité de lecture de ces documents. L'écriture cursive de ces documents historiques est difficilement déchiffrable, car la forme des tracés, le style d'écriture, ainsi que la structure des phrases ont considérablement évolué au fil des siècles.

En outre, les dégradations de l'encre du papier empêchent la lecture.

Toutes ces difficultés rendent la lecture de ces documents très laborieuse. En conséquence, il est particulièrement coûteux et chronophage d'obtenir des annotations sur la mise en page ou le contenu. Pourtant, la mise en place et l'évaluation d'un prototype de reconnaissance automatique de documents nécessite une base de données labellisée, c'est-à-dire associée à des annotations. En particulier, les méthodes de reconnaissance les plus performantes se basent sur *l'apprentissage supervisé*, et nécessite un grand nombre d'exemples pour apprendre à reconnaître la structure des documents ou bien leur contenu textuel.

Or, il n'existe pas à ce jour de base de données annotée de registres paroissiaux. Si certains généalogistes amateurs ont réalisé des transcriptions, celles-ci sont souvent partielles, ou différentes du texte original (corrections orthographiques ou grammaticales, abréviations). Pour être exploitables, ces transcriptions doivent donc être corrigées et normalisées. Des associations de généalogie ont, quant à elles, réalisé des relevés d'information à partir des registres. En particulier, les bénévoles ont extrait les noms, prénoms, date, et autres informations pertinentes dans des actes. Mais ces relevés ne sont pas directement exploitables pour apprendre un modèle d'extraction d'information, car la transcription complète n'est pas disponible. Les informations sur la mise en page des documents ne sont quant à elles jamais annotées par les généalogistes.

## Contributions

Dans cette thèse, nous proposons quatre contributions pour la reconnaissance automatique de registres paroissiaux :

— une méthode de reconnaissance de mise en page de ces registres. Notre approche, combinant des réseaux de neurones et des règles logiques, est capable de reconnaître la structure de ces registres : localisation des pages, des actes, des lignes de texte, des signatures ;

— l'adaptation d'une architecture pour la reconnaissance d'écriture. L'architecture sélectionnée est basée sur un réseau de neurones séquence à séquence avec mécanisme d'attention. Elle permet de reconnaître le texte à partir d'images de lignes de texte ;

— une méthode pour localiser et extraire les informations pertinentes dans ces registres. En particulier, nous comparons les approches séquentielles et conjointes

pour l'extraction d'information dans les actes ;
— une chaîne de traitement pour générer des actes synthétiques annotés, ce qui permet de limiter le besoin en transcriptions manuelles.

Ce travail a également nécessité l'annotation de registres paroissiaux. En effet, ces annotations sont nécessaires pour l'évaluation de nos contributions, ainsi que pour l'apprentissage de certains modèles de reconnaissance.

# Plan du manuscrit

La suite de ce document comporte six chapitres.

**Le premier chapitre** présente l'historique des archives démographiques françaises, ainsi que leur intérêt pour la recherche généalogique. Plus précisément, nous décrivons l'importance des registres paroissiaux, ainsi que les caractéristiques de ces documents. Enfin, nous introduisons un protocole d'annotation applicable à ces registres, et présentons les ensembles de données annotés pendant ce projet. Nous mettons ainsi en lumière la contrainte principale de cette thèse : le manque d'annotations associées aux registres paroissiaux.

**Le deuxième chapitre** expose l'état de l'art dans le domaine de la reconnaissance automatique de documents manuscrits ou historiques. Nous présentons les différentes tâches adressées par la communauté scientifique, ainsi que les approches permettant la reconnaissance de structure, la reconnaissance d'écriture, et l'extraction d'information. Nous décrivons également les bases de données publiques permettant d'évaluer les systèmes de reconnaissance automatique de documents. Enfin, nous discutons des stratégies les plus pertinentes pour la reconnaissance de registres paroissiaux, à la lumière des objectifs et des verrous scientifiques présentés dans le premier chapitre.

**Le troisième chapitre** adresse la question de la reconnaissance de mise en page des registres paroissiaux. En premier lieu, nous décrivons la structure de ces documents, ainsi que les indices visuels permettant la localisation d'éléments graphiques et textuels. À la lumière de cette analyse, nous présentons quatre contributions pour la localisation des actes dans les images de registres. Enfin, nous comparons ces approches sur des registres paroissiaux et sur une base de données publique.

**Le quatrième chapitre** présente l'approche envisagée pour la reconnaissance d'écriture manuscrite. Nous adaptons une architecture neuronale séquence à séquence avec mécanisme d'attention pour la reconnaissance optique de caractères manuscrits. Nous dé-

crivons les expérimentations réalisées sur cette architecture ainsi que sur les paramètres d'apprentissage, puis nous évaluons l'apport de cette approche sur quatre bases de données publiques. Enfin, nous étudions l'impact de l'apprentissage par transfert, afin de spécialiser cette architecture sur l'écriture particulière des registres paroissiaux, et ce à partir de peu de données d'apprentissage spécialisées.

**Le cinquième chapitre** expose nos contributions pour l'extraction d'informations pertinentes dans ces registres. Nous comparons deux stratégies d'analyse : une approche séquentielle, pour laquelle la phase de reconnaissance d'écriture intervient avant la phase de classification en entités nommées, et une approche combinée, pour laquelle ces deux tâches sont effectuées en même temps. Nous évaluons l'apport des réseaux séquence à séquence pour cette stratégie combinée, puis nous proposons deux nouvelles approches basées sur cette stratégie. Enfin, nous évaluons nos contributions sur une base de données publique.

**Le sixième chapitre** s'intéresse à la génération de documents synthétiques. Nous présentons une chaîne de traitement pour générer automatiquement des actes réalistes, à partir de polices manuscrites. Pour cela, nous modélisons les structures de phrase récurrentes dans les registres paroissiaux, et utilisons des relevés d'informations réalisés par le Cercle Généalogique de l'Est de l'Ille-et-Vilaine (CGE35). Dans un second temps, nous détaillons les transformations réalisées sur les images générées, afin d'augmenter artificiellement la variabilité des écritures et les dégradations. Puis, nous étudions les stratégies d'apprentissage par transfert et évaluons l'apport de cette contribution pour la reconnaissance d'écriture et l'extraction d'informations dans les registres paroissiaux.

Enfin, nous résumons le travail réalisé pendant cette thèse, et discutons des perspectives intéressantes pour la suite du projet.

Chapitre 1

# Présentation des documents

Dans ce chapitre, nous introduisons des éléments contextuels et historiques concernant les registres de population français, ainsi que leur intérêt en généalogie. Plus précisément, nous présentons les registres paroissiaux et mettons en lumière leurs caractéristiques ainsi que la difficulté de traitement de ces documents. Nous définissons ensuite les objectifs de reconnaissance pour ces documents, et introduisons la notion de *base de données annotée*, essentielle pour le développement et l'évaluation de systèmes de reconnaissance automatique. Enfin, nous présentons les bases de données mises au point pour ce travail, en décrivant les méthodes d'échantillonnage des registres ainsi que les annotations réalisées.

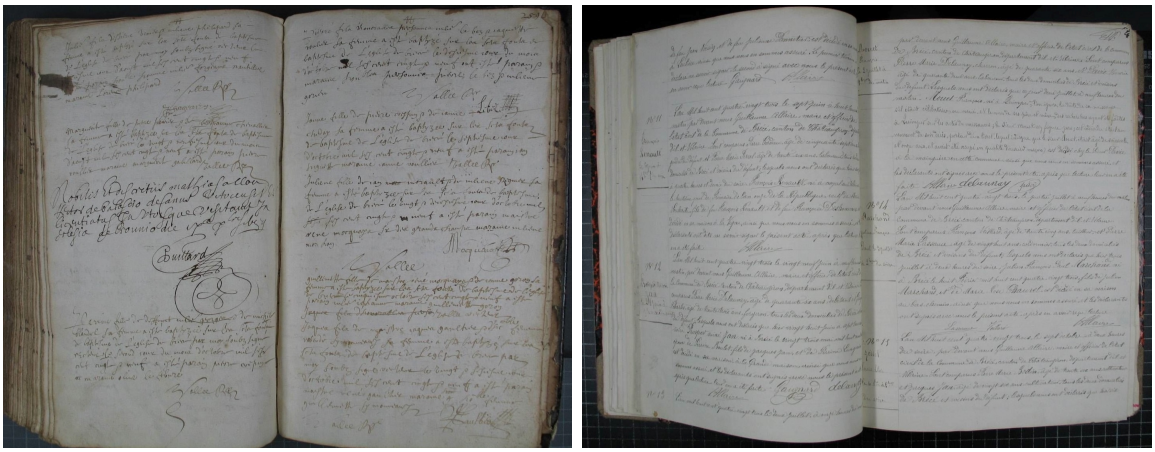
## 1.1 Les registres de population en France

Dans cette section, nous présentons les registres paroissiaux et d'état civil français. Ces documents ont un intérêt majeur pour les généalogistes souhaitant retracer leur histoire familiale.

### 1.1.1 Contexte historique

En France, les registres paroissiaux, dits BMS (Baptême Mariage Sépulture), ont été instaurés par l'ordonnance de Villers-Cotterêts, en 1539. Celle-ci fait du français la langue officielle en France, et rend obligatoire l'enregistrement des actes de baptême des membres de la paroisse par les curés. Puis, en 1579, l'ordonnance de Blois impose aux prêtres d'inscrire également les actes de mariage et de sépulture dans les registres paroissiaux. À partir de cette période, les registres paroissiaux compilent donc les actes nominatifs et datés pour trois types de cérémonie : les baptêmes, les mariages religieux, et les sépultures. En 1667, l'ordonnance de Saint-Germain-en-Laye, prescrit une tenue obligatoire des registres en deux exemplaires : l'original est conservé dans la paroisse tandis que la copie est déposée chaque année au greffe du bailliage. Cette ordonnance impose également aux témoins de

signer les actes. Enfin, en 1736, l'ordonnance du Chancelier d'Aguesseau prescrit la tenue de deux registres originaux, c'est-à-dire signés, et non plus une simple copie. La Révolution Française initie une laïcisation des registres de population. En 1792, la tenue des registres est confiée aux maires, avec l'instauration de l'état civil que nous connaissons aujourd'hui. Les registres d'état civil, dits NMD (Naissance Mariage Décès), compilent tous les actes de naissance, mariage civil, et décès. Les mairies sont également tenues de rédiger des tables décennales, qui reprennent par ordre alphabétique tous les actes d'état civil enregistrés dans une même commune pendant dix ans. Des exemples de registres paroissiaux et d'état civil sont présentés dans la figure 1.1.



**(a)** Registre paroissial (1641)
Cote 10 NUM 35039 1

**(b)** Registre d'état civil (1881)
Cote 10 NUM 35001 752

**Figure 1.1** – Exemples de registres paroissiaux et d'état civil. Ces documents proviennent de la paroisse/commune de Brécé et sont conservés aux Archives Départementales d'Ille-et-Vilaine.

En 1996, le ministère de la Culture lance le plan national de numérisation, qui encourage le développement de projets de numérisation dans les bibliothèques, musées et services d'archives [France 2008]. Le but de ce projet est de faciliter l'accès des documents au plus grand nombre, tout en garantissant une conservation optimale des documents originaux. Aujourd'hui, la numérisation des registres de population français touche à sa fin. Selon le site FranceArchives [1], seuls deux départements n'avaient pas intégralement numérisé leurs collections en 2018 [2]. Les documents numérisés par les Archives Départementales incluent les registres d'état civil et paroissiaux, mais également des registres matricules, plans cadastraux, photographies, cartes postales, affiches ou encore journaux anciens. Toutes ces

1. `https://francearchives.fr/map/b2b076144a3c4392a34b14bb5e364c95`
2. Le Gard et le Jura. Il n'existe pas de données plus récentes.

ressources sont désormais consultables physiquement en salle de lecture, ou en ligne sur les sites des Archives Départementales (AD) [3].

La figure 1.2 synthétise l'historique des archives paroissiales et d'état civil en France, de leur instauration au début de la numérisation.

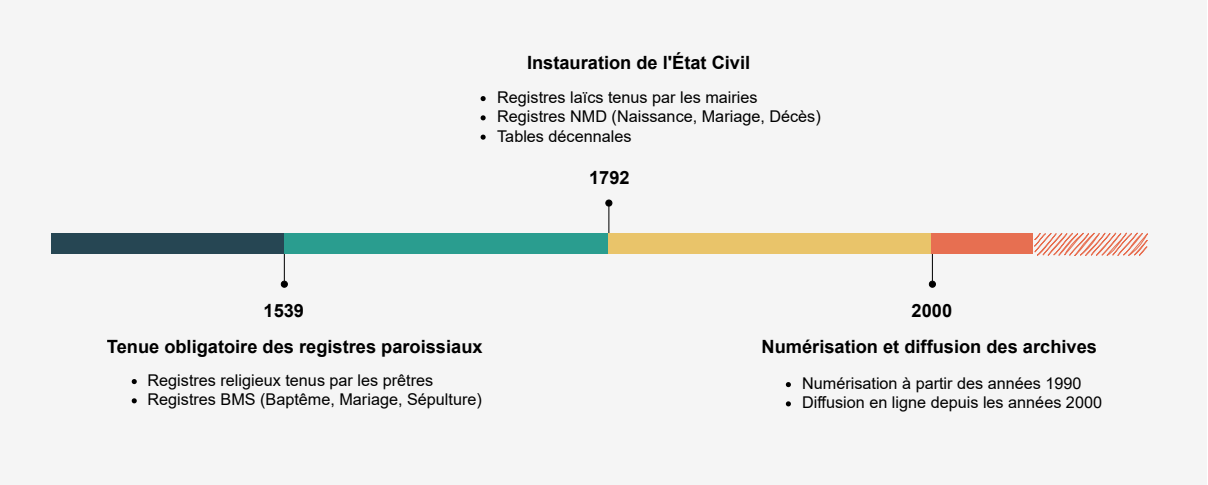



**Figure 1.2** – Historique des registres paroissiaux et d'état civil en France. Les registres de population étaient tenus par des prêtres du jusqu'à la Révolution Française, puis par les mairies. Ces documents, désormais conservés aux Archives Départementales, ont été numérisés puis mis en ligne à partir des années 2000.

### 1.1.2 Recherche généalogique

Les registres de population ont un intérêt particulier en généalogie. En effet, les généalogistes amateurs et professionnels se basent sur les informations contenues dans ces documents pour retrouver leurs ancêtres et établir des liens de filiation entre eux. La diffusion des archives en ligne a considérablement facilité l'accès à ces informations, car les généalogistes n'ont plus besoin de se déplacer physiquement en salle de lecture pour consulter les documents.

Mais si des millions d'images de registres sont désormais consultables en ligne, il n'existe pas d'accès direct à ces images par le contenu. Par exemple, il n'est pas possible d'effectuer une recherche par nom pour retrouver un document. Pour autant, ces registres numérisés ont été indexés par deux informations : le lieu de rattachement (paroisse ou commune) et la date. Il est donc nécessaire de connaître précisément ces deux informa-

3. Par exemple pour l'Ille-et-Vilaine : `https://archives.ille-et-vilaine.fr/fr`

tions pour retrouver un registre. La recherche d'un acte s'effectue ensuite en parcourant le registre page par page. Il est important de noter qu'un registre couvre souvent plusieurs années et contient des milliers d'actes inscrits par ordre chronologique. Par conséquent, la recherche généalogique reste extrêmement chronophage.

Dans les registres NMD, les actes apparaissent dans l'ordre chronologique, mais la recherche d'un acte est simplifiée par la présence de tables décennales par paroisse. Les généalogistes peuvent effectuer une recherche par ordre alphabétique dans ces tables afin de retrouver la date exacte de l'acte recherché. L'acte peut ensuite être retrouvé plus facilement dans le registre original. Une grande partie des actes d'état civil est à ce jour indexée sur des sites généalogiques ou sur les sites d'associations généalogiques.

En revanche, la recherche d'actes dans les registres BMS est beaucoup plus laborieuse, car il n'existe généralement pas de tables pour ces registres. Ainsi, la recherche d'un acte nécessite la connaissance précise de la paroisse et de la date de l'acte. Il faut ensuite parcourir le registre page par page jusqu'à repérer les noms et prénoms de la personne recherchée. En France, des associations généalogiques ont effectué des relevés nominatifs partiels afin de générer des tables. Cependant, ces relevés ne sont pas complets, et ne sont pas toujours liés aux images originales. C'est à ces registres BMS que nous nous intéressons dans cette thèse. Plus spécifiquement, nous nous concentrons sur la période antérieure à la Révolution Française. Si des registres paroissiaux existent après cette période, ils n'intéressent pas particulièrement les généalogistes, qui préfèrent effectuer leur recherche dans les actes d'état civil.

Tous les relevés d'information sont actuellement effectués manuellement, souvent par des généalogistes bénévoles. Il n'existe pas à ce jour de système de reconnaissance automatique de documents en production sur les sites de généalogie.

## 1.2 Les registres paroissiaux

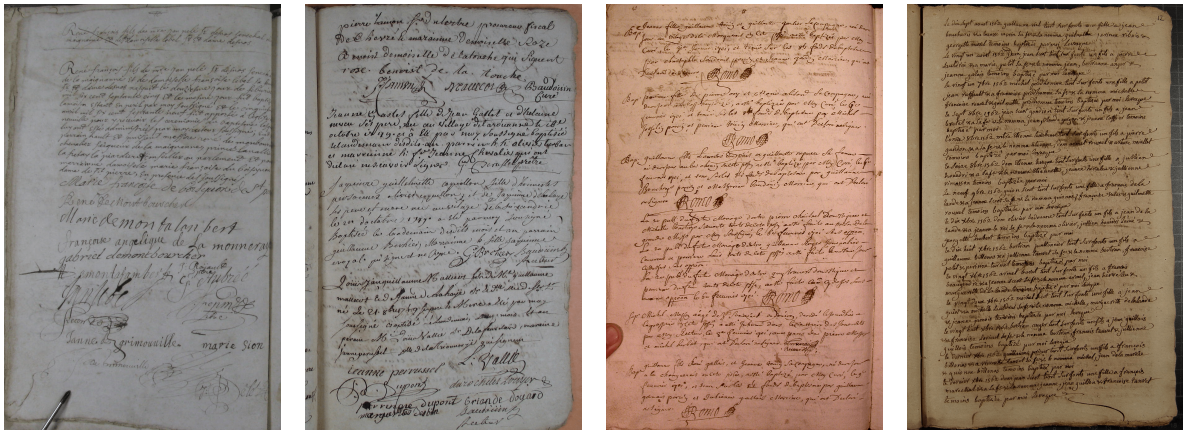
Dans cette section, nous présentons quelques caractéristiques des registres paroissiaux afin d'illustrer la difficulté de lecture, de compréhension, et d'interprétation de ces documents.

### 1.2.1 Mise en page

Les registres paroissiaux sont des documents faiblement structurés : les pages sont organisées en actes successifs, mais souvent sans la présence de séparateurs physiques nets. S'il existe certains éléments visuels permettant d'identifier le début ou la fin d'un acte, ils n'apparaissent pas systématiquement. C'est, par exemple, le cas des signatures à la fin des actes, des annotations marginales à gauche du texte, ou bien des majuscules en début d'acte.

De plus, la longueur des actes n'est pas uniforme. Par exemple, les actes de mariage sont généralement plus longs que les actes de baptême ou sépulture, car ils contiennent plus d'informations (nom et prénoms des mariés et de leurs parents, présence de nombreux témoins et signatures). Mais deux actes d'un même type peuvent contenir différents niveaux de détails, en fonction du niveau de détails fourni par le prêtre et du nombre de personnes présentes à la cérémonie. Par conséquence, certains actes prennent plus d'une page, quand d'autres ne prennent qu'une ou deux lignes (annonces de fiançailles, publication des bans).

Pour toutes ces raisons, il est parfois difficile d'identifier et de séparer visuellement les différents actes présents dans une même page. Dans ces cas-là, c'est la lecture du texte qui permet de repérer la structure logique du document. La figure 1.3 illustre la variabilité des mises en page des registres paroissiaux.



**(a)** Page avec 1 acte Cote 10 NUM 35003 32 — **(b)** Page avec 4 actes Cote 10 NUM 35031 214 — **(c)** Page avec 7 actes Cote 10 NUM 35347 68 — **(d)** Page avec 15 actes Cote 10 NUM 35121 1

**Figure 1.3** – Variabilité des mises en page des registres paroissiaux. Le nombre et la longueur des actes, ainsi que la largeur de la marge et la place occupée par les signatures varient sensiblement selon les pages.

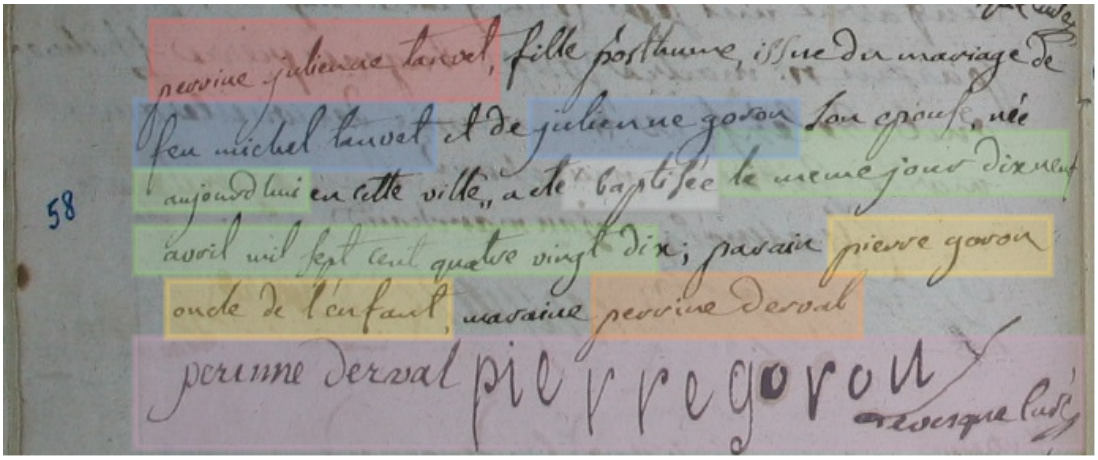
### 1.2.2 Informations contenues dans les actes

Les registres paroissiaux contiennent des actes de baptême, mariage, et sépulture, mais également des bans et annonces de fiançailles. Les actes contiennent plus ou moins de détails selon les époques, les types de cérémonies et les prêtres. Nous présentons ici les informations présentes dans la plupart des actes.

**Actes de baptême** Un exemple d'acte de baptême est présenté dans la figure 1.4. Pour ce type d'acte, le nom et prénom de la personne baptisée, les noms et prénoms des parents, ainsi que la date et le lieu du baptême sont généralement inscrits. L'enfant est fréquemment baptisé le jour de sa naissance, ou bien le lendemain. La plupart des actes contiennent les nom et prénom du parrain et de la marraine, avec éventuellement leur lien de parenté avec l'enfant. Certains actes sont plus détaillés, et contiennent la profession des parents, leur domicile, ou des informations sur leur statut social (par exemple : « honnête homme » ou « honnête femme »). Il est également parfois précisé si l'enfant est issu d'une union légitime ou non, ou si l'un des parents est décédé. Enfin, on retrouve souvent la signature du prêtre et celles des témoins du baptême.

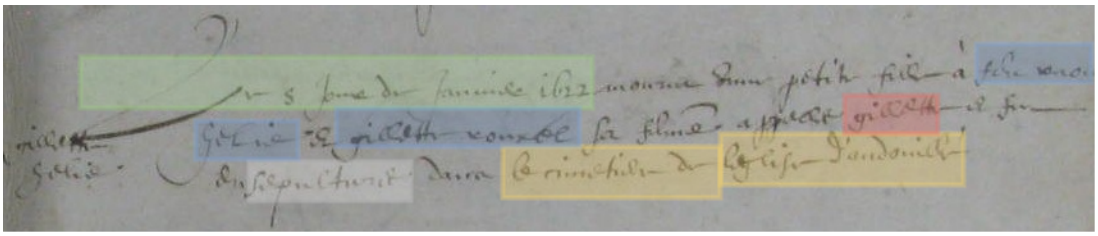
**Actes de sépulture** Un exemple d'acte de sépulture est présenté dans la figure 1.5. Un acte de sépulture contient les nom et prénom du défunt, ainsi que son âge estimé. La date et le lieu du décès et de l'inhumation sont souvent précisés. Si le défunt est un enfant, on retrouve souvent le nom et prénom de ses parents. Si c'est un adulte, le nom du conjoint ou sa situation familiale sont généralement inscrits. Enfin, on retrouve fréquemment la liste des témoins et leur signature, ainsi que celle du prêtre.

**Actes de mariage** Un exemple d'acte de mariage est présenté dans la figure 1.6. L'acte de mariage contient des informations sur les deux époux et leur famille respective, il est donc privilégié par les généalogistes. De plus, il est souvent précédé des annonces de fiançailles et des bans, ce qui permet de fiabiliser certaines informations lues sur l'acte de mariage. Sur les actes de mariage, les noms, prénoms, âges et lieux de naissance des époux sont généralement indiqués. Il est également fréquent de voir mentionné leur situation (par exemple : « majeur », « mineur » ou « veuf »). Les noms, prénoms et lieux d'origine des parents sont souvent mentionnés. Le curé précise également lorsque l'un des parents est décédé. Enfin, le prêtre inscrit généralement la liste des personnes présentes à la cérémonie. L'acte est ensuite signé par les époux, leurs témoins et le prêtre.



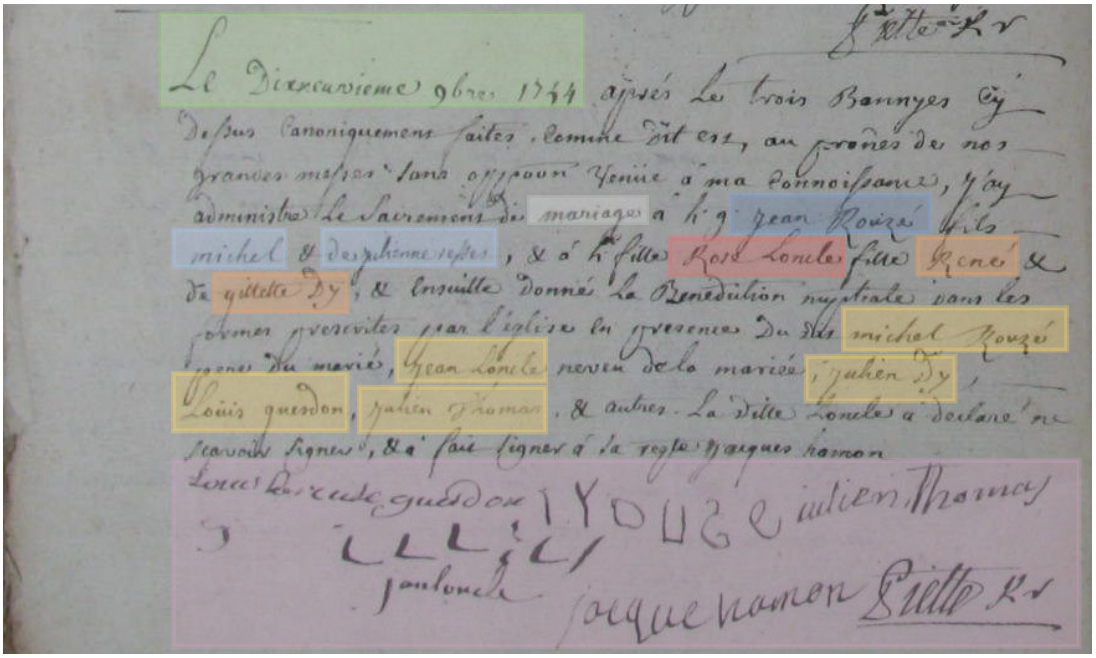
*perrine julienne tanvet, fille posthume, issue du mariage de feu michel tanvet et de julienne goron son epouse, née aujourd'hui en cette ville„ a eté baptisée le meme jour dix neuf avril mil sept cent quatre vingt dix ; parain pierre goron oncle de l'enfant, maraine perrine derval*

**Figure 1.4** – Un acte de baptême et sa transcription. Légende : nom et prénom de l'enfant, noms et prénoms des parents (père décédé), type d'acte, date, nom et prénom du parrain (lien familial), nom et prénom de la marraine, signatures.



*Le 5 jour de janvier 1622 mourut une petite fille à feu raoul helie et gillette rouxel sa feme appelee gillette et fut en sepulture dans le cimetiere de leglise d'andouillé*

**Figure 1.5** – Un acte de sépulture et sa transcription. Légende : date, type d'acte, noms et prénoms des parents (père décédé), prénom de l'enfant, lieu de la sépulture.



*Le Dixneuvième 9bre 1744 après les trois Bannyes cy dessus canoniquement faites, comme dit est, au prônes de nos grandes messes sans oppoon venüe à ma connoissance, j'ay administre le sacrement de mariage à h. g. Jean Rouzé fils michel et de julienne reffet, & à h. fille Rose Loncle fille René & de gillette dy, & ensuitte donné la benediction nuptiale dans les formes prescrites par l'eglise en presence du dit michel Rouzé pere du marié, jean loncle neveu de la mariée, julien Dy, Loüis guesdon, julien Thomas, & autres. La ditte Loncle a declaré ne scavoir signer, & a fait signer à la regle jacques hamon*

**Figure 1.6** – Un acte de mariage et sa transcription. Légende : date, type d'acte, nom et prénom du marié, nom et prénom des parents du marié, nom et prénom de la mariée, nom et prénom des parents de la mariée, nom et prénom des témoins présents, signatures. Abréviations : « oppoon » : « opposition », « h.g. » : « honnête garçon », « h. : honnête ».

### 1.2.3 Lecture des actes et paléographie

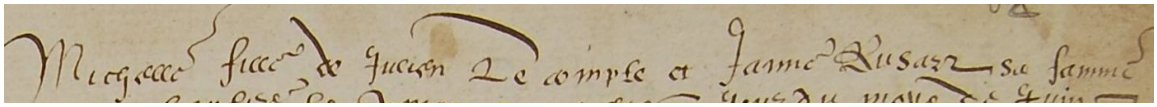
L'écriture cursive de ces documents historiques est très difficile à déchiffrer pour une personne peu habituée à la lecture des documents anciens. Nous détaillons ici quelques difficultés liées la lecture des écritures anciennes [Tarn 2013]. Un exemple de cette évolution est illustré sur la figure 1.7.

**La forme des lettres et les ligatures** Les registres paroissiaux présentent de nombreux styles d'écriture manuscrite variés, avec des tendances qui évoluent selon les années et les lieux. Ainsi, certains actes sont très difficiles à lire sans expertise paléographique, car les façons de tracer les lettres ont considérablement évolué au fil des siècles. De plus, le tracé de certaines lettres dépend parfois de leur localisation dans le mot. C'est le cas du `s` qui peut être écrit de différentes façons par un même scripteur, comme illustré sur la figure 1.8. Enfin, les ligatures entre les lettres compliquent la lisibilité des actes.

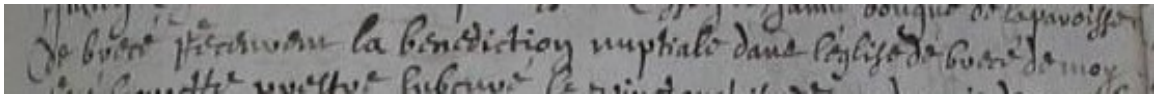
**L'orthographe** On trouve dans ces actes de nombreux mots dont l'écriture a évolué. Parmi les exemples les plus fréquents, on peut citer les mots « baptize » (« baptisé »), « mesme » (« même »), ou « moy » (« moi »). Cette évolution peut compliquer la lecture, car il est plus difficile d'interpréter des mots qui ne font plus partie du français actuel. Enfin, la ponctuation est souvent absente, et l'utilisation des majuscules est irrégulière. Ces éléments sont illustrés sur la figure 1.8.

**Les abréviations** De nombreuses abréviations sont utilisées dans les actes, ce qui complique leur lecture. Certains mots sont abrégés par suspension, ce qui signifie que le mot n'est pas fini. Un exemple fréquent est « h.h. » qui signifie « honnête homme » ou « lab. » pour « laboureur ». D'autres mots sont abrégés par contraction, ce qui signifie que le mot est tronqué des lettres du milieu. Certains prêtres écrivent « oppon » pour « opposition », ou « ptre » pour « prêtre ». Les dates, et en particulier les mois, sont également souvent abrégés : « 7bre » pour « septembre », « 8bre » pour « octobre », « 9bre » pour « novembre », « 10bre » pour « décembre ». Par exemple, dans l'acte de la figure 1.8, deux abréviations sont utilisées. Certains généalogistes, archivistes et paléographes ont entrepris de lister ces abréviations au fil de leur recherche. Une liste non exhaustive des abréviations fréquentes dans les actes est disponible sur le site Geneafinder [4].

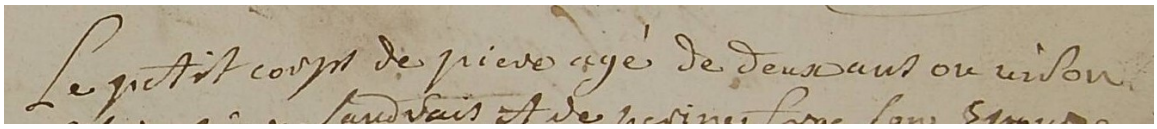
4. `https://geneafinder.com/blog?id=13%3A113`



**(a)** Écriture issue d'un acte de 1585.



**(b)** Écriture issue d'un acte de 1663.



**(c)** Écriture issue d'un acte de 1742.

**Figure 1.7** – Complexité et hétérogénéité des écritures manuscrites sur les registres paroissiaux. Les styles d'écriture évoluent selon les époques.

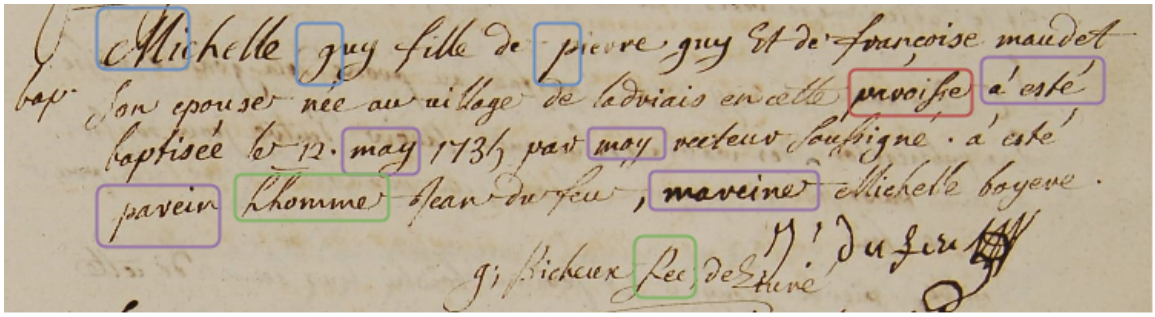
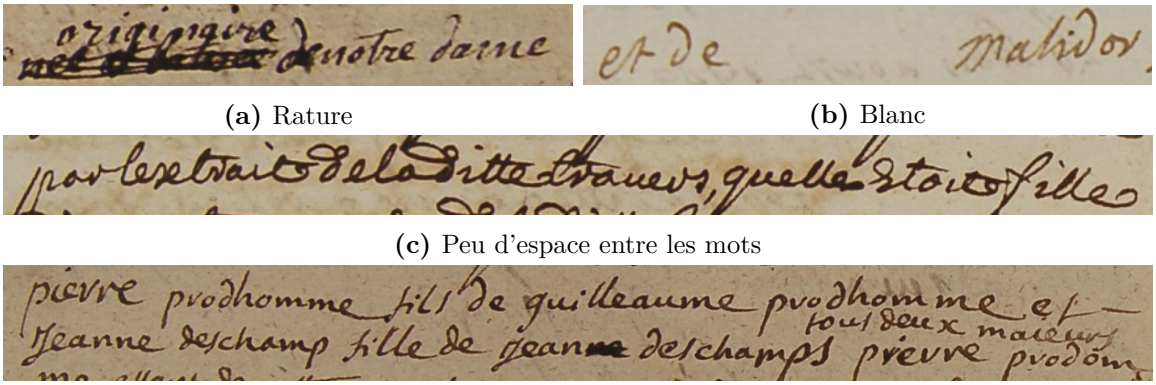



**Figure 1.8** – Un acte de 1734 présentant certaines difficultés de lecture : l'utilisation irrégulière des majuscules (en bleu), les différentes façons d'écrire la lettre `s` (en rouge dans le mot « paroisse »), les différences orthographiques (en violet, « may » pour « mai », « moy » pour « moi », « à esté » pour « a été », « parein » pour « parrain » et « mareine » pour « marraine ») et les abréviations (en vert : « hhomme » pour « honnête homme », « Rec » pour « Recteur »).

**Artefacts** Enfin, d'autres éléments sont facilement interprétables par un lecteur humain, mais peuvent perturber un système de reconnaissance automatique. C'est le cas des ratures et des corrections, des blancs, des annotations interligne, ou encore des lignes contenant peu d'espace entre chaque mot. Certains de ces éléments sont illustrés sur la figure 1.9

### 1.2.4 Numérisation des documents

La numérisation de ces registres a été réalisée par les Archives Départementales, souvent en partenariat avec des associations généalogiques locales. Un guide de numérisation



**(a)** Rature **(b)** Blanc

**(c)** Peu d'espace entre les mots

**(d)** Mentions supplémentaires dans l'interligne

**Figure 1.9** – Éléments perturbant la lecture automatique des actes : les ratures, blancs, annotations

a été défini en 2008 par la direction des Archives de France [France 2008] afin de normaliser les conditions de numérisation des documents historiques. Ce guide préconise la numérisation des documents avec les résolutions suivantes :

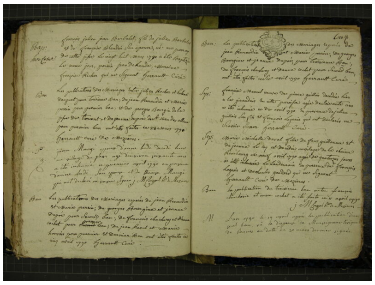
— 300 dpi pour les fichiers de conservation. Le fichier de conservation est une reproduction du document la plus fidèle possible à l'original, créée à des fins de conservation ;
— 150 dpi pour les fichiers de diffusion. Le fichier de diffusion est celui qui est destiné à être diffusé au public ;
— 72 dpi pour les fichiers de visualisation. Le fichier de visualisation est destiné à une consultation d'ensemble à l'écran.

En pratique, la numérisation des archives a débuté avant la publication de ce guide. Les conditions de numérisation sont variables, car différents appareils photos ont été utilisés. À titre d'exemple, aux Archives Départementales d'Ille-et-Vilaine, la plupart des documents ont été photographiés avec une résolution de 180 dpi. Aussi, si une telle résolution est adaptée pour la lecture humaine, la reconnaissance optique des caractères nécessite des résolutions plus élevées. Les leaders du marché (ABBY [5], Google Vision [6]) ou la Bibliothèque Nationale de France [7] préconisent une résolution à 300 ou 400 dpi pour la reconnaissance de caractères sur des documents numérisés.

Outre la faible résolution des images, les conditions d'éclairage et d'exposition, ainsi

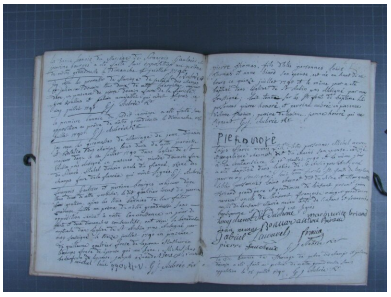
5. https://support.abbyy.com/hc/en-us/articles/360017733239
6. https://cloud.google.com/vision/docs/supported-files
7. https://www.bnf.fr/fr/formats-et-techniques-de-numerisation-en-mode-image

que la balance des blancs sont également variables. En effet, les documents peuvent avoir une teinte bleutée, jaunie ou orangée, selon les photographies. De plus, les supports des documents sont également hétérogènes (table en bois, tissu coloré, fond quadrillé, autre document), et peuvent ainsi perturber les méthodes automatiques de localisation des pages. D'autre part, le cadrage n'est pas uniforme. Bien que les prises de vue soient majoritairement en format paysage, certains documents ont été photographiés sous un autre angle. Par exemple, certaines pages ont été prises en mode portrait, quand d'autres ont été photographiées avec un effet de zoom. Les figures 1.10 et 1.11 illustre les variabilités de prise de vue et d'éclairage.



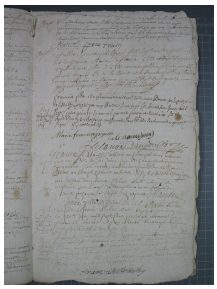
**(a)** Lumière jaune
Cote 10 NUM 35178 122

**(b)** Lumière rose
Cote 10 NUM 35347 207



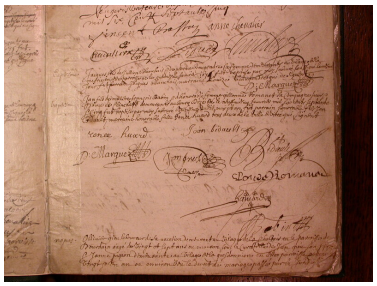
**(c)** Lumière bleue
Cote 10 NUM 3521 236

**Figure 1.10** – Variabilité de la luminosité et des balances de blancs lors la prise de vue des registres paroissiaux sur des images provenant des Archives Départementales d'Ille-Et-Vilaine.



**(a)** Portrait
Cote 10 NUM 35072 93

**(b)** Paysage
Cote 10 NUM 35192 154

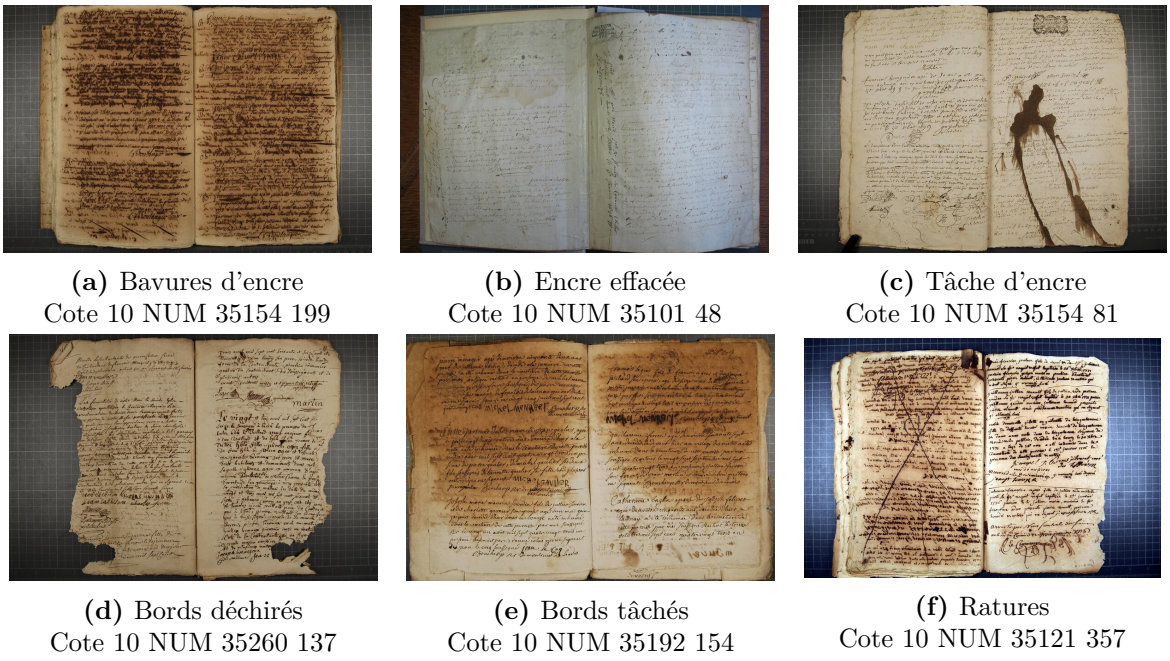


**(c)** Zoom
Cote 10 NUM 35260 137

**Figure 1.11** – Variabilité des conditions de numérisation des registres paroissiaux au niveau des conditions d'éclairage et des prises de vue, sur des images provenant des Archives Départementales d'Ille-Et-Vilaine.

### 1.2.5 Dégradations

Malgré des conditions de préservation optimales aux Archives Départementales, ces documents ont fait face à l'épreuve du temps. La qualité du papier et de l'encre sont deux sources de dégradations majeures. Sur de nombreux documents, l'encre s'est éclaircie, ce qui la rend difficilement lisible. Sur d'autres documents, une trop grande quantité d'encre a été utilisée. Dans ces cas, le texte est difficilement lisible, car l'encre a bavé. De plus, les documents étant rédigés sur les deux faces, l'encre du verso est généralement visible en transparence. En outre, le papier est parfois tâché, déchiré ou troué. Par conséquent, certains actes sont incomplets, ou ne peuvent pas être lus correctement. Certaines de ces dégradations peuvent être observées sur la figure 1.12.



**(a)** Bavures d'encre
Cote 10 NUM 35154 199

**(b)** Encre effacée
Cote 10 NUM 35101 48

**(c)** Tâche d'encre
Cote 10 NUM 35154 81

**(d)** Bords déchirés
Cote 10 NUM 35260 137

**(e)** Bords tâchés
Cote 10 NUM 35192 154

**(f)** Ratures
Cote 10 NUM 35121 357

**Figure 1.12** – Principales dégradations observées sur les registres paroissiaux provenant des Archives Départementales d'Ille-Et-Vilaine.

### 1.2.6 Erreurs dans ces documents

Goubert [Goubert 1954] alerte sur les différentes erreurs qu'il est fréquent de rencontrer dans les registres paroissiaux. Dans son ouvrage, l'auteur insiste sur la nécessité de croiser les sources lors de l'analyse des informations contenues dans ces documents. À travers l'étude de registres paroissiaux issus de différents départements, l'auteur conclut

que les différentes ordonnances royales ont été suivies de façon très inégale selon les paroisses. En effet, certaines paroisses ont appliqué les ordonnances avec plusieurs années de retard, quand d'autres paroisses les ont simplement ignoré. L'auteur précise que peu de contrôles visaient à faire respecter ces directives. En second lieu, il rappelle que les prêtres sont maîtres de leurs registres, ainsi ceux-ci sont tenus avec plus ou moins de soin. La copie pour le bailliage est également une source d'erreur, car les prêtres n'effectuent pas nécessairement les copies eux-même. Par erreur, ils versent parfois le registre original, ou bien les deux exemplaires, au bailliage. L'auteur prévient également que pendant une grande partie du XVII$^{e}$ siècle, les curés inscrivaient principalement les sépultures des décès des notables de la paroisse. Les morts d'enfants, pourtant fréquentes, ne sont pas toutes enregistrées. Enfin, il rappelle que les étrangers et les personnes non catholiques sont exclues de ces registres. Ainsi, les statistiques qui reposent sur l'analyse de ces documents sont à interpréter avec précaution. Par ailleurs, de nombreux généalogistes constatent que l'orthographe des prénoms, des noms, ou des lieux sont approximatifs, car souvent basés sur la phonétique.

En plus de toutes ces imprécisions, il faut ajouter qu'un système de reconnaissance automatique risque d'ajouter une couche d'erreur supplémentaire aux informations extraites.

## 1.3 Les documents traités dans cette thèse

Dans cette section, nous définissons les objectifs de reconnaissance envisagés et présentons les bases de données que nous avons annotées dans le cadre de cette thèse. En effet, l'évaluation d'un prototype de reconnaissance automatique de documents nécessite la mise en place d'une base de données labellisée, c'est-à-dire des images de documents associées à des annotations sémantiques réalisées par humain, qui décrivent la mise en page du document et son contenu textuel. Le système peut alors être évalué en comparant les informations prédites avec la vérité établie par un annotateur humain. En outre, les méthodes récentes les plus performantes sont basées sur un mécanisme d'*apprentissage statistique* : elles ont donc besoin d'analyser un grand nombre d'exemples pour *apprendre* à reconnaître la structure des documents ou bien leur contenu textuel.

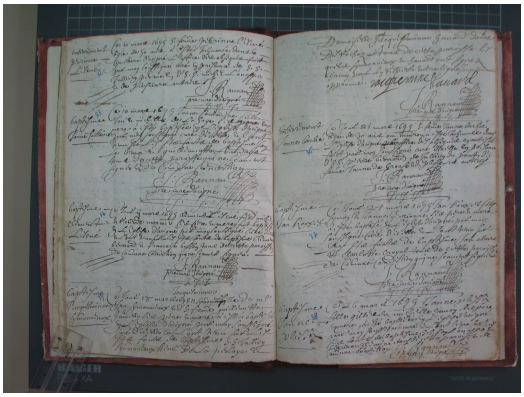
### 1.3.1 Objectifs de reconnaissance des registres paroissiaux

Nous définissons les objectifs de reconnaissance pour la mise en page et le contenu des registres paroissiaux. Pour cela, nous réalisons des annotations sur quelques images de documents afin de confectionner des bases de données pour le développement et l'évaluation de nos modèles de reconnaissance.

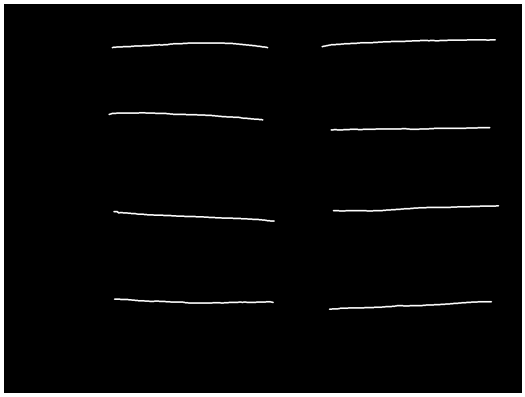
#### 1.3.1.1 Reconnaissance de la mise en page

Dans un premier temps, nous souhaitons reconnaître la structure logique des registres : localisation des pages, des actes, des lignes de texte. Nous avons établi un protocole d'annotation afin de représenter les informations qui nous paraissent intéressantes pour l'analyse de ces images. Les annotations réalisées sont illustrées dans la figure 1.13 et résumées dans la table 1.1. Elles ont été réalisées grâce à l'outil d'annotation VGG Annotator [Dutta 2019].

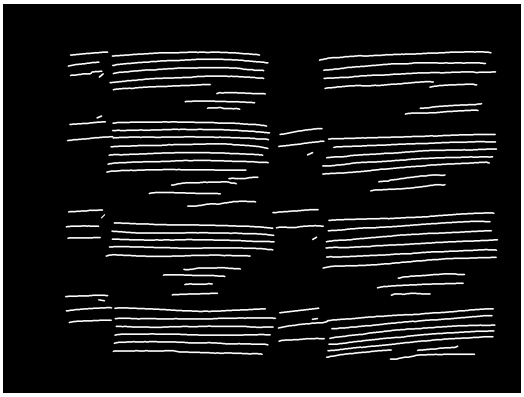
**Localisation des indices visuels (lignes de texte, signatures, pages)** La représentation des lignes de texte se fait souvent par un tracé des lignes de base du texte. La ligne de base désigne la ligne sur laquelle la plupart des lettres sont situées, au-dessous de laquelle s'étendent les jambages. Pour annoter ces lignes de base, nous avons tracé des lignes polygonales en suivant la courbure de chaque ligne. En outre, chaque ligne dessinée est associée à l'une des deux classes suivante : première ligne d'un acte, autre ligne. La connaissance des lignes de base permet de créer les images de lignes en calculant l'interligne et en redressant les lignes courbées.
La représentation des signatures peut être envisagée de différentes façons : polygone englobant, ligne de base, ou bien classification de chaque pixel de l'image. L'annotation la plus précise est celle à l'échelle des pixels, mais ce type d'annotation est extrêmement chronophage. Ainsi, nous proposons de tracer des polygones englobants autour des signatures, de façon à englober tous les pixels des signatures, en excluant dans la mesure du possible les pixels correspondant à du texte. Une pseudo-annotation au niveau pixel peut ensuite être obtenue en binarisant l'image originale, puis en appliquant le masque des polygones annotés.
Pour annoter les pages, nous avons dessiné les polygones englobants autour des bords des pages du registre.



**(a)** Image originale

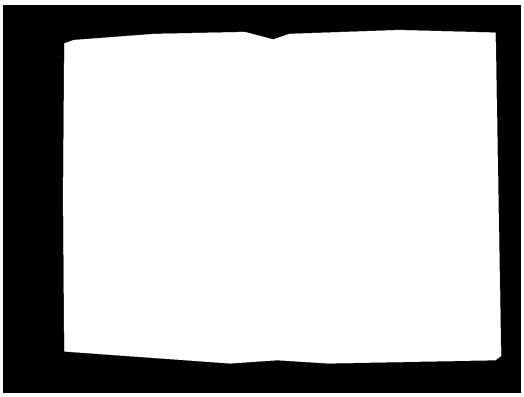


**(b)** Premières lignes de texte de chaque acte

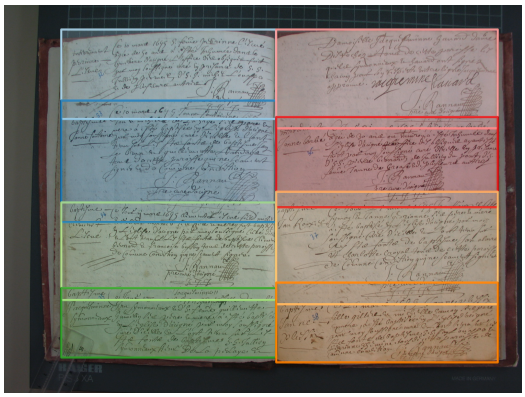


**(c)** Autres lignes de texte

**(d)** Signatures



**(e)** Bords de page



**(f)** Actes

**Figure 1.13** – Exemple d'annotation pour la mise en page. Les indices visuels sont segmentés et identifiés par des pixels blancs, alors que les actes sont identifiés par rectangles englobants qui peuvent se superposer.

**Localisation des actes** Pour délimiter les actes, nous proposons de dessiner les rectangles englobants des actes. Ces rectangles doivent englober tous les pixels correspondant au texte, en incluant les annotations marginales et les signatures. Il est à noter que ces rectangles peuvent se superposer en cas d'interaction texte/signature, ou bien en cas de courbure de l'écriture. Nous avons également choisi de normaliser les bordures des actes sur la largeur des pages.

**Table 1.1** – Synthèse des modes de représentation des objets de mise en page.

| Objet | Mode de représentation |
|---|---|
| Première ligne de chaque acte | Ligne polygonale |
| Ligne de texte | Ligne polygonale |
| Signatures | Polygone englobant + binarisation → pixels |
| Actes | Rectangle englobant |
| Pages | Polygone englobant |

#### 1.3.1.2 Objectifs de reconnaissance du contenu textuel

Dans un second temps, nous souhaitons reconnaître le texte associé à l'image de document, ainsi que les informations les plus intéressantes pour les généalogistes : noms, prénoms, dates ou lieux. Pour cela, nous avons effectué une transcription complète des actes. Nous avons également défini des tags associés à chaque mot transcrit, afin d'identifier son importance pour les généalogistes. L'annotation du texte a été réalisée grâce à un outil [8] développé par Ivan Leplumey et LabelStudio [9].

**Classification des actes** Un premier label permet de typer l'acte suivant le classement *fiançailles/promesses*, *publication/bans*, *mariage*, *baptême* et *sépulture*. Un second label permet de préciser si l'acte est complet, multiple ou tronqué. Un *acte multiple* correspond au cas où plusieurs événements sont décrits dans un seul acte (regroupement de bans de plusieurs mariages ou naissances multiples). Un *acte tronqué* désigne un acte qui tient sur plusieurs pages.

**Transcription du texte** L'étape de transcription consiste à transcrire le texte tel qu'il apparaît sur l'image, sans corriger les abréviations, la ponctuation ou l'orthographe. Les

8. Basé sur Python et Qt
9. `https://labelstud.io/`

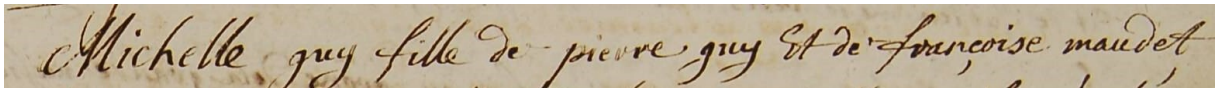
transcriptions sont découpées en trois champs : la marge, le texte principal et les signatures. Le retour à la ligne est symbolisé par le symbole « \n ». Les lettres qui ne peuvent pas être lues correctement (cachées par la pliure ou une tâche) sont annotées par le symbole « _ », et les mots qui ne peuvent pas être lus (illisibles, cachés) sont annotés par le symbole « [...] ». Les abréviations sont transcrites telles quelles, mais localisée par des crochets, afin d'enrichir un dictionnaire d'abréviations.

**Annotation d'entités nommées** L'annotation d'entités nommées consiste à associer à chaque mot transcrit un ou plusieurs labels. Nous avons identifié six catégories sémantiques à annoter dans ces documents : le *prénom*, le *nom*, l'*âge*, le *date*, le *lieu*, et le *métier*. Par ailleurs, nous souhaitons également associer à chaque mot une classe désignant le rôle de la personne à laquelle le mot se réfère : la personne principale, son père, sa mère, son conjoint, sa conjointe, son parrain et sa marraine. Les classes choisies sont volontairement simples et génériques, afin d'avoir suffisamment d'exemples dans chaque classe. Ces labels sont synthétisés dans les tables 1.2 et un exemple d'annotation est également présenté dans la figure 1.14

**Table 1.2** – Liste des catégories sémantiques (gauche) et des rôles des personnes (droite) annotés

| Catégorie | Symbole |
|---|---|
| Prénom | [name] |
| Nom | [surname] |
| Age | [age] |
| Date | [date] |
| Lieu | [loc] |
| Métier | [occ] |

| Personne | Symbole |
|---|---|
| Personne principale | [main-M] / [main-F] |
| Conjoint/conjointe | [spouse-M] / [spouse-F] |
| Père/mère | [parent-M] / [parent-F] |
| Parrain/marraine | [godp-M] / [godp-F] |



| *Michelle* | *guy* | *fille* | *de* | *pierre* | *guy* | *et* | *de* | *françoise* | *maudet* |
|---|---|---|---|---|---|---|---|---|---|
| [name] | [surname] | - | - | [name] | [surname] | - | - | [name] | [surname] |
| [main-F] | [main-F] | - | - | [parent-M] | [parent-M] | - | - | [parent-F] | [parent-F] |

**Figure 1.14** – Exemple d'annotation du texte et des entités nommées. Pour chaque mot, nous identifions sa catégorie sémantique ainsi que le rôle de la personne à laquelle il se rapporte.

### 1.3.2 Bases de données annotées

Nous présentons ici les bases d'images pour lesquelles nous avons réalisé des annotations manuelles.

#### 1.3.2.1 Mise en page

Deux bases d'images ont été annotées pour apprendre et évaluer un système de reconnaissance de structure de document.

**DLA-BMS-1** Les images ont été échantillonnées à partir de collections issues de 50 paroisses d'Ille-Et-Vilaine. Pour chaque paroisse, nous avons sélectionné quatre registres datant de quatre périodes distinctes : 1675, 1715, 1755, 1790. La septième page de chaque registre sélectionné a été extraite arbitrairement, afin d'éviter les pages de couvertures et les pages blanches. Au total, 200 images ont été échantillonnées, ce qui correspond à 1565 actes. Ces documents ont été numérisés entre 2003 à 2009 à l'aide de huit modèles d'appareil photo. En ce qui concerne la résolution des images, 167 images ont été numérisées à 180 dpi, 30 à 300 dpi, et 3 à une résolution inconnue. Pour chacune de ces 200 images, nous avons extrait les lignes de texte à l'aide du réseau ARU-Net [Grüning 2018], puis nous avons corrigé ces prédictions afin d'obtenir la vérité. Nous avons ensuite annoté les premières lignes de texte, les signatures, les pages et les actes.

**DLA-BMS-2** Les images de cette base sont issues des mêmes 50 paroisses d'Ille-Et-Vilaine. En revanche, l'échantillonnage est fait de manière à couvrir plus de périodes temporelles. La méthodologie de sélection est la suivante : pour chaque période de 25 ans, nous sélectionnons les paroisses disposant de documents de cette période. Pour ces paroisses, nous sélectionnons un registre datant de cette période, puis une page tirée aléatoirement dans le registre. Au total, nous disposons pour cette base de 3 images datant d'avant 1550, 11 pages datant de 1550 à 1575, 21 pages datant de 1575 à 1600, puis 25 images par période de 25 ans jusqu'à 1775. Au total 209 images ont été échantillonnées, ce qui correspond à 2143 actes. Ce lot d'images est plus hétérogène, car les actes étaient généralement plus compacts aux XVI$^{e}$ et XVII$^{è}$ siècles. Ces documents ont été numérisés entre 2003 à 2009 à l'aide de huit modèles d'appareil photo. En ce qui concerne la résolution des images, 173 images ont été numérisées à 180 dpi, 34 à 300 dpi, et 2 à une résolution inconnue. Pour ces images, nous avons annoté uniquement les actes.

#### 1.3.2.2 Contenu

Peu d'actes ont été transcrits ou annotés, car la lecture de ces documents est difficile et nécessite une certaine expertise.

**IE-BMS** Cette base contient quelques actes intégralement transcrits, dans lesquelles les entités nommées sont également annotées. Les transcriptions ont été réalisées par Ricarda Leroux et Ivan Leplumey. Ricarda Leroux nous a fourni des transcriptions réalisées dans le cadre de sa généalogie personnelle, à partir de documents issus de la Mayenne et de la Sarthe. Une partie de ses transcriptions (55 actes) a été corrigée afin qu'elles soient conformes aux textes inscrits sur les documents. Ivan Leplumey a quant à lui annoté 92 actes de la base DLA-1, en suivant le protocole de transcription défini ci-dessus. Nous avons ensuite annoté les entités nommées à partir de ces transcriptions. Finalement, 147 actes (721 lignes) ont été annotés. Il n'a pas été possible d'annoter plus de documents en raison du temps considérable que représente le processus de lecture et de transcription.

**HTR-paleo** Cette base contient 20 pages de documents issus de cours de paléographie réalisées par l'École des Chartres et les Archives Départementales du Tarn. Les documents datent du XVI$^{e}$ au XVIII$^{e}$ siècle. Ce sont principalement des documents administratifs, tels que des décrets royaux, des mises en paiement, des avis d'exécution, ou des annulations de mariage. Ils comportent 1180 lignes de texte transcrites. L'extraction des lignes de texte a été réalisée par un logiciel basé sur le réseau ARU-Net [Grüning 2018] et l'alignement des transcriptions sur les lignes extraites a été réalisé manuellement.

**INFO-CGE35** Des relevés d'informations, réalisés par Philippe Hervagault, nous ont été fournis par le Cercle Généalogique de l'Est de l'Ille-et-Vilaine (CGE35) [10]. Ils sont structurés sous la forme de fichiers CSV contenant des noms, prénoms, métiers et lieux issus de 15 paroisses. Les relevés contiennent des informations issues de 9619 actes de mariage, 1950 actes de sépulture et 2400 actes de baptême, soit environ 14000 actes issus de 15 paroisses de l'Ille-Et-Vilaine. Les informations contenues dans ces relevés incluent le type d'acte (baptême, mariage ou sépulture), le code postal de la commune ainsi que le permalien vers l'image originale. Pour un baptême, le relevé contient des informations sur la personne baptisée : nom, prénom, qualité du père, prénom du père, nom et prénom mère, parrain, marraine. Pour un mariage, le relevé contient des informations sur chacun

10. `https://www.cge35.fr/index.php`

des époux : nom, prénom, âge, prénom du père, prénom de la mère, nom de la mère. Il contient également la liste des témoins. Enfin, pour une sépulture, le relevé contient des informations sur la personne décédée : nom, prénom, qualité, âge, lieu-dit, prénom du père, prénom de la mère, nom de la mère, prénom du conjoint, nom du conjoint. Ces relevés sont associés à l'image originale, mais la transcription complète n'est pas disponible, ni la localisation de l'acte dans l'image. Ainsi, cette base n'est pas utilisable pour l'apprentissage d'un système de reconnaissance. Cependant, ces relevés sont utiles pour définir des dictionnaires de noms, prénoms ou lieux, ou pour évaluer notre système à grande échelle. Au total, 14011 relevés nous ont été fournis, dont 9632 correspondent à des actes de mariage, 1964 à des actes de sépulture, et 2415 à des actes de baptême.

## 1.4 Conclusion du chapitre

Dans ce chapitre, nous avons introduit le contexte applicatif de cette thèse. Nous avons présenté les registres paroissiaux, leur histoire, leurs caractéristiques et leur intérêt pour la recherche généalogique. Nous avons aussi illustré les difficultés propres à ces documents, en particulier les difficultés de lecture et d'interprétation des actes, la qualité de numérisation variable, ainsi que les dégradations fréquentes. Nous avons également présenté les objectifs de reconnaissance de cette thèse, notamment pour l'analyse automatique de la mise en page et du contenu textuel. Enfin, nous avons introduit les différentes bases de données mises en place et utilisées pendant cette thèse, dont une synthèse est présentée dans la table 1.3.

**Table 1.3** – Synthèse des bases de données annotées disponibles pour ce travail.

| Nom | Objectifs | Annotations | Documents | Actes | Lignes |
|---|---|---|---|---|---|
| DLA-BMS-1 | Mise en page | Actes + indices visuels | 200 | 1565 | - |
| DLA-BMS-2 | Mise en page | Actes | 209 | 2143 | - |
| IE-BMS | Contenu | Texte + entités nommées | - | 147 | 721 |
| HTR-PALEO | Contenu | Texte | - | - | 1180 |
| INFO-CGE35 | Contenu | Entités nommées | - | 14011 | - |

Ce travail de recherche se place dans un contexte industriel avec des contraintes fortes : les images que nous souhaitons reconnaître sont très hétérogènes, de résolution variable et difficiles à lire. Le processus d'annotation étant extrêmement couteux en temps, peu

d'annotations ont pu être réalisées. À terme, Doptim envisage de développer une plateforme d'annotation collaborative afin de récupérer un certain nombre de transcriptions. Cependant, le déploiement d'un tel outil est coûteux, et n'a pas pu être réalisé pendant ce projet. Notre objectif est donc de réaliser un prototype de reconnaissance, à partir de peu de données d'apprentissage. Le système pourra alors être enrichi et spécialisé avec les annotations et transcriptions issues de la plateforme collaborative.

Dans le prochain chapitre, nous dressons un état de l'art du domaine de la reconnaissance automatique de documents numérisés.

Chapitre 2

# État de l'art pour la reconnaissance de documents

Dans ce chapitre, nous introduisons la chaîne de traitement classique pour la reconnaissance automatique de documents imprimés ou manuscrits. Nous présentons une revue des articles les plus influents et les plus pertinents pour l'analyse de documents numérisés. Dans un premier temps, nous présentons les stratégies développées pour l'analyse de la mise en page de documents. Puis, nous détaillons les méthodes utilisées pour reconnaître le contenu textuel à partir d'images de zones de texte. Enfin, nous proposons un aperçu des principaux outils, bases de données et méthodes d'évaluation utilisés dans les nombreux articles et compétitions du domaine de recherche.

## 2.1 Introduction

La reconnaissance d'images de documents consiste à extraire automatiquement de l'information à partir de l'analyse de l'image d'un document numérisé.

Ce domaine de recherche a connu un essor important ces dernières années. D'une part, un intérêt croissant pour la numérisation des documents s'est développé dans les entreprises, les administrations et les institutions culturelles. Celles-ci ont donc cherché à développer des outils automatiques pour traiter rapidement un grand nombre de documents. D'autre part, les progrès scientifiques dans les domaines de l'apprentissage automatique et de la vision par ordinateur ont permis d'améliorer considérablement l'efficacité des systèmes d'analyse automatique de documents, ce qui a largement contribué au développement de ce domaine de recherche. Dans le même temps, de nombreuses bases de données ont été créées afin d'améliorer l'efficacité et la robustesse de ces systèmes [Marti 2002 ; Augustin 2006 ; Brunessaux 2014]. Cet engouement pour le traitement automatique de documents a renouvelé l'intérêt de la communauté scientifique pour de nombreuses thématiques de recherche, dont un aperçu est représenté sur la figure 2.1.

Une première thématique de recherche concerne la classification automatique de documents numérisés. L'enjeu de cette tâche est de parvenir à trier automatiquement des documents selon des critères définis, ou bien de regrouper les documents similaires. Cette thématique a de nombreuses applications industrielles, notamment le tri automatique du courrier et la classification de documents administratifs (documents d'identité, fiches de paie, factures, contrats...). Pour les archivistes, ces systèmes de classification sont également une aide précieuse pour retrouver la langue ou la période d'un manuscrit historique.

Une seconde thématique concerne l'analyse de la mise en page. En effet, la mise en page d'un document permet d'en apprendre plus sur son contexte, son type, sa date, ainsi que l'ordre de lecture. La reconnaissance de mise en page est un enjeu particulièrement important pour le traitement automatique de documents structurés, tels que les formulaires administratifs ou les documents d'identité. En effet, la localisation des champs est une étape préliminaire essentielle au traitement automatique de ces documents. Pour les documents non structurés, la localisation des zones de texte homogènes est également une étape clé nécessaire à la reconnaissance du texte.

Une troisième thématique majeure est la reconnaissance de texte ou de symboles contenus dans les documents. En effet, l'analyse de l'image peut permettre la reconnaissance du texte imprimé ou manuscrit, sur des documents modernes ou historiques. Plus généralement, la reconnaissance de symboles s'applique pour de nombreux types de documents. En particulier, la reconnaissance des chiffres a un intérêt pour le tri automatique de courrier, la reconnaissance de symboles musicaux pour l'analyse de partitions, ou encore la reconnaissance de symboles mathématiques pour la reconnaissance d'équations.

Une quatrième thématique majeure concerne la compréhension du texte contenu dans les documents. Cette thématique est à cheval entre la vision par ordinateur et le traitement du langage naturel. L'analyse plus approfondie du langage permet d'aller au-delà d'une simple reconnaissance de texte, en analysant le langage utilisé. En particulier, il est possible de corriger la transcription automatique en utilisant des caractéristiques statistiques de la langue, de reconnaître les mots appartenant des catégories sémantiques, de créer un résumé synthétique du document, ou encore une traduction automatique du texte.

Enfin, une dernière thématique concerne la sécurisation des documents. Ce domaine de recherche se concentre sur la vérification de l'authenticité de documents, et ce, par l'analyse forensique de l'encre, la vérification des signatures présentes, ou encore l'analyse du style linguistique de l'auteur.


**Figure 2.1** – Les différentes tâches du domaine de la reconnaissance automatique de documents

Dans cette thèse, nous nous intéressons à la reconnaissance de registres paroissiaux, qui sont des documents historiques, semi-structurés, complètement textuels et manuscrits. L'analyse automatique de ces documents passe par plusieurs étapes :

1. La numérisation du document pour obtenir une image numérique ;
2. L'analyse de la mise en page et la localisation des zones de texte ;
3. La reconnaissance du texte contenu dans ces zones d'intérêt ;
4. L'extraction des informations importantes.

Ainsi, nous concentrons cette analyse bibliographique sur les systèmes qui permettent l'analyse de la mise en page, la reconnaissance de texte manuscrit et l'extraction automatique d'informations pertinentes à partir d'images de documents. Dans la suite, nous présentons les méthodes de l'état de l'art les plus pertinentes pour chacune de ces tâches.

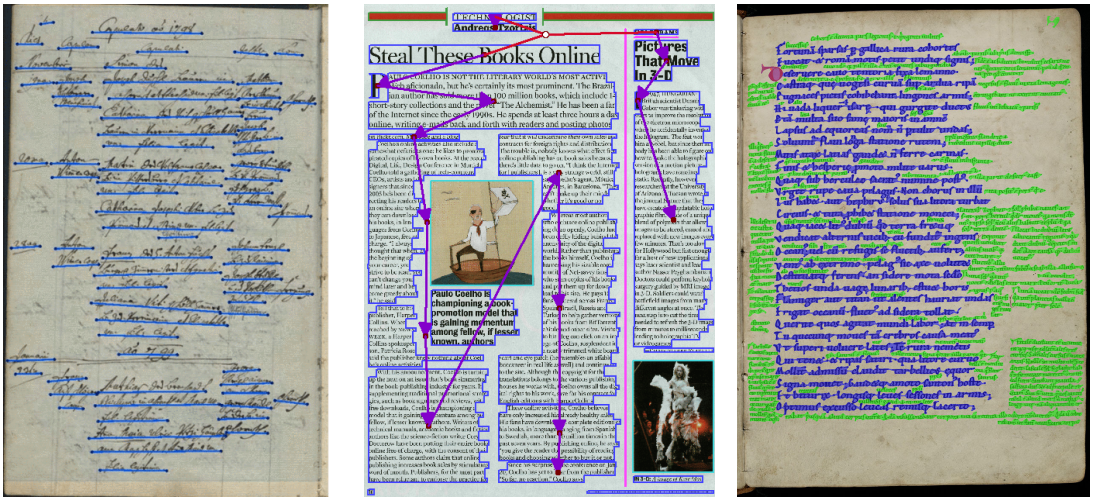
## 2.2 Reconnaissance de la mise en page

La reconnaissance automatique de structure de documents consiste à reconnaître l'organisation logique des pages. C'est une étape essentielle à la reconnaissance automatique de document. D'une part, comprendre l'organisation du document facilite son interprétation sémantique. En effet, la mise en page amène parfois des connaissances sur le contenu et le contexte du texte. C'est le cas des documents structurés tels que les factures, les formulaires, les articles de journaux, ou encore les courriers, pour lesquels la localisation du texte permet de déduire des informations sur son contexte. D'autre part, la localisation de blocs de texte est une étape essentielle pour appliquer des moteurs de reconnaissance d'écriture. En effet, ces moteurs prennent généralement en entrée des images de mots, de lignes, ou plus récemment de paragraphes. Ainsi, il est nécessaire de procéder à une analyse des pages afin d'isoler les zones de texte d'intérêt.

La grande variabilité des collections de documents, ainsi que la complexité de certaines mises en page font que le domaine de recherche de la reconnaissance de structure est encore très actif [Eskenazi 2016 ; Mehri 2015 ; Binmakhashen 2019]. Dans cette section, nous présentons les différentes tâches adressées par la communauté scientifique pour la reconnaissance de la mise en page. Puis, nous proposons un aperçu des méthodes qui ont été développées pour localiser des zones homogènes dans des documents. Enfin, nous analysons l'intérêt potentiel de ces approches pour la reconnaissance de structure des registres paroissiaux.

### 2.2.1 Tâches de localisation et de détection

Certains documents complexes sont composés de différents types d'objets graphiques, tels que de zones de texte, des structures tabulaires, ou encore des illustrations. Pour ce type de document, il est intéressant de localiser et de classifier chaque zone homogène, afin d'isoler les zones de texte. C'est notamment le cas pour les articles scientifiques, les affiches publicitaires, ou encore les bandes dessinées. À l'inverse, pour les documents principalement textuels, le travail d'analyse de la mise en page permet essentiellement de segmenter les zones de texte homogènes, en particulier les lignes de texte et les paragraphes. Dans cette section, nous présentons les tâches les plus populaires pour la reconnaissance de mise en page de documents. La figure 2.2 illustre trois tâches majeures pour la reconnaissance de structure de documents : la localisation de lignes de texte, la détection et classification de zones, et la segmentation de pages.




**(a)** Localisation des lignes de texte **(b)** Détection et classification de zones **(c)** Segmentation de page

**Figure 2.2** – Illustration des tâches adressées par les compétitions de reconnaissance de structure des documents. Les images sont issues des compétitions suivantes : a) cBAD2017 [Diem 2017], b) RDCL2019 [Clausner 2019], c) HBA2019 [Mehri 2019]

**Localisation de lignes de texte** Cette tâche consiste à localiser les pixels correspondants à la base de chaque ligne de texte, comme illustré sur la figure 2.2a. C'est une étape essentielle avant la reconnaissance automatique d'écriture, car les systèmes OCR prennent en entrée des images de lignes. La difficulté de ce problème vient de la variabilité des écritures, des langages et des mises en page. Ce problème occupe la communauté scientifique depuis de nombreuses années, en particulier dans les compétitions proposées lors des conférences [Murdock 2015 ; Diem 2017 ; Diem 2019b].

**Détection et classification de zones** Cette tâche consiste à localiser les différentes zones du document, et à attribuer une classe à chacune d'entre-elles. Ce type de tâche est particulièrement adapté aux documents complexes, car il est ainsi possible de séparer les zones textuelles des zones graphiques ou tabulaires. Par exemple, dans la figure 2.2b, chaque zone est délimitée par une boite englobante à laquelle est attribuée une classe parmi : table, texte, titre, figure. En outre, ces approches permettent de reconnaître l'ordre de lecture du document. De nombreuses compétitions ont permis de développer de nouvelles stratégies pour résoudre ce problème [Antonacopoulos 2015 ; Clausner 2017 ; Clausner 2019].

**Segmentation de page** Cette tâche consiste à attribuer une classe à chaque pixel de l'image. Par exemple, dans la figure 2.2c, les pixels bleus correspondent au texte principal, les pixels verts correspondent à des commentaires, les pixels rouges à des décorations, et les autres pixels appartiennent au fond du document. La segmentation de documents est particulièrement utiles pour le traitement de documents anciens [Simistira 2017 ; Clausner 2018 ; Mehri 2019].

### 2.2.2 Méthodologies pour la reconnaissance de mise en page

De nombreuses stratégies ont été proposées par la communauté scientifique pour résoudre ces tâches de segmentation ou de détection. Nous distinguons ici quatre grands types de stratégie d'analyse :

- **Stratégies ascendantes** (*Bottom-up* ou *data-driven*). Les algorithmes ascendants partent des éléments les plus fins de l'image et cherchent à regrouper ceux qui se ressemblent. Ils sont rapides et efficaces, mais très sensibles au bruit présent sur des documents à faible résolution.
- **Stratégies descendantes** (*Top-down* ou *model-driven*). Ces méthodes partent de la page entière et cherchent à la partitionner en zones homogènes. Elles sont généralement très rapides, mais ne conviennent pas pour des collections dont la mise en page est complexe ou variable.
- **Stratégies hybrides**. Ces stratégies combinent des analyses ascendantes et descendantes. Elles sont à la fois robustes et rapides et peuvent être utilisées sur une grande variété de documents, y compris complexes. En revanche, elles sont difficiles à mettre en place, car de nombreux paramètres doivent être ajustés.
- **Stratégies basées sur des réseaux de neurones**. Les réseaux de neurones permettent de segmenter efficacement des documents complexes, mais nécessitent des annotations qui sont coûteuses à obtenir. Ces approches arrivent en tête de la plupart des compétitions récentes d'analyse d'images de documents.

#### 2.2.2.1 Stratégies ascendantes

Ces algorithmes agglomèrent les éléments tels que les pixels, composantes connexes, lettres, mots ou lignes de texte pour créer des zones homogènes. Ils ont l'avantage de pouvoir s'appliquer à une grande variété de documents, sans nécessiter d'information a priori sur leur mise en page. En pratique, ils sont surtout appliqués à des documents

dont les zones sont simples ou clairement délimitées. En revanche, l'inconvénient de ces méthodes est qu'elles sont très sensibles au bruit, et donc à la qualité de numérisation. Trois types de stratégies ascendantes existent.

**Filtres et morphologies mathématiques** Les filtres morphologiques sont couramment utilisés en reconnaissance de formes. En effet, l'application de filtres peut permettre de mettre en lumière des zones d'intérêt de taille et de formes variables, en se basant sur des caractéristiques de texture, de contraste, de couleur ou d'intensité. Ces stratégies ont été utilisées pour localiser les zones graphiques et les zones de texte dans des documents contemporains Bockholt et al. [Bockholt 2011], Bukhari et al. [Bukhari 2011b] et Bukhari et al. [Bukhari 2011c]. Une autre approche proposée par Ferilli et al. [Ferilli 2009] permet également de localiser les zones de texte, et a été validée sur des documents variés à la mise en page complexe, ainsi que sur des documents nativement numériques. Finalement, les méthodes à bases de filtres sont également populaires pour la détection des lignes de texte manuscrites [Bukhari 2011c ; Lemaitre 2011 ; Tang 2014]. En pratique, les méthodes basées sur des filtres sont appliquées sur des documents modernes et de bonne qualité, qui ont souvent été binarisés en amont de l'analyse.

**Partitionnement** Les approches de partitionnement ou de *clustering* sont populaires pour la reconnaissance de mise en page. Elles consistent à grouper des éléments de l'image à partir de caractéristiques particulières, de façon à former des zones homogènes. Certains algorithmes se basent sur des caractéristiques génériques telles que l'orientation locale pour segmenter des lignes de texte [Kumar 2010 ; Ziaratban 2010]. Carel et al. [Carel 2015] identifient les composantes connexes puis effectuent un partitionnement à différentes résolutions sur chaque couche de couleur. D'autres approches se basent sur des caractéristiques issues de la géométrie du document. Certaines calculent des caractéristiques géométriques telles que la distance, l'aire, la densité avant d'effectuer un partitionnement pour obtenir les lignes de texte [Diem 2013 ; Yin 2009] ou bien effectuent une coloration de graphe pour obtenir des zones homogènes dans le document [Gaceb 2008]. Enfin, certains algorithmes se basent sur des caractéristiques issues de la texture du document, en particulier pour traiter des documents anciens [Mehri 2013b ; Mehri 2013a ; Journet 2008]. En particulier, Journet et al. [Journet 2008] proposent d'extraire cinq caractéristiques de texture locale à différentes résolutions et parviennent à identifier les zones principales des documents sans aucune connaissance a priori.

**Classification** Les algorithmes de classification sont également très populaires. Ils permettent de classer des éléments (pixels, mots, lignes...) à partir de caractéristiques apprises, de façon à former des zones homogènes (lettres, lignes, paragraphes...). Certains algorithmes partent d'une image binaire : Pinson et al. [Pinson 2011] et Peng et al. [Peng 2011] pour détecter des annotations manuelles sur des documents imprimés, Benjlaiel et al. [Benjlaiel 2014] pour l'extraction lignes de texte, Hebert et al. [Hebert 2011] et Fernández et al. [Fernández 2012] pour analyser la structure de documents anciens. D'autres partent d'une image en niveau de gris : Garz et al. [Garz 2011] parviennent à détecter des lettrines dans des documents anciens. Diem et al. [Diem 2011] tentent quant à eux de classifier chaque mot comme étant imprimé, manuscrit, ou simplement du bruit. Enfin, d'autres méthodes partent de l'image couleur : Baechler et al. [Baechler 2013] pour extraire des lignes de texte dans des manuscrits anciens, Chen et al. [Chen 2015a ; Chen 2014], Fischer et al. [Fischer 2014] et Wei et al. [Wei 2013] pour la reconnaissance de structure dans ces mêmes manuscrits.

#### 2.2.2.2 Stratégies descendantes

Ces stratégies sont basées sur une forte connaissance a priori sur la structure d'un type de documents dont la structure est invariante et codifiée (ex : formulaires, lettres). Elles ne sont pas robustes à une modification de la mise en page.

**Description de structure** La description de structure de document a été initialement proposée en 2006 par Coüasnon [Coüasnon 2006a] qui a développé dans la méthode DMOS un langage de description de documents : EPF. La description se fait sous forme de règles logiques et permet la segmentation et la classification des zones d'intérêt. La méthode DMOS a été appliquée avec succès à de nombreux types de documents structurés : lettres de correspondances, documents administratifs ou anciens, journaux. En 2008, Shafait et al. [Shafait 2008] ont proposé une autre approche descriptive qui nécessite une interaction avec l'utilisateur. Celui-ci définit un découpage approximatif de la mise en page des documents, puis le système renvoie les zones les plus probables de chaque document, à partir d'une analyse de l'image et de la structure a priori.

**Profils de projection** Ces méthodes sont principalement utilisées pour analyser les lignes de texte. Elles sont très efficaces sur des documents ne contenant que du texte, mais sont difficilement applicables à des documents complexes (dégradations, structure

complexe, éléments graphiques...). Ouwayed et al. [Ouwayed 2011] proposent d'effectuer un pavage de rectangle d'orientations différentes sur le texte pour permettre d'extraire des lignes penchées. La direction des rectangles est déterminée en utilisant la distribution de Wigner-Ville sur l'histogramme du profil de projection. Papavassiliou et al. [Papavassiliou 2010] utilisent l'algorithme de Viterbi pour trouver les lignes de texte en localisant les successions de texte et de blanc à l'intérieur de zones verticales. Ils parviennent également à segmenter le texte en mots en utilisant un SVM qui sépare les composantes connexes successives. Cette méthode a gagné la compétition ICDAR 2007 sur la segmentation de texte.

**Identification de séparateurs verticaux et horizontaux** Cette stratégie consiste à identifier les frontières physiques des zones d'intérêt ou les espaces blancs qui séparent des éléments textuels. À partir d'une image binaire, Louloudis et al. [Louloudis 2008] découpent les composantes connexes en paragraphes. Puis ils appliquent la transformation de Hough sur ces paragraphes pour obtenir les lignes de texte. Cette méthode ne fonctionne que sur des lignes de texte horizontales. Chen et al. [Chen 2013] proposent quant à eux d'analyser les espaces blancs pour séparer les colonnes dans des journaux anciens. Leur méthode a gagné les deux compétitions de segmentation à ICDAR 2013.

**Analyse de fonction** Ce type de méthode est basé sur l'optimisation d'une fonction spécialement conçue pour résoudre un problème donné. Bukhari et al. [Bukhari 2011a] proposent une méthode basée sur le principe des contours actifs, dans laquelle des contours ouverts (« snakes ») viennent délimiter les lignes de texte par le haut et le bas. La méthode est appliquée avec succès à des documents manuscrits sur des lignes de texte extrêmement courbées. Une autre méthode proposée par Ryu et al. [Ryu 2014] se base sur une fonction d'énergie conçue de telle manière que sa minimisation renvoie les lignes de texte. Un autre type de stratégie consiste à appliquer des modèles probabilistes : Yin et al. [Yin 2009] estiment le nombre de lignes de texte à l'aide d'un filtre flou puis appliquent une méthode bayésienne variationnelle pour segmenter les lignes de texte, tandis que Cruz et al. [Cruz 2014] proposent d'appliquer un mélange de gaussiennes aux différentes régions de la page pour obtenir la distribution logique du document.

#### 2.2.2.3 Stratégies hybrides

La plupart de ces méthodes combinent les avantages des deux types stratégies décrites précédemment. En revanche, elles sont complexes à implémenter, car de nombreux paramètres doivent être optimisés pour chaque type de document. Elles sont souvent appliquées à des documents contemporains très complexes (journaux, affiches publicitaires), mais quelques publications s'intéressent aussi à des documents manuscrits ou anciens.

En particulier, Bulacu et al. [Bulacu 2007] localisent les lignes de texte à l'aide d'un profil de projection, puis tirent profit de la mise en page uniforme de leur collection de documents anciens pour localiser les paragraphes. Une stratégie similaire est mise en place par Lemaitre et al. [Lemaitre 2011] pour localiser les mots dans des documents manuscrits : les distances entre les mots et entre les lettres sont calculées à l'aide d'un graphe Delaunay, et une connaissance a priori sur la structure du texte, notamment grâce à la ponctuation est utilisée.

Wei et al. [Wei 2014] proposent une sélection hybride de caractéristiques textuelles et s'intéressent à l'analyse de structure de manuscrits anciens. Barlas et al. [Barlas 2014] appliquent successivement une approche ascendante pour détecter les mots manuscrits et imprimés dans des formulaires administratifs, puis une approche descendante pour segmenter en blocs de texte. Clausner et al. [Clausner 2012] proposent de combiner une analyse des composantes connexes (ascendante) à des règles logiques (descendante) pour obtenir les lignes de texte. Ils montrent que leur approche hybride surpasse l'approche uniquement ascendante ou descendante. Finalement, Asi et al. [Asi 2015] s'intéressent à des documents anciens contenant des lignes de texte très courbées, et parviennent à extraire les lignes de texte grâce à une approche hybride, puis à les lisser en utilisant une transformation géométrique non linéaire.

#### 2.2.2.4 Stratégies basées sur des réseaux de neurones

Récemment, les réseaux de neurones, et en particulier les réseaux convolutifs, se sont révélés être particulièrement efficaces pour la segmentation de page et la détection d'objets dans des documents numérisés, qu'ils soient récents ou anciens [Diem 2019b ; Mehri 2019].

Les réseaux de neurones récurrents ont notamment permis une avancée majeure pour la localisation de lignes de texte. Moysset et al. [Moysset 2015] ont été parmi les premiers à développer un réseau récurrent pour localiser les lignes de texte dans les documents hétérogènes de la base de données Maurdor [Brunessaux 2014]. Ce réseau apprend à

modéliser les paragraphes comme une séquence de lignes de texte et d'interlignes. Un autre modèle proposé par la société A2IA [Murdock 2015] est basé sur un réseau convolutif couplé à des couches LSTM a atteint la seconde place de la compétition ICDAR 2015 pour la détection de lignes de texte.

Plus récemment, les réseaux complètement convolutifs ont gagné en popularité pour la détection de lignes de texte et la segmentation de pages [Diem 2019b ; Mehri 2019]. De nombreuses architectures ont été proposées pour localiser les lignes de texte dans des documents manuscrits. Grüning et al. [Grüning 2018] ont proposé ARU-Net, un U-net avec des couches récurrentes et couplé à un réseau d'attention, qui obtient les meilleurs résultats pour la détection des lignes de texte sur la base de données cBAD [Diem 2019b]. Renton et al. [Renton 2018] ont également proposé un réseau entièrement convolutif avec des convolutions dilatées, ce qui leur permet d'obtenir des résultats compétitifs pour la détection de lignes de texte. Enfin, Oliveira et al. [Oliveira 2018] ont proposé dhSegment, un U-Net dont la partie contractante est composée d'un ResNet-50. Les auteurs ont démontré la capacité de dhSegment à traiter différentes tâches sur des documents anciens, en particulier la localisation de lignes de texte, la détection d'ornements, et la segmentation de page. De nombreuses autres architectures basées sur des réseaux convolutifs ont été proposées par la suite [Renton 2017 ; Quirós 2018 ; Boillet 2021a].

D'autres approches s'intéressent à la segmentation sémantique de page [Chen 2015b ; Chen 2017a ; Xu 2017] Alberti et al. [Alberti 2019] ont utilisé le framework DeepDIVA [Alberti 2018] pour annoter précisément les documents afin d'obtenir une segmentation sémantique de qualité pour extraire les lignes de texte. Une autre approche basée sur des réseaux siamois a été proposée par Alaasam et al. [Alaasam 2019], pour effectuer la classification de chaque pixel dans des manuscrits arabes anciens.

Enfin, les réseaux de détection d'objets sont régulièrement utilisés pour la détection et classification de zones dans des documents récents. De nombreux articles utilisent des réseaux de neurones pour localiser des blocs de texte, des équations, des illustrations, ou des tableaux dans des documents imprimés complexes [Saha 2019 ; Yi 2017 ; Oliveira 2017]. Plus récemment, ces réseaux ont été appliqués pour localiser différentes instances d'objets dans des documents anciens [Prusty 2019 ; Biswas 2021a]. Mais si ces réseaux sont capables de reconnaître des mises en page variées avec une grande précision, ils nécessitent un grand nombre de documents d'apprentissage ($> 1000$ images de documents), ont des capacités de généralisation limitées.

### 2.2.3 Logiciels existants

Certains logiciels permettent de reconnaître la structure des documents. Par exemple, dhSegment [Oliveira 2018] et ARU-Net [Grüning 2018] sont deux logiciels open-source permettant la localisation de lignes de texte dans des images. dhSegment permet également la segmentation sémantique des documents, la localisation des pages et des illustrations, ainsi que la détection d'ornements. Deux autres outils propriétaires permettent de reconnaître la structure des documents : Transkribus [1] et Arkindex [2]. Ces systèmes parviennent à localiser correctement les lignes de texte sur les registres paroissiaux. En revanche, la localisation des actes n'est pas satisfaisante. De plus, leur utilisation est payante.

### 2.2.4 Discussion

Si un nombre considérable d'approches a été proposé par la communauté scientifique au fil des années, peu d'entre elles s'adaptent à différents types de mise en page. Aucune des méthodes présentées dans cette section n'est directement capable de reconnaître la structure particulière des registres paroissiaux.

Les méthodes ascendantes ne sont pas envisageables, car elles nécessitent une mise en page clairement apparente, et des documents de bonne qualité. Or, nous avons montré dans le chapitre 1 que ce n'est pas le cas des registres paroissiaux. Leur structure n'apparait pas clairement : le texte est serré, avec peu d'espace interligne entre deux actes successifs. De plus, les lignes de texte sont souvent courbées, avec une interaction fréquente entre les signatures en fin d'acte et texte de l'acte suivant. Enfin, la qualité de la numérisation est variable, et certaines pages sont fortement dégradées.

Les méthodes descendantes ne sont pas non plus applicables, car la structure des pages n'est pas homogène. Les documents sont photographiés en simple ou double page et chaque page contient un nombre variable d'actes, avec des styles d'écritures très hétérogènes. Certains indices visuels permettent de repérer la localisation des actes, en particulier les annotations marginales, les signatures en fin d'acte, les alinéas ou majuscules en début d'acte, ou encore l'espacement vertical entre les actes. Cependant, ces indices ne sont pas présents dans tous les documents.

Mais d'autres stratégies présentées dans cet état de l'art semblent applicables aux registres paroissiaux. En particulier, les méthodes neuronales, elles ont prouvé être capables

1. `https://readcoop.eu/transkribus/`
2. `https://teklia.com/solutions/arkindex/`

de traiter efficacement des documents dégradés, complexes, ou hétérogènes. Elles sont capables de reconnaître une grande variété de mises en page à partir de l'apprentissage de caractéristiques pertinentes. En particulier, les réseaux encodeur-décodeur convolutifs (U-Net) sont désormais capables de localiser les lignes de texte, les pages ou encore les décorations dans des documents historiques et manuscrits [Grüning 2018 ; Oliveira 2018]. D'autres types de réseaux, les réseaux de détection d'objets, permettent de localiser des instances d'une même classe, même quand celles-ci se chevauchent. Ce type de réseau est principalement utilisé pour localiser des zones de texte, des tableaux, des équations ou des figures dans des documents imprimés [Oliveira 2017 ; Yi 2017 ; Saha 2019], mais plus récemment, ils ont été appliqués avec succès à des documents anciens [Prusty 2019 ; Trivedi 2021 ; Biswas 2021a]. La limite majeure des approches neuronales est qu'elles nécessitent un grand nombre d'exemples, ou *données d'apprentissage*, pour apprendre à extraire des caractéristiques pertinentes pour la tâche à résoudre. Or, nous disposons de très peu d'annotations sur des registres paroissiaux.

Les méthodes hybrides peuvent être un compromis intéressant pour dépasser ce problème. En effet, nous avons vu que la localisation des actes ne peut pas se baser entièrement sur la présence d'indices visuels, car ceux-ci n'apparaissent pas systématiquement dans les pages. En revanche, lorsque ceux-ci apparaissent, ils peuvent permettent de guider l'analyse. La combinaison de plusieurs indices visuels (lignes de texte, signatures, annotations marginales, mots clés) pourrait également fiabiliser cette analyse. En particulier, la localisation des indices visuels par des réseaux de neurones, couplée à des règles logiques de construction d'actes, pourraient permettre d'obtenir un système robuste de localisation d'actes. Un autre avantage majeur est la stabilité des indices visuels. En effet, les motifs stables peuvent être reconnus à partir de peu d'exemples, ce qui limite les problématiques de généralisation des réseaux de neurones.

Le chapitre 3 aborde le problème de la reconnaissance de structure des registres paroissiaux. Nous y présentons les deux types d'approches envisagées pour la reconnaissance de pages de registres paroissiaux : une approche complètement neuronale et une approche hybride.

## 2.3 Reconnaissance de texte manuscrit

La reconnaissance de contenu consiste à reconnaître et comprendre le texte présent dans un document numérisé. De nombreux systèmes grand public de reconnaissance op-

tique de caractères (OCR [3]), tels que Tesseract [4], ABBY Finereader [5], ou Google Cloud Vision [6] sont actuellement capables de reconnaître le texte de certains documents numérisés. Les OCR représentent un enjeu industriel important pour différents domaines, notamment pour le tri du courrier, la reconnaissance chèque, la reconnaissance de papiers d'identité, le traitement de formulaires administratifs, ou la reconnaissance de documents d'archive [Plamondon 2000 ; Plötz 2009]. Cependant, la performance de ces systèmes chute considérablement lorsqu'ils sont appliqué à des documents à faible résolution, dégradés, ou manuscrits [Romero 2013]. En conséquence, la tâche de la reconnaissance de texte est toujours un problème ouvert et un sujet de recherche actif. Les travaux de recherche actuels se focalisent sur la robustesse de tels systèmes sur des documents manuscrits, dégradés, ou anciens.

Dans la suite, nous considérons que la mise en page du document à reconnaître a été simplifiée, grâce aux méthodes décrites dans la section précédente. En effet, les systèmes de reconnaissance d'écriture prennent généralement en entrée des images de mots, de lignes, ou de paragraphes. Nous détaillons le fonctionnement des systèmes de reconnaissance, avec un accent sur le script latin, l'écriture manuscrite et les documents historiques.

### 2.3.1 Méthodologies pour la reconnaissance de texte

La reconnaissance de contenu à partir d'une image de texte manuscrit ou imprimé est complexe. En effet, la reconnaissance s'appuie sur la modélisation du script, en se basant sur des caractéristiques issues des images d'apprentissage (modélisation de la forme des lettres, apparence de l'écriture). Les différents styles d'écriture, propres à chaque scripteur, augmentent considérablement la variabilité interclasse de chaque lettre : une même lettre s'écrit de multiples façons. Mais les performances de ces systèmes peuvent être améliorés par une modélisation linguistique, en utilisant des caractéristiques séquentielles sur l'ordre d'apparition des caractères dans une certaine langue ou dans un vocabulaire fermé.

#### 2.3.1.1 Prétraitement

Plusieurs prétraitements sont couramment appliqués à l'image avant l'application de moteurs de reconnaissance d'écriture. Ces traitements ont pour but de simplifier l'image

3. *Optical Character Recognition*
4. `https://github.com/tesseract-ocr/tesseract`
5. `https://pdf.abbyy.com/fr/`
6. `https://cloud.google.com/vision`

d'entrée, ou bien d'en augmenter la qualité. Certains algorithmes nécessitent une binarisation ou une augmentation conséquente du contraste, afin de distinguer les pixels du signal de ceux du fond [Plamondon 2000 ; Bluche 2013]. Ensuite, il est fréquent de corriger les inclinaisons verticale et horizontale du texte, ainsi que l'alignement entre les mots [Bluche 2013 ; LeCun 1995 ; España-Boquera 2011]. Dans certains cas, la hauteur des caractères est normalisée en fixant la taille des ascendants et des descendants [España-Boquera 2011]. Dans la plupart des systèmes, l'image est également normalisée et débruitée [Vinciarelli 2002 ; Plamondon 2000]. Enfin, la parallélisation du traitement des images nécessite souvent un redimensionnement de la taille des images vers une taille fixe. L'utilisation du padding est également fréquente afin de maintenir le ratio de forme initial de l'image [Bluche 2013 ; LeCun 1995].

#### 2.3.1.2 Segmentation et classification des caractères ou mots

Les premières méthodes de reconnaissance d'écriture consistaient à localiser les mots ou les caractères présents sur l'image, et à les classifier. Cette stratégie peut donc s'appliquer à deux niveaux :

— À l'échelle des *mots*, elle nécessite la localisation des mots dans l'image, puis permet de reconnaître des mots issus d'un vocabulaire réduit, en classifiant chaque image de mots à partir de caractéristiques. Les classes à reconnaître correspondent aux mots issus d'un dictionnaire réduit.

— À l'échelle des *caractères*, elle nécessite la localisation des caractères dans l'image, puis permet la classification de chaque caractère à partir de caractéristiques. Les classes à reconnaître sont alors les lettres de l'alphabet, avec éventuellement des chiffres ou caractères spéciaux.

**Segmentation des mots et des caractères** La segmentation des mots ou des caractères imprimés peut être réalisée par l'analyse de pixels blancs, l'analyse de projection, ou encore l'analyse des composantes connexes [Casey 1996]. En effet, les lettres et les mots sont séparés de leurs voisins par un espace de taille fixe. La segmentation des mots manuscrits est également possible par ces techniques, bien que l'espace inter-mots varie selon les scripteurs. La segmentation des caractères manuscrits est quant à elle bien plus complexe, en particulier pour l'écriture cursive. Certains algorithmes parviennent à segmenter les caractères par à une analyse des contours, en localisant les ligatures entre les lettres, ou en détectant les ascendants et descendants [Casey 1996]. D'autres se basent

sur la programmation dynamique pour chercher la segmentation la plus optimale [Wang 1994]. Cependant, la localisation des caractères cursifs peut introduire des erreurs, avec des caractères fréquemment fusionnés ou séparés en deux. Certaines heuristiques ont été proposées pour corriger ces erreurs, soit en se basant sur des caractéristiques linguistiques, soit en utilisant le résultat de la classification pour valider ou non la segmentation [Britto 2001].

**Extraction des caractéristiques** La phase d'extraction de caractéristiques consiste à représenter l'image sous la forme d'un vecteur caractéristique. Ce vecteur doit être construit de façon à concentrer les informations qui sont discriminantes pour la phase de classification. Cette étape est donc cruciale pour assurer des performances de classification optimales. De nombreuses stratégies d'extraction de caractéristiques ont été développées pour l'analyse d'image de caractères et de mots. La *sélection manuelle* des caractéristiques consiste à définir une combinaison linéaire de plusieurs descripteurs de l'images, qui peuvent être locaux ou globaux [Kumar 2014]. Les caractéristiques globales représentent l'image dans son ensemble, alors que les caractéristiques locales sont extraites à partir de zones dans l'image, le plus souvent à l'aide d'une grille [Camastra 2007]. Les caractéristiques statistiques permettent de résumer les informations issues de la distribution statistique du signal, comme la distance entre les pixels d'une certaine intensité, le nombre de pixels du premier plan, le nombre de transitions entre pixels blancs et noirs sur une même colonne ou ligne [Camastra 2007 ; Plötz 2009 ; Bluche 2013]. Des caractéristiques géométriques peuvent être extraites. On peut citer par exemple les caractéristiques issues de la courbure, des contours, des projections sur les axes x et y [Bluche 2013]. D'autres transformations plus complexes permettent également d'obtenir des informations sur l'image. En particulier, les transformations de Fourier, de Hough, de Gabor, ou de Sobel sont fréquemment utilisées pour l'extraction de caractéristiques dans des images, ainsi que les descripteurs SIFT, SURF ou HoG [Das 2012 ; Dhaka 2015]. Enfin, le vecteur final peut être optimisé à l'aide de transformations analytiques, telles que l'analyse linéaire discriminante (LDA [7]) et l'analyse en composantes principales (PCA [8]) [Nopsuwanchai 2003 ; Plötz 2009]. La sélection manuelle de caractéristiques pertinentes est extrêmement longue et complexe à réaliser. En effet, celles-ci doivent être suffisamment variées pour représenter l'image dans le cadre de la tâche à résoudre. Mais l'utilisation

7. Pour *Linear Discriminant Analysis* en anglais
8. Pour *Principle Component Analysis* en anglais

d'un trop grand nombre de caractéristiques peut faire chuter les performances du modèle de classification [Kumar 2014].
Pour dépasser ces limites, certains travaux se sont focalisés sur la *sélection automatique* de caractéristiques. LeCun et al. [LeCun 1995] ont été parmi les premiers à proposer un réseau convolutif (CNN) capable d'apprendre une représentation vectorielle caractéristiques de l'image pour la classification d'images de chiffres. L'avantage de cette approche est que le réseau apprend à sélectionner les caractéristiques les plus utiles pour discriminer les différentes classes. En revanche, elle nécessite un grand nombre d'images labellisées, c'est-à-dire dont la classe est connue, car les paramètres du réseau sont appris à partir d'exemples. Une dernière contribution intéressante a été proposée par Masci et al. [Masci 2011] : les réseaux auto-encoders. Ces réseaux sont non supervisés, c'est-à-dire qu'ils sont capables d'extraire des caractéristiques discriminantes sans labels associés aux images. Un *encodeur* convolutif apprend une représentation vectorielle de l'image, puis un *decodeur* convolutif reconstruit l'image initiale à partir du vecteur caractéristique. Une fois l'apprentissage réalisé, l'*encodeur* permet d'obtenir une représentation vectorielle à la fois compacte et discriminante de l'image initiale. Dans les méthodes les plus récentes, l'extraction des caractéristiques dans des images est majoritairement effectuée à l'aide de réseaux convolutifs. De nombreuses bases de données ont également été proposées [LeCun 2010 ; Deng 2009 ; Lin 2014] pour apprendre les modèles à extraire des caractéristiques générales à partir d'images.

**Méthodes de classification** Les vecteurs caractéristiques obtenus servent à alimenter les modèles de classification. Ces modèles appartiennent à la famille des *algorithmes d'apprentissage supervisés*, car ils apprennent à estimer la classe associée à chaque vecteur à partir d'exemples labellisés. De nombreux algorithmes d'apprentissage pour la classification supervisée ont été proposés [Caruana 2006] par la communauté scientifique. Les méthodes les plus couramment utilisées pour la classification de mots ou de caractères sont la méthode des plus proches voisins [LeCun 1995] les machines à vecteurs de support (SVM [9]) [Camastra 2007 ; Nasien 2010 ; Ayyaz 2016], les modèles de Markov cachés (HMM [10]) [Britto 2001 ; Nopsuwanchai 2003 ; Das 2012], ou encore les réseaux de neurones complètement connectés (perceptrons) [LeCun 1995].

9. Pour *Suport Vector Machine* en anglais
10. Pour *Hidden Markov Models* en anglais

**Limites de ces approches** L'approche à l'échelle des *mots*, qui consiste à segmenter puis classer directement un mot dans un vocabulaire réduit, ne permet pas de traiter des mots hors du vocabulaire appris. Elle est aussi vulnérable à la variabilité des longueurs des mots, car des mots de taille variable doivent être représentés par un vecteur de taille fixe. Ainsi, cette approche est loin d'être satisfaisante. L'approche à l'échelle des *caractères* consiste à segmenter puis classifier les caractères. Contrairement à l'approche basée sur les *mots*, cette approche permet de reconnaître des mots hors du vocabulaire d'apprentissage. En revanche, la segmentation en caractères pose problème dans un contexte industriel. En effet, une mauvaise segmentation est une source d'erreur sur la reconnaissance finale, comme illustré dans la figure 2.3. Or, la segmentation en caractères est difficile et coûteuse à réaliser, car elle nécessite une optimisation manuelle qui dépend du style d'écriture. En outre, cette approche est difficilement applicable aux écritures manuscrites cursives, dans lesquelles les tracés et caractères sont liés.


**Figure 2.3** – Ambiguïtés liées à la segmentation en caractères : différentes façons de segmenter le mot ont un impact sur la reconnaissance (figure issue de [Plötz 2009])

#### 2.3.1.3 Reconnaissance séquentielle des caractères

Pour dépasser ces limites, les méthodes les plus récentes s'orientent vers des approches qui ne nécessitent pas de segmentation des caractères. En effet, ces méthodes effectuent une segmentation implicite pour permettre de reconnaître une séquence de caractères issue d'images de mots ou de lignes de texte. Dans un premier temps, les modèles de Markovs Cachés ont été utilisés avec un mécanisme de fenêtre glissante pour reconnaître une séquence de caractères. Plus récemment, les approches basées sur des réseaux de neurones profonds ont connu un essor considérable, car ceux-ci apprennent à extraire les caractéristiques nécessaires à la reconnaissance. De plus, certaines architectures permettent de modéliser un modèle de langue implicite.

**Modèles de Markov Cachés (HMM)** Les HMM sont des méthodes statistiques utilisées pour la modélisation de séquence, avec des applications dans les domaines de

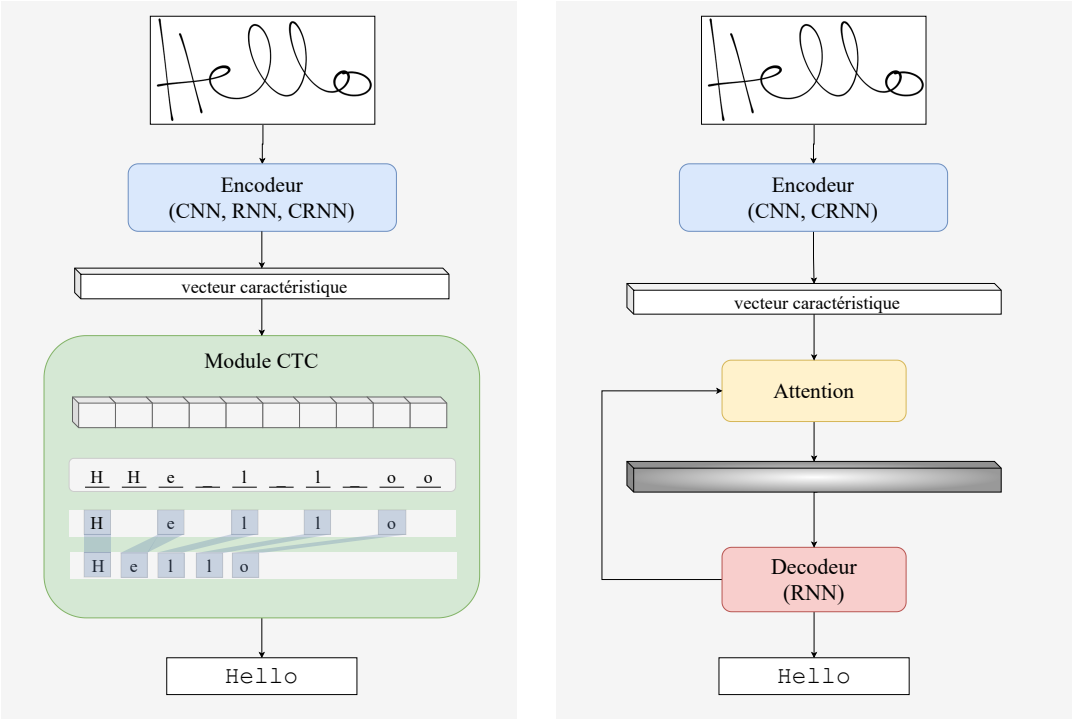
la reconnaissance de la parole, ou de la reconnaissance de formes [Fink 2007]. Ils ont été largement utilisés pour la reconnaissance d'écriture sans segmentation explicite, en se basant sur une fenêtre glissante [Vinciarelli 2002 ; Plötz 2009 ; Kaltenmeier 1993 ; Schwartz 1996]. La fenêtre glissante est appliquée sur l'image de gauche à droite pour suivre le sens de l'écriture. Sur chaque zone de la fenêtre, des caractéristiques sont extraites grâce aux techniques détaillées plus haut. En plus des caractéristiques usuelles, certains ajoutent des caractéristiques heuristiques spécifiques à l'écriture manuscrites, telles les ligatures entre les lettres, les ascendants, les descendants, les boucles, ou encore les blancs entre les mots [Plötz 2009 ; Vinciarelli 2002]. Une fois le modèle appris, la phase de décodage permet d'obtenir la séquence de caractères prédite à partir de l'image initiale. L'utilisation de modèles de langue statistiques a permis une première amélioration des systèmes de reconnaissance. Ces modèles permettent d'estimer des distributions de probabilité sur des séquences de lettres ou de mots. Bazzi et al. [Bazzi 1999] ont été les premiers à utiliser ce type d'approche pour corriger la séquence produite par le HMM. Mais les modèles de langue sont encore plus efficaces lorsqu'ils sont directement intégrés dans la fonction de décodage du HMM [Marti 2001 ; Toselli 2004 ; Vinciarelli 2004]. Enfin, l'hybridation des modèles de Markov cachés et des réseaux de neurones [Bluche 2013 ; España-Boquera 2011] a également permis d'augmenter l'efficacité des systèmes de reconnaissance.

**Réseaux de neurones et CTC** La fonction de coût Classification Temporelle Connexionniste (CTC) [Graves 2006] a initié une nouvelle ère dans le domaine de la reconnaissance d'écriture. L'utilisation de la fonction CTC permet aux réseaux de neurones récurrents de générer une séquence textuelle à partir d'une image de mots ou de ligne de texte, sans nécessiter de segmentation explicite des caractères. Une première architecture basée sur un réseau récurrent (RNN), appris avec la fonction de coût CTC, a été proposée par Graves et al. [Graves 2009a ; Graves 2009b]. Cette architecture est souvent appelée RNN-CTC. Le réseau récurrent, composé de couches Long Short-Term Memory (LSTM), prend en entrée des caractéristiques de l'image [Graves 2009a], ou directement les pixels de l'image [Graves 2009b]. Ce nouveau modèle a permis une réduction de 40% du taux d'erreur, comparé aux approches basées sur les HMM. Menasri et al. [Menasri 2012] et Moysset et al. [Moysset 2014] ont ensuite introduit des architectures CRNN-CTC, en alternant des couches récurrentes et des couches de convolution. L'architecture CRNN-CTC a ensuite été massivement reprise et améliorée, en explorant de nouvelles couches de convolution [Bluche 2017 ; Dutta 2018], ou des couches de récurrences sur différentes

dimensions [Puigcerver 2017]. Plus récemment, certains travaux ont cherché à se passer des couches de récurrence [Yousef 2018 ; Ptucha 2019 ; Coquenet 2020], car celles-ci rallongent considérablement le temps de calcul. Ces études ont montré qu'une architecture CNN-CTC peut être aussi performante que CRNN-CTC, mais bien plus rapide. La dernière grande avancée s'est faite avec l'intégration d'un modèle de langue à ces architectures [Liu 2015a ; Liu 2017]. D'autres contributions ont été proposées pour améliorer la partie récurrente de l'architecture, notamment avec l'ajout de dropout comme méthode de régularisation [Pham 2013] et l'intégration des couches Gated Recurrent Units (GRU) [Chen 2017b]. Enfin, certains travaux réseaux ont amélioré cette approche pour permettre la reconnaissance de lignes de texte courbées [Wan 2019 ; Wigington 2018a] ou la reconnaissance de paragraphes [Wigington 2018a ; Coquenet 2021].

**Réseaux de neurones avec mécanisme d'attention** Plus récemment, les réseaux Séquence à séquence (seq2seq) avec mécanisme d'attention ont instauré un nouvel état-de-l'art pour des domaines connexes : la traduction automatique [Bahdanau 2015], la reconnaissance de la parole [Chorowski 2015a], et le sous-titrage d'images [Xu 2016]. Ces réseaux ont donc naturellement été adaptés pour la reconnaissance de texte [Chowdhury 2018 ; Michael 2019]. L'architecture est composée d'un réseau convolutif ou convolutif récurrent, appelé *encodeur*, qui extrait les caractéristiques images, et d'un réseau récurrent, appelé *decodeur*, qui prédit la séquence textuelle caractère par caractère. Un réseau d'*attention* apprend un alignement entre les pixels de l'image et la séquence de texte correspondante. Ainsi, le vecteur de caractéristiques est pondéré par le réseau d'*attention*, afin de faciliter la tâche du *decodeur* qui peut se concentrer sur une zone pertinente du vecteur de caractéristiques pour prédire chaque caractère. Une autre force de cette architecture vient du *decodeur*, qui permet d'apprendre un modèle de langue implicite à partir des transcriptions d'apprentissage. Enfin, cette architecture est également modulable. En particulier, il est possible de connecter un *encodeur* à plusieurs décodeurs, ce qui permet de traiter des tâches multiples [Luong 2015a]. Diaz et al. [Diaz 2021] ont également mené une étude approfondie des différentes combinaisons d'*encodeur* et *décodeur*. Leur analyse met en lumière l'intérêt du mécanisme d'auto-attention, utilisé dans les Transformers, associée à un *décodeur* CTC et à un modèle de langue explicite. Cette architecture Transformer, proposée par Vaswani et al. [Vaswani 2017], repose sur un mécanisme d'attention sans aucune couche de récurrence, et a récemment établi un nouvel état de l'art dans les domaines du traitement du langage naturel et la reconnaissance de la parole. L'avantage

de ces réseaux est qu'ils permettent de paralléliser tous les calculs durant l'apprentissage. Quelques tentatives de reconnaissance d'écriture ont été faites avec cette architecture [Kang 2020a ; Diaz 2021], mais l'architecture nécessite un très grand nombre d'exemples d'apprentissage pour dévoiler tout son potentiel.




**(a)** Architecture CRNN-CTC.

**(b)** Architecture Encodeur-Decodeur avec mécanisme d'attention.

**Figure 2.4** – Illustration des réseaux de neurones pour la reconnaissance d'écriture. Dans ces deux architectures, la phase d'extraction de caractéristiques est généralement faite avec un encodeur CRNN. Dans l'architecture Encodeur-CTC, le vecteur de caractéristiques est décodé grâce au CTC, alors que dans l'architecture Encodeur-Decodeur, le vecteur de caractéristiques est décodé grâce à un réseau récurrent couplé à un réseau d'attention.

### 2.3.2 Logiciels existants

De nombreux logiciels propriétaires ou open-source existent pour reconnaître automatiquement des documents numérisés. Certains systèmes permettent d'effectuer une reconnaissance optique de caractères. La plupart des OCR propriétaires sont capables de reconnaître du texte imprimé. C'est le cas de ABBY FineReader [11]. D'autres outils open-source

11. `https://pdf.abbyy.com/fr/`

ne fonctionnent que sur du texte imprimé, comme Tesseract [12] ou OCR4all [13], spécialisé dans la reconnaissance de documents anciens imprimés. Google Vision permet également de reconnaître du texte imprimé ou manuscrit. Enfin, deux outils propriétaires permettent de localiser les lignes de texte, de saisir des transcriptions manuelles, ou bien d'appliquer un modèle de reconnaissance existant : Transkribus [14] et Arkindex [15]. Cependant, ces outils ne sont pas performants sur des registres paroissiaux, en particulier pour ce qui concerne la reconnaissance d'écriture. En effet, la reconnaissance de texte est souvent spécialisée sur de l'écriture moderne, ou sur un style particulier d'écriture ancienne. Ce résultat peut être observé sur la figure 2.5 : les transcriptions réalisées par Google Vision et Transkribus ne sont pas exploitables. Ainsi, il est nécessaire de réaliser un apprentissage spécialisé sur des documents français, manuscrits, et anciens, dont le style d'écriture ressemble à ceux présents dans les registres paroissiaux.

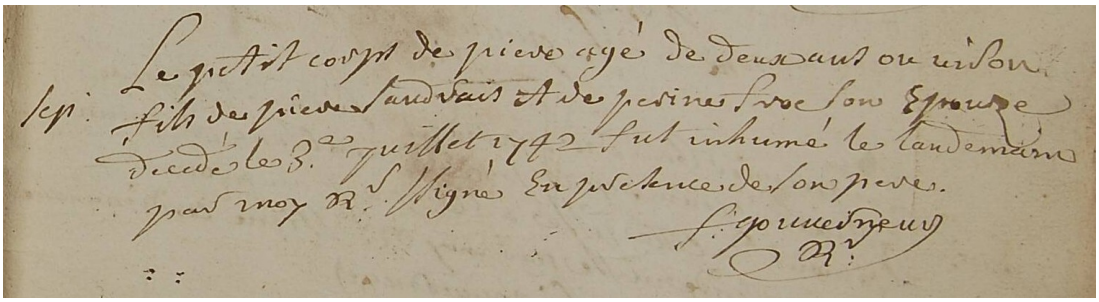
### 2.3.3 Discussion

Les approches nécessitant la segmentation des caractères ou des mots ne sont plus les plus performantes. Elles ont été dépassées par les approches neuronales, plus génériques, qui permettent de traiter directement des images de mots ou de lignes. La tendance actuelle s'oriente d'ailleurs vers une reconnaissance de texte plus complète, à partir d'images de paragraphes [Coquenet 2021] ou de pages [Wigington 2018a], lorsque cela est possible. En effet, la modélisation du langage à cette échelle permet d'intégrer un contexte linguistique plus large à l'échelle des phrases.

Mais en pratique, la plupart des systèmes de reconnaissance de texte actuels s'appliquent à des images de lignes de texte, afin de bénéficier d'un apprentissage par transfert sur les bases de données publiques, qui sont principalement constituées d'images de lignes. De la même manière, la taille de notre jeu de données ne nous permet pas d'envisager une reconnaissance à partir des actes.

Les systèmes basés sur les réseaux de neurones sont actuellement les plus performants : CRNN-CTC, seq2seq, ou Transformers. En particulier, les systèmes basés sur un mécanisme d'attention semblent prometteurs. En effet, le mécanisme d'attention permet au réseau de se focaliser sur différentes zones de l'image, ce qui lui permet d'être plus

12. `https://github.com/tesseract-ocr/tesseract`
13. `https://github.com/OCR4all`
14. `https://readcoop.eu/transkribus/`
15. `https://teklia.com/solutions/arkindex/`



**(a)** Image

*le seEit cospt de piese ege de denx ans ou*
*fils de srese sandvais et de pesine Pereon 2*
*de cedés le 3er. quillet 1743 sut inhume le 2e*
*pas moy Xe. Signe En jesétence de Son per*

**(b)** Transkribus - modèle 1 (Intendants)

*e pitit corpt de pieveage de dennant ov*
*Liss der prieoe Sandact A de perines Pvoc Con*
*dicede les 3 quillet 1743 fut inhume leste*
*pad mox R. signe expucte de Comper*

**(c)** Transkribus - modèle 2 (CITLAB)

*Ail compides prieve age Je Jeusaur ou wilon*
*Tils de sucre surddais At de periner free fon Ehr*
*Decadelzuidet ?y92 fut inainen groupe*
*par mox Reigne Capdetence de son pere.*

**(d)** Google Vision

*Le petit corps de pierre agé de deux ans ou environ*
*fils de pierre tardvais et de perine froc son epouze*
*decedé le 3e juillet 1472 fut inhumé le lendemain*
*par moy Rt signé en presence de son pere*

**(e)** Transcription manuelle

**Figure 2.5** – Évaluation qualitative des moteurs de transcription automatique appliqué à un acte du XVIII$^{e}$ siècle.

robuste aux déformations (potentiellement aux blancs, aux ratures, aux annotations interligne...). De plus, contrairement aux CRNN-CTC, les réseaux seq2seq et Transformers apprennent un modèle de langue implicite, ce qui s'adapte tout particulièrement aux actes qui contiennent un vocabulaire limité et des structures de phrases récurrentes. Si les performances théoriques des Transformers sont remarquables, il nous semble inenvisageable de les utiliser dans cette thèse, car ils nécessitent une grande quantité de données annotées. Ainsi, il nous semble intéressant d'explorer les réseaux seq2seq pour la reconnaissance d'écriture.

Le chapitre 4 présente l'architecture basée sur un mécanisme d'attention que nous avons adaptée pour la reconnaissance automatique d'écriture manuscrite.

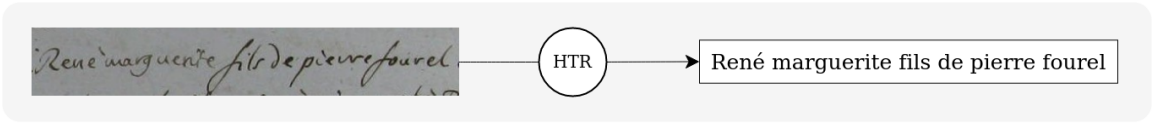
## 2.4 Extraction d'informations pertinentes

La tâche d'extraction d'information consiste à aller au-delà d'une simple reconnaissance de caractères, en associant des catégories sémantiques à chaque mot. Concrètement, cette tâche consiste à reconnaître le texte issu d'un document, mais surtout à associer à chaque mot une catégorie sémantique : lieu, date, nom, prénom...

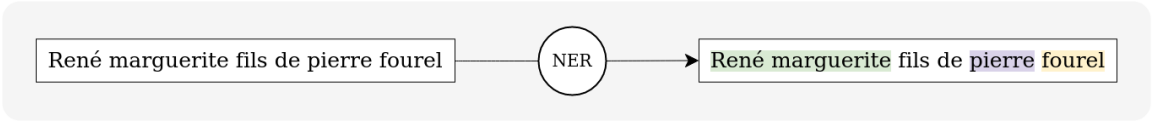
### 2.4.1 Tâches d'extraction d'information

La tâche d'extraction d'information (IE) consiste donc à combiner dans une même chaîne de traitement les tâches de reconnaissance de texte (HTR) et de reconnaissance d'entités nommées (NER). Ces trois tâches sont illustrées sur la figure 2.6, et détaillées dans la suite du chapitre.

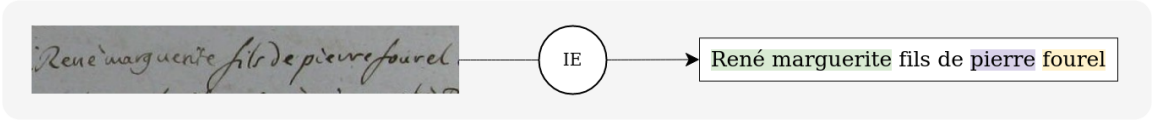
Dans les documents structurés, tels que les tableaux, les formulaires, ou les factures, l'information sémantique peut être dérivée de la localisation des mots. Pour ces documents, il est fréquent d'utiliser des modèles unifiés pour la localisation des mots, leur transcription et leur analyse sémantique [Palm 2019 ; Yu 2021]. En revanche, pour les documents semi-structurés, comme des actes, la connaissance du contexte doit être déduit de l'analyse linguistique des phrases ou des paragraphes. Par exemple, un nom est généralement placé après un prénom. La compétition ICDAR20217 sur l'extraction d'information dans des actes de mariage [Fornés 2017] a largement contribué au développement de travaux de recherche sur cette thématique.




**(a)** Reconnaissance d'écriture (HTR) à partir d'une image




**(b)** Reconnaissance d'entités nommées (NER) à partir d'un texte




**(c)** Extraction d'information (IE) à partir d'une image

**Figure 2.6** – Présentation des trois tâches pour l'analyse du contenu de documents manuscrits. Légende : prénom du fils, prénom du père, nom du père

### 2.4.2 Méthodologies pour l'extraction d'information

Trois stratégies ont été considérées pour l'extraction d'information dans ces actes de mariages. La stratégie traditionnellement utilisée est une stratégie séquentielle : le texte est reconnu par un modèle de reconnaissance d'écriture, puis une classification des mots est réalisée grâce à un modèle de reconnaissance d'entités nommées (NER) [Fornés 2017]. Une autre stratégie consiste à classer chaque image de mot dans une catégorie sémantique, avant de procéder à la reconnaissance de caractères [Toledo 2019]. Enfin, une dernière stratégie est une approche conjointe, dans laquelle les deux tâches sont adressées par un modèle unifié de reconnaissance de texte est d'entités nommées [Carbonell 2018 ; Carbonell 2020].

#### 2.4.2.1 Reconnaissance de texte avant la classification sémantique

L'approche la plus utilisée repose sur l'utilisation d'un modèle HTR pour prédire une transcription à l'échelle des caractères, puis sur un modèle de traitement du langage naturel pour classer chaque mot dans une ou plusieurs catégories sémantiques, grâce à des caractéristiques textuelles. L'inconvénient majeur de cette approche est le manque d'information sémantique lors de la transcription. Le contexte ne peut être utilisé qu'en post-processing, à l'aide de modèle de langue basé sur des catégories sémantiques. De nombreuses méthodes ont été proposées dans la compétition ICDAR20217 sur l'extraction d'information dans des documents structurés [Fornés 2017]. Le modèle de base (*Baseline-*

*HMMs*) utilise des modèles de Markov Cachés HMM couplé à un modèle de langue n-gram par catégorie. Une technique d'inférence grammaticale est ensuite utilisée pour fiabiliser la reconnaissance d'entités nommées. L'équipe de recherche Hitsz-ICRC a développé une approche capable d'extraire des informations à partir d'images de mots. Un modèle convolutif modélisant les bi-grams permet de reconnaître les caractères, sans modèle de langue explicite. Puis, les mots sont classés grâce à un champ aléatoire conditionnel (CRF [16]). Trois variantes d'une même méthode ont été proposées par le groupe de recherche CITlab ARGUS. Ils proposent d'utiliser un réseau CRNN-CTC pour la reconnaissance optique de caractères, puis des expressions régulières afin d'extraire les catégories sémantiques. La première variante (*CITlab-ARGUS-1*) fiabilise la reconnaissance d'écriture avec un vocabulaire fermé. Les deux autres variantes (*CITlab-ARGUS-2* et *CITlab-ARGUS-3*) utilisent un vocabulaire ouvert, mais avec quelques variations dans l'architecture pour la troisième variante. Plus récemment, une approche a été proposée par Prasad et al. [Prasad 2018]. Un réseau de neurones CNN-BLSTM est appris avec une fonction de coût CTC pour la reconnaissance d'écriture. Puis, une couche BLSTM est appliquée sur le vecteur de caractéristique, afin de classer chaque mot dans des catégories sémantiques.

#### 2.4.2.2 Classification sémantique avant la reconnaissance de texte

Cette nouvelle approche a été proposée par Toledo et al. [Toledo 2019]. Elle consiste à reconnaître la classe sémantique de chaque mot avant de chercher à le transcrire. Les auteurs montrent que le système de reconnaissance d'écriture tire avantage de cette connaissance sémantique a priori. Par exemple, si le réseau s'attend à transcrire un nom masculin, il aura tendance à transcrire le mot « John » plutôt que « born », et ce même si le mot ressemble plus au mot « born ». Cette stratégie est intéressante pour reconnaître des images de mots, car chaque image peut être classifiée avant d'être transcrite.

#### 2.4.2.3 Transcription et classification sémantique conjointes

Cette dernière approche a été proposée par Carbonell et al. [Carbonell 2018]. Leur méthode consiste à enrichir la transcription avec des tags qui permettent d'identifier les catégories et les personnes associées aux mots. Les tags sont localisés juste avant les mots et sont traités par le modèle de la même façon qu'un caractère. Par exemple, le mot `former` peut être enrichi en `<occupation_WF>former` afin de signifier que le mot est

16. Pour *Conditional Random Field*, en anglais

un métier (*occupation*) qui se réfère au père de la mariée (*WF* pour `wife's father`). Les mots neutres, qui ne correspondent à aucune catégorie sémantique intéressante ou à aucune personne, ne sont pas associés à des tags. Les auteurs utilisent un modèle CRNN-CTC pour effectuer la transcription automatique. Cette approche est intéressante, car le réseau apprend à extraire des caractéristiques qui sont pertinentes pour la reconnaissance de caractères et la classification sémantique. Ainsi, ces deux types d'informations sont implicitement encodés dans le vecteur caractéristique. Le même auteur a proposé un second modèle unifié capable de produire la transcription, la catégorie sémantique, et la boite englobante de chaque mot dans des images de documents [Carbonell 2020]. Cette approche unifiée est capable de traiter des images de pages, mais nécessite un ensemble d'apprentissage dans lequel la boite englobante de chaque mot est connue.

### 2.4.3 Logiciels existants

Peu de logiciels permettent d'effectuer l'extraction d'informations dans des documents. En revanche, il existe quelques logiciels open source pour effectuer la reconnaissance d'entités nommées à partir de l'analyse du texte.

BERT et ses différentes variantes (CamenBERT, RoBERTa) sont des modèles Transformers qui sont pré-entraînés pour différentes tâches de modélisation du langage [Devlin 2019 ; Martin 2020 ; Liu 2019].

FLAIR [Akbik 2019] est également un modèle pré-entraîné sur différentes langues, qui permet de modéliser le langage. Il permet par exemple de classer un texte ou de reconnaître certaines entités nommées.

### 2.4.4 Discussion

L'approche traditionnelle séquentielle, qui consiste à reconnaître le texte, puis à reconnaître les entités nommées dans la transcription, paraît envisageable dans notre contexte. En effet, les réseaux HTR et NER peuvent chacun bénéficier d'un apprentissage par transfert, ce qui limite le besoin en données annotées. L'inconvénient majeur de cette approche est que les erreurs de transcription peuvent se propager lors de la phase de reconnaissance d'entités nommées. Au contraire, la seconde approche séquentielle, qui consiste à reconnaître les entités nommées avant la transcription, ne nous parait pas applicable. En effet, elle est difficilement applicable à des images de lignes, car elle nécessite de connaître les boites englobantes de chaque mot.

La reconnaissance conjointe des caractères et des entités nommées nous semble intéressante. L'extraction de caractéristiques communes à ces deux tâches pourraient profiter à chacune des tâches individuelles. En effet, la connaissance de l'entité nommée donne un indice a priori important sur la transcription d'un mot. Réciproquement, la connaissance de la transcription simplifie la reconnaissance de l'entité nommée. Cependant, l'apprentissage de cette architecture nécessite un certain nombre de données annotées. Elle nous semble donc intéressante, mais difficile à appliquer dans notre contexte.

Les architectures basées sur un mécanisme d'attention nous semblent également pertinentes pour l'extraction d'information. En effet, le réseau d'attention peut apprendre à se focaliser sur différentes zones de l'image. Ainsi, il pourrait apprendre des caractéristiques visuelles locales et globales pour reconnaître le texte ainsi que les entités nommées à partir des images de lignes. De plus, la capacité de ces réseaux à traiter des problèmes multiples en parallèle [Luong 2015a] nous semble intéressante à explorer pour une reconnaissance conjointe d'écriture et d'entités nommées.

Le chapitre 5 présente nos contributions pour la tâche d'extraction d'information à partir d'actes de registres paroissiaux.

## 2.5 Stratégies d'apprentissage avec peu de données spécialisées

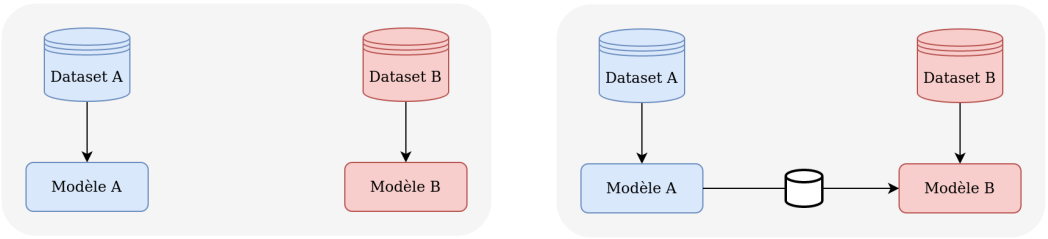
Dans cette section, nous donnons un aperçu des outils disponibles pour réduire le besoin en données spécialisées lors de l'apprentissage de modèles de reconnaissance de documents.

### 2.5.1 Apprentissage par transfert

L'apprentissage par transfert vise à transférer des connaissances apprises sur une tâche source vers une tâche cible. Il peut être vu comme la capacité d'un système à reconnaître et appliquer des connaissances et des compétences, apprises à partir d'un domaine $A$, vers un nouveau domaine $B$.

Classiquement, la plupart des architectures de base pour l'analyse d'image (classification, reconnaissance de formes ou d'objet, segmentation sémantique, reconnaissance d'instances) sont d'abord apprises sur une large base de données, telle que ImageNet [17]

17. `https://www.image-net.org/`




**(a)** Apprentissage classique **(b)** Apprentissages par transfert

**Figure 2.7** – Illustration du processus d'apprentissage par transfert. Les connaissances apprises sur une tâche $A$ sont transférés pour aider à l'apprentissage d'une tâche $B$.

[Deng 2009] ou COCO [18] [Lin 2014]. Ces deux bases de données sont composées de millions d'images naturelles annotées (paysages, personnes, animaux, objets), ce qui permet d'extraire des caractéristiques générales sur les formes, les couleurs, et les textures. Dans un second temps, les systèmes sont spécialisés sur un nouveau domaine (par exemple, la reconnaissance d'écriture), à l'aide d'une base de données plus petite. La spécialisation peut se faire en ré-entraînant tout le réseau, ou bien seulement quelques couches, sur le nouveau domaine.

L'apprentissage par transfert peut également s'effectuer sur le même domaine. Par exemple, un réseau de localisation des lignes de texte peut être pré-appris sur la base de données cBAD [Diem 2017], puis être spécialisé pour la même tâche sur peu d'exemples venant d'une collection différente. De la même manière, un système de reconnaissance d'écriture peut être pré-entraîné sur des bases de données publiques, puis spécialisé sur un petit corpus.

La figure 2.7 schématise ce processus d'apprentissage par transfert.

### 2.5.2 Bases de données publiques

Nous proposons ici un aperçu des bases de données publiques pouvant être utilisées pour un transfert de connaissance.

**Pour la reconnaissance de mise en page** De nombreuses bases de données sont disponibles pour apprendre ou évaluer des modèles de reconnaissance de structure de documents. La plupart d'entre elles sont composées de documents récents et imprimés, souvent avec une mise en page complexe (articles de journaux ou magasine, articles scientifiques).

18. `https://cocodataset.org/`

Certaines de ces bases sont complètement synthétiques [Antonacopoulos 2009], tandis que d'autres sont composées de documents réels [Zhong 2019 ; Li 2020b ; Li 2020a]. La base de données MAURDOR [Brunessaux 2014] contient quant à elle des documents manuscrits récents et multilingues, sur lesquels les zones de texte ont été annotées. Quelques bases de données composées de documents historiques ont également été proposées. Il existe trois bases de manuscrits médiévaux, avec des annotations sur les lignes de base [Hazem 2020] ou la segmentation de page [Simistira 2016]. La base de données READ-BAD [Grüning 2017] contient 3021 images de pages de documents anciens, sur lesquelles les lignes de base ont été manuellement annotées. Cette base de données sert de comparaison pour la compétition cBAD [Murdock 2015 ; Diem 2017 ; Diem 2019a]. Enfin, la base NewsEye [Michael 2021] est composée de 40 pages de journaux historiques imprimés, sur lesquelles la localisation des lignes de texte et des zones de texte est connue.

**Pour la reconnaissance de contenu** Quatre bases de données manuscrites peuvent nous intéresser. Quelques exemples issus de ces bases de données sont présentées dans la figure 2.8 Certaines bases de données historiques ont été crées afin d'apprendre des modèles de reconnaissance d'écriture manuscrite ancienne :

— La base de données READ [Strauß 2018], qui contient des documents allemands manuscrits datant du XV$^{e}$ au XIX$^{è}$ siècles. L'écriture est très difficile à déchiffrer.

— La base de données Esposalles [Romero 2013], qui contient des actes de mariage catalans datant du XVIII$^{e}$. Les entités nommées sont également annotées sur cette base [Fornés 2017]. En revanche, un seul style d'écriture est représenté (un seul scripteur sur une courte période).

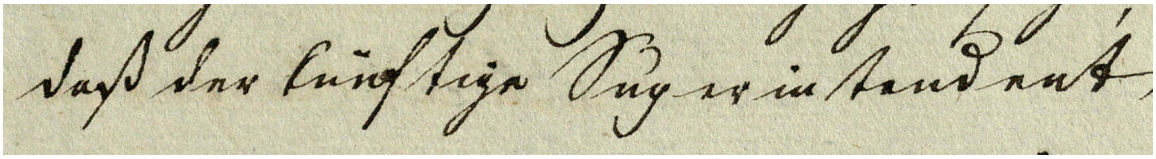
D'autres bases de données modernes sont fréquemment utilisées pour évaluer et comparer des systèmes de reconnaissances :

— La base de données IAM [Marti 2002] contient des documents manuscrits anglaise.

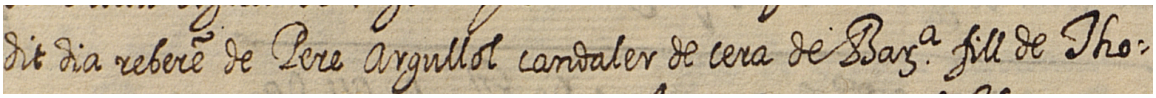
— RIMES [Augustin 2006] contient des documents manuscrits français.

### 2.5.3 Augmentation et génération de documents

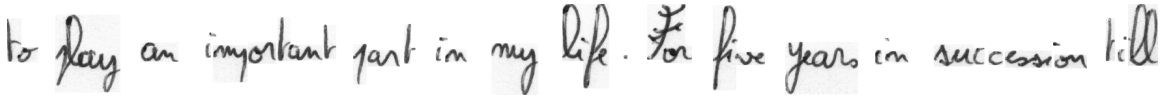
L'apprentissage d'un réseau de neurones profond sur un petit nombre d'images favorise le sur-apprentissage. Cela signifie que le réseau extrait des caractéristiques pertinentes uniquement pour le jeu d'apprentissage. Dans ce cas, les performances mesurées sur l'ensemble d'apprentissage sont bonnes, mais chutent lorsque le réseau est appliqué à un échantillon de test.



**(a)** Image issue de la base READ [Strauß 2018]



**(b)** Image issue de la base Esposalles [Romero 2013]



**(c)** Image issue de la base IAM [Marti 2002]

**(d)** Image issue de la base RIMES [Augustin 2006]

**Figure 2.8** – Illustration des bases de données pour la reconnaissance de texte manuscrit

**Méthodes de régularisation** L'une des solutions pour remédier à ce problème est d'appliquer des méthodes de régularisation, comme le *dropout*, qui consiste à ignorer certains neurones de façon aléatoire pendant l'apprentissage.

**Augmentation de données** Une autre solution consiste à augmenter artificiellement la taille de l'ensemble d'apprentissage, en appliquant des transformations aux images. Les transformations les plus utilisées en traitement d'image incluent des déformations géométriques, des changements de perspectives, des modifications de couleurs ou de contraste, des rotations, des translations, du rognage et l'ajout de bruit. Cet aspect a été abordé par certains chercheurs pour la reconnaissance de structure de documents [Journet 2017 ; Grüning 2018] ou la reconnaissance de texte [Wigington 2017 ; Atienza 2021].

**Génération de données synthétiques** Enfin, la dernière solution consiste à générer des images synthétiques pour augmenter la taille de l'ensemble d'apprentissage. Certaines approches ont été proposées pour créer des pages de documents synthétiques. D'une part, les documents modernes peuvent être générés facilement en format PDF natifs [Biswas 2021b]. JADLI et al. [JADLI 2020] ont utilisé un GAN [19] pour générer des documents

19. Pour *Generative Adversarial Networks* en anglais

imprimés. Journet et al. [Journet 2017] ont proposé DocCreator [20], un outil permettant de créer une page de document historique par couche successive. L'utilisateur peut choisir un fond, ajouter des zones de texte ou d'illustration, contrôler le contenu du texte, sa police et sa taille. Il peut également appliquer des dégradations, comme l'ajout de tâches, de caractères fantômes. Enfin, des augmentations peuvent également être appliquées, comme le changement de perspective. Plus récemment, des approches permettant de générer des images de texte manuscrit ont également été développées. Kang et al. [Kang 2020a] ont utilisé des polices manuscrites auxquelles ils ont appliqué des déformations et augmentations. D'autres chercheurs ont tenté d'utiliser des GANs pour générer des images de mots manuscrits [Kang 2020b ; Mattick 2021]. Ces approches sont encourageantes pour synthétiser de l'écriture moderne, mais elles ne sont pas encore capables de générer de l'écriture historique réaliste.

## 2.6 Conclusion du chapitre

La reconnaissance automatique de documents passe souvent par trois étapes : la reconnaissance de la structure, la reconnaissance du texte et l'extraction des informations pertinentes. Nous avons présenté dans ce chapitre les méthodes de l'état de l'art qui permettent d'aborder chacune de ces tâches. Si de nombreuses innovations ont eu lieu dans les domaines de la vision par ordinateur et du traitement automatique de la langue, la reconnaissance de document demeure extrêmement complexe.

L'une des tendances actuelles pour le traitement de documents est d'effectuer une reconnaissance complète à partir d'une image de page. En effet, certains systèmes de reconnaissance sont désormais capables de localiser et transcrire les lignes à partir d'images de paragraphes [Coquenet 2021] ou à partir de pages [Wigington 2018a]. Le système de reconnaissance d'écriture apprend alors à segmenter les lignes de manière implicite. Cette dualité entre structure et contenu est également étudiée pour la reconnaissance de documents structurés ou semi-structurés. En effet, certains systèmes tentent désormais de reconnaître dans le même temps la mise en page et le contenu textuel des documents [Boillet 2021b ; Yu 2021 ; Zhang 2021]. Cette approche est intéressante, car la mise en page et le contenu du texte peuvent être liés sémantiquement, en particulier pour des documents comme des journaux, des registres de population, des tableaux, ou encore des courriers. Mais en pratique, peu de bases de données sont intégralement annotées et

20. `https://doc-creator.labri.fr/`

transcrites à l'échelle des pages, ce qui rend le développement de ces systèmes unifiés très complexe.

Dans cette thèse, nous nous intéressons à des documents particulièrement difficiles : ils sont manuscrits, anciens, et présentent de nombreuses difficultés de lecture et dégradations. Ainsi, aucun système actuel ne permet de reconnaître les registres paroissiaux. Pourtant, de nombreuses stratégies d'analyse sont encourageantes. En particulier, les réseaux de neurones convolutifs et récurrents ont prouvé être capables de reconnaître des mises en page et des contenus textuels difficiles. Mais ces réseaux nécessitent de nombreuses images d'apprentissage. Or, peu de registres paroissiaux ont été annotés, car leur annotation est très chronophage et nécessite une expertise en généalogie et paléographie. De plus, il n'existe aucune base de données publique qui ressemble suffisamment aux registres paroissiaux, c'est-à-dire composée de documents historiques, français et datant du XVI$^{e}$ au XVIII$^{e}$ siècle.

Notre travail sera donc axé sur l'apprentissage de réseaux neuronaux avec peu de registres paroissiaux annotés. Dans la suite de ce document, nous présentons nos contributions pour la reconnaissance de registres paroissiaux. Dans le prochain chapitre, nous proposons d'étudier la reconnaissance de la mise en page de ces registres. Dans le chapitre 4 et 5, nous étudions une architecture basée sur un mécanisme d'attention pour la reconnaissance de caractères et d'entités nommées. Enfin, dans le chapitre 6, nous proposons un protocole de génération d'actes synthétiques, afin de réduire le besoin en données annotées.

Chapitre 3

# Analyse de structure des registres paroissiaux

Ce chapitre adresse la question de la reconnaissance de mise en page de registres paroissiaux. Nous proposons de comparer deux types d'approches pour la localisation d'actes dans ces documents. La première approche est entièrement basée sur des réseaux de neurones et permet de localiser les actes par apprentissage. Nous comparons trois architectures pour cette tâche, ainsi que différents scénarios d'apprentissage. La seconde approche que nous proposons est une approche hybride, dans laquelle la localisation des actes est guidée par la détection par apprentissage d'indices visuels structurels, comme les lignes de texte et les signatures. Nous proposons trois variantes de cette méthode hybride. Nous comparons ces deux types d'approches sur des registres paroissiaux français et des registres de mariage catalans. Enfin, nous évaluons leur capacité à apprendre et généraliser à partir de peu de données d'apprentissage. Les résultats montrent que l'approche hybride permet de réduire la dépendance en données d'apprentissage, car les motifs à localiser sont plus simples que les actes. Les travaux présentés dans ce chapitre ont permis la publication de deux articles [Tarride 2019 ; Tarride 2021c].

## 3.1 Introduction

L'analyse de structure ou de mise en page de documents consiste à séparer les éléments de différentes natures, comme les zones de texte, les zones graphiques et les zones tabulaires. Mais lorsque l'on s'intéresse à des documents complétement textuels, l'analyse de structure consiste principalement à localiser et à séparer les zones de texte cohérentes, c'est-à-dire les lignes de texte ou les paragraphes. Dans le cas des registres paroissiaux, il est particulièrement intéressant de localiser les actes, car ils sont indépendants en termes de contenu. En effet, chaque acte contient des informations relatives à une cérémonie religieuse. Notre objectif est donc d'identifier les actes dans les registres paroissiaux. Pour

cela, nous souhaitons localiser les boites englobantes des actes, comme illustré dans la figure 3.1.

**(a)** Image originale **(b)** Boites englobantes des actes

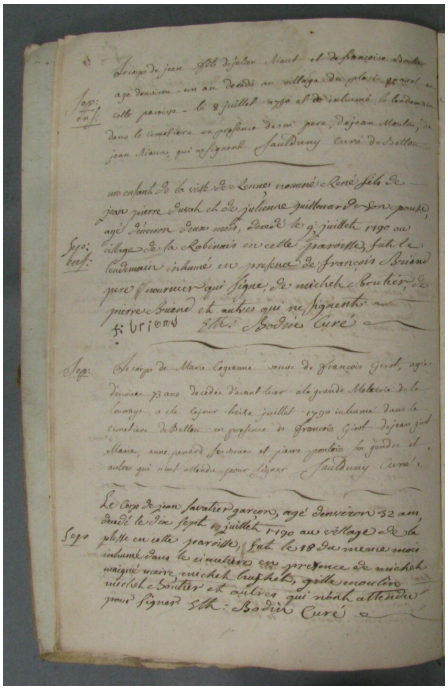
**Figure 3.1** – Exemple d'une page de registre et la localisation des actes. Les actes sont représentés par des rectangles qui englobent le texte et les signatures correspondants. Ainsi, les actes peuvent se superposer, notamment lorsque les signatures chevauchent le texte de l'acte suivant. La largeur des rectangles est normalisée sur la largeur de la page.

La localisation automatique des actes dans ces registres a des applications immédiates pour Doptim. D'une part, localiser automatiquement les zones de texte est intéressant dans le cadre de la mise en place d'une interface de visualisation et de transcription collaborative des registres. Dans ce scénario, les utilisateurs n'ont qu'à sélectionner une zone de texte pour corriger sa position et sa taille, ou bien pour saisir la transcription correspondante. D'autre part, la localisation des paragraphes ou des lignes de texte est une étape préalable à l'application de moteurs de reconnaissance de texte. De plus, la connaissance des frontières d'un acte permet d'effectuer des corrections linguistiques au niveau des actes.

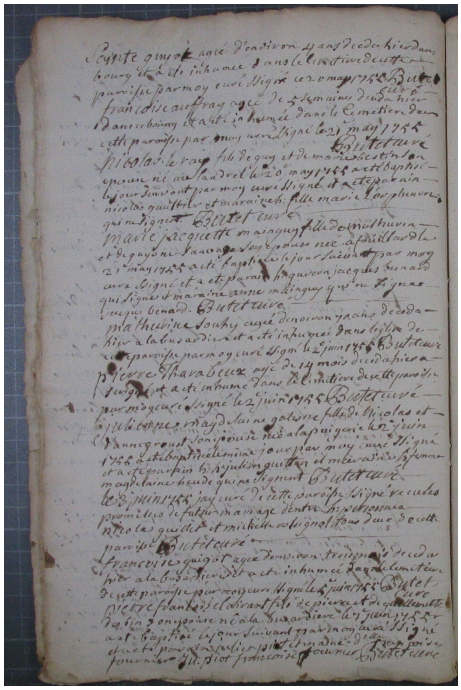
Cependant, la détection automatique des actes présente de nombreuses difficultés, dont certaines ont été détaillées dans le chapitre 1. Pour cette tâche, la difficulté majeure est la variabilité des mises en pages, qui est illustrée dans la figure 3.2. En effet, les

caractéristiques visuelles permettant de délimiter les actes sont faibles et inconsistantes entre les pages. Certaines pages, comme celle de la figure 3.2a, présentent des indices visuels clairs et stables pour séparer les actes, comme des espacements verticaux entre les lignes de texte appartenant à des actes différents, des séparations physiques entre les actes, la présence de signatures, ou encore des alinéas. Au contraire, sur d'autres registres, comme celui de la figure 3.2b, le texte est écrit avec un espace vertical constant. Seuls les éventuels retours à la ligne permettent de localiser les actes, mais il est parfois nécessaire de lire le texte pour en comprendre l'organisation logique. Les actes sont également des objets structurels complexes, qui sont hétérogènes en termes de longueur et de style d'écriture. De plus, ils n'ont pas de frontière facilement identifiable. Par conséquent, ils sont difficiles à reconnaître par un apprentissage automatique réalisé avec peu de données.
En revanche, certains indices visuels nous paraissent utiles pour localiser les actes dans ces registres. En particulier, la présence de signatures à la fin des actes est un indice régulier qui permet de repérer la fin d'un acte. D'autres indices visuels apparaissent de façon plus hétérogène, comme la présence d'alinéa, d'espacement vertical avant chaque acte, de majuscule en début d'acte, ou bien des annotations marginales sur la gauche de certains actes. La localisation de certains indices visuels nous semble donc une piste intéressante pour fiabiliser la localisation des lignes de texte, car elle permettrait de guider la reconnaissance des actes. Ces indices visuels sont également plus simples à localiser, car ils sont plus homogènes.

Nous proposons, dans ce chapitre, d'étudier deux stratégies pour la localisation automatique d'actes. La première approche, présentée dans la section 3.2, est basée sur un apprentissage automatique des actes par des réseaux de neurones pour la détection d'objets. Ces réseaux sont capables d'apprendre les caractéristiques de différents objets apparaissant dans des images. Nous souhaitons évaluer la capacité d'un tel réseau à sélectionner les caractéristiques visuelles pertinentes pour la localisation d'actes, avec peu d'exemples. Trois architectures sont comparées dans la suite de ce chapitre : YOLOv3 [Redmon 2018], RetinaNet [Lin 2018] et Mask R-CNN [He 2017]. Nous étudions également différents scénarios d'apprentissage et l'impact d'une connaissance a priori sur la localisation des lignes de texte. Si ces réseaux sont généralement très rapides et performants, ils nécessitent un grand nombre d'images d'apprentissage, notamment pour apprendre à reconnaître des objets complexes et hétérogènes comme les actes. La seconde approche que nous proposons dans la section 3.3 est une approche hybride, guidée par l'analyse d'indices visuels que nous avons sélectionnés. L'idée de cette approche est de tirer parti de l'efficacité des



**(a)** Les quatre actes de cette page apparaissent clairement grâce à la présence de traits séparateurs, d'annotations marginales, et d'un espace vertical conséquent entre les actes.



**(b)** Les dix actes de cette page sont difficilement visibles en raison de la densité du texte : il est nécessaire de lire le texte pour comprendre l'organisation de la page.

**Figure 3.2** – Deux registres paroissiaux dont la mise en page apparaît plus ou moins clairement. Sur la page de gauche, la mise en page est identifiable sans lire le texte. Sur la page de droite, il est nécessaire de lire le texte pour identifier les actes.

réseaux de neurones pour localiser des indices visuels, comme les lignes de texte et les signatures. Ces indices sont des objets plus simples à reconnaître pour un réseau de neurones. La localisation de ces indices nous permet de reconstruire les actes en utilisant des règles logiques établies grâce à une connaissance a priori sur l'organisation logique des registres. Cette approche hybride est plus complexe et plus lente, mais permet de fiabiliser la reconnaissance. Enfin, dans la section 3.4, nous comparons ces deux approches sur des registres paroissiaux français et des registres de mariage catalans, puis nous évaluons leur capacité à apprendre et généraliser à partir de peu de données d'apprentissage.

## 3.2 Réseaux de détection d'objets pour la localisation d'actes

La bibliographie présentée dans le chapitre 2 montre l'efficacité des réseaux de neurones profonds pour l'analyse de la mise en page des documents. En particulier, des réseaux de neurones de détection d'objets ou de segmentation sémantiques permettent de localiser des objets. La localisation de zones de texte peut s'effectuer à l'aide de réseaux de détection d'objets, qui apprennent à prédire les rectangles englobants des objets. Ces réseaux sont classiquement utilisés pour localiser des zones de texte, graphiques ou tabulaires. Lorsque suffisamment de données d'apprentissage sont disponibles, ils parviennent à reconnaître des mises en page complexes de documents imprimés et manuscrits, avec des performances remarquables [Oliveira 2017 ; Zhong 2019]. Depuis peu, cette approche est également envisagée pour la reconnaissance de structure de documents historiques et manuscrits [Prusty 2019 ; Biswas 2021a].

### 3.2.1 Principe

La localisation des actes peut être envisagée par des réseaux de neurones de *détection d'objets*. Ces réseaux apprennent à identifier et localiser des objets dans des images en estimant les coordonnées de leurs boites englobantes, ainsi que les probabilités associées à chaque classe. Dans le domaine de la reconnaissance de documents, ces réseaux ont été utilisés pour localiser des images, zones de texte, pages dans des documents.

Les actes sont des zones de texte qui partagent des similarités, tant au niveau de la structure (espacement vertical et horizontal, annotations marginales, signatures) que du contenu (structures de phrase et mots-clés récurrents). Ces caractéristiques propres aux actes peuvent être apprises par un réseau de neurones à partir d'exemples vus pendant la phase d'apprentissage. Une fois le modèle appris, il doit être en mesure de reconnaitre et localiser ces caractéristiques sur de nouveaux documents. Le défi d'une telle stratégie est d'apprendre à reconnaître des caractéristiques pertinentes pour les actes en utilisant peu de données d'apprentissage. La difficulté est renforcée par le fait que les actes sont des objets structurels complexes et hétérogènes, dont les rectangles englobants peuvent se chevaucher.

Il existe deux types d'approches pour la détection d'objets dans des images :

— Les approches à deux phases [1] : une première phase permet de sélectionner des régions d'intérêt, puis une seconde phase permet de localier les objets de manière plus fine, et d'effctuer la classification. Ces approches sont généralement très précises. Les réseaux RetinaNet [Lin 2018], Fast RCNN [Girshick 2015], Faster R-CNN [Ren 2015], Mask R-CNN [He 2017] appartiennent à cette catégorie.
— Les approches unifiées [2] : la localisation et la classification des objets sont faites en une seule étape. En conséquence, ces approches sont extrêmement rapides, mais souvent moins précises. Les réseaux SSD [Liu 2015b] et YOLO [Redmon 2015 ; Redmon 2016 ; Redmon 2018] appartiennent à cette catégorie.

### 3.2.2 Architectures sélectionnées

Nous présentons chacune des architectures que nous comparons dans ce travail pour la localisation d'actes : Mask R-CNN [He 2017], RetinaNet [Lin 2018] et YOLOv3 [Redmon 2018]. Ces réseaux étaient à l'état de l'art sur la base publique COCO [Lin 2014] en 2019, à l'époque où ces travaux ont débuté. Il nous semble donc logique d'évaluer leur performance sur la tâche de localisation des actes.

#### 3.2.2.1 YOLOv3

YOLO est un réseau unifié, dont l'architecture de base a été proposé en 2016 par Redmon et al. [Redmon 2015]. Des versions améliorées ont ensuite été progressivement proposées [Redmon 2016 ; Redmon 2018]. L'architecture YOLO aborde la tâche de détection d'objet comme un problème de régression, afin de séparer spatialement les boîtes englobantes des objets. Le modèle est capable d'apprendre de façon unifiée grâce à une fonction de coût complexe qui pénalise à la fois la tâche de classification, la tâche de localisation, ainsi que le score de confiance associé à chaque boîte englobante. Cette architecture est simple, précise et peut être utilisée pour une localisation en temps réel. Dans YOLOv3, quelques améliorations sont proposées sur le calcul de la fonction de coût et l'architecture du réseau. En particulier, un réseau convolutif pyramidal est utilisé pour extraire les caractéristiques à trois échelles. La limite de cette architecture est liée à l'utilisation d'une grille : l'image d'entrée est divisée en une grille, où chaque cellule estime la présence d'un seul objet. Ainsi, YOLO impose une contrainte spatiale forte, car il n'est

1. Aussi appelées *two-shots*
2. Aussi appelées *one-shot*

pas capable de localiser des objets denses et très proches.

#### 3.2.2.2 RetinaNet

RetinaNet [Lin 2018] est un détecteur à une phase. L'architecture du réseau est composée d'un réseau convolutif pyramidal qui calcule des cartes de caractéristiques sur l'ensemble de l'image. Un premier sous-réseau est utilisé pour la classification des objets et un second pour la régression de la boîte englobante. L'amélioration majeure de RetinaNet vient de l'introduction d'une nouvelle fonction de coût, appelée *focal loss*, qui gère le déséquilibre des classes entre les objets d'intérêt et le fond. Cette fonction de coût, basée sur l'entropie croisée, assigne des poids faibles aux exemples correctement classifiés, pour mieux se concentrer sur les zones de l'image les plus difficiles à prédire. RetinaNet a montré sa capacité à égaler la vitesse des détecteurs à une phase, tout en surpassant la précision de nombreux détecteurs à deux phases, notamment Faster R-CNN [Ren 2015].

#### 3.2.2.3 Mask R-CNN

L'architecture R-CNN [Girshick 2013] est une approche à deux phases qui a connu de nombreuses améliorations avec Fast R-CNN [Girshick 2015], Faster R-CNN [Ren 2015], puis Mask R-CNN [He 2017]. La première phase de sélection des zones d'intérêt est réalisée par un réseau convolutif pyramidal, qui permet d'extraire des caractéristiques visuelles à différentes échelles de l'image du document, et d'identifier des zones d'intérêt. Ce réseau de sélection de zones d'intérêt est relié à une branche de *détection d'objets* qui effectue la localisation et la classification des zones d'intérêt. Mask R-CNN possède également une autre branche de *segmentation d'instance* permettant d'obtenir une segmentation précise des objets. Au-delà de la segmentation d'instances, Mask R-CNN introduit un processus d'alignement des zones d'intérêt, qui conduit une amélioration significative des performances pour les deux branches. Ainsi, Mask R-CNN dépasse les performances de Faster-RCNN, même lorsque l'on n'utilise pas la branche de segmentation.

### 3.2.3 Protocole d'apprentissage et d'évaluation

Nous présentons ici la méthodologie pour l'apprentissage de ces trois architectures. Notre but est de pouvoir comparer ces trois réseaux pour la détection d'actes.

#### 3.2.3.1 Protocole d'apprentissage

**Pre-processing et augmentation** En premier lieu, les images d'entrée sont mises à l'échelle de sorte que le plus grand côté soit égal à 1000 pixels. Des augmentations de données sont effectuées en appliquant des transformations miroir, du bruit gaussien, du flou gaussien. Les images d'entrée sont tirées aléatoirement, ainsi que l'ordre des actes dans les images.

**Apprentissage** Ces trois réseaux ont été pré-entraînés sur la base publique COCO [Lin 2014]. L'apprentissage est ensuite réalisé sur la base de données DLA-BMS-1, présentée dans le chapitre 1, en utilisant une validation croisée à cinq feuilles. Ainsi, le système peut être évalué sur les 200 images (1565 actes) de cette base.
L'arrêt précoce est utilisé pendant l'entraînement afin de régulariser l'apprentissage du modèle : l'apprentissage est arrêté lorsque les performances de validation se dégradent. Les trois réseaux sont implémentés avec Keras, et l'apprentissage est effectué sur une carte graphique NVIDIA RTX 2080 Ti.

**Post-traitement** Pour le post-traitement, les prédictions avec un faible score de confiance ($< 0.5$) sont rejetées. Puis, la largeur des actes est ensuite normalisée pour correspondre à la largeur de la page détectée.

#### 3.2.3.2 Métriques d'évaluation

La localisation des actes est évaluée grâce à deux métriques.

**Average Precision** Chaque modèle est évalué en utilisant la mesure d'Average Precision (AP), très courante en détection d'objet. L'AP mesure la qualité de détection d'une classe en particulier. La mAP correspond à la moyenne des AP pour chaque classe. Dans notre cas, l'AP et la mAP sont égales, car une seule classe existe. Les zones prédites sont comparées aux zones vérité en utilisant un score de recouvrement, calculé par l'intersection sur l'union (IoU [3]), aussi appelé indice de Jaccard. Le seuil $\tau$ généralement choisi est $\tau = 0.5$. L'IoU entre deux zones $A$ et $B$ se calcule de la façon suivante :

$$IoU = \frac{|A \cap B|}{A \cup B} \geq \tau$$

3. Pour Intersection over Union

Celles qui sont correctement détectées, c'est-à-dire qui ont un score de recouvrement supérieur à un seuil fixé, sont considérées comme des zones vraies positives (TP). Les autres zones détectées sont des fausses positives (FP). Enfin, les zones vérité qui n'ont pas été associées à une zone prédite sont appelées les vraies négatives (FN). Ces valeurs permettent de calculer la précision et le rappel.

Toutes les zones prédites sont ensuite classées par score de confiance décroissant. La précision moyenne est ensuite calculée en fonction des labels. Elle correspond à l'aire sous la courbe de la courbe précision $p$ en fonction du rappel $r$.

$$AP = \int_0^0 p(r)dr$$

Dans ce travail, nous évaluons l'AP avec deux seuils de recouvrement, l'un à 0.5 ($AP$@0.50) l'autre, plus strict, à 0.75 ($AP$@0.75).

**ZoneMap** La métrique ZoneMap [Galibert 2015] a été construite spécifiquement pour la détection et la classification de zones dans des documents scannés, dans le cadre de la campagne d'évaluation Maurdor [Brunessaux 2014]. Les zones références et les zones hypothèses sont associées dans différentes configurations :

- Parfait : une zone référence associée à une zone hypothèse ;
- Oubli : une zone référence n'est associée à aucune zone hypothèse ;
- Faux positif : une zone hypothèse n'est associée à aucune zone référence ;
- Division :une zone référence est associée à plusieurs zones hypothèses ;
- Fusion : une zone hypothèse est associée à plusieurs zones références.

Pour chaque configuration, le score ZoneMap caractérise une erreur de surface sur les pixels du premier plan, qui correspondent au texte. Un score ZoneMap parfait correspond à un score de 0. En revanche, ce même score peut devenir arbitrairement grand si des zones larges sont manquées ou faussement détectées. En ce sens, la métrique est difficile à interpréter de manière absolue, mais reste utile pour comparer différentes méthodes sur un même jeu de données.

### 3.2.4 Évaluation des architectures de détection d'objets

Nous comparons ici les performances des trois architectures YOLOv3, RetinaNet et Mask R-CNN. Nous évaluons ensuite certaines configurations d'apprentissage pour la meilleure architecture.

#### 3.2.4.1 Comparaison des trois architectures

Dans un premier temps, nous comparons les différents réseaux de neurones de détection d'objets : Mask R-CNN [He 2017], Retina-Net [Lin 2018] and YOLOv3 [Redmon 2018].

Les performances des trois architectures sont présentées dans la table 3.1. Nous observons que YOLOv3 ne parvient pas à apprendre des caractéristiques pertinentes pour la localisation des actes. Le réseau RetinaNet, appris avec sa fonction de coût particulière, obtient des résultats supérieurs mais qui restent insatisfaisants. Le réseau Mask R-CNN surpasse de loin RetinaNet et YOLOv3. La supériorité du Mask R-CNN peut provenir de son mécanisme de proposition des régions d'intérêt. Puis, la branche de localisation et de classification traite ces régions candidates plus finement.

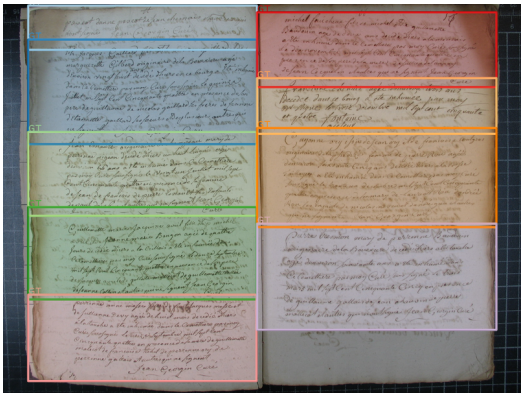
**Table 3.1** – Évaluation quantitative de chaque réseau pour la localisation des actes sur la base DLS-BMS-1 (200 images, 1565 actes)

| Model | ZoneMap ↓ | AP@.50 ↑ | AP@.75 ↑ |
|---|---|---|---|
| Mask R-CNN | **31.9** | **86.8** | **66.1** |
| RetinaNet | 47.4 | 68.9 | 36.5 |
| YOLOv3 | 76.3 | 24.3 | 0.8 |

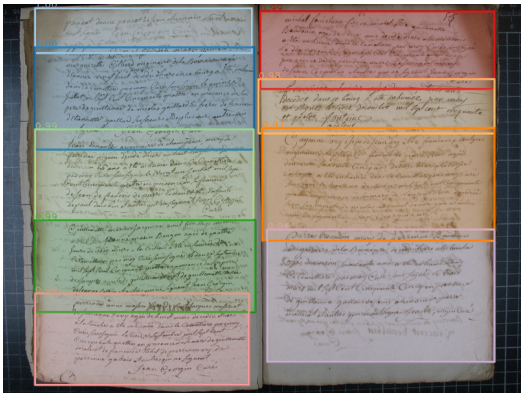
La figure 3.3 montre une illustration des résultats qualitatifs sur une même image pour chaque réseau. La sortie de Mask R-CNN semble proche de la vérité terrain puisque tous les actes sont correctement trouvés. Cependant, le résultat est pénalisé par de petites erreurs de surface, principalement dues à un débordement dans le dernier acte. Cette petite erreur peut être expliquée par la présence de signatures de la page suivante, visibles par transparence. RetinaNet produit des zones de largeur correcte sur les deux pages, mais elles sont très imprécises, avec de grands chevauchements entre deux actes successifs. Il produit également de nombreuses erreurs de fusion. YOLOv3 a du mal à trouver des zones pertinentes, surtout sur la page de droite. La largeur des zones n'est pas cohérente avec les frontières des pages, et de nombreux actes sont manqués ou fusionnés. De plus, les frontières des actes sont très imprécises.

#### 3.2.4.2 Expériences supplémentaires sur Mask R-CNN

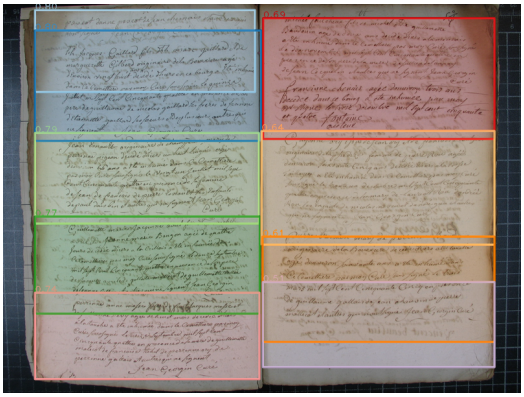
Les résultats montrent que Mask R-CNN surpasse les autres architectures pour la tâche de localisation des actes. Nous effectuons quelques expériences supplémentaires sur cette architecture pour trouver les conditions optimales de détection des actes. Trois fac-



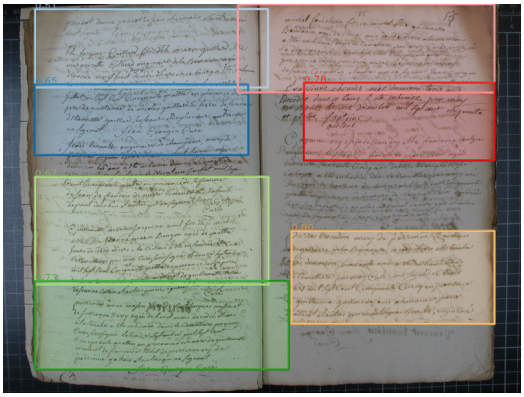
**(a)** Vérité terrain



**(b)** Mask R-CNN
ZoneMap=19.8



**(c)** Retina-Net
ZoneMap=32.2



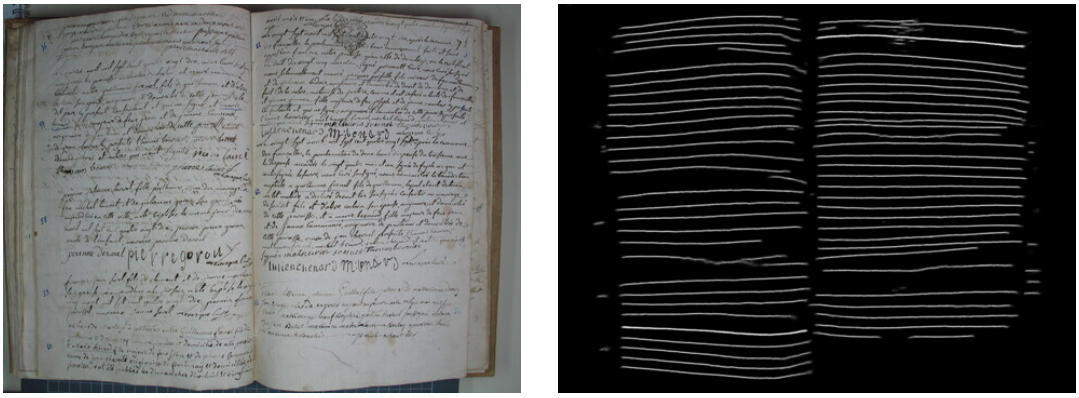
**(d)** YOLOv3
ZoneMap=76.7

**Figure 3.3** – Évaluation qualitative de la localisation des actes par des réseaux de détection d'objets sur une même image de la base DLA-BMA-1. Chaque acte est illustré par un rectangle de couleur et le score ZoneMap est associé à chaque figure.

teurs sont comparés dans le tableau 3.2.

Dans un premier temps, nous comparons les résultats obtenus par Mask R-CNN avec deux architectures de base pré-apprises sur COCO : ResNet-50 ou ResNet-101. Les performances observées montrent que ResNet-101 permet d'obtenir un petit gain sur les scores ZoneMap et AP. Nous comparons ensuite deux types de stratégies d'augmentation de données. La stratégie *simple* consiste à appliquer des transformations miroirs et du flou gaussien aux images. La stratégie plus *avancée* consiste à combiner aléatoirement des transformations miroirs, du bruit et flou gaussien, des variations de couleurs et de contraste, des transformations affines et des découpages de l'image. Les résultats montrent que la stratégie *simple* suffit pour obtenir de bons résultats, alors que la stratégie *avancée* dégrade fortement les résultats.

Enfin, nous réalisons également des expérimentations sur l'impact des connaissances a priori sur les performances du réseau. Pour cela, nous comparons les résultats obtenus à partir des images brutes (RGB) ou des images contenant les informations sur la localisation des lignes de texte un canal supplémentaire (RGBL), comme illustré sur la figure 3.4. Cette idée est motivée par la structure des actes qui dépend fortement des lignes de texte. Dans l'ensemble, nous constatons que l'utilisation des lignes de texte prédites élimine systématiquement les erreurs récurrentes et réduit le nombre d'actes faux positifs. Cela permet principalement d'éviter des erreurs lorsque le texte de la page précédente ou suivante est visible par transparence.



**(a)** Image originale (RGB). **(b)** Lignes de texte prédites avec ARU-Net (L).

**Figure 3.4** – L'image d'entrée (RGB) est concaténée avec le masque de probabilité des lignes de texte (L) localisées par ARU-Net [Grüning 2018].

**Table 3.2** – Évaluation quantitative des performances de Mask R-CNN pour la détection d'actes sur la base DLA-BMS-1. La meilleure configuration est formatée en gras.

| Paramètres d'apprentissage | | | Scores | | |
|---|---|---|---|---|---|
| Architecture de base | Entrée | Augmentation | ZoneMap ↓ | AP@.50 ↑ | AP@.75 ↑ |
| ResNet-50 [4] | RGB | Simple | 31.9 | 86.8 | 66.1 |
| ResNet-50 | RGBL | Simple | 30.1 | **91.9** | 73.5 |
| ResNet-101 | RGB | Simple | 29.6 | 88.5 | 70.6 |
| **ResNet-101** | **RGBL** | **Simple** | **29.1** | 89.6 | **73.9** |
| ResNet-101 | RGB | Avancé | 39.2 | 81.8 | 32.2 |
| ResNet-101 | RGBL | Avancé | 35.9 | 87.1 | 38.8 |

### 3.2.5 Discussion

Cette étude comparative des différents réseaux de détection d'objets met en lumière la supériorité de Mask R-CNN pour la localisation d'actes. Les expérimentations menées sur cette architecture indiquent que l'utilisation d'une architecture profonde permet d'augmenter les scores de reconnaissance. De plus, la connaissance a priori sur la localisation des lignes de texte permet également de réduire les erreurs de segmentation. Mais si les performances quantitatives de Mask R-CNN sont convaincantes, la qualité des résultats ne nous convient pas tout à fait. En effet, les boites englobantes prédites par le réseau sont peu précises, avec des chevauchements importants entre les actes, même lorsque le réseau bénéficie d'une connaissance a priori sur les lignes de texte. De plus, un problème de généralisation se pose : il n'est pas garanti que le réseau parvienne à reconnaître des registres paroissiaux issus d'autres communes ou d'autres époques. Dans la suite, nous proposons une approche hybride qui vise à dépasser les limites principales des approches entièrement neuronales.

## 3.3 Approche hybride pour la reconnaissance de structure

Dans cette section, nous présentons notre contribution principale pour la localisation des actes dans les registres paroissiaux : l'approche hybride. Cette approche est composée de deux phases : en premier lieu, des indices visuels sont localisés par des réseaux de neurones, en second lieu, ces indices visuels sont groupés par des règles logiques, définies selon la structure des registres.

### 3.3.1 Principe

L'équipe Intuidoc a développé un système de reconnaissance de documents : la méthode DMOS (Description et MOdification de la Segmentation) [Coüasnon 2006b]. La méthode DMOS a été largement testée et validée sur des collections de document variées, comme des listes de prix boursiers [Guerry 2019] ou encore des cartes historiques [Lemaitre 2021]. Cette méthode se découpe en deux étapes. La première étape consiste à localiser les indices visuels dans les documents : segments verticaux or horizontaux, composantes connexes... La seconde étape consiste à décrire la mise en page des documents

à partir de ces indices visuels, en définissant des règles logiques décrites dans le langage EPF.

Afin d'appliquer cette approche à des registres paroissiaux, nous identifions plusieurs indices visuels importants pour décrire la mise en page des registres : les bords de page, les lignes de texte, les premières lignes de texte et les signatures. Nous proposons de localiser ces indices visuels à l'aide de réseaux de neurones. Enfin, nous définissons des règles logiques qui décrivent l'organisation de la mise en page. Trois variantes de cette méthode sont envisagées et comparées dans la suite de ce chapitre :

— Une variante basée sur la seule localisation des **signatures**, des lignes de texte, et des bords de page. Dans cette variante, la localisation des signatures permet de repérer la fin des actes ;

— Une variante basée sur la seule localisation des **premières lignes de texte** de chaque acte, des lignes de texte, et des bords de page. Dans cette variante, la localisation des premières lignes permet de localiser le début des actes ;

— Une variante **mixte**, basée sur la localisation des premières lignes de texte, des signatures, des lignes de texte et des bords de page. Dans cette variante, les informations sur la localisation des premières lignes de texte et des signatures sont combinées afin de fiabiliser la détection des actes. Un aperçu de cette variante mixte est proposé dans la figure 3.5.

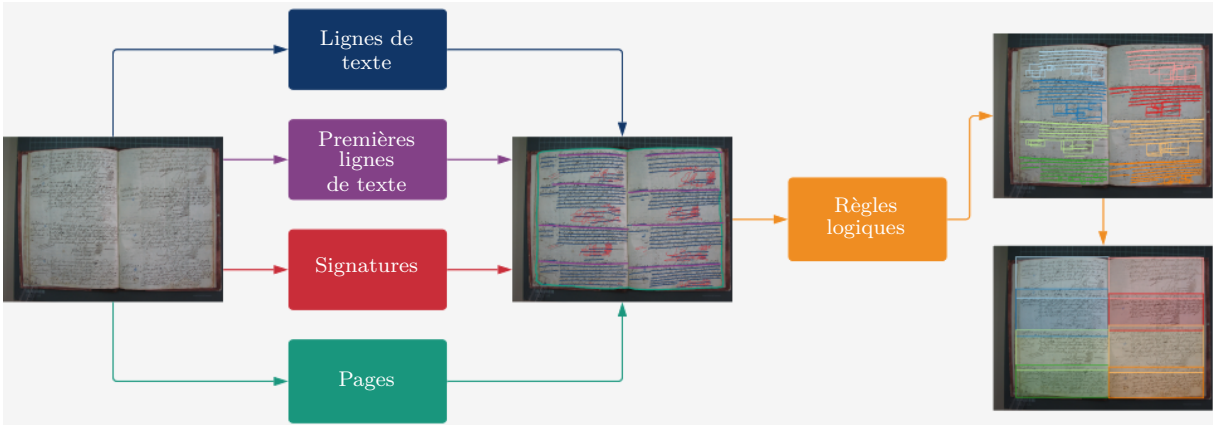



**Figure 3.5** – Aperçu de méthode hybride **mixte** pour la localisation des actes. Des réseaux de neurones sont utilisés pour localiser des indices visuels : lignes de texte, premières ligne de texte, signatures et les bords de page. Puis, des règles logiques sont appliquées pour grouper les indices qui appartiennent à un même acte. Les frontières des actes sont ensuite extraites en calculant la boite englobante de chaque acte.

Dans la suite, nous présentons l'intérêt des indices visuels sélectionnés, nous détaillons le processus de localisation des indices visuels, puis nous définissons les règles logiques utilisées pour décrire la mise en page des registres paroissiaux.

### 3.3.2 Extraction des indices visuels

Nous présentons les indices utiles pour détecter les actes dans des registres paroissiaux : les bords de page, les lignes de texte, les signatures, et les premières lignes de chaque acte.

#### 3.3.2.1 Sélection des indices visuels pertinents

Nous détaillons l'intérêt de chacun de ces indices visuels pour la localisation des actes. Nous présentons également le protocole d'apprentissage pour leur localisation, ainsi que les potentielles difficultés de reconnaissance.

**Lignes de texte** Les lignes de texte sont la principale composante structurelle des actes, car un acte est une suite de lignes de texte. Leur extraction est donc une étape clé dans la tâche de localisation des actes. Le regroupement des lignes appartenant à un même acte permet de calculer le rectangle englobant de l'acte. La localisation des lignes de texte peut s'effectuer avec des réseaux de neurones de segmentation sémantique. Les lignes sont alors représentées par leur ligne de base, et ces réseaux apprennent à classifier chaque pixel comme appartenant à une ligne de base ou non.
Pour localiser les lignes de base, nous avons sélectionné le réseau ARU-Net, car il permet d'obtenir les meilleurs résultats sur la base de données cBAD [Diem 2017]. Le post-traitement introduit par Oliveira et al. [Oliveira 2018] est ensuite utilisé : les cartes de probabilité sont filtrées à l'aide d'un filtre gaussien et un seuillage par hystérésis est appliqué. Les composantes connexes correspondant aux lignes de base sont ensuite vectorisées en lignes polygonales.
Cette tâche ne présente pas de difficultés majeures, car de nombreux documents historiques hétérogènes sont disponibles dans la base de données cBad [Diem 2019a] pour l'apprentissage de ces modèles. Par conséquent, les systèmes automatiques de reconnaissance de lignes de texte sont généralement très pertinents.

**Bords de pages** Les images des registres paroissiaux n'ont pas été découpées de façon à ne garder que le document. Ainsi, les supports sur lesquels sont posés les registres sont

visibles. Ce support est variable selon les documents : table en bois, papiers, quadrillage, autre registre... La localisation des bords de page dans ces images est donc utile, car elle permet de restreindre la recherche des actes à l'intérieur d'une page. Les bords de pages sont également utilisés pour normaliser la largeur de chaque acte.
La localisation des bords de page peut s'effectuer de différentes façons. Une première façon de faire consiste à utiliser un seuillage sur les gradients verticaux et horizontaux de l'image afin de mettre en évidence les transitions entre le fond et la page. Une seconde façon consiste à utiliser un réseau de neurones ou de classification sémantique des pixels, comme dhSegment [Oliveira 2018] appris sur la base DLA-BMS-1. En sortie du réseau, chaque pixel correspond à sa probabilité d'appartenir à une page. Les cartes de probabilité sont ensuite seuillées et la localisation du plus petit rectangle englobant permet d'obtenir les coordonnées des bords de page.
La localisation des bords est une tâche relativement simple. En revanche, certains documents sont placés sur un fond de papier, comme sur la figure 3.6, et peuvent ainsi poser des difficultés. En effet, le système doit être capable de localiser les pages principales sur l'image.

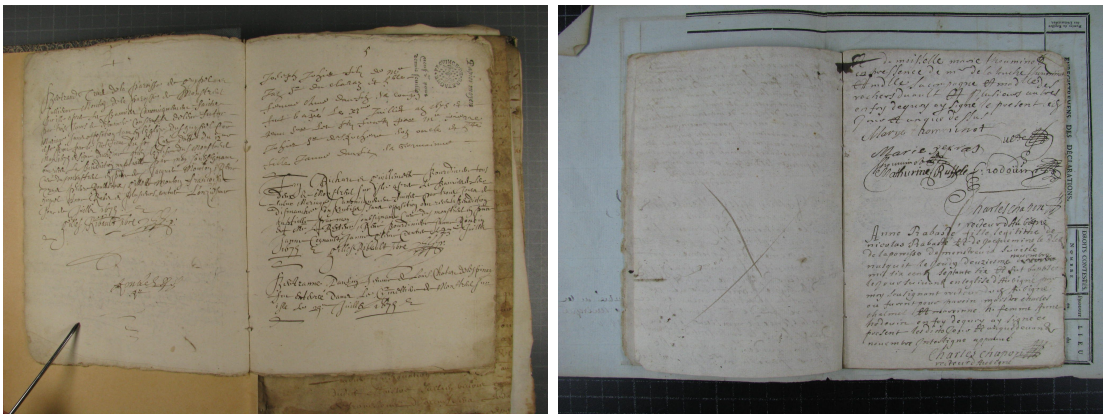
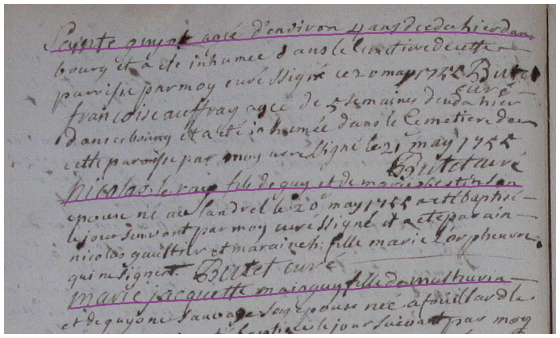


**Figure 3.6** – Registres pouvant poser des difficultés pour la localisation de bords de page : les documents sont posés sur des supports en papier

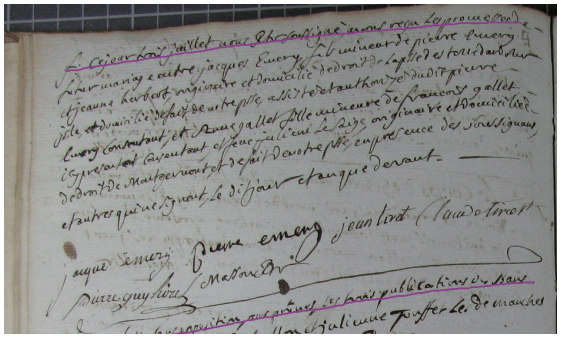
**Premières lignes de texte** Les premières lignes de chaque acte constituent un indice visuel majeur dans la compréhension de la mise en page des registres paroissiaux, car elles permettent de localiser le début de chaque acte. Or, les premières lignes partagent des caractéristiques visuelles qui permettent de les identifier en utilisant un réseau de neurones. Ces lignes présentent fréquemment des mots, expressions ou structures de phrase

récurrentes. En particulier, l'acte commence généralement par une date ou un prénom, et la première ligne commence souvent par une majuscule. Enfin, des éléments de mise en page permettent de localiser ces lignes : la présence d'alinéa, d'un espace interligne important, de signatures au-dessus, ou d'annotations marginales sur la gauche. D'autre part, certains mots, expressions ou structures de phrase récurrentes sont fréquemment utilisées par les prêtres dans la première ligne de texte.
La localisation des premières lignes peut s'effectuer par un apprentissage sur la base DLA-BMS-1, de la même façon qu'un système classique de reconnaissance de lignes de texte. Nous comparons trois architectures de réseaux de neurones pour cette tâche : dhSegment, ARU-Net, LARU-Net. Deux stratégies d'apprentissage peuvent être utilisées : la première consiste à localiser uniquement les premières lignes, la seconde consiste à localiser et classifier toutes les lignes. Dans la suite, nous comparons ces deux scénarios d'apprentissage, ainsi que différents algorithmes de post-traitement.
La localisation des premières lignes est complexe à automatiser en raison des confusions possibles entre les différentes lignes. La figure 3.7 donne un aperçu de quelques configurations particulières pour la localisation des lignes de texte. La principale difficulté vient du fait que les premières lignes de texte peuvent facilement être confondues avec les autres lignes de texte. La difficulté est particulièrement importante sur les documents présentant des mises en page serrées, comme dans la figure 3.7a. Dans cet exemple, il n'y a pratiquement aucun espacement vertical entre deux actes successifs, ce qui complique la localisation des premières lignes. Dans la figure 3.7b, le groupe de signatures présente les mêmes caractéristiques qu'une première ligne de texte, avec notamment un espace vertical important au-dessus. Dans la figure 3.7c, la première ligne de texte peut également être manquée à cause de la présence d'un sceau au milieu de la ligne. Enfin, dans la figure 3.7d, une première ligne de texte est barrée, puis réécrite.

**Signatures** L'un des motifs qui apparaît régulièrement dans les registres est la signature du prêtre, et éventuellement des témoins, à la fin des actes. En effet, à partir de 1667, chaque acte est théoriquement signé par le prêtre. Par conséquent, l'extraction des signatures peut donc permettre de localiser la fin de chaque acte. Cette extraction est possible, car les signatures partagent des caractéristiques communes. En effet, les signatures sont souvent localisées à la fin d'un acte, soit à la suite du texte (donc plutôt sur la droite de la page), soit sur une nouvelle ligne. Certaines signatures stylisées se repèrent facilement avec la présence de boucles. D'autres signatures plus simples comprennent le

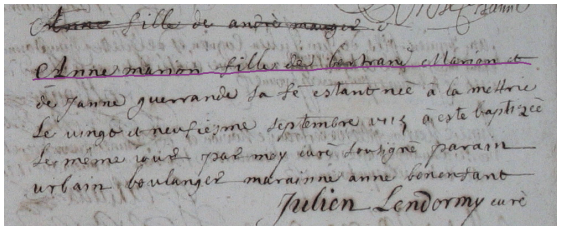


**(a)** L'écriture est très compacte.



**(b)** Le groupe de signatures ressemble à des premières lignes (majuscules, interligne)

**(c)** La première ligne de texte est écrite autour d'un sceau.



**(d)** La première ligne de texte est raturée, puis réécrite.

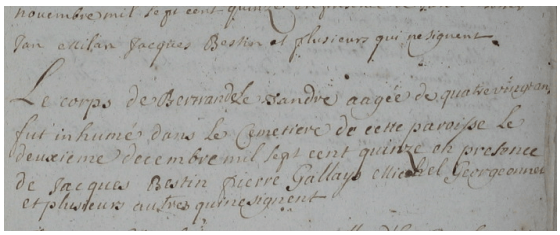
**Figure 3.7** – Exemples d'actes difficiles pour la localisation de premières lignes de texte. Les premières lignes sont localisées par des lignes en violet. Dans l'acte (a), les premières lignes de texte sont difficiles à localiser à cause d'un espace interligne uniforme. Dans l'acte (b), le groupe de signatures peut être faussement détectées comme une première ligne de texte. L'acte (c) montre une première ligne coupée en deux qui peut être oubliée par le système. Dans l'acte (d), la première ligne de texte est barrée, mais pourrait être reconnue malgré tout.

nom du prêtre suivi de la mention « curé » ou « prêtre »/« ptre ».

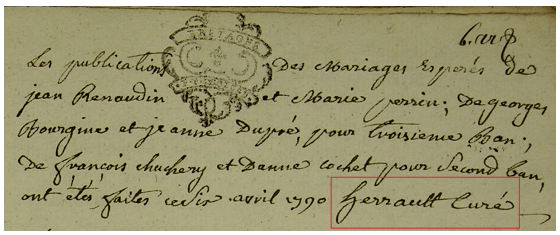
La localisation de signatures peut se faire à l'aide de réseaux de neurones de segmentation sémantique appris sur la base DLA-BMS-1. Nous comparons cinq architectures pour cette tâche : dhSegment, U-Net, RU-Net, ARU-Net et LARU-Net. Pour le post-traitement, les masques de probabilité sont seuillés, et les petites composantes connexes sont retirés.

La localisation des signatures simplifie grandement la tâche de localisation des actes. En revanche, certaines configurations semblent difficiles à traiter, en raison des interactions et des confusions possibles entre le texte et les signatures. Le premier cas évident est l'absence de signature, comme illustré dans la figure 3.8a. Une autre difficulté vient du fait que les signatures sont parfois très similaires à des mots issus du texte principal. Par exemple, dans la figure 3.8b, la signature ressemble beaucoup au texte principal. Par opposition, le réseau peut également produire un faux positif sur des actes où les mots du texte principal sont plus stylisés que les signatures, comme sur la figure 3.8c. Enfin,

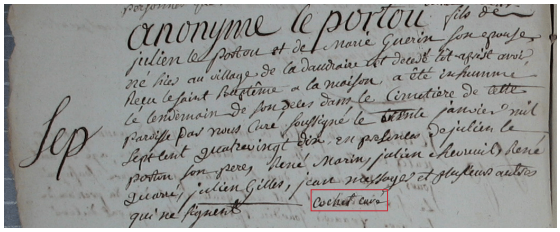
il existe une forte interaction entre les signatures et le texte principal de l'acte suivant, comme sur la figure 3.8d.



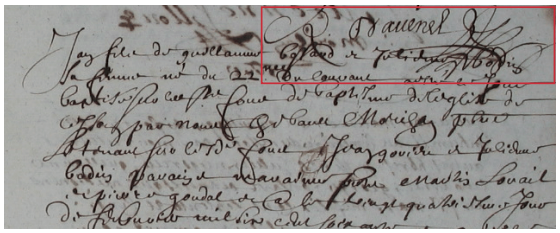
**(a)** Pas de signature sur cet acte.



**(b)** Les signatures ressemblent au texte principal.



**(c)** Les noms (en haut) sont plus stylisés que les signatures.



**(d)** Le texte et les signatures se superposent.

**Figure 3.8** – Exemples d'actes difficiles pour la reconnaissance de signatures. Les signatures sont localisées par des rectangles englobants rouges. L'acte (a) n'est pas signé. Les signatures de l'acte (b) ne sont pas stylisés et peuvent être manqués. Les noms et annotations marginales de l'acte (c) sont élaborés et peuvent être faussement détectés conne des signatures. Le texte de l'acte (d) et les signatures se superposent

#### 3.3.2.2 Métriques d'évaluation

L'extraction des indices visuels peut être évalué grâce à une matrice de confusion présentant le nombre de pixels vrais positifs TP, vrais négatifs TN, faux positifs FP et faux négatifs FN. Ces termes sont illustrés par la figure 3.9

Ces valeurs permettent également de mesurer la précision, le rappel, et le score F1 d'un système.

**Précision** La précision mesure le pourcentage d'éléments classés positifs qui sont réellement positifs. Une précision élevée signifie que la plupart des éléments classés comme positifs sont bien des positifs.

$$p = \frac{TP}{TP + FP}$$

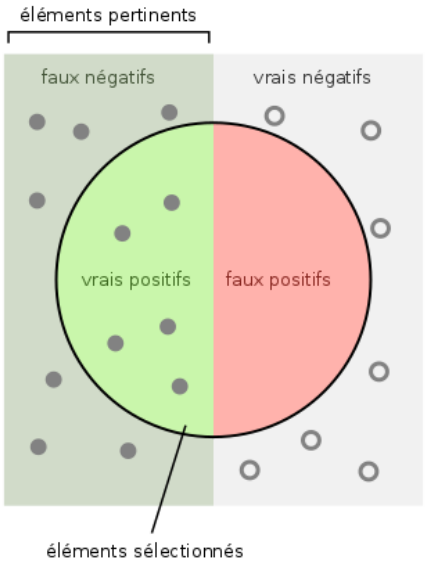



**Figure 3.9** – Illustration des erreurs de classification : les échantillons positifs correctement classifiés correspondent aux vrais positifs TP, les échantillons négatifs correctement classifiés correspondent aux vrais négatifs TN, les échantillons négatifs incorrectement classifiés correspondent aux faux positifs FP et les échantillons positifs incorrectement classifiés correspondent aux faux négatifs FN

**Rappel** Le rappel mesure le pourcentage d'éléments positifs qui ont été correctement identifiés comme positifs par le système. Un rappel élevé signifie que le système oublie peu d'éléments positifs.

$$r = \frac{TP}{TP + FN}$$

**F1-score** Le F1-score permet de résumer ces deux mesures et permet une bonne évaluation globale de la performance d'un système.

$$f = 2 \times \frac{p \times r}{p + r}$$

#### 3.3.2.3 Évaluation des indices visuels

Les indices visuels ont été annotés sur la base de donnés DLA-BMS-1. Ainsi, les résultats sont présentés sur cette base de données. L'apprentissage est effectué en utilisant une validation croisée, afin de pouvoir évaluer les modèles sur toutes les images.

**Lignes de texte** Nous avons choisi ARU-Net pour extraire les lignes de texte dans les images, car ce réseau surpasse dhSegment sur la base publique cBAD [Diem 2017]. Puisque la vérité terrain de la base DLA-BMS-1 a été établie à partir des prédictions d'ARU-Net, nous avons introduit un biais en la faveur d'ARU-Net. Ainsi, nous ne comparons pas ces deux architectures pour cette tâche.

**Bords de page** Nous comparons deux approches pour la localisation des bords de page : une méthode basée sur un seuillage développée dans l'équipe Intuidoc, et le réseau de neurones dhSegment. Nous présentons des mesures de précision et de rappel, le score F1 et l'IoU de ces deux méthodes dans la table 3.3. Des exemples de résultats sont également présentés dans la figure 3.10. Le réseau dhSegment permet une localisation des bords de page fiable et robuste. Nous retenons donc cette méthode pour localiser les bords de page.

**Table 3.3** – Évaluation quantitative de la détection des bords de page sur la base de données DLA-BMS-1. Le réseau est appris sur 120 images en utilisant la validation croisée à 5 feuilles.

| Modèle | Précision ↑ | Rappel ↑ | F1 score ↑ | IoU ↑ |
|---|---|---|---|---|
| dhSegment [5] | **0,99** | **0,99** | **0,99** | **0,98** |
| Seuillage | 0,92 | 0,96 | 0,93 | 0,88 |

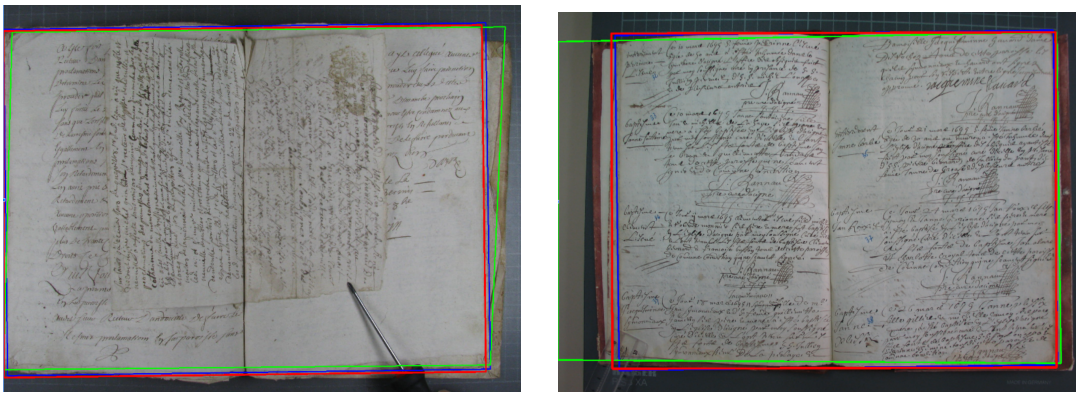
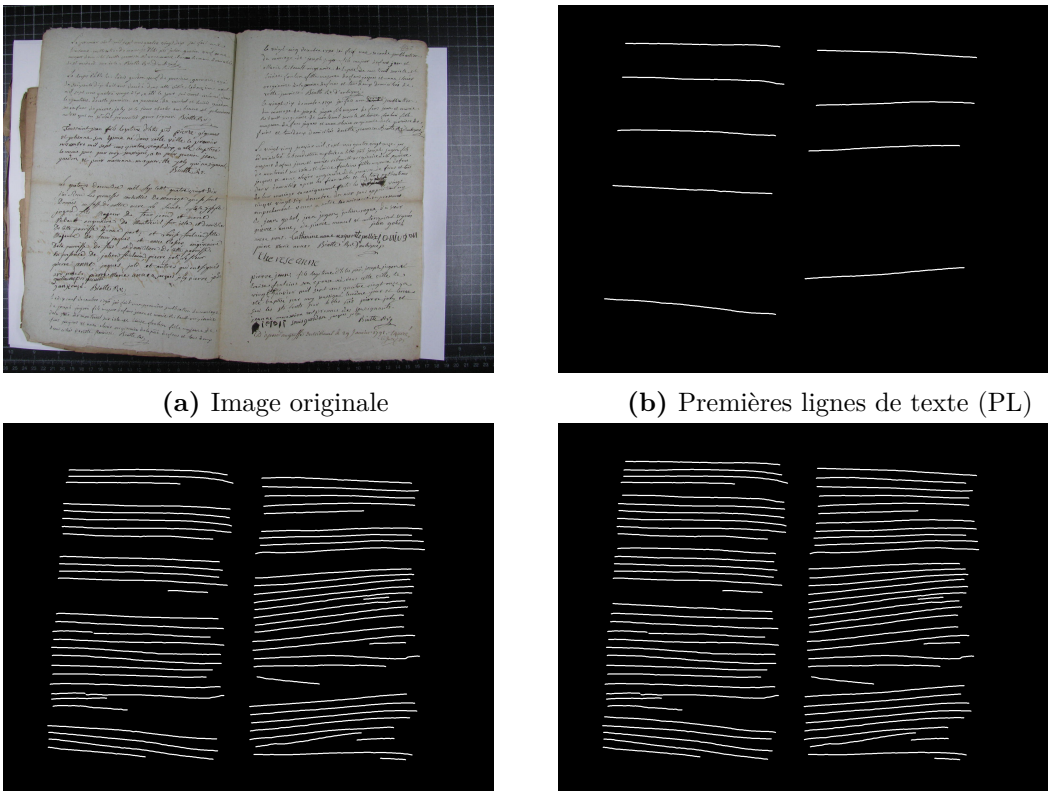


**Figure 3.10** – Illustration de la détection des bords de page. Les bords en rouge correspondent à la vérité, les bords bleus et verts aux prédictions de dhSegment et de la méthode de seuillage.

**Premières lignes de texte** L'apprentissage des premières lignes de texte est effectué à partir de l'échantillon d'apprentissage de DLA-BMS-1 en utilisant la validation croisée

afin que chaque image soit évaluée.
Les premières lignes de texte sont évaluées grâce au protocole d'évaluation READ-BAD [6] [Grüning 2017]. Ce protocole a été conçu pour évaluer des systèmes de localisation de lignes de bases. Pour chaque image trois scores sont calculés : une pseudo Précision (P-value), un pseudo rappel (R-value), et un pseudo F1 score (F-value). Ces scores sont ensuite moyennés sur toutes les images. Les résultats sont présentés sur la table 3.4.
Nous avons comparé trois architectures pour cette tâche, à savoir dhSegment [Oliveira 2018], ARU-Net et LARU-Net [Grüning 2018]. Les résultats indiquent une supériorité nette des réseaux ARU-Net et LARU-Net pour la localisation des premières lignes de texte. Le réseau LARU-Net est retenu pour la suite des expériences, car il obtient le meilleur score.
Sur cette architecture, plusieurs stratégies d'apprentissage ont été comparées. En effet, nous souhaitons mesurer si la localisation des premières lignes bénéficie d'un apprentissage des autres lignes en simultané. Ainsi nous comparons plusieurs protocoles. Dans la première expérience, le réseau apprend à localiser uniquement les premières lignes. Dans la seconde expérience, le réseau apprend à localiser toutes les lignes et à leur associer un label parmi deux classes exclusives : première ligne (PL) et autre ligne (AL) de texte. La dernière expérience est similaire, mais avec deux classes non exclusives : première ligne (PL) et toutes lignes (TL) de texte. Ces classes sont illustrées dans la figure 3.11. Dans tous les scénarios, seules les performances de la classe correspondant aux premières lignes est évaluée. Les résultats montrent que le premier scénario d'apprentissage surpasse les deux autres.
Enfin, nous évaluons l'impact de différentes méthodes de post-traitement de ces masques de probabilité. La première méthode consiste à seuiller la carte de probabilité par hystérésis. La seconde consiste à extraire les lignes à partir de l'image floue du masque de probabilités [Lemaitre 2014]. La dernière consiste à effectuer une combinaison des deux en retirant les pixels peu probables par hystérésis, puis en effectuant l'extraction des lignes dans l'image floue.
La configuration retenue correspond à l'architecture LARU-Net appris avec une seule classe (PL) dont les masques de probabilités sont post-traités à l'aide de la méthode des lignes floues.

6. https ://github.com/Transkribus/TranskribusBaseLine EvaluationScheme



**(a)** Image originale

**(b)** Premières lignes de texte (PL)

**(c)** Autres lignes de texte (AL)

**(d)** Toutes les lignes de texte (TL)

**Figure 3.11** – Illustration des différentes classes apprises. Les premières lignes correspondent aux premières lignes de texte de chaque acte. Les autres lignes correspondent aux lignes de texte qui ne sont pas des premières lignes de texte.

| Paramètrs d'apprentissage | | | Métriques | | |
|---|---|---|---|---|---|
| Architecture | Classes | Post-traitement | Précision ↑ | Rappel ↑ | F1 score ↑ |
| dhSegment | PL | Hystérésis | 0.58 | 0.78 | 0.66 |
| ARU-Net | PL | Hystérésis | 0.79 | 0.93 | 0.85 |
| LARU-Net | PL | Hystérésis | 0.81 | 0.94 | 0.87 |
| LARU-Net | PL / TL | Hystérésis | 0.77 | 0.74 | 0.75 |
| LARU-Net | PL / AL | Hystérésis | 0.81 | 0.91 | 0.86 |
| LARU-Net | PL / AL | Hystérésis | 0.81 | 0.94 | 0.87 |
| LARU-Net | PL / TL | Lignes floues | **0.91** | **0.89** | **0.90** |
| LARU-Net | PL / AL | Combinaison | 0.86 | 0.93 | 0.89 |

**Table 3.4** – Évaluation des premières lignes de texte. Les classes correspondent à PL : premières lignes, F fond, TL : toutes les lignes, AL autres lignes

**Signatures** L'apprentissage automatique des signatures peut également se faire à l'aide d'un réseau de segmentation sémantique [Tarride 2019]. L'apprentissage est effectué à partir de l'échantillon d'apprentissage de DLA-BMS-1 en utilisant la validation croisée afin que chaque image soit évaluée en test. Le post-traitement consiste à seuiller ces cartes de probabilité, et à retirer les petites composantes connexes. Les masques de signatures sont ensuite évalués à l'échelle des pixels. Les configurations d'apprentissage sont évaluées et comparées dans la table 3.5.

Nous avons comparé différents réseaux de neurones pour cette tâche, en particulier dh-

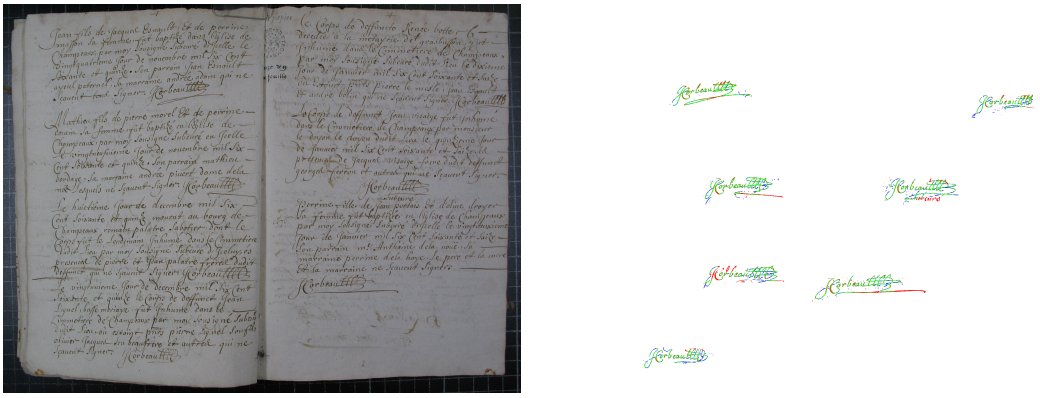
| Paramètres d'apprentissage | | | | Métriques | | |
|---|---|---|---|---|---|---|
| | Architecture | Classes | Augmentation | Précision ↑ | Rappel ↑ | F1-score ↑ |
| E1 | dhSegment | 2 | S+R | 0.45 | 0.22 | 0.28 |
| | U-Net | 2 | S+R | 0.37 | 0.45 | 0.38 |
| | RU-Net | 2 | S+R | 0.32 | 0.48 | 0.34 |
| | ARU-Net | 2 | S+R | 0.59 | 0.43 | 0.46 |
| | LARU-Net | 2 | S+R | 0.59 | 0.43 | 0.47 |
| E2 | LARU-Net | 3 | S+R | 0.63 | 0.48 | 0.52 |
| E3 | LARU-Net | 3 | S | 0.66 | 0.45 | 0.51 |
| | LARU-Net | 3 | S+A | **0.67** | **0.51** | **0.55** |
| | LARU-Net | 3 | S+R+E | 0.61 | 0.48 | 0.52 |
| | LARU-Net | 3 | S+R+A | 0.62 | 0.50 | 0.52 |

**Table 3.5** – Évaluation quantitative de différentes configurations pour la segmentation de signatures. L'expérimentation E1 compare les architectures des réseaux de neurones, et met en évidence LARU-Net. L'expérimentation E2 souligne l'intérêt d'utiliser trois classes pour l'apprentissage (trait de la signature, zone polygonale de la signature, et autre) au lieu de deux (trait de la signature, autre). L'expérimentation E3 compare les stratégies d'augmentation (S redimensionnement, R rotation, A transformation affine, E transformation élastique).

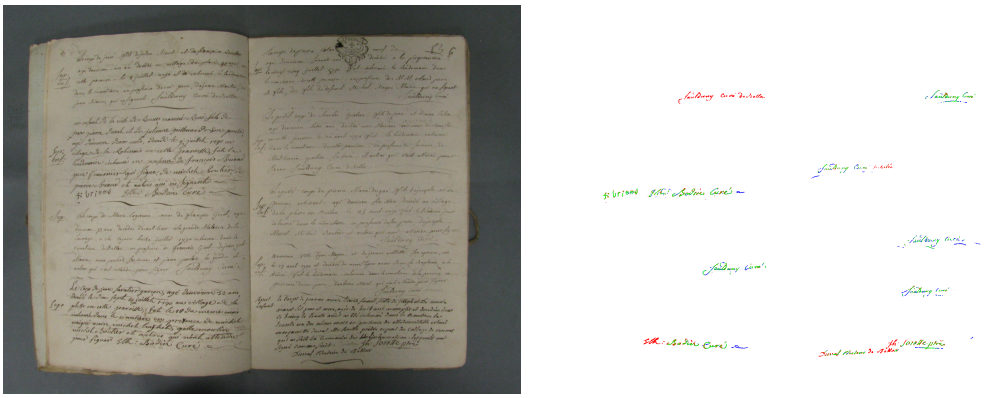
Segment, U-Net, ARU-Net et LARU-Net [Diem 2017]. Les résultats indiquent clairement que les réseaux ARU et LARU-Net sont les plus adaptés pour cette tâche.

Nous comparons également différentes représentations des signatures. En particulier, une signature peut être représentée par une segmentation fine de l'encre, ou bien par son polygone englobant. Nous étudions l'intérêt de combiner ces deux représentations lors de l'apprentissage (3 classes : *trait*, *polygone*, *fond*) ou pas (2 classes : *trait*, *fond*). Nous constatons que l'ajout d'une classe *polygone* permet d'améliorer les résultats. Notre intuition est que le réseau apprend à localiser grossièrement les zones d'intérêt, puis apprend à les segmenter plus finement.

Enfin, nous évaluons différentes stratégies d'augmentation de données. En particulier,



**(a)** Image pour laquelle aucune signature n'est oubliée ou faussement détectée. En revanche, des erreurs au niveau pixel existent et les scores sont donc moyens ($p = 0.68$, $r = 0.72$)



**(b)** Image pour laquelle une signature est complètement oubliée. La précision est donc moyenne, et le rappel bas ($p = 0.68$, $r = 0.54$).

**Figure 3.12** – Analyse des erreurs pour la localisation des signatures au niveau pixel. Les images originales sont situées à gauche, et les masques de prédiction des signatures à droite. Ces masques contiennent des pixels verts, qui correspondent aux vrais positifs TP, des pixels rouges qui correspondent aux faut négatifs FN, des pixels bleus qui correspondent aux faux positifs FP, et des pixels blancs qui correspondent aux vrais négatifs TN.

nous étudions l'impact du redimensionnement des images (S), des rotations (R), des transformations affines (A) et élastiques (E). Différentes combinaisons de ces transformations sont comparées. La meilleure stratégie d'apprentissage consiste à effectuer un redimensionnement et des transformations affines.

Globalement, les résultats au niveau pixels sont faibles, en particulier le rappel. Cela signifie que le réseau ne retrouve pas tous les pixels correspondant aux signatures. En revanche, ceux qu'il trouve correspondent plutôt bien à des signatures. Ces résultats faibles s'expliquent par le bruit introduit lors de la création de la vérité terrain, qui nécessite une

phase de binarisation. En particulier, certaines signatures contenant des taches d'encre, des éléments visibles par transparence, ou bien des contrastes bas, contiennent des erreurs de segmentation. Cette observation se quantifie : plus de la moitié des pixels faux positifs sont localisés dans le polygone englobant de la signature correspondante. Ainsi, ces pixels faux positifs viennent d'imprécisions autour de la signature, et ne correspondent pas à des signatures faussement détectées. En conséquence, ces erreurs n'ont pas d'impact négatif lors de l'application des règles logiques. Nous illustrons ce point sur la figure 3.12.

### 3.3.3 Règles logiques des actes

À partir des indices visuels extraits, des règles logiques sont appliquées pour grouper ceux qui appartiennent à un même acte. Les règles logiques sont définies grâce au formalisme EPF [Coüasnon 2017].

Le système doit s'adapter à des documents en simple-page ou double-page. La première étape consiste donc à délimiter chaque page dans les documents, à partir des bords de page. Pour cela, le système tente de localiser la pliure entre les pages, par filtrage, ou grâce à l'alignement des lignes de texte. Si aucun séparateur n'est trouvé, le système fait l'hypothèse que l'image ne contient qu'une page.

Des règles logiques sont ensuite appliquées à chaque page pour décrire la mise en page. Une page est définie comme une succession d'actes. Mais un acte peut être défini de différentes façons, selon la variante de la méthode utilisée. Nous décrivons ci-dessous les trois variantes proposées pendant cette thèse.

#### 3.3.3.1 Variante basée sur les signatures

À première vue, il est possible d'envisager une localisation des actes uniquement basée sur la localisation des signatures. En effet, la plupart des actes sont signés, la signature peut donc permettre de repérer la fin des actes. La règle principale pour cette variante stipule qu'un acte est composé d'un groupe de lignes de texte suivi d'une ou plusieurs signatures. La construction d'un acte se fait en localisant une signature `S`, et en groupant toutes les lignes de texte `TL` localisées au-dessus de cette signature Le pseudo-code de cette règle principale est fourni ci-dessous.

```
record := AT(topPage) &&
          signature S &&
          AT(above S) &&
```

```
          textLines TL.
```

Pour chaque page, l'analyse s'effectue de haut en bas. Chaque groupe de signature marque la fin d'un acte : toutes les lignes situées au-dessus de ce groupe de signature correspondent à un même acte. Le rectangle englobant ces lignes de texte et la signature de fin permet de localiser l'acte, comme le montre la figure 3.5. Ces indices visuels sont alors retirées de la liste des indices à traiter pour le reste de l'analyse.

D'autres règles sont également définies pour traiter les actes incomplets :

— En début de page, on peut trouver un acte incomplet composé uniquement d'un groupe de signatures `S` (le texte de l'acte étant sur la page précédente) ;

— En fin d'acte, on peut trouver un acte composé uniquement de lignes de texte `TL` (le reste du texte ainsi que la signature étant sur la page suivante).

Cette variante a deux limites principales. D'une part, tous les actes ne sont pas signés, en particulier avant le milieu du XVII$^{e}$ siècle. D'autre part, les potentielles erreurs de segmentation des signatures se propagent au niveau des actes : une signature manquée produit une fusion de deux actes, une signature faussement détectée produit un acte coupé en deux.

#### 3.3.3.2 Variante basée sur les premières lignes

Une deuxième façon de faire consiste à se baser uniquement sur la localisation des premières lignes de texte. La règle principale stipule qu'un acte est composé d'une première ligne de texte `FTL` suivi d'un groupe de lignes de texte `TL` qui s'étend jusqu'à la première ligne de texte suivante `FTL_next`. Le pseudo-code de cette règle principale est fourni ci-dessous :

```
record := AT(topPage) &&
          firstTextLine FTL &&
          AT(under FTL) &&
          firstTextLine FTL_next &&
          AT(between FTL FTL_next) &&
          textLines TL.
```

La plus haute première ligne de texte `FTL` permet de délimiter le haut d'un acte. Le système recherche alors la première ligne suivante `FTL_next`, afin de localiser la fin de l'acte. Toutes les lignes de texte `TL` situées entre ces deux éléments sont groupées, et leur boite englobante forme l'acte courant.

D'autres règles sont également définies pour traiter les actes incomplets :

— En début de page, on peut trouver un acte incomplet composé uniquement de lignes de texte (la première ligne étant sur la page précédente) ;

— En fin d'acte, on peut trouver un acte avec uniquement une première ligne (le reste du texte étant sur la page suivante).

La principale limite de cette approche vient des potentielles erreurs de segmentation des premières lignes de texte. Une première ligne manquée produit une fusion de deux actes, une première ligne faussement détectée produit un acte coupé en deux.

#### 3.3.3.3 Variante mixte

La dernière façon de faire consiste à combiner les informations sur la localisation des premières lignes de texte et des signatures pour localiser la séparation entre les actes. Dans ce cas, la règle principale stipule qu'un acte est composé d'une première ligne de texte suivi d'un groupe de lignes de texte et d'un groupe de signatures. Le pseudo-code de cette règle principale est fourni ci-dessous :

```
record := AT(topPage) &&
          firstTextLine FTL &&
          AT(under FTL) &&
          signature S &&
          AT(between FTL S) &&
          textLines TLS.
```

Tous les autres cas particuliers sont traités à l'aide de règles secondaires. En particulier, ces règles permettent de corriger les erreurs fréquentes de prédiction des signatures et des premières lignes de texte. Ces règles ont été conçues pour obtenir un compromis entre les erreurs de division et de fusion. Ces règles se basent principalement sur les premières lignes de texte pour délimiter les actes, car elles sont prédites avec plus de précision que les signatures. Cependant, dans certaines configurations, les signatures sont également utilisées pour trouver la fin de l'acte.

Dans cette section, nous avons présenté notre contribution originale : la méthode hybride. Dans la suite, nous évaluons cette méthode sur deux bases de données, puis nous comparons cette approche avec une méthode complètement neuronale : Mask-RCNN.

### 3.3.4 Évaluation de l'approche hybride

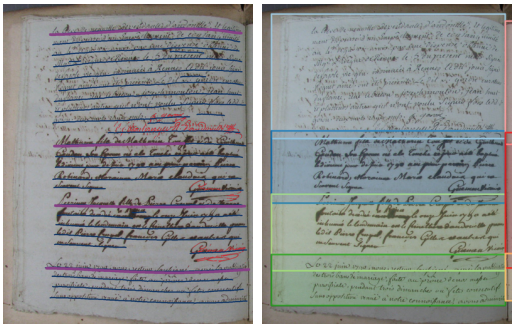
Dans cette section, nous évaluons les trois variantes hybrides proposées. Dans un premier temps, nous évaluons la reconnaissance automatique des indices visuels, puis nous comparons les trois variantes de règles pour la reconnaissance des actes.

#### 3.3.4.1 Évaluation de la localisation d'actes

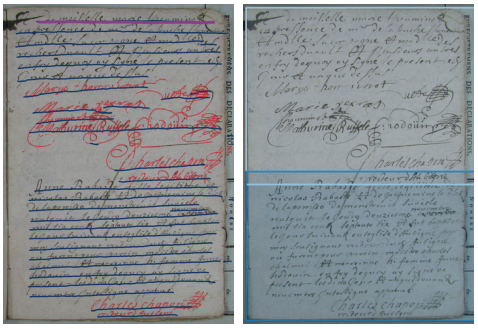
Dans cette section, nous évaluons les différentes variantes de la méthode hybride et nous montrons l'intérêt d'utiliser la variante mixte pour fiabiliser la localisation des actes sur la base DLA-BMS-1.

Nous avons présenté au début de ce chapitre quelques exemples pour lesquels la combinaison des signatures et des lignes de texte est pertinente. En effet, combiner ces indices visuels permet de dépasser deux limites principales. La première limite vient du fait que certains actes ne sont pas signés, ainsi les signatures ne suffisent pas à séparer les actes. La seconde limite vient des potentielles erreurs de reconnaissance : les signatures et premières lignes de texte peuvent être oubliées ou faussement détectées par le réseau de neurones. La figure 3.13 illustre quelques cas où la combinaison des indices visuels permet de séparer correctement les actes. La figure 3.13a présente une configuration parfaite : chaque acte commence par une première ligne de texte et finit par une signature ou la fin de la page. Tous les indices visuels sont correctement identifiés. Dans cette configuration, les trois variantes permettent de localiser les actes de manière efficace. Dans la figure 3.13b, la première ligne de texte du deuxième acte est oubliée. Dans ce cas, l'acte est détecté grâce à la signature du premier acte qui est suffisamment grande pour être considérée comme fiable. Ainsi, la variante mixte et celle basée sur les signatures permettent de séparer les actes de cette page. En revanche, la variante basée sur les premières lignes ne localise qu'un acte (erreur de fusion). Dans la figure 3.13c, les actes ne sont pas signés par le prêtre. Dans ce cas, les premières lignes de texte sont utilisées pour délimiter les actes. Ainsi la variante mixte et celle basée sur les premières lignes de texte permettent de localiser correctement les actes, contrairement à celle basée sur les signatures. Enfin, dans la figure 3.13d, une signature est faussement détectée au milieu à droite du dernier acte. Dans ce cas, la variante mixte permet d'ignorer cette signature, car elle n'est pas suffisamment fiable et est ignorée.

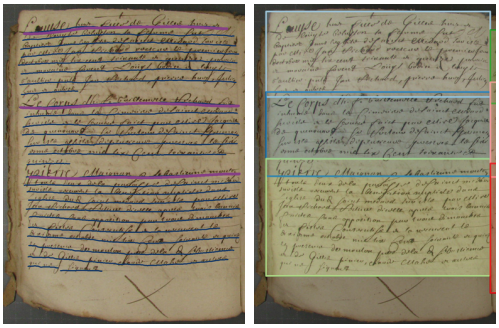
Les résultats présentés dans la table 3.6 confirment l'intérêt de combiner ces deux indices visuels pour la localisation des actes, car la variante mixte surpasse les autres



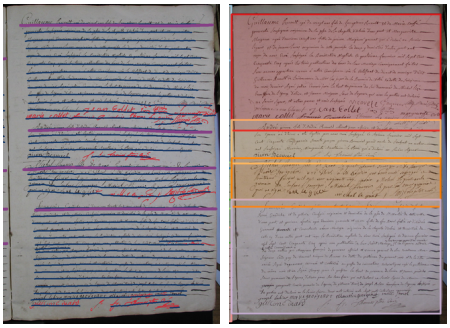
**(a)** Configuration principale : un acte est composé d'une première ligne de texte, de lignes de texte, et d'une signature.



**(b)** La première ligne du deuxième acte est oubliée : la signature est utilisée pour localiser la fin du premier acte.



**(c)** Signature manquante : les actes ne sont pas signés, ils sont alors localisés par les premières lignes de texte.



**(d)** Signature faussement détectée : la signature est ignorée en raison de sa petite taille et de l'absence de première ligne.

**Figure 3.13** – Illustration des configurations principales. Pour chaque sous-figure, l'image de gauche présente les indices visuels localisés par les réseaux de neurones (premières lignes de texte en violet, lignes de texte en bleu, signatures en rouge). L'image de droite présente les actes localisés par les règles logiques.

**Table 3.6** – Évaluation quantitative des différentes variantes de la méthode hybride pour la localisation d'acte lorsque 120 images sont utilisées pour l'apprentissage.

| Variante | ZoneMap ↓ | AP@0.50 ↑ | AP@0.75 ↑ |
|---|---|---|---|
| Signatures | 32.1 | 78.8 | 49.1 |
| Premières lignes | 28.4 | 87.2 | 54.4 |
| **Mixte** | **27.1** | **89.5** | **69.8** |

variantes. Si l'on regarde la métrique $AP@0.75$, le gain apporté par la variante mixte atteint 28% par rapport à la variante basée sur les premières lignes, et 42% par rapport à la variante basée sur les signatures.

## 3.4 Comparaison entre l'approche hybride et les approches neuronales

Nous comparons à présent les deux types de stratégies pour la localisation des actes : les réseaux de neurones de détection d'objets et l'approche hybride mixte. Ces deux approches sont comparées sur trois bases de données : les bases DLA-BMS-2, et la base publique Esposalles [Romero 2013]. L'évaluation sur ces différentes bases de données permet d'évaluer les capacités de généralisation des deux stratégies.

### 3.4.1 DLA-BMS-1

Dans cette section, nous focalisons la comparaison de ces approches sur la base de données DLA-BMS-1, présentée dans le chapitre 1. L'utilisation de la validation croisée nous permet de présenter des résultats en test sur les 200 images.

Les résultats permettant de comparer les deux stratégies sont présentés dans la table 3.7. Les deux méthodes produisent des résultats acceptables sur cette base de données. Si l'on regarde les erreurs de surface, la méthode hybride obtient un meilleur score ZoneMap, mais Mask R-CNN obtient un meilleur score d'AP. L'analyse de ces chiffres ne permet pas de déterminer la meilleure approche. En revanche, lorsque l'on s'intéresse aux erreurs de mise en correspondance des actes hypothèses et références, nous constatons que la méthode hybride obtient un nombre de mise en correspondance très correct (1401 pour la méthode hybride contre 1293 pour Mask R-CNN, sur un total de 1565 actes références). Mask R-CNN semble produire plus d'erreurs de fusions, en particulier sur les actes les plus petits. A l'opposé, la méthode hybride semble produire plus d'erreurs de division, ce qui peut être lié à des faux positifs au niveau des indices visuels. Pour les deux méthodes, les fausses détections apparaissent principalement sur des pages où les prêtres rédigeaient des paragraphes pour décrire les registres : ces paragraphes ne sont pas annotés comme des actes, mais ils y ressemblent fortement. Enfin, certains actes écrits sur la page suivante ou précédente sont parfois faussement détectés, car ils apparaissent en transparence. Si les deux stratégies obtiennent une précision acceptable, la méthode hybride obtient un

meilleur rappel, ce qui la place légèrement au-dessus de Mask R-CNN.

**Table 3.7** – Évaluation quantitative des résultats pour la localisation d'actes sur la base de données DLA-BMS-1 (1565 actes) lorsque 120 images sont utilisées pour l'apprentissage.

**(a)** Erreurs de surface

| | Mask R-CNN | Méthode hybride |
|---|---|---|
| ZoneMap | 29.1 | **27.1** |
| AP@0.5 | **89.6** | 89.5 |
| AP@0.75 | **73.9** | 69.8 |

**(b)** Erreurs de mise en correspondance

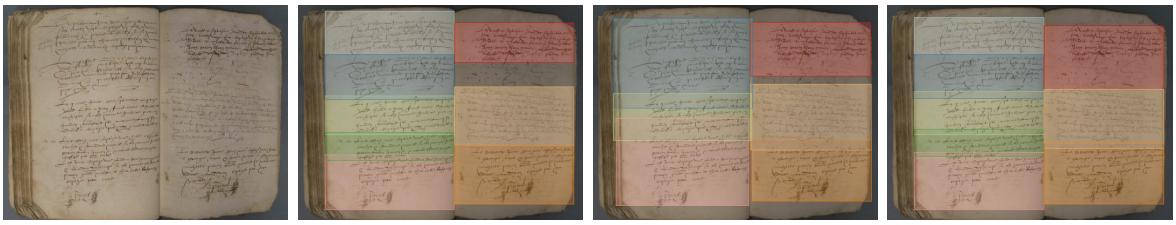
| | Mask R-CNN | Méthode hybride |
|---|---|---|
| Correct | 1293 | **1401** |
| Division | **40** | 71 |
| Fusion | 107 | **42** |
| Fausse détection | 22 | **16** |
| Oubli | 2 | **1** |
| Précision | 0.86 | **0.87** |
| Rappel | 0.83 | **0.90** |
| F1-score | 0.84 | **0.88** |

### 3.4.2 DLA-BMS-2

Dans cette section, les deux approches sont apprises sur la base DLA-BMS-1, puis évaluées sur la base DLA-BMS-2. Ces deux bases sont présentées dans le chapitre 1. Cette évaluation permet d'évaluer les systèmes sur des documents provenant de périodes temporelles non représentées dans l'ensemble d'apprentissage. En particulier, les registres les plus anciens sont difficiles à reconnaître, car les actes sont plus courts et la présence de signatures n'est pas systématique. De plus, le style d'écriture et la langue évoluent avec le temps, ainsi ces actes apparaissent différents.

La figure 3.14 permet de comparer les prédictions faites par ces deux méthodes de façon qualitative. Ce registre date du XVI$^{e}$ siècle, période non représentée dans la base d'apprentissage. Les rectangles prédits par Mask R-CNN sont peu précis : les deux premiers et les deux derniers actes de la page de gauche sont fusionnés. Les actes correctement détectés ont des boites englobantes trop hautes, avec une large superposition entre les actes

successifs. En revanche, la méthode hybride localise correctement tous les actes. Seul le premier acte de la page de droite est rallongé en raison du texte visible par transparence.



**(a)** Image orginale **(b)** Vérité terrain **(c)** Mask R-CNN **(d)** Méthode hybride

**Figure 3.14** – Comparaison qualitative des systèmes sur un registre du XVI$^{e}$ siècle provenant de la base de données DLA-BMS-2.

Nous avons également comparé les performances de ces systèmes avec un nombre de documents d'apprentissage variable. L'objectif est d'identifier la méthode permettant de généraliser correctement à partir de peu d'exemples d'apprentissage, car c'est la principale difficulté rencontrée dans notre contexte industriel. Les systèmes sont évalués sur la base DLA-BMS-2, lorsqu'ils sont entraînés sur 60, 120 ou 180 documents de la base DLA-BMS-1. Les résultats sont présentés dans le tableau 3.8.

**Table 3.8** – Évaluation des systèmes sur les 2143 actes de la base DLA-BMS-2. Pour chaque système, les performances sont évaluées avec un nombre de documents d'apprentissage variable.

| | | Mask R-CNN | | | Méthode hybride | | |
|---|---|---|---|---|---|---|---|
| Images d'apprentissage (DLA-BMS-1) | | 60 | 120 | 180 | 60 | 120 | 180 |
| Surface | ZoneMap | 23.0 | 20.2 | 17.9 | 14.7 | 14.1 | **13.7** |
| | AP@0.5 | 80.1 | 77.4 | 82.0 | 83.3 | **84.0** | 83.1 |
| | AP@0.75 | 52.3 | 59.0 | **63.8** | 59.6 | 62.8 | 61.1 |
| Correspondance | Correct | 1576 | 1547 | 1653 | 1762 | 1787 | **1794** |
| | Division | 54 | **14** | 21 | 80 | 56 | 67 |
| | Fusion | 205 | 224 | 183 | 122 | 121 | **105** |
| | Fausse détection | **2** | **2** | 4 | 4 | 3 | 5 |
| | Oubli | 5 | 5 | 4 | **1** | **1** | 4 |
| | Précision | 0.83 | 0.86 | **0.88** | 0.85 | **0.88** | **0.88** |
| | Rappel | 0.74 | 0.72 | 0.77 | 0.82 | 0.83 | **0.84** |
| | F1-score | 0.78 | 0.78 | 0.82 | 0.84 | **0.86** | **0.86** |

Une baisse importante des performances est observée sur le système de détection d'objets : le F1-score diminue de 0.84 sur la base DLA-BMS-1 à 0.78 sur la base DLA-BMS-2,

en utilisant 120 images d'entraînement. Une explication possible est que les documents de la base DLA-BMS-2 sont plus anciens et présentent des mises en page plus compactes. Par conséquent, les actes successifs sont souvent fusionnés. Si l'augmentation du nombre d'exemples d'entraînement augmente un peu les performances, elle n'est pas suffisante pour rivaliser avec celles obtenues par la méthode hybride. Cela dit, il est très probable que Mask-RCNN s'améliorerait si davantage de documents étaient annotés.

En revanche, les performances de la méthode hybride tendent à rester stables sur les deux bases de données. La principale différence est que le système produit plus d'erreurs de fusion que d'erreurs de division sur la base DLA-BMS-2, alors qu'il produit plus d'erreurs de division que d'erreurs de fusion sur la base DLA-BMS-1. Cette différence peut être liée aux mises en page plus compactes de la base DLA-BMS-2. Les performances augmentent lorsque le nombre de documents d'apprentissage augmente.

Lorsqu'il est entraîné sur le même ensemble d'apprentissage que Mask R-CNN, la méthode hybride produit en moyenne 12% de plus de configurations de correspondance tout en réduisant l'erreur de surface ZoneMap de 30 %. Lorsqu'il est entraîné sur trois fois moins de données que Mask R-CNN, la méthode hybride parvient à produire 7% de plus de bonne mise en correspondance, et permet de réduire le score de ZoneMap de 18%. Ainsi, il serait plus facilement applicable pour le traitement massif des registres paroissiaux français.

### 3.4.3 Esposalles

Enfin, nous évaluons ces deux systèmes sur les registres catalans de la base publique Esposalles [Romero 2013]. Cette base de données est composée de registres de mariages catalans des Archives de la Cathédrale de Barcelone. Chaque image présente un document manuscrit d'une page contenant des actes de mariage. Les actes sont rédigés en catalan ancien et proviennent du XVII$^{e}$ siècle. La base de données est composée de 125 pages : 75 pour l'apprentissage, 75 pour la phase de validation, et 25 pour la phase d'évaluation. Ces registres présentent une disposition différente de celle des registres paroissiaux français : les actes sont plus petits et la langue est différente. De plus, il n'y a pas de signatures, mais un symbole de taxe peut permettre de localiser la fin des actes. Par conséquent, les deux systèmes doivent être appris sur cette base de données. Nous étudions l'influence de la taille de l'ensemble d'apprentissage sur les performances de chaque système. Les résultats des deux stratégies sont ensuite comparés et discutés.

Les deux systèmes sont appris sur des sous-ensembles de la base de données Esposalles,

en utilisant 10, 25, 50 ou 75 exemples. Les scores détaillés sont présentés dans le tableau. 3.9. Ils montrent que de très bonnes performances peuvent être obtenues en utilisant peu de données d'apprentissage sur cette base de données. En effet, les résultats suggèrent que les deux systèmes deviennent efficaces à partir de 25 documents d'apprentissage, ce qui correspond à environ 250 actes. Cet apprentissage rapide peut être expliqué par l'homogénéité des actes, puisqu'il n'y a qu'un seul scripteur. Pour la méthode hybride, les expériences montrent que les symboles de taxe sont correctement détectés à partir de 10 images d'apprentissage, tandis que les premières lignes de texte sont correctement détectées à partir de 25 images d'apprentissage. La figure 3.15 illustre les prédictions des deux méthodes sur une seule image dans cette condition.

**(a)** Image originale **(b)** Vérité terrain **(c)** Mask R-CNN **(d)** Méthode hybride

**Figure 3.15** – Comparaison des deux systèmes sur la base de données Esposalles avec différentes tailles d'apprentissage.

Si les deux systèmes parviennent à reconnaître les documents avec précision, Mask R-CNN parvient à produire des boîtes qui correspondent très bien à la vérité terrain. L'erreur de surface ZoneMap diminue systématiquement lorsque la taille de l'ensemble d'apprentissage augmente, alors qu'elle reste constante pour la méthode hybride. Une explication probable est que la méthode hybride produit des boîtes de délimitation contraintes par des règles. Par conséquent, de petites erreurs de surface subsistent, même avec des prédictions de modèle parfaites. Par opposition, les réseaux de détection d'objets apprennent à s'adapter à chaque acte.

L'évaluation de la mise correspondance montre que les erreurs se produisent sur un petit nombre d'actes. En effet, la grande majorité des actes sont correctement détectés. La méthode hybride produit seulement trois faux positifs, qui correspondent aux trois actes annulés qui apparaissent dans l'ensemble de test. Nous observons très peu d'erreurs de fusion ou d'oublis sur l'ensemble de test. En revanche, il y a plusieurs erreurs de

**Table 3.9** – Évaluation sur les 253 actes de test de la base de données Esposalles.

| | | Mask R-CNN | | | | Méthode hybride | | | |
|---|---|---|---|---|---|---|---|---|---|
| Images d'apprentissage | | 10 | 25 | 50 | 75 | 10 | 25 | 50 | 75 |
| Surface | ZoneMap | 18.5 | 14.6 | 12.8 | **11.9** | 17.1 | 13.9 | 14.3 | 14.1 |
| | AP@0.5 | 96.1 | 98.3 | 98.7 | **99.1** | 95.1 | 97.6 | 97.2 | 98.2 |
| | AP@0.75 | 70.0 | 85.6 | 89.5 | **89.6** | 74.2 | 82.3 | 82.6 | 81.4 |
| Correspondance | Correct | 241 | 249 | **252** | 251 | 240 | 248 | 248 | 249 |
| | Division | 4 | 2 | **1** | 2 | 5 | 6 | 5 | 4 |
| | Fusion | 4 | 1 | **0** | **0** | 1 | **0** | **0** | **0** |
| | Fausse détection | 2 | 5 | 3 | **1** | 3 | 3 | 3 | 3 |
| | Oubli | **0** | **0** | **0** | **0** | 6 | **0** | **0** | **0** |
| | Précision | 0.95 | 0.98 | 0.98 | 0.98 | 0.95 | 0.95 | 0.95 | 0.96 |
| | Rappel | 0.95 | 0.98 | 1.00 | 0.99 | 0.95 | 0.98 | 0.98 | 0.98 |
| | F1-score | 0.95 | 0.98 | 0.99 | 0.98 | 0.95 | 0.96 | 0.96 | 0.97 |

division. Pour Mask R-CNN, il y a également quelques erreurs de division et de fusion, mais elles tendent à disparaître lorsque la taille de l'ensemble d'apprentissage augmente. Les trois actes qui ont été annulés posent problème aux deux systèmes, même lorsqu'ils sont entraînés avec le nombre maximal d'images.

## 3.5 Conclusion du chapitre

Dans ce chapitre, nous avons comparé deux stratégies pour la localisation d'actes : les réseaux de détection d'objets et les méthodes hybrides. Nous avons étudié leur applicabilité dans un contexte de traitement massif de documents, à partir de peu de données d'apprentissage. Pour cela, nous avons comparé différents scénarios d'apprentissage avec un échantillon d'apprentissage de taille variable. Nous avons également évalué ces systèmes sur une base de données complexe de registres paroissiaux contenant des documents issus de périodes temporelles non représentées dans la base d'apprentissage. Enfin, ces méthodes permettent également de reconnaître d'autres types de registres, comme les registres catalans de la base Esposalles [Romero 2013]. Les résultats sont synthétisés sur la figure 3.16

Les réseaux de détection d'objets obtiennent de bonnes performances sur la base de données Esposalles, car les actes sont très homogènes (même scripteur, même longueur, même structure de phrase). En revanche, la performance chute considérablement sur les

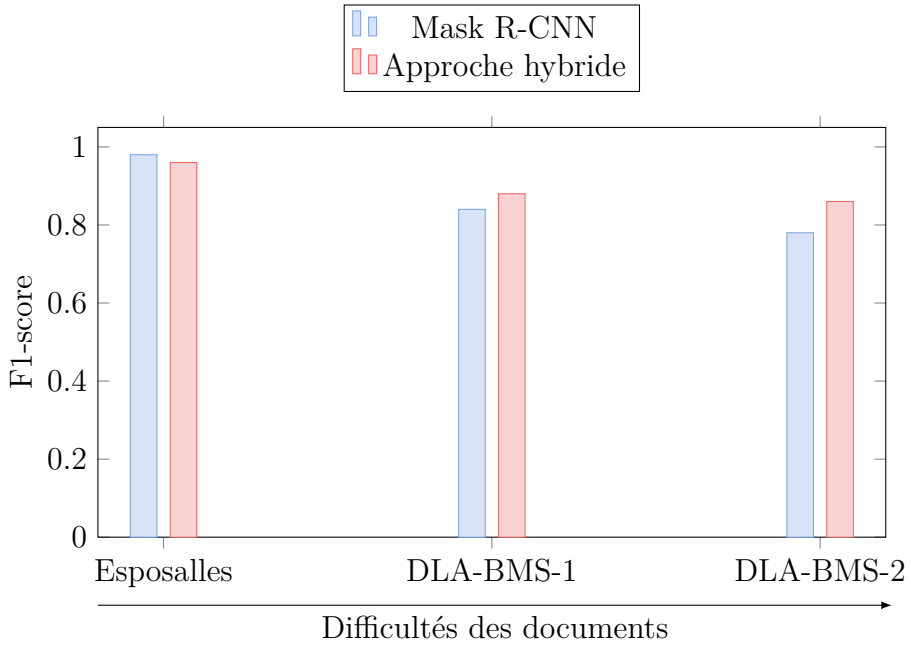



**Figure 3.16** – Le F1-score des deux approches pour chaque base de données. Les deux approches sont apprises sur 25 documents de la base Esposalles, ou 120 documents de la base DLA-BMS-1. La difficulté des documents augmente de gauche à droite. Les résultats montrent que la méthode hybride généralise mieux que Mask R-CNN.

deux bases de données de registres paroissiaux. Les actes les plus courts sont les plus problématiques pour ces réseaux, en particulier lorsque l'écriture est resserrée. Les rectangles englobants prédits sont peu précis, avec une grande superposition entre les actes successifs. Le réseau produit des erreurs récurrentes sur les registres paroissiaux : les actes courts sont fusionnés ou manqués et les grands actes sont coupés en deux. Le modèle Mask R-CNN nécessite plus d'images d'apprentissage afin de capturer la variabilité des actes, notamment avec des registres issus de différentes paroisses et époques. Un point crucial est la variabilité des époques, car les actes les plus anciens sont généralement moins détaillés, moins signés, et donc plus courts.

À l'opposé, les méthodes hybrides reposent sur la localisation d'objets plus simples, comme les signatures et les lignes de texte. Ces objets peuvent être appris à partir d'un nombre limité d'images. Ainsi, cette stratégie est plus robuste pour la détection d'acte. La limite principale de cette méthode vient de sa complexité, car elle nécessite deux étapes d'analyse. De plus, les boites englobantes sont contraintes par des règles, et ne peuvent pas s'adapter à de nouvelles mises en pages. La méthode hybride est applicable sans limites pour la détection d'actes dans les registres. L'utilisation de motifs structurants

aide à simplifier le problème de localisation des actes. En effet, les actes sont des objets bien plus complexes et hétérogènes que les lignes de texte ou les signatures. Nous avons également montré que la combinaison de plusieurs indices visuels permet de fiabiliser la reconnaissance. L'avantage notable de cette approche est qu'elle nécessite peu de données, comparé à Mask R-CNN. En effet, elle permet d'obtenir 12% de configurations correctes supplémentaires, avec trois fois moins de données d'apprentissage. Enfin, cette méthode permet une description de la structure plus riche qu'une simple localisation des actes. En effet, la composition structurelle de chaque acte est également connue : première ligne de texte, lignes de texte. Les règles ont également été enrichies afin de séparer les lignes de texte appartenant à la marge, au texte principal, et aux signatures.

De nombreuses pistes restent à explorer pour la reconnaissance de structure de documents. En particulier, il serait intéressant d'intégrer le contenu textuel des documents dans la segmentation des actes [Boillet 2021a]. Ces mots clés pourraient également permettre de classer chacun des actes selon son type : baptême, mariage, sépulture.

Chapitre 4

# Reconnaissance d'écriture manuscrite

Dans ce chapitre, nous adaptons un réseau de neurones basé sur un mécanisme d'attention pour la reconnaissance d'écriture manuscrite. Cette architecture nous semble intéressante, car elle permet d'extraire des caractéristiques pertinentes pour la prédiction des caractères et la modélisation implicite du langage. Dans un premier temps, nous décrivons les différentes composantes de cette architecture ainsi que son fonctionnement. Nous détaillons ensuite les expérimentations et optimisations réalisées sur l'architecture et les paramètres d'apprentissage. L'architecture retenue est ensuite évaluée sur quatre bases de données publiques, et comparée aux autres approches de l'état de l'art. Enfin, nous discutons de l'applicabilité de cette architecture aux registres paroissiaux, dans un contexte où peu de données d'apprentissage sont disponibles.

## 4.1 Introduction

Les réseaux de neurones s'appuyant sur des mécanismes d'attention [Xu 2016] ont rencontré un succès majeur dans de nombreux domaines connexes à la vision par ordinateur ou au traitement automatique de la langue. Les mécanismes d'attention permettent aux réseaux de neurones d'apprendre à focaliser leur attention sur des caractéristiques d'intérêt pour la décision. L'idée défendue par [Xu 2016] est que ce mécanisme permet de simuler le système visuel humain [Rensink 2000 ; Corbetta 2002]. Ce mécanisme permet de mettre en avant certaines caractéristiques importantes de l'image de façon dynamique, lorsque cela est utile au système. À l'inverse, les systèmes sans mécanisme d'attention compressent l'image en une représentation vectorielle compacte et statique.

Le fonctionnement d'un mécanisme d'attention est illustré dans la figure 4.1, dans le cadre d'une tâche de sous-titrage d'images. Les mécanismes d'attention font partie des innovations majeures en deep learning apparues ces dernières années. En particulier, ils

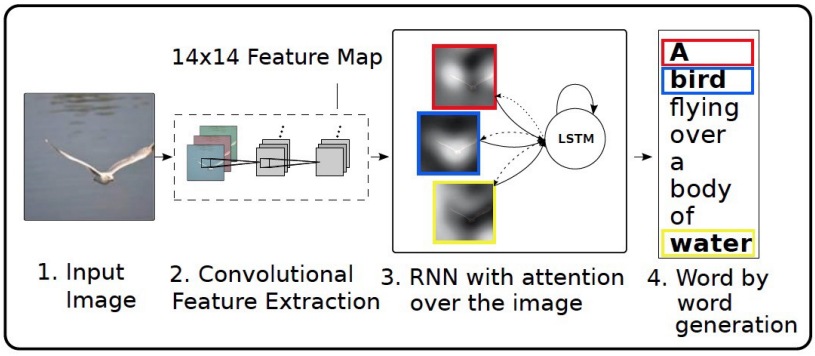



**Figure 4.1** – Illustration du mécanisme d'attention pour le sous-titrage d'image proposé par Xu et al. [Xu 2016]. L'image est encodée par un encodeur convolutif, puis le mécanisme d'attention permet au décodeur de se focaliser sur certaines parties de l'image pour prédire chaque mot. L'attention se focalise sur l'oiseau pour générer le mot « bird », et sur le fond de l'image pour générer le mot « water ».

ont révolutionné les domaines de la traduction automatique [Bahdanau 2014], la reconnaissance de la parole [Chorowski 2015b], le sous-titrage d'images [Xu 2016], ou encore la segmentation sémantique de scènes naturelles [Fu 2019] ou d'images de documents [Soullard 2020]. Outre les améliorations quantitatives, ces mécanismes rendent également les réseaux de neurones plus explicables, car la visualisation des cartes d'attention permet de mieux comprendre le mécanisme de décision.

Les mécanismes d'attention semblent pertinents pour la reconnaissance automatique d'écriture, et plus généralement pour l'extraction d'information dans des images de documents. En effet, ils permettent au modèle de se focaliser sur les zones de l'image présentant le plus d'intérêt pour la prise de décision. Par exemple, un modèle de reconnaissance d'écriture pourra se concentrer sur la zone de l'image correspondant à la lettre à reconnaitre. Cette focalisation du modèle sur les caractéristiques importantes peut également permettre d'atténuer l'impact des éléments de l'image pouvant potentiellement perturber la reconnaissance, comme les ratures, les taches d'encre ou encore les éléments décoratifs. Enfin, une dernière propriété intéressante de cette architecture est la présence d'un décodeur récurrent, qui permet une modélisation implicite d'un langage.

De premières architectures s'appuyant sur des mécanismes d'attention ont récemment été proposées pour la reconnaissance d'écriture manuscrite. Certains travaux de recherche ont exploré l'intérêt d'un mécanisme d'attention à travers des convolutions à portes [1] pour l'extraction de caractéristiques pertinentes [Bluche 2017 ; Yousef 2018 ; Coquenet 2020].

1. *Gated convolutions* en anglais

Plus récemment, d'autres modèles intègrent un mécanisme d'attention au sens de Xu et al. [Xu 2016] pour la reconnaissance d'écriture à partir d'images de mots [Kang 2019b], de lignes [Poulos 2017 ; Michael 2019] ou de paragraphes [Bluche 2016].

Dans ce chapitre, nous adaptons une architecture avec mécanisme d'attention pour la reconnaissance d'écriture manuscrite à partir de lignes de texte.

## 4.2 Description de l'architecture

Nous présentons ici une architecture, adaptée de celle proposée par Xu et al. [Xu 2016], pour la reconnaissance d'écriture manuscrite. L'implémentation est effectuée en utilisant le framework Pytorch. Dans la suite du chapitre, nous menons des expérimentations sur les différentes composantes du modèle afin d'aboutir à l'architecture la plus adaptée à la reconnaissance d'écriture.

### 4.2.1 Architecture

L'architecture seq2seq est un réseau encodeur-décodeur. L'encodeur permet d'encoder l'image dans un vecteur de caractéristique de taille fixe. Le décodeur est ensuite utilisé pour générer le texte associé à l'image, caractère par caractère. Pour chaque pas de temps, le mécanisme d'attention met en évidence une partie du vecteur de caractéristique sur laquelle le décodeur se concentre pour prédire le caractère correspondant. L'architecture est schématisée sur la figure 4.2.

#### 4.2.1.1 Entrée du réseau

Le réseau de neurones prend des images RGB de lignes de texte en entrée. Les images sont redimensionnées afin de permettre un traitement par batch, et normalisées entre 0 et 1. Pour cela, une image noire de taille 1400x80 pixels est créée. L'image de ligne est redimensionnée en conservant son ratio originel, normalisée, puis collée dans l'image noire. Certaines transformations sont ensuite appliquées aléatoirement afin d'augmenter artificiellement la variabilité des images. La résolution, la luminosité, et le contraste de l'image sont modifiés. Une modification lourde peut également être appliquée aléatoirement. Cette transformation est tirée parmi trois types de transformations : un changement de perspective, une transformation affine, ou une déformation élastique sont appliqués.

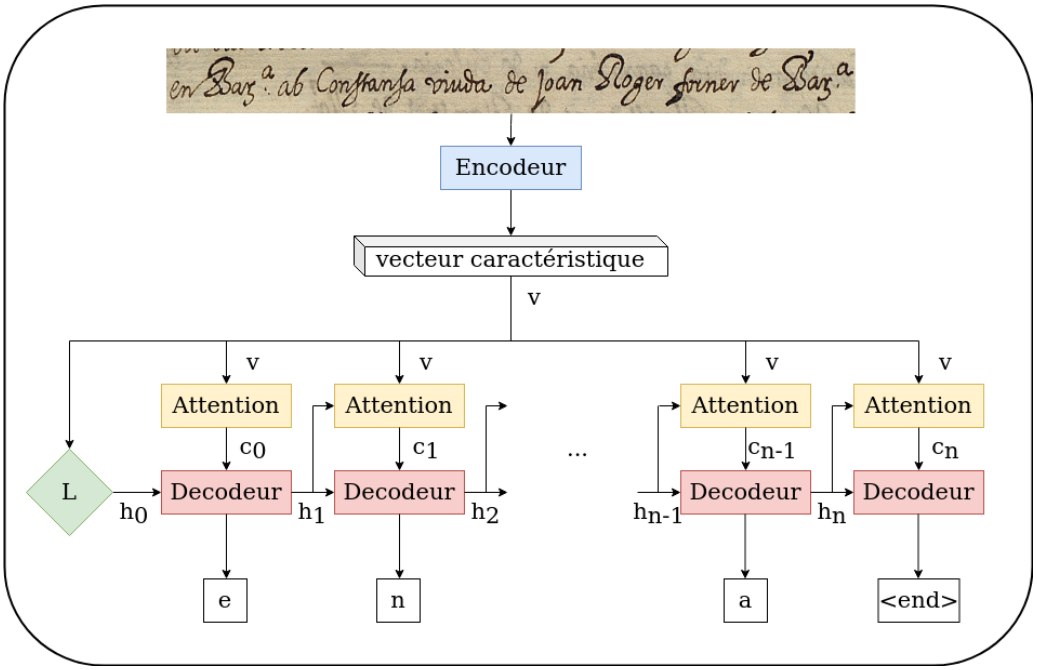



**Figure 4.2** – Architecture seq2seq avec mécanisme d'attention. $L$ est une fonction linéaire, $(c_i)$ est le vecteur de contexte du décodeur et $(h_i)$ est l'état caché du décodeur.

Enfin, des dilatations et érosions sont utilisées afin de grossir ou affiner la taille du trait. Chacune de ces transformations est appliquée avec une probabilité $p = 0.2$.

Le modèle est supervisé grâce au texte associé à l'image. Le texte correspondant à l'ensemble d'apprentissage est analysé afin de calculer le nombre d'occurrences de chaque caractère du dictionnaire. Les caractères apparaissant moins de 10 fois dans l'ensemble d'apprentissage sont remplacés par le token `<unk>`. Deux tokens sont ajoutés à chaque séquence : un premier token `<start>` et un dernier token `<end>`. Chaque séquence est ensuite complétée avec le token `<pad>` jusqu'à être de taille fixe $N$. Cette taille est choisie de façon à être supérieure à la taille de la plus grande séquence de caractères de l'ensemble d'apprentissage ($N = N_{max\_app} + 25$), afin d'être en mesure de générer des séquences suffisamment longues. Le texte est ensuite encodé grâce à une couche d'embedding de dimension 32.

#### 4.2.1.2 Encodeur

Le rôle de l'encodeur consiste à extraire des caractéristiques visuelles à partir des pixels de l'image d'entrée. Le papier original utilise une architecture VGG [Simonyan 2015], pré-entraînée sur la base de données ImageNet, dont les couches linéaires sont retirées afin de

ne garder que la partie convolutive.

Cette architecture étant très profonde, nous choisissons d'entraîner uniquement les dernières couches sur cette nouvelle tâche. Le vecteur de caractéristiques final a une taille 3x45 avec 512 canaux. Enfin, une couche de pooling adaptatif est utilisée afin de changer la taille du vecteur de caractéristiques. Cette étape a un impact sur les performances du système, et notamment sur l'attention. Dans la suite de ce chapitre, nous évaluons d'autres architectures pour l'encodeur. Nous étudions également l'intérêt d'ajouter des couches de récurrences dans la phase d'extraction des caractéristiques.

#### 4.2.1.3 Mécanisme d'attention

Le mécanisme d'attention est situé entre l'encodeur et le décodeur. Il permet au décodeur de se concentrer sur la partie du vecteur de caractéristiques la plus pertinente pour la prise de décision. L'utilisation du mécanisme d'attention améliore la performance de ces réseaux. Nous utilisons ici le modèle d'attention proposée par Bahdanau et al. [Bahdanau 2014]. D'autres modèles sont comparés dans la suite de ce chapitre.

Le mécanisme d'attention permet de calculer un vecteur de contexte $c_t$ pour chaque pas de temps. Ce vecteur est donné en entrée du décodeur, comme illustré sur le schéma 4.2. Ce vecteur de contexte correspond au vecteur caractéristique $v$ pondéré par les scores d'attention et est calculé de la façon suivante, pour chaque pas de temps $t$ :

$$c_t = \sum_{i=1}^{n} \alpha_{t,i} v_i$$

Le score $\alpha_{t,i}$ correspond à la pertinence d'un pixel du vecteur de caractéristique $v_i$ pour la décision prise au pas de temps $t$. Concrètement, une carte de probabilité est calculée afin de pondérer le vecteur de caractéristiques, ce qui permet de mettre en évidence les zones de l'image les plus pertinentes. Pour chaque pas de temps $t$, un score d'alignement est calculé en tout point du vecteur de caractéristique $v$. La carte de probabilité est calculée en passant le score d'alignement dans une fonction softmax.

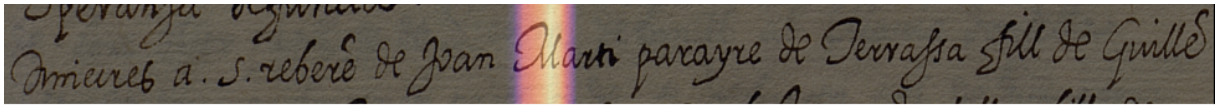
$$\alpha_{t,i} = \text{softmax}(\text{align}(h_{t-1}, v_i) = \frac{\exp(\text{align}(h_{t-1}, v_i))}{\sum_{k=1}^{T_X} \exp(\text{align}(h_{t-1}, v_k))}$$

$$\text{align}(h_{t-1}, v_i) = V_a^T \tanh(W_1 h_{t-1} + W_2 v_i))$$

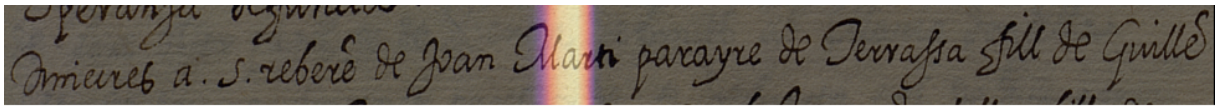
où $W_1$, $W_2$ et $V_a^T$ sont des matrices de poids apprises de façon conjointe avec le modèle

de reconnaissance.

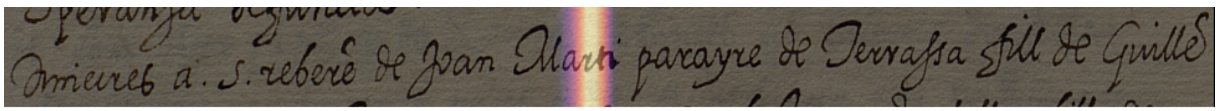
La figure 4.3 illustre une carte d'attention lors de la génération de la séquence prédite. Dans cet exemple, l'attention est focalisée sur la bonne zone de l'image pour prédire la lettre a. Plus généralement, les cartes d'attention se concentrent sur des zones cohérentes pour la génération des caractères. Cet alignement entre l'image et la séquence prédite n'est pas supervisé : aucune segmentation préalable des lettres n'est nécessaire.




**(a)** Étape $t = 29$




**(b)** Étape $t = 30$




**(c)** Étape $t = 31$

**Figure 4.3** – Visualisation des scores d'attention sur une image de la base de données Esposalles [Romero 2013] sur trois étapes consécutives. Le texte en bleu correspond à la transcription vérité complète. Le texte en rouge correspond à la transcription prédite par le réseau à l'étape $t$ - l'attention se concentre sur la zone mise en évidence dans l'image pour prédire la dernière lettre de cette séquence.

#### 4.2.1.4 Décodeur

Le rôle du décodeur est de générer la séquence des caractères à partir du vecteur de caractéristiques pondéré par la carte d'attention. Le décodeur est un réseau récurrent. Pour chaque pas de temps, la sortie du décodeur est passée dans une couche softmax afin d'obtenir les probabilités respectives de chaque caractère de l'alphabet. Le décodage s'appuie sur l'algorithme *beam search*, avec une taille de faisceau de 5. Ainsi, les 5 séquences les plus probables sont sélectionnées à chaque pas de temps $t$. À la fin du décodage, la meilleure séquence parmi les 5 retenues est sélectionnée.

### 4.2.2 Paramètres d'apprentissage

**Fonction de coût** L'architecture proposée par Xu et al. [Xu 2016] est entraînée en minimisant la fonction de coût d'entropie croisée $\mathcal{L}_{CE}$, ou Cross Entropy (CE).

Il a été démontré que l'intégration de la fonction de coût de classification connexionniste temporelle, Connectionnist Temporal Classification (CTC), accélère la convergence du système et améliore ses performances, notamment pour des tâches de reconnaissance de la parole [Kim 2016] et de reconnaissance d'écriture [Michael 2019]. Nous utilisons donc la fonction de coût CTC $\mathcal{L}_{CTC}$ afin d'affiner les poids de l'encodeur.

Un terme de régularisation est ajouté au niveau du mécanisme d'attention, afin d'encourager le modèle à se concentrer sur toutes les zones de l'image durant la génération

$$\mathcal{L}_{reg} = -\lambda \sum_{i}^{L} (1 - \sum_{t}^{T} \alpha_{ti}^2)$$

où $T$ est la taille de la séquence, $L$ le nombre de pixels du vecteur caractéristique et $\alpha_{ti}$ le poids de l'attention du pixel $i$ de la carte d'attention généré au pas de temps $t$.

La fonction de coût finale est une combinaison convexe de la fonction CE et CTC, à laquelle s'ajoute le terme de régularisation du mécanisme d'attention.

$$\mathcal{L} = \beta \mathcal{L}_{CE} + (1 - \beta)\mathcal{L}_{CTC} + \mathcal{L}_{reg}$$

**Convergence** La descente de gradient est effectuée à l'aide de l'algorithme Adam. L'apprentissage s'arrête lorsque la fonction de coût de validation ne diminue pas pendant 20 epochs. Les poids de la meilleure epoch en validation sont retenus pour l'inférence.

**Teacher forcing** Durant l'apprentissage, le *teacher forcing* est utilisé afin d'accélérer et stabiliser la convergence. Au moment de la prédiction de l'étape $t$, la prédiction à l'étape $t-1$, la vérité terrain $t-1$ est utilisée. L'inconvénient de cette approche est qu'elle simplifie grandement le problème : elle empêche ainsi le décodeur d'apprendre à partir de ses propres erreurs, ce qui provoque un biais au moment de l'inférence. Pour résoudre ce problème, nous utilisons du *teacher forcing* aléatoirement pendant l'apprentissage, avec une probabilité $p = 0.5$.

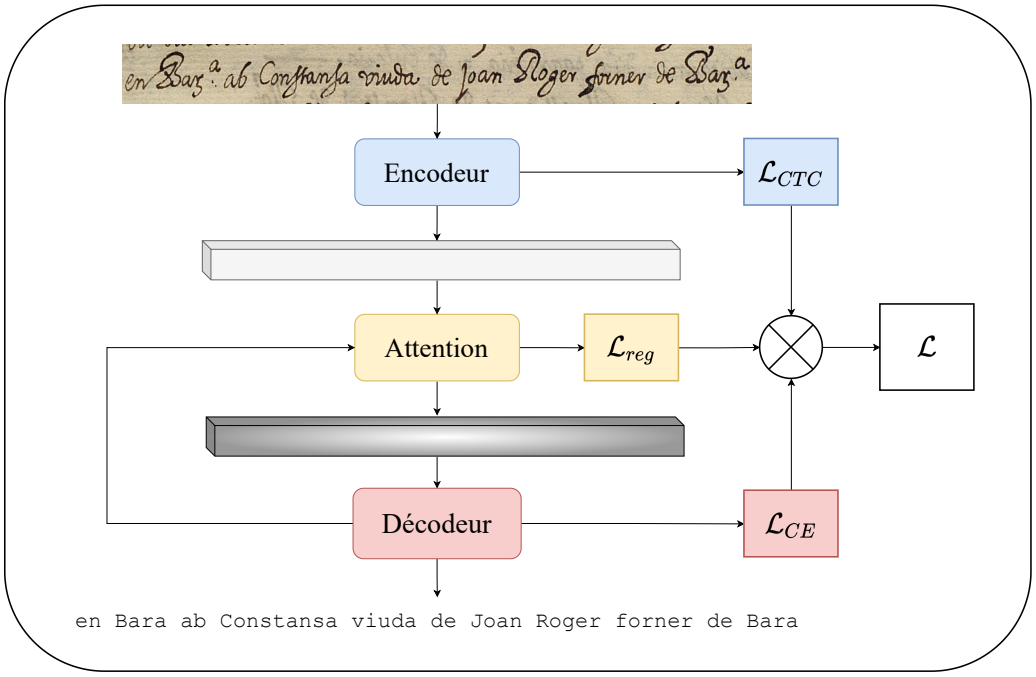



**Figure 4.4** – Fonction de coût hybride pour l'apprentissage du modèle seq2seq. La fonction est une combinaison des fonctions de coût CE et CTC, à laquelle s'ajoute un terme de régularisation de l'attention.

**Dropout** Enfin, le *dropout* est utilisé comme méthode de régularisation dans l'encodeur et dans le décodeur.

## 4.3 Expérimentations

Dans cette section, nous réalisons certaines expérimentations sur la base de données IAM [Marti 2002] afin d'améliorer l'architecture et les paramètres d'apprentissage pour la tâche de reconnaissance d'écriture manuscrite.

### 4.3.1 Architecture

Dans un premier temps, nous réalisons des optimisations sur l'architecture du réseau.

#### 4.3.1.1 Encodeur

Trois encodeurs génériques pré-entraînés sur ImageNet [Deng 2009] sont comparés : ResNet-50, ResNet-101 et VGG-16. Seules les dernières couches de convolutions sont spé-

**Table 4.1** – Expérimentations menées sur l'architecture de l'encodeur. La base de l'encodeur est un réseau pré-entraîné sur la base ImageNet [Deng 2009]. Seules quelques couches de convolution sont entraînées sur la base de données IAM [Marti 2002] afin de réduire le nombre de paramètres. L'impact des couches récurrentes est également évalué : elles permettent d'améliorer les scores de reconnaissance.

**(a)** Paramètres des différents modèles de base

| Encodeur | Paramètres (total) | Paramètres (entraînables) |
|---|---|---|
| ResNet-50 | 25M | 2.5M |
| ResNet-101 | 44M | 2.5M |
| VGG-16 | 138M | 2.4M |

**(b)** Scores des différentes architectures d'encodeurs

| Encodeur | CER (%) ⬇ | WER (%) ⬇ |
|---|---|---|
| VGG-16 | 48.83 | 98.05 |
| ResNet-50 | 10.34 | 35.59 |
| ResNet-101 | 7.10 | 26.99 |
| ResNet-101 + GRU (2 couches) | 7.49 | 29.25 |
| ResNet-101 + Bi-GRU (2 couches) | 5.52 | 23.14 |
| ResNet-101 + LSTM (2 couches) | 8.28 | 30.82 |
| ResNet-101 + Bi-LSTM (2 couches) | **5.22** | **21.71** |
| ResNet-101 + Bi-LSTM (1 couche) | 6.63 | 26.36 |
| ResNet-101 + Bi-LSTM (3 couches) | 8.03 | 30.88 |

cialisées pour la reconnaissance d'écriture, afin de réduire le nombre de paramètres à apprendre à environ 2.5M. Pour le réseau VGG, seule la dernière couche est spécialisée. Pour les deux architectures ResNet, le dernier bloc de convolution est spécialisé, ce qui correspond aux trois dernières couches de convolution. Les résultats du tableau 4.1 montrent que l'architecture ResNet-101 permet d'améliorer considérablement les performances du modèle. Deux facteurs peuvent expliquer ce gain considérable. D'une part, seule une couche de l'architecture VGG est spécialisée : les caractéristiques sont donc probablement peu adaptées à la reconnaissance d'écriture. D'autre part, ResNet permet un encodage plus riche, avec 2018 canaux au lieu de 512 pour le vecteur de caractéristiques.

Nous étudions également l'impact des blocs de récurrence ajoutés après le réseau convolutif. Au préalable, le vecteur de caractéristique est aplati en utilisant un *average pooling* adaptatif. Les résultats montrent que l'ajout d'un réseau BLSTM à la suite de ResNet-101

améliore les résultats : un gain de 2% sur le taux d'erreur caractère, et de 5% sur le taux d'erreur mot. Cette configuration est sélectionnée pour la suite de ce travail.

#### 4.3.1.2 Décodeur

Nous étudions également l'architecture du décodeur, en particulier nous comparons l'utilisation de cellule LSTM ou GRU, dont une comparaison détaillée a été proposée par Chung et al. [Chung 2014]. Leur étude met en évidence que les GRUs et les LSTMs permettent d'obtenir des résultats comparables. Dans le cas de notre architecture, les résultats du tableau 4.2 montrent une amélioration non négligeable des résultats avec la cellule LSTM. C'est donc cette configuration qui est retenue pour la suite de notre étude.

**Table 4.2** – Expérimentations sur l'architecture du décodeur sur la base de données IAM

| Décodeur | CER (%) ↓ | WER (%) ↓ |
|---|---|---|
| LSTM | **5.22** | **21.71** |
| GRU | 6.91 | 27.26 |

### 4.3.2 Mécanisme d'attention

En plus de l'attention *additive*, proposée par [Bahdanau 2014], nous étudions les scores d'alignement proposés par Luong et al. [Luong 2015b] et Chorowski et al. [Chorowski 2015b]. Nous proposons également un score pénalisé pour la reconnaissance d'écriture.

**Attention multiplicative** L'attention multiplicative de Luong et al. [Luong 2015b] a été proposée pour une tâche de traduction automatique, et a depuis été reprise dans les réseaux Transformers [Vaswani 2017]. Le score d'alignement mesure une similarité entre le vecteur de caractéristique et l'état caché du décodeur :

$$\text{align}_L(h_t, v_i) = h_t^T v_i$$

**Attention hybride** L'attention hybride de Chorowski et al. [Chorowski 2015b] a été proposée pour une tâche de reconnaissance de la parole. Elle étend la fonction d'alignement de Bahdanau et al. [Bahdanau 2014] en ajoutant une information sur la localisation de l'attention à l'étape précédente. Pour cela, le score d'attention de l'étape précédente est

convolué par une matrice $F$ dont les poids sont appris par le réseau.

$$f_{i,t} = F \times \alpha_{i-1}$$

Le score d'alignement est ensuite défini de la façon suivante :

$$\text{align}_C(s_t, h_i) = w^T \tanh(Wh_{t-1} + Vv_i + Uf_{i,t} + b)$$

où $w$ et $b$ sont des vecteurs et $W$, $V$ sont des matrices.

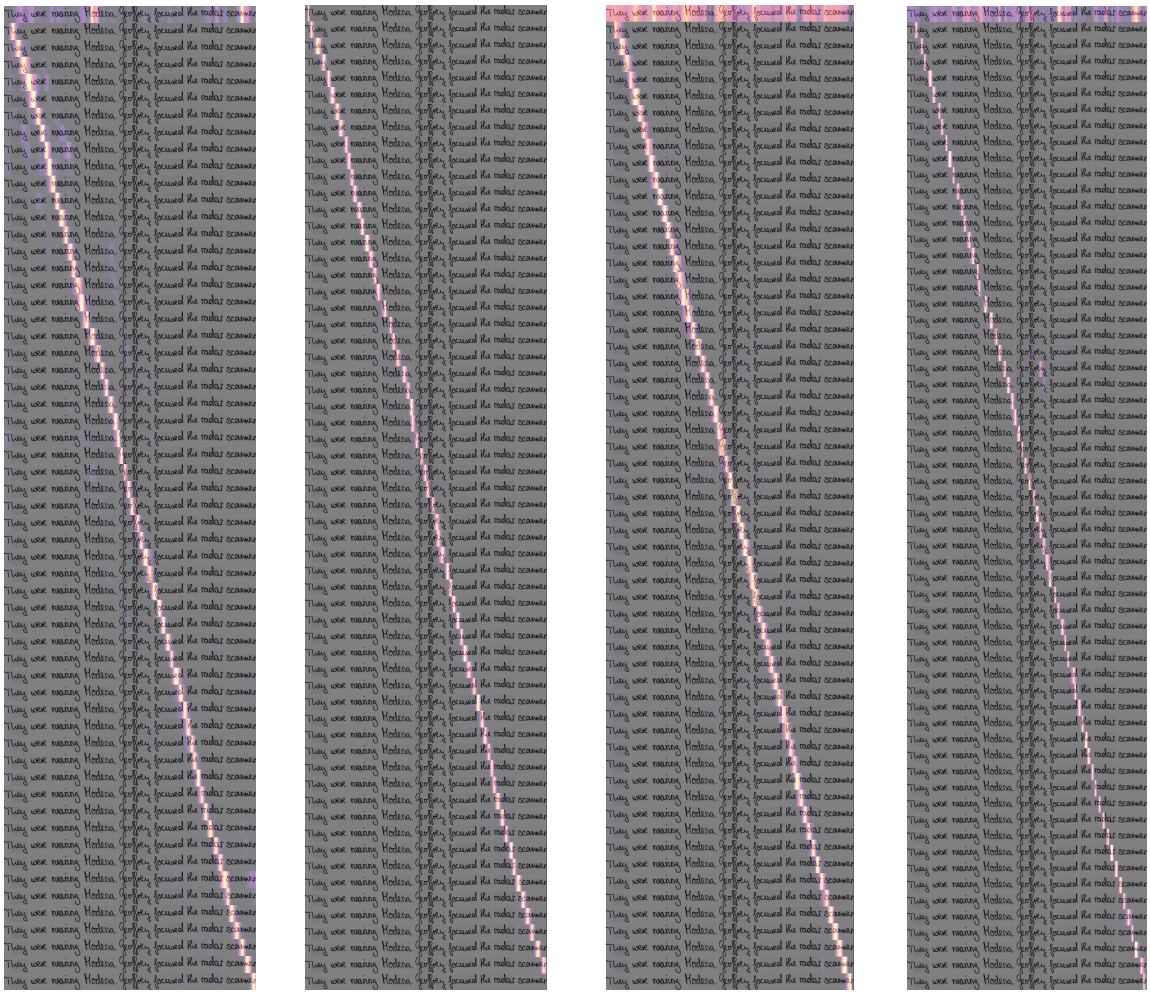
**Attention hybride pénalisée** Nous proposons également une variante d'attention *pénalisée*, adaptée à la reconnaissance d'écriture. La pénalisation force l'attention à bouger de sa position précédente. Nous proposons un mécanisme d'attention pénalisée pour la reconnaissance d'écriture latine, basé sur l'attention de Chorowski et al. [Chorowski 2015b]. Puisque l'écriture latine se lit de gauche à droite, nous proposons de pénaliser fortement le texte qui a déjà été lu par le réseau, afin de le forcer à se concentrer sur une zone qui n'a pas encore été vue.

$$\alpha_{t,i} = \text{softmax}(\text{align}_C(s_t, h_i) \times (1 - \text{align}_C(s_{t-1}, h_i)))$$

**Comparaison des mécanismes d'attention** Les résultats quantitatifs sont présentés dans la table 4.3. Le mécanisme d'attention le moins effectif semble être l'attention multiplicative. La convergence de l'architecture est fortement perturbée par ce mécanisme d'attention. L'observation des cartes d'attention montre que le réseau se focalise sur des zones incohérentes.
Au contraire, l'attention hybride permet d'obtenir les meilleurs résultats, avec un gain de 1% sur le CER par rapport à l'attention additive. Les cartes d'attention pour les attentions hybride, additive et hybride pénalisée se ressemblent : le réseau se concentre sur des zones cohérentes.
Enfin, l'attention hybride pénalisée n'apporte rien de plus au réseau : l'attention se déplace de façon cohérente de gauche à droite, et ce, sans pénalisation.

### 4.3.3 Fonction de coût hybride

Dans le tableau **??**, nous avons souhaité comparer l'architecture seq2seq avec l'architecture CRNN-CTC correspondante. Pour cela, il suffit de fixer $\alpha = 0$ dans la fonction de



**(a)** Additive **(b)** Multiplicative **(c)** Hybride **(d)** Hybride pénalisée

**Figure 4.5** – Visualisation des différents types d'attention. La séquentialité de l'analyse est représentée de haut en bas. Pour chaque caractère, les zones d'attention apparaissent en rose. La diagonale montre que le réseau est capable de focaliser son attention sur des zones cohérentes au fur et à mesure de l'analyse.

**Table 4.3** – Comparaison des différents types de mécanismes d'attention sur la base de données IAM

| Type d'attention | CER (%) ⬇ | WER (%) ⬇ |
|---|---|---|
| Multiplicative [Luong 2015b] | 7.15 | 30.19 |
| Additive [Bahdanau 2014] | 6.15 | 25.43 |
| Hybride [Chorowski 2015b] | **5.39** | **23.00** |
| Hybride pénalisée | 5.50 | 23.39 |

coût hybride, afin d'entraîner et évaluer l'encodeur uniquement.
Nous évaluons également l'architecture seq2seq apprise avec la fonction de coût hybride et $\alpha = 0.5$. Dans cette configuration, l'encodeur et le décodeur peuvent chacun être évalués indépendamment [Michael 2019]. Enfin, nous évaluons cette architecture sans utiliser la fonction de coût CTC, soit avec $\alpha = 1$.

Il est intéressant d'observer que l'évaluation de l'encodeur permet d'obtenir un meilleur taux de reconnaissance caractère, alors que l'évaluation du décodeur permet d'obtenir un meilleur taux de reconnaissance au niveau des mots. Dans les deux configurations, le modèle seq2seq appris avec la fonction de coût hybride est plus performant que la configuration CRNN-CTC, qui correspond à l'architecture classique pour la reconnaissance d'écriture. Cela est probablement dû au mécanisme d'attention et au modèle linguistique implicite appris par le décodeur.

### 4.3.4 Intégration d'un modèle de langue

La dernière expérimentation menée sur cette architecture concerne l'intégration d'un modèle de langue explicite. Pour cela, nous avons évalué l'architecture avec et sans la connaissance d'un modèle de langue 3-gram. La fusion des connaissances du modèle seq2seq et du modèle de langue 3-gram se fait au niveau dans la phase de décodage.

D'une part, le décodeur produit un vecteur de probabilité par caractères $p_{dec}$. D'autre part, le modèle 3-gram modélise la fréquence des triplets de caractères observés dans l'ensemble d'apprentissage. Il est donc capable de produire un vecteur de probabilités par caractère $p_{gram}$, en fonction des deux derniers caractères prédits par le réseau. Ces probabilités sont fusionnées $p = (1 - \beta_{lm}) \times p_{dec} + \beta_{lm} \times p_{lm}$, à l'exception de celle du token `<end>` qui est très peu probable dans le modèle N-gram.

Les résultats, présentés dans la table 4.5, ne sont pas concluants. En effet, le réseau seq2seq obtient de meilleures performances sans modèle de langue explicite. Ces résul-

**Table 4.4** – Évaluation quantitative de la reconnaissance d'écriture manuscrite sur quatre bases de données, sans post-processing. Le tableau permet de comparer les performances d'un CRNN-CTC et d'un seq2seq basés sur le même encodeur. Le seq2seq appris avec la loss hybride peut être évalué au niveau de l'encodeur (CTC) ou du décodeur (CE).

| Base de données | IAM | READ | RIMES | Esposalles |
|---|---|---|---|---|
| | | CER (%) ⬇ | | |
| Encodeur (CTC) | 6.06 | 14.53 | 5.18 | **1.61** |
| Décodeur (CE) | 8.72 | 20.38 | 7.59 | 2.62 |
| Encodeur (hybrid) | **5.13** | **12.10** | **3.29** | 2.36 |
| Décodeur (hybrid) | 5.22 | 12.70 | 4.35 | 2.35 |
| | | WER (%) ⬇ | | |
| Encodeur (CTC) | 27.68 | 48.02 | 27.94 | **6.75** |
| Decoder (CE) | 32.90 | 57.01 | 23.73 | |
| Encodeur (hybrid) | 24.30 | 43.42 | 20.33 | 11.33 |
| Décodeur (hybrid) | **23.00** | **41.66** | **18.93** | 7.90 |

tats ont également été observés par Kang et al. [Kang 2019a]. Dans leur article, les auteurs proposent trois façons de fusionner les informations d'un modèle seq2seq et du modèle de langue. Si les deux premières méthodes proposées ne parviennent pas à améliorer les résultats, leur troisième proposition « Candidate fusion » permet un gain de performance. Cependant, ce gain reste négligeable. En revanche, ils montrent que le score peut être amélioré de façon plus significative par l'utilisation d'un dictionnaire fermé en post-processing.

| Coefficient $\beta_{lm}$ | N-gram | CER (%) ⬇ | WER (%) ⬇ |
|---|---|---|---|
| 0 | - | **5.22** | **23.00** |
| 0.5 | 3-gram | 8.38 | 34.84 |
| 0.3 | 3-gram | 7.39 | 30.49 |
| 0.1 | 3-gram | 7.16 | 29.05 |

**Table 4.5** – Résultats quantitatifs de l'intégration d'un modèle de langue à l'architecture seq2seq. Les résultats montrent que l'architecture seq2seq est plus performante sans modèle de langue explicite.

### 4.3.5 Comparaison à l'état de l'art

L'architecture retenue est composée d'un encodeur convolutif récurrent (ResNet-100 combiné à un réseau Bi-LSTM), d'un mécanisme d'attention hybride et d'un décodeur LSTM. L'apprentissage se fait avec une fonction de coût hybride. Dans un premier temps, nous démontrons l'intérêt de cette architecture seq2seq pour la reconnaissance d'écriture. L'évaluation est réalisée sur quatre bases de données publiques au niveau des lignes : IAM [Marti 2002] (moderne, anglais) et RIMES [Augustin 2006] (moderne, français). Nous comparons notre architecture avec d'autres méthodes à l'état de l'art, sans post-traitement ni modèle de langue, dans le tableau. 4.6. Notre architecture finale obtient des résultats compétitifs sur la base de données IAM.

**Table 4.6** – Comparaison avec les méthodes de l'état de l'art pour la reconnaissance d'écriture (CER %) à l'échelle des lignes de texte, sur l'échantillon de test, et sans modèle de langue ou post-processing.

| System | Method | IAM | RIMES |
|---|---|---|---|
| CRNN-CTC | Wigington et al. [Wigington 2018b] | 6.4 | **2.1** |
| CRNN-CTC | Puigcerver [Puigcerver 2017] | 5.8 | 2.3 |
| CRNN-CTC | Dutta et al. [Dutta 2018] | 5.2 | 5.1 |
| CRNN-CTC | Ours | 6.1 | 5.2 |
| Seq2seq | Poulos et al. [Poulos 2017] | 16.6 | 12.1 |
| Seq2seq | Chowdhury et al. [Chowdhury 2018] | 8.1 | 3.5 |
| Seq2seq | Bluche [Bluche 2016] | 7.9 | 2.9 |
| Seq2seq | Michael et al. [Michael 2019] | **5.2** | - |
| Seq2seq | Ours | **5.2** | 4.4 |
| Transformers | Kang et al. [Kang 2020a] | 7.6 | - |

## 4.4 Applications aux registres paroissiaux

Nous souhaitons à présent étudier l'applicabilité de ce modèle à la reconnaissance du texte des registres paroissiaux. Pour cela, nous réalisons quelques expérimentations de transfert de connaissance, avec très peu ou pas de données de spécialisation.

Nous étudions également, sur les bases publiques, l'impact de la taille de l'ensemble d'apprentissage sur les performances du système. Cette étude nous permet d'estimer un nombre de documents à transcrire afin d'obtenir un modèle utilisable.

### 4.4.1 Stratégie d'apprentissage

Nous disposons de 700 lignes de texte de registres paroissiaux, qui correspondent à la base HTR-BMS. Ces lignes sont réparties dans les ensembles d'apprentissage, de validation et de test selon trois découpages. Nous créons également une base de données **HTR-générique** qui regroupe les bases de données IAM, RIMES, Esposalles, READ et HTR-Paleo.

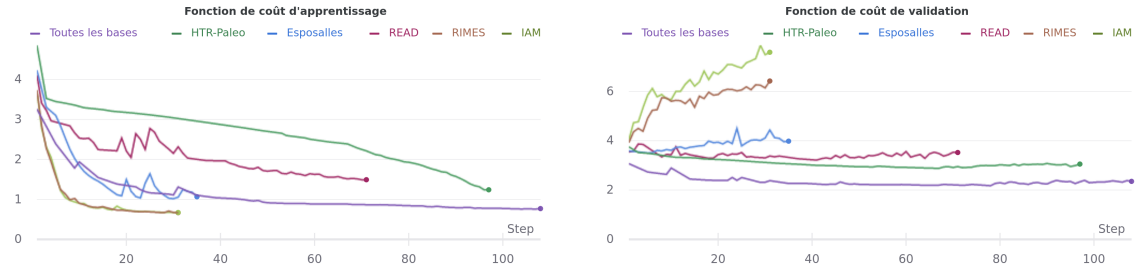
**Évaluation avec spécialisation** Dans cette première expérimentation, la base HTR-BMS est découpée en 200 images d'apprentissage, 200 images de validation et 300 images de test. Nous spécialisons le modèle, pré-entraîné sur la base HTR-générique, en utilisant ces 200 images d'apprentissage et 200 images de validation. Nous constatons que le modèle diverge lors de la spécialisation, ce qui indique que le nombre d'images d'apprentissage n'est pas suffisant pour envisager une spécialisation En règle générale, il est souvent déconseillé de spécialiser un réseau sur si peu d'échantillons. En effet, l'utilisation d'un faible nombre d'images tend à favoriser le sur-apprentissage, ce qui dégrade les performances du modèle sur l'échantillon de test.

**Évaluation sans spécialisation** Dans cette deuxième expérimentation, toutes les images de la base HTR-BMS sont utilisées en test. Le modèle est appris sur la base HTR-générique et évaluée sur la base HTR-BMS, sans aucune spécialisation sur ces documents. Les résultats sont faibles : le taux d'erreur caractère est d'environ 50%.

**Évaluation avec spécialisation uniquement dans l'ensemble de validation** Dans cette dernière expérimentation, aucune ligne de la base HTR-BMS n'est utilisée pour l'apprentissage, 300 sont utilisées en validation et 400 en test. L'apprentissage est effectué sur la base de données HTR-générique, en utilisant 300 lignes de registres paroissiaux comme échantillons de validation. Les taux de reconnaissance s'améliorent, avec un CER à environ 34%, mais restent inexploitables. Nous effectuons la même expérience avec le réseau de neurones pré-entraîné sur une base de données publique individuelle. Les résultats de la table 4.7 montrent que le pré-entraînement sur READ et HTR-Paleo permettent d'obtenir les meilleurs résultats. À l'inverse, lorsque l'architecture est pré-entraînée sur Esposalles, RIMES ou IAM, les scores chutent. Cela peut s'expliquer par le fait que READ et HTR-Paleo sont les plus proches des registres paroissiaux. Au contraire, RIMES, Esposalles et IAM sont plutôt homogènes.

**Table 4.7** – Résultats du transfert de connaissance sur la base de données HTR-BMS. L'utilisation d'une base générique en apprentissage, et de registres paroissiaux en validation aide à améliorer les performances du réseau. Cependant, les taux d'erreur restent trop élevés.

| Base d'apprentissage | Base de validation | Base de test | CER (%) ⬇ | WER (%) ⬇ |
|---|---|---|---|---|
| HTR-generique | HTR-generique | HTR-BMS | 49.95 | 95.42 |
| Esposalles | HTR-BMS | HTR-BMS | 99.14 | 129.49 |
| READ | HTR-BMS | HTR-BMS | 54.92 | 100.02 |
| RIMES | HTR-BMS | HTR-BMS | 67.52 | 97.88 |
| IAM | HTR-BMS | HTR-BMS | 65.20 | 119.36 |
| HTR-Paleo | HTR-BMS | HTR-BMS | 57.48 | 142.86 |
| HTR-generique | HTR-BMS | HTR-BMS | **33.87** | **82.84** |




**(a)** Fonction de coût d'apprentissage

**(b)** Fonction de coût de validation

**Figure 4.6** – Problèmes de convergence des différents modèles lorsque la validation est effectuée sur la base de données HTR-BMS. Une même architecture est apprise sur IAM (vert clair), READ (rose), RIMES (marron), Esposalles (bleu), HTR-Paleo (vert foncé), ou sur l'union de toutes ces bases (violet). L'apprentissage est validé sur HTR-BMS. Les bases de données READ, HTR-Paleo et HTR-générique sont les seules à permettre la convergence du modèle.

La figure 4.6 illustre les problèmes de convergence rencontrés lors de cette configuration de transfert de connaissance. Lorsque l'architecture est apprise sur des images très différentes des registres paroissiaux, la fonction de coût de validation diverge en raison de la différence trop importante entre l'ensemble d'apprentissage et de validation. En revanche, lorsque le modèle apprend sur une base plus hétérogène ou proche des registres paroissiaux, comme les bases HTR-générique, READ, ou HTR-Paleo, l'apprentissage est possible. La figure 4.7 présente un exemple de carte d'attention obtenue sur une ligne de registre BMS, avec le modèle appris sur la base de données HTR-générique. Le mécanisme d'attention s'adapte bien à ce nouveau style d'écriture, mais le réseau produit de nombreuses erreurs à l'échelle des caractères.

**Figure 4.7** – Exemple de carte d'attention sur la base de données HTR-BMS. L'attention est cohérente avec la génération des lettres. En revanche, le taux d'erreur est élevé.
Vérité : *riebaut originaire de montreuil sur isle et domicilier*
Prédiction : *nubant eiquelire de montrevil sur ister et demilier*

### 4.4.2 Estimation du nombre d'images nécessaires

Les résultats présentés sur la base HTR-BMS confirment que le manque de données est un point bloquant pour la reconnaissance de registres paroissiaux. Nous souhaitons à présent estimer le nombre de registres qu'il est nécessaire de transcrire afin d'obtenir des performances compétitives.

Pour cela, nous avons comparé les taux d'erreur caractère pour IAM, RIMES et Esposalles, en fonction du nombre d'images d'apprentissage. Nous avons choisi d'ignorer la base de données READ, car elle contient des documents issus de collections variées. La figure 4.8 synthétise les résultats. Pour Esposalles, environ 400 images de lignes suffisent à obtenir un CER à 10%. Cela correspond à environ 100 actes ou 15 pages de documents transcrits. Ce résultat s'explique par l'homogénéité de la base de données Esposalles : elle ne contient qu'un style d'écriture, et les documents sont de très bonne qualité. Les bases de données IAM et RIMES obtiennent des résultats comparables : environ 6500 lignes sont nécessaires pour obtenir 10% de taux d'erreur caractère. Ces documents sont plus hétérogènes, mais sont de bonne qualité et ne présentent pas de dégradations. Enfin, sur la base READ, l'utilisation de tous les documents (>10000 lignes) ne permet pas d'obtenir un CER inférieur à 10%. Les images de cette base sont très hétérogènes, mais de bonne qualité. Pour ces raisons, nous pensons qu'il est nécessaire d'annoter environ 10000 lignes de registres paroissiaux pour obtenir un CER avoisinant les 10% sur la base HTR-BMS. Cela correspond à environ 1500 actes, soit 200 pages de documents rien que pour l'apprentissage. En tenant compte des ensembles de validation et de test, ce chiffre monterait à environ 2000 actes.

### 4.4.3 Discussion

Les expérimentations menées dans ce chapitre montrent clairement l'impact du manque de transcriptions de registres paroissiaux sur les performances du système.

Nous sommes en mesure d'affirmer que la limite de notre approche vient du manque

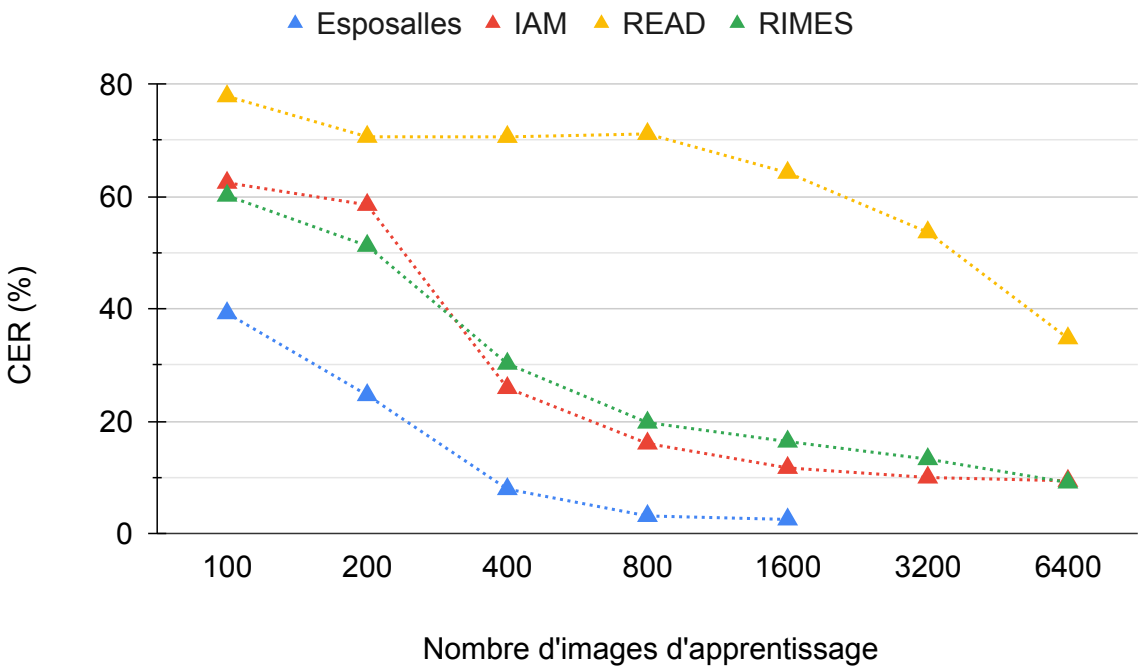



**Figure 4.8** – Impact du nombre d'images d'apprentissage sur les performances du système.

d'annotations, et pas de l'architecture. En effet, le réseau de neurones obtient un CER de 12% sur READ, qui est d'une difficulté comparable à la base HTR-BMS. Or, un score proche de 10% est acceptable sur des documents anciens, car la correction des erreurs nécessite peu d'effort [Strauß 2018].

L'annotation de ces documents n'a pas été envisageable pendant ce travail de thèse en raison de la difficulté de lecture et du temps nécessaire à la réalisation d'une transcription de qualité. Pour surmonter ce manque de données, nous introduisons dans le dernier chapitre une méthode de génération d'actes synthétiques qui vise à réduire la dépendance en transcriptions.

## 4.5 Conclusion du chapitre

Dans ce chapitre, nous avons proposé une architecture basée sur un mécanisme d'attention pour la reconnaissance d'écriture manuscrite.

Cette architecture est intéressante, car elle permet d'extraire des caractéristiques pertinentes pour la prédiction des caractères et la modélisation implicite du langage. De plus,

le mécanisme d'attention rend la prise de décision plus explicable, car il met en évidence les zones de l'image les plus utiles pour générer la séquence.

Nous avons réalisé de nombreuses expérimentations sur cette architecture, et nous avons comparé plusieurs mécanismes d'attention pour la reconnaissance d'écriture manuscrite. Les expérimentations menées ont également souligné l'intérêt de la fonction de coût hybride, proposée par Michael et al. [Michael 2019]. Enfin, nous avons montré son intérêt par rapport à l'architecture CRNN-CTC, classiquement utilisée pour la reconnaissance d'écriture.

Nous avons ensuite discuté de l'applicabilité de cette architecture aux registres paroissiaux, dans un contexte où peu de données d'apprentissage sont disponibles. Plusieurs stratégies ont été envisagées afin de transférer les connaissances apprises sur des bases de données publiques vers les registres paroissiaux. La stratégie la plus efficace consiste à apprendre le modèle sur une base de données générique, combinant cinq bases de données publiques, et à utiliser 200 lignes de registres paroissiaux pour la validation du modèle. Cependant, les résultats restent loin d'être satisfaisants. L'analyse des résultats sur des bases données publiques nous a également permis d'estimer que la transcription d'environ 2000 actes est nécessaire pour l'apprentissage de réseaux de neurones sur des registres paroissiaux.

Dans le prochain chapitre, nous abordons la question de l'extraction d'information à partir d'images de documents. Enfin, dans le dernier chapitre, nous présentons une méthodologie de génération d'actes synthétiques, afin de réduire la dépendance en annotations.

Chapitre 5

# Extraction d'informations pertinentes

Dans ce chapitre, nous présentons nos contributions pour la tâche d'extraction d'informations dans des documents historiques semi-structurés. Nous comparons objectivement les deux grandes approches proposées pour adresser cette tâche : l'*approche séquentielle*, pour laquelle la phase de reconnaissance d'écriture intervient avant la phase de classification des entités nommées, et l'*approche combinée*, pour laquelle ces deux tâches sont effectuées en même temps. Nous proposons également d'étudier l'intérêt des réseaux de neurones avec mécanisme d'attention pour l'approche combinée. Nous explorons également de nouvelles stratégies d'apprentissage conjointes basées sur des configurations multi-tâches et multi-échelles. Enfin, nous évaluons nos contributions sur la base de données Esposalles [Romero 2013] et discutons de l'applicabilité de ces méthodes pour la reconnaissance de registres paroissiaux. Les travaux présentés dans ce chapitre ont été présentés au Doctoral Consortium d'ICDAR 2021 [Tarride 2021a] et un article est en cours de soumission pour la conférence DAS 2022 [Tarride 2021b].

## 5.1 Introduction

L'extraction d'information consiste à reconnaître automatiquement les mots importants à partir de l'image d'un document numérisé. L'intérêt de cette tâche de reconnaissance s'est développé avec la numérisation massive des registres de population en Europe, comme évoqué dans le chapitre 2. L'enjeu est d'extraire massivement des informations structurées (nom, prénom, âge, métier...) à partir de l'analyse de ces registres. Cette extraction permettra, d'indexer ces documents afin de faciliter leur recherche en ligne, mais également d'envisager des analyses statistiques et macroscopiques des données extraites.

La recherche dans ce domaine a connu un essor important grâce à la publication de la base de données publique Esposalles [Romero 2013] et l'organisation de la compétition

IEHHR (Information Extraction in Historical Handwritten Records), proposée lors de la conférence ICDAR 2017 [Fornés 2017]. Le but de cette compétition est d'identifier les mots importants dans des actes de mariages catalans du XVII$^{e}$ siècle. Ces mots importants correspondent à certaines catégories sémantiques (nom, prénom, lieu, métier...) ainsi que les personnes auxquelles ils se rapportent (époux, épouse, parents...). Cette compétition a permis de développer de premières stratégies de reconnaissance encourageantes.

L'approche classique pour aborder ce problème est séquentielle, et consiste à effectuer la transcription automatique du document, puis la classification des mots reconnus dans des entités nommées. Une seconde approche, proposée plus récemment, consiste à combiner les deux tâches de reconnaissance de texte et d'entités nommées, en partageant des caractéristiques contextuelles pour ces deux tâches. Cette seconde approche nous semble particulièrement intéressante pour fiabiliser l'extraction d'information. En effet, les actes possèdent un vocabulaire réduit, qui est encore plus limité au sein d'une même catégorie sémantique. Par exemple, on retrouve souvent les mêmes prénoms, les mêmes lieux et les mêmes métiers dans un registre. En outre, la structure des phrases, les informations présentes, ainsi que leur ordre d'apparition dans l'acte sont récurrents. La connaissance de la catégorie sémantique doit aider à transcrire le mot de façon plus précise, et la connaissance du mot doit permettre de déduire sa catégorie sémantique. Pour ces raisons, nous pensons que les tâches de reconnaissance de caractères et d'entités nommées sont interdépendantes, et doivent être traités simultanément. Nous souhaitons vérifier cette intuition en comparant équitablement ces deux stratégies. Par ailleurs, les réseaux de neurones avec mécanismes d'attention se prêtent particulièrement bien à cette approche, car ils permettent de se focaliser sur des caractéristiques pertinentes pour chaque tâche. Pourtant, l'intérêt de ces réseaux n'a pas été étudié en détail dans un contexte d'extraction d'information.

Dans ce chapitre, nous proposons donc d'étudier deux axes autour des approches combinées pour l'extraction d'information. En premier lieu, nous souhaitons mesurer l'apport d'une approche combinée par rapport à une approche séquentielle. Pour cela, nous comparons ces deux approches dans des conditions similaires : même architecture, mêmes données d'apprentissage, sans post-processing ni modèle de langue. En second lieu, nous souhaitons évaluer l'impact des architectures basées sur un mécanisme d'attention pour la reconnaissance conjointe de l'écriture et des entités nommées. Enfin, nous proposons plusieurs variations de l'architecture seq2seq afin de permettre des analyses multi échelles ou multi tâches, en modulant un ou plusieurs encodeurs à un ou plusieurs décodeurs.

Ces expérimentations sont évaluées sur la base de données Esposalles, et comparées aux résultats soumis à la compétition IEHHR. Nous discutons également de l'applicabilité de ces approches pour l'extraction d'informations dans les registres paroissiaux.

## 5.2 Quelle stratégie pour l'extraction d'information ?

L'objectif de cette première section est de comparer objectivement les deux approches couramment utilisées pour l'extraction d'information : l'approche *séquentielle* et l'approche *combinée*. En effet, si l'approche combinée nous semble intéressante et particulièrement adaptée à la reconnaissance de documents semi-structurés, son intérêt n'a pas été démontré. Nous souhaitons ainsi réaliser une comparaison objective de ces deux approches, en utilisant la même architecture et les mêmes conditions d'apprentissage. L'objectif est de mesurer le gain potentiel amené par la connaissance du contexte, aussi bien au niveau de la reconnaissance d'écriture que de la reconnaissance des catégories sémantiques. Nous souhaitons également identifier les forces et les faiblesses de chaque approche.

### 5.2.1 Présentation des deux approches

Nous présentons ici les caractéristiques des deux types d'approches envisagées pour l'extraction d'information :

— L'*approche séquentielle*, qui consiste à effectuer la reconnaissance d'écriture, puis la classification sémantique des mots ;
— L'*approche combinée*, qui consiste à effectuer la reconnaissance d'écriture en même temps que la classification sémantique des mots ;

#### 5.2.1.1 Approche séquentielle

L'approche séquentielle consiste à appliquer deux modèles disjoints pour les tâches de reconnaissance d'écriture et de reconnaissance de catégories sémantiques. Un premier modèle effectue la reconnaissance d'écriture, ce qui consiste à prédire la séquence textuelle de caractères visibles sur l'image. Puis, un second modèle est utilisé pour la tâche de reconnaissance d'entités nommées, qui consiste à classer chaque mot à partir de la transcription prédite par le premier modèle. Cette stratégie est la plus populaire pour l'extraction d'information : les quatre méthodes soumises à la compétition en 2017 étaient basées sur cette stratégie [Fornés 2017]. Celles-ci ont été détaillées dans le chapitre 2.

Nous proposons de reproduire cette approche avec le réseau seq2seq, présenté dans le chapitre précédent, pour la modélisation des caractères. Puis, nous utilisons le modèle FLAIR [Akbik 2019] avec différents *embeddings* pour reconnaître les catégories sémantiques. Un schéma de cette stratégie est présenté dans la figure 5.1a.

#### 5.2.1.2 Stratégie combinée

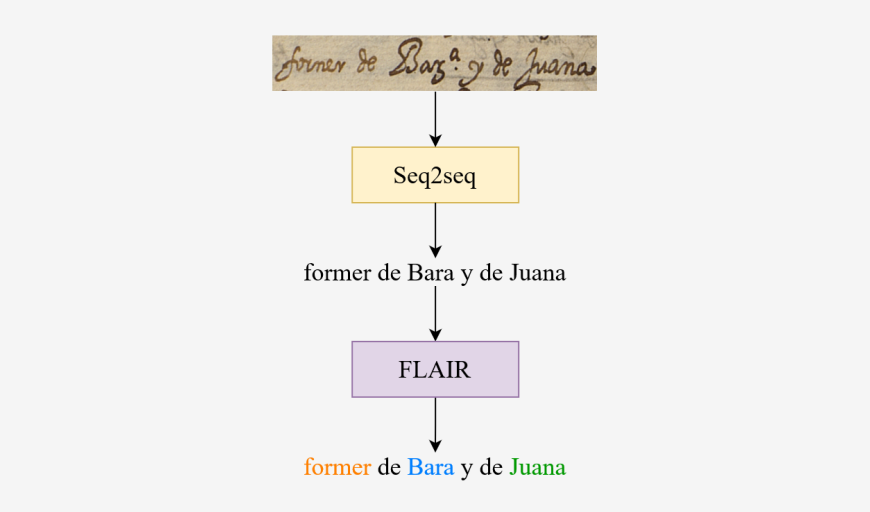
La stratégie conjointe consiste à utiliser un seul modèle pour effectuer les deux tâches de reconnaissance d'écriture et de catégories sémantiques. Son fonctionnement a été détaillé dans le chapitre 2. Leur méthode consiste à enrichir la transcription avec des tags qui permettent d'identifier les catégories et les personnes associées aux mots. Les tags sont localisés juste avant les mots et sont traités par le modèle de reconnaissance de la même façon qu'un caractère. Cette approche, initiallement proposée par Carbonell et al. [Carbonell 2018] en 2018, a éte fait en utilisant un réseau CRNN-CTC. En 2021, elle a été reprise par Rouhou et al. [Rouhou 2021] en utilisant un réseau Transformer.

Nous proposons de reproduire cette méthodologie avec une architecture seq2seq simple. En effet, l'utilisation d'un mécanisme d'attention permet au réseau de se concentrer sur les caractéristiques les plus pertinentes pour les caractères et pour les tags. Intuitivement, il semble pertinent d'avoir différents niveaux d'attention : une attention localisée sur la zone correspondante pour la prédiction des caractères et une attention plus contextuelle pour la prédiction des tags. Le mécanisme d'attention permet cette distinction entre les deux types de tokens. Un autre atout de cette architecture vient des couches récurrentes du décodeur, qui peuvent apprendre un modèle de langue implicite. Cette modélisation linguistique peut avoir un intérêt pour modéliser les deux niveaux de la langue : la modélisation de l'enchaînement des caractères et la modélisation de l'enchaînement des catégories sémantiques. La méthode que nous proposons est synthétisée dans la figure 5.1b.

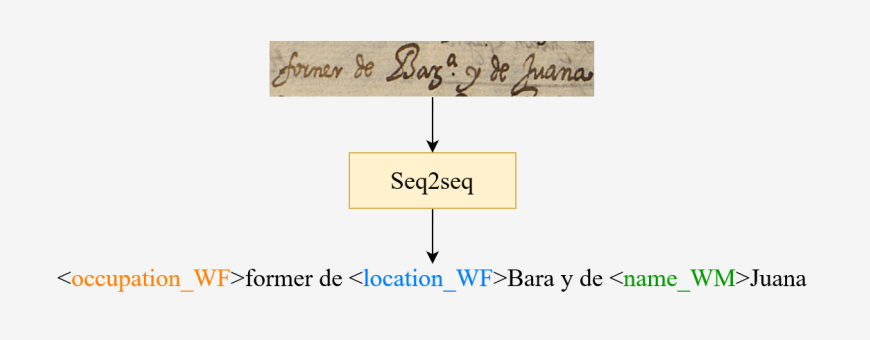
### 5.2.2 Protocole expérimental pour une comparaison objective

Nous proposons un protocole d'apprentissage et d'évaluation unifié pour évaluer objectivement ces deux stratégies, à architecture et données égales. Nous présentons également les métriques d'évaluation utilisées dans la compétition IEHHR.

**Architecture et apprentissage** Nous utilisons une même architecture pour les deux stratégies. Dans les deux cas, la base de l'encodeur est pré-entraînée sur ImageNet. Nous




**(a)** *Stratégie séquentielle* : un réseau seq2seq prédit les caractères, puis un réseau FLAIR est utilisé pour classer chaque mot dans des catégories sémantiques.




**(b)** *Stratégie combinée* : un réseau seq2seq prédit à la fois les caractères et les tags correspondant aux catégories sémantiques.

**Figure 5.1** – Illustration des deux approches proposées pour l'extraction d'information. Légende : métier du père de la mariée, lieu du père de la mariée, nom de la mère de la mariée.

n'utilisons ni post-processing, ni modèle de langue. Nous choisissons d'évaluer la qualité de l'extraction d'information ainsi que la qualité de la reconnaissance d'écriture, indépendamment des tags.

**Base de données** La base de données Esposalles [Romero 2013] est une collection de registres de mariage catalans du XVII$^{e}$ siècle, issue des Archives de la Cathédrale de Barcelone. Chaque acte contient des informations concernant les mariées : leurs noms, métiers, origines, ainsi que des informations sur leurs parents. Un exemple est illustré sur la figure 5.2. L'ensemble d'apprentissage est divisé en deux échantillons, de façon à garder 441 images de lignes pour la phase de validation et 2629 images de lignes pour l'apprentissage. Les modèles appris sont ensuite évalués sur l'ensemble de test qui contient 757 images de lignes.

**Métriques d'évaluation pour l'extraction d'information** L'évaluation est effectuée en suivant le protocole d'évaluation défini pendant la compétition IEHHR [Fornés 2017]. Deux catégories sémantiques sont associées à chaque mot : la catégorie et la personne.
Un premier score *basique* ($S_{basique}$) permet de mesurer la qualité de la reconnaissance des catégories parmi le prénom (*name*), le nom (*surname*), le métier (*occupation*), le lieu (*location*), le statut civil (*civil state*) ou autre (*other*) :

$$S_{basic} = \begin{cases} 100-\text{CER}, & \text{si la catégorie est correctement détectée} \\ 0, & \text{sinon} \end{cases}$$

Le score complet ($S_{complet}$) permet d'évaluer la qualité de reconnaissance des catégories sémantiques ainsi que des personnes auxquelles se rapportent chaque mot. La personne identifiée peut être la mariée (*wife*), le marié (*husband*), le père de la mariée (*wifes_father*), la mère de la mariée (*wifes_mother*), le père du marié (*husbands_father*), la mère du marié (*husbands_mother*), une autre personne (*other_person*), ou aucune personne (*none*).

$$S_{complet} = \begin{cases} 100-\text{CER}, & \text{si la catégorie et la personne sont correctement détectées} \\ 0, & \text{sinon} \end{cases}$$

Cette évaluation ne tient pas compte des mots neutres, c'est-à-dire classés dans la catégorie *other* ou se rapportant à aucune personne *none*.


**Figure 5.2** – Base de données Esposalles proposée dans le cadre de la compétition IEHHR. La transcription de chaque image de ligne est connue. En outre, chaque mot est associé à deux types de classes : la catégorie et la personne.

**Table 5.1** – Nombre de mots par catégorie et personne sur la base Esposalles. Certaines classes sont moins représentées que d'autres, et certaines combinaisons (catégorie + personne) sont très marginales.

**(a)** Mots importants de l'ensemble d'apprentissage (13012 autres mots sont classés *other* et *none*)

| | *name* | *surname* | *occupation* | *location* | *state* | *total* |
|---|---|---|---|---|---|---|
| *wife* | 897 | 30 | 237 | 411 | 800 | 2375 |
| *husband* | 876 | 870 | 1039 | 1971 | 187 | 4943 |
| *wife_mother* | 628 | 12 | 0 | 0 | 0 | 640 |
| *wife_father* | 648 | 657 | 762 | 1023 | 0 | 3090 |
| *husband_mother* | 494 | 16 | 0 | 4 | 0 | 514 |
| *husband_father* | 525 | 503 | 546 | 446 | 0 | 2020 |
| *other_person* | 199 | 205 | 0 | 1 | 0 | 405 |
| *total* | 4267 | 2293 | 2584 | 3856 | 987 | 13987 |

**(b)** Mots importants de l'ensemble de test (3791 autres mots sont classés *other* et *none*)

| | *name* | *surname* | *occupation* | *location* | *state* | *total* |
|---|---|---|---|---|---|---|
| *wife* | 276 | 4 | 84 | 144 | 252 | 760 |
| *husband* | 268 | 264 | 321 | 639 | 65 | 1557 |
| *wife_mother* | 186 | 0 | 0 | 0 | 0 | 186 |
| *wife_father* | 193 | 198 | 224 | 285 | 0 | 900 |
| *husband_mother* | 153 | 1 | 0 | 0 | 0 | 154 |
| *husband_father* | 169 | 158 | 168 | 19 | 0 | 514 |
| *other_person* | 67 | 69 | 0 | 0 | 2 | 138 |
| *total* | 1312 | 694 | 797 | 1087 | 319 | 4209 |

**Métriques d'évaluation pour la reconnaissance d'écriture** Pour compléter cette métrique, nous évaluons également la qualité de la reconnaissance d'écriture complète, indépendamment de la classification sémantique. Pour cela, nous mesurons les taux d'erreur caractère (CER) et taux d'erreur mot (WER).

### 5.2.3 Résultats

Les résultats obtenus par chacune des approches sont étudiés individuellement, puis comparés en détail.

#### 5.2.3.1 Approche séquentielle

Pour l'approche séquentielle, la première phase de reconnaissance optique des caractères est effectuée avec le réseau de neurones seq2seq. Les résultats obtenus sont tout à fait acceptables, avec un taux d'erreur caractère égal à 2.82%, et un taux d'erreur mot égal à 8.33%. Concrètement, cela signifie qu'environ huit mots sur dix contiennent au moins une erreur. Ces erreurs peuvent être des insertions (caractère en trop), des suppressions (caractère en moins) ou des substitutions (caractère remplacé par un autre). Ces résultats sont obtenus sans post-processing ou modèle de langue explicite.

Pour la phase de reconnaissance d'entités nommées, nous avons comparé différents modes de représentation du texte (*embeddings)*. Ces méthodes permettent d'apprendre une représentation de mots à partir d'un apprentissage sur un corpus. En particulier, nous avons étudié différents embeddings pré-entraînés issus de la librairie FastText [1]. Ces embeddings au niveau caractère (*CharEmb*), des embeddings au niveau mot (*CatalanWordEmb*), ainsi que des embeddings FLAIR (*FlairEmb*) proposés par Akbik et al. [Akbik 2018]. Différentes configurations sont évaluées à partir de la vérité terrain, et sont présentées dans le tableau 5.2. La meilleure représentation consiste à utiliser une combinaison d'embeddings FLAIR couplés à des embeddings appris sur des mots catalans. C'est donc cette configuration que nous retenons pour la suite.

Nous comparons ensuite les résultats obtenus par le réseau FLAIR à partir de différents types de texte en entrée. Quatre configurations sont comparées dans le tableau 5.3 :

— Vérité/lignes : Le modèle FLAIR apprend à reconnaître les catégories et personnes à partir du texte correspondant aux lignes de texte. Il est évalué sur les transcriptions réelles correspondant aux lignes.

1. `https://fasttext.cc/`

**Table 5.2** – Comparaison de différentes méthodes d'encodage du texte pour la reconnaissance des catégories et personnes. Différentes combinaisons d'embeddings sont évaluées à partir de la transcription vérité des actes.

| Vectorisation (embeddings) | $S_{basic}$ ↑ | $S_{complet}$ ↑ |
|---|---|---|
| CharEmb | 95.5 | 95.5 |
| CatalanWordEmb | 97.5 | 97.7 |
| FlairEmb | 97.9 | 97.7 |
| CharEmb + FLAIREmb | 98.1 | 98.1 |
| CharEmb + CatalanWordEmb | 98.1 | 98.1 |
| CatalanWordEmb + FlairEmb | **98.4** | **98.4** |
| CharEmb + CatalanWordEmb + FlairEmb | 98.2 | 98.3 |

— Vérité/actes : Le modèle apprend à reconnaître les catégories et personnes à partir du texte correspondant aux actes. Il est évalué sur les transcriptions réelles correspondant aux actes.

— Seq2seq/lignes : Le modèle apprend à reconnaître les catégories et personnes à partir du texte correspondant aux lignes. Il est évalué sur les transcriptions des lignes prédites par le réseau seq2seq.

— Seq2seq/actes : Le modèle apprend à reconnaître les catégories et personnes à partir du texte correspondant aux actes. Les transcriptions prédites par le réseau seq2seq sont collées de façon à reconstituer le texte des actes complets, sur lequel le réseau FLAIR est évalué.

**Table 5.3** – Résultats obtenus par l'approche séquentielle.

| HTR | NER | $S_{basic}$ (%) ↑ | $S_{complet}$ (%) ↑ | CER (%) ↓ | WER (%) ↓ |
|---|---|---|---|---|---|
| Vérité | FLAIR (lignes) | 95.3 | 90.7 | 0.00 | 0.00 |
| Vérité | FLAIR (actes) | 98.4 | 98.4 | 0.00 | 0.00 |
| Seq2seq | FLAIR (lignes) | 91.2 | 86.7 | 2.82 | 8.33 |
| Seq2seq | FLAIR (actes) | 93.5 | 93.3 | 2.82 | 8.33 |

Plusieurs conclusions peuvent être tirées du tableau 5.3. D'une part, lorsque la transcription prédite est utilisée, le score basique obtenu par le modèle FLAIR diminue d'environ 5% pour les actes, et 4% pour les lignes. Cela signifie que les éventuelles erreurs de transcriptions faites par le réseau seq2seq perturbent le réseau FLAIR. Un mot erroné pourra alors être classé dans la mauvaise catégorie. Ce tableau permet également

de montrer que FLAIR n'est pas parfait : au mieux, il permet d'obtenir un score complet de 98.4% à partir des actes, et 90.7% à partir des lignes. Cette comparaison illustre l'intérêt du contexte provenant de l'acte complet sur les performances : le score baisse d'environ 8% si la reconnaissance est effectuée uniquement à partir des lignes. En effet, la connaissance du texte associé à l'acte améliore les performances, car le système bénéficie alors d'un contexte plus important. En particulier, la reconnaissance des personnes bénéficie largement du contexte présent dans l'acte : le score complet chute considérablement lorsque FLAIR n'a accès qu'au texte des lignes. Au contraire, quand FLAIR a accès à l'acte entier, il y a peu de différence entre les deux scores.

#### 5.2.3.2 Approche combinée

L'approche combinée effectue les deux tâches, reconnaissance de caractères et reconnaissance d'entités nommées, en même temps.

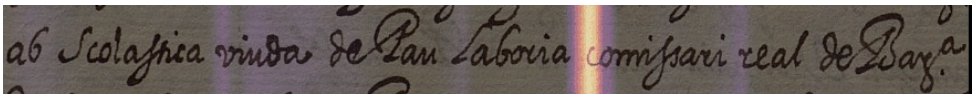
Le réseau apprend à générer une transcription enrichie, dans laquelle chaque mot d'intérêt est précédé d'un tag qui caractérise sa catégorie sémantique et la personne à laquelle il se réfère. Nous avons étudié deux modes de représentation pour encoder ces informations dans des tags. La première représentation *mixte séparée* consiste à associer deux tags à chaque mot d'intérêt : un tag pour la catégorie et un tag pour la personne. Les tags sont sous la forme : `<occupation><wifes_father>pages`. Cette représentation est complexe, car le réseau doit produire deux tags avant chaque mot d'intérêt. En revanche, elle permet une meilleure représentation de chaque classe individuelle. La seconde représentation *mixte jointe* consiste à créer un tag unique contenant les deux informations. Le tag est sous la forme `<occupation_wifes_father>pages`. Cette représentation est plus simple, mais certaines configurations sont peu fréquentes dans l'ensemble d'apprentissage, ce qui entraîne un biais. Le tableau 5.4 montre que les tags combinés permettent d'obtenir les meilleurs résultats.

**Table 5.4** – Résultats obtenus par l'approche conjointe.

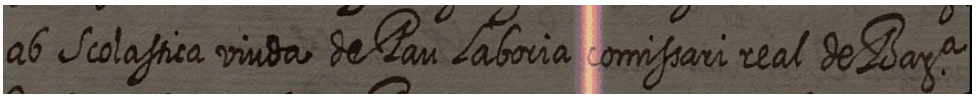
| Tags | $S_{basic}$ (%) ↑ | $S_{complet}$ (%) ↑ | CER (%) ↓ | WER (%) ↓ |
|---|---|---|---|---|
| *mixte séparée* | 27.52 | 24.56 | 4.81 | 12.29 |
| *mixte jointe* | **94.66** | **94.40** | **1.81** | **8.33** |

Nous avons ensuite analysé l'attention afin de vérifier la cohérence du modèle. La carte d'attention semble cohérente avec notre intuition : lorsque le réseau prédit des tags, il a

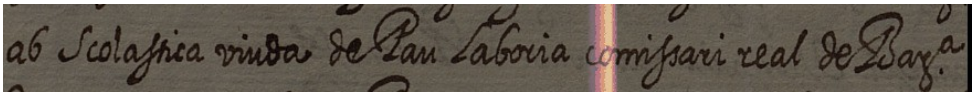
tendance à se concentrer sur des zones pertinentes pour retrouver le contexte du mot. Par exemple, dans la figure 5.3, l'attention se focalise aussi sur le mot « viuda » (*veuve*) lors de la génération du tag, ce qui permet de déduire que le mot se rapporte à la mariée.



**(a)** ...<other_person_surname>Pau <wife_occupation>



**(b)** ...<other_person_surname>Pau <wife_occupation>c



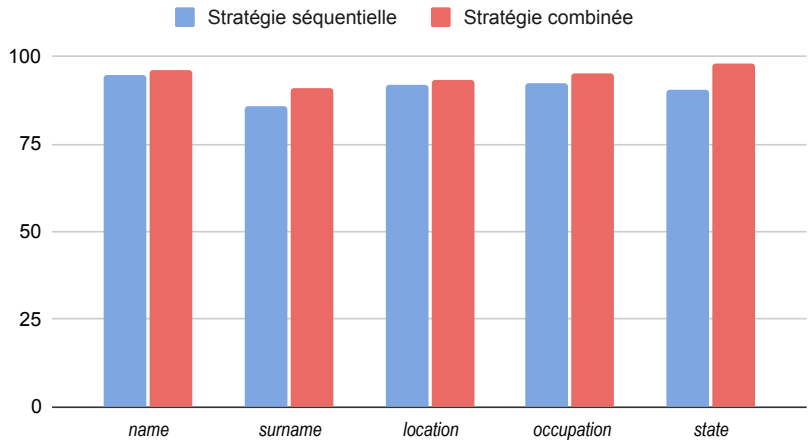
**(c)** ...<other_person_surname>Pau <wife_occupation>co

**Figure 5.3** – Visualisation des cartes d'attention pour la stratégie combinée. Les images correspondent à trois étapes consécutives de la transcription automatique. La transcription prédite à chaque étape apparait en dessous de chaque image. Les trois cartes d'attention correspondent aux zones de l'image choisies par le réseau pour prédire les tokens colorés en orange. Pour prédire un tag, l'attention se concentre sur la zone locale, mais aussi sur d'autres mots du texte, avec un intérêt pour les mots « viuda » et « real ». En revanche, pour prédire les caractères, l'attention est uniquement focalisée sur les caractères.

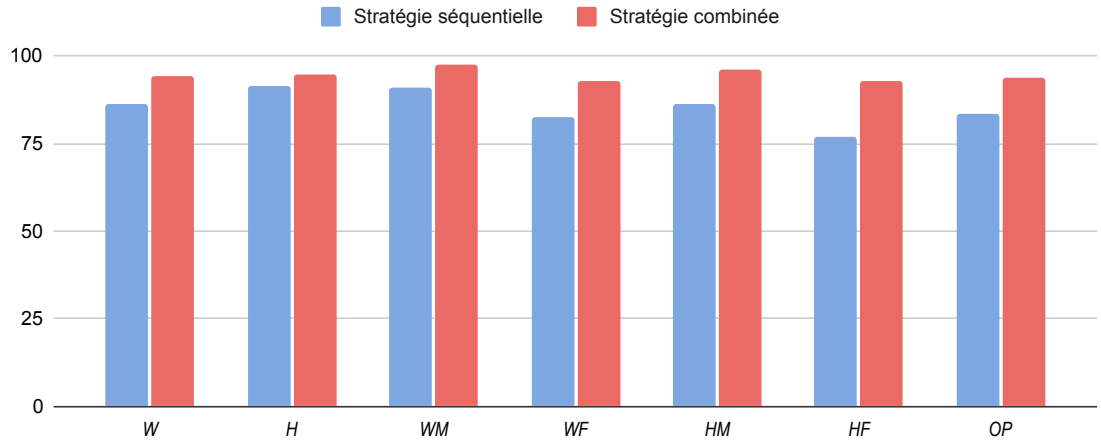
#### 5.2.3.3 Comparaison des approches

Nous proposons à présent de comparer ces deux approches. Le modèle FLAIR sur les lignes individuelles est retenu afin de comparer les deux approches dans les mêmes conditions. La comparaison entre ces deux stratégies est présentée dans le tableau 5.5. Les résultats soulignent l'intérêt de la stratégie combinée pour la tâche d'extraction d'information, car elle permet d'obtenir un gain d'environ 3% sur le score basique et de 8% sur le score complet de la compétition IEHHR [Fornés 2017]. Les résultats suggèrent donc que l'approche combinée est plus adaptée à la tâche d'extraction d'information. En particulier, le score de l'approche séquentielle chute d'environ 5% lorsque l'évaluation des personnes est prise en compte, ce qui suggère que cette tâche pose particulièrement problème au modèle FLAIR, surtout à l'échelle des lignes. Mais l'approche combinée devance également l'approche séquentielle quand FLAIR a accès au texte des actes complets.

Pour affiner cette analyse, les scores par catégorie et par personne sont présentés dans les figures 5.4. Ces scores sont à mettre en perspective avec le tableau 5.1 qui décompte




**(a)** Comparaison des stratégies pour la reconnaissance des catégories




**(b)** Comparaison des stratégies pour la reconnaissance des personnes (*W* pour la classe *wife*, *H* pour *husband*, *WF* pour *wifes_father*, *WM* pour *wifes_mother*, *HF* pour *husbands_father*, *HM* pour *husbands_mother*, *OP* pour *other_person*)

**Figure 5.4** – Comparaison détaillée des deux stratégies en fonction des différentes classes. Les scores de reconnaissance présentés correspondent au mode de calcul du score basique.

**Table 5.5** – Comparaison des approches séquentielle et combinée pour l'extraction d'informations à partir de lignes. L'approche combinée augmente les taux de reconnaissance pour l'extraction d'information. Elle permet également une baisse des taux d'erreurs pour la reconnaissance d'écriture.

| Stratégie | $S_{basique}$ (%) ↑ | $S_{complet}$ (%) ↑ | CER (%) ↓ | WER (%) ↓ |
|---|---|---|---|---|
| Seq2seq sequentiel | 91.2 | 86.7 | 2.82 | 8.33 |
| **Seq2seq combiné** | **94.7** | **94.0** | **1.81** | **6.10** |

la fréquence de chaque classe dans les ensembles d'apprentissage et de test.

La stratégie combinée est légèrement supérieure sur la reconnaissance des catégories. Nous observons que la catégorie la plus difficile à reconnaître est le nom de famille (*surname*). Cela peut s'expliquer par la grande variabilité des mots appartenant à cette catégorie : contrairement aux prénoms, aux métiers et aux lieux, les noms sont uniques. Il semble donc plus difficile de bénéficier d'un modèle de langue implicite pour cette catégorie.

L'écart entre les deux approches se creuse lorsque l'on s'intéresse à la reconnaissance des personnes. En particulier, l'approche séquentielle rencontre des difficultés sur les mots correspondant à la mariée. Cette observation peut être liée à la sous-représentation de la classe *wife* dans l'ensemble d'apprentissage. Cette baisse des résultats est surtout liée aux classes *wife_surname* et *wife_location*, pour lesquelles l'approche séquentielle obtient respectivement 16.67% (moyenne sur 4 mots) et 67.37% (moyenne sur 144 mots) de reconnaissance. Pourtant, l'approche combinée obtient des taux de reconnaissance similaires pour les deux époux. L'approche séquentielle peine à reconnaître les pères des époux, alors que ces personnes sont mieux représentées dans l'ensemble d'apprentissage que les mères des époux. Les résultats par catégorie mixte montrent que ce sont surtout les catégories *occupation* et *location* qui posent problème pour ces personnes. Or, ces catégories ne sont jamais associées aux mères des époux.

Outre les scores d'extraction d'informations, les taux d'erreur aux niveaux caractère et mot sont réduits avec l'approche combinée. Ainsi, la stratégie combinée bénéficie de la connaissance des catégories sémantiques et des personnes pour affiner la transcription des mots. A l'inverse, la stratégie séquentielle amplifie les erreurs : des erreurs de transcription provoquent des erreurs de labellisation des catégories et personnes. L'apprentissage disjoint des deux réseaux font que FLAIR n'apprend pas à s'adapter à des mots mal orthographiés.

### 5.2.4 Discussion

L'approche combinée permet la reconnaissance conjointe des catégories sémantiques, des personnes et des caractères. Les résultats montrent que cette combinaison impacte positivement les scores de reconnaissance, non seulement pour l'extraction d'information, mais aussi pour la reconnaissance de caractères. Ainsi, les deux tâches bénéficient du partage de caractéristiques contextuelles. Ce résultat peut s'expliquer par la forte homogénéité de certaines catégories sémantiques. Par exemple, certains prénoms sont très fréquents dans ces actes, comme « Catherina », « Elisabeth » ou « Joan ». C'est également le cas des métiers (« pages », « parayre », « texidor »), des lieux (« Bara », « Andreu », « Villafranca »), ou du statut civil (« donsella », « viuda », « viudo »). La connaissance contextuelle doit permettre au décodeur de modéliser des modèles de langue implicite par catégorie. En outre, le mécanisme d'attention permet de se focaliser à différents niveaux dans l'image. En particulier, nous observons que l'attention se concentre sur des zones locales pour prédire les caractères, et sur des zones plus globales et contextuelles pour prédire les tags.

L'approche séquentielle présente également des avantages. En particulier, la séparation des tâches permet à chaque réseau d'être optimisé séparément, en utilisant le maximum de données disponibles. Ainsi, chaque réseau peut être pré-entraîné sur une base publique, et optimisé avec un modèle de langue spécifique. Une autre caractéristique intéressante est la possibilité de travailler à différents niveaux : la phase de reconnaissance d'écriture peut s'effectuer au niveau des lignes, et la phase de reconnaissance d'entités au niveau des actes, ce qui permet d'introduire un contexte supplémentaire. Cependant, même dans ce scénario avantageux, l'approche séquentielle reste moins performante que l'approche combinée. Une piste intéressante serait d'appliquer l'approche combinée directement au niveau des actes. Mais cette stratégie complique la tâche de reconnaissance, tout en diminuant le nombre d'images disponibles pour l'apprentissage. Elle ne semble donc pas applicable en l'état actuel et nécessiterait la création d'une base de données plus grande que Esposalles.

## 5.3 Stratégies d'apprentissage pour les stratégies combinées

Les modèles à base d'attention semblent adaptés à cette stratégie combinée. Cependant, il nous semble intéressant d'explorer différentes manières d'intégrer l'information

contextuelle. L'idée de ce travail est motivé par les observations de Luong et al. [Luong 2015a]. Dans leur article, les auteurs explorent de nouvelles stratégies d'apprentissage pour de la traduction automatique avec une architecture seq2seq multi-tâche.

— *one-to-many* - l'encodeur est partagé entre plusieurs décodeurs, ce qui permet de faire de la traduction automatique d'un langage source à plusieurs langages cibles.
— *many-to-one* - le décodeur est partagé entre plusieurs encodeurs, ce qui permet de faire de la traduction automatique de plusieurs langages sources vers un langage cible.
— *many-to-many* - plusieurs encodeurs et décodeurs sont partagés, ce qui permet d'effectuer une traduction automatique de plusieurs langages sources vers plusieurs langages cibles.

Leurs résultats montrent que le partage d'encodeurs ou de décodeurs dédiés à différentes tâches peut profiter au modèle. Nous souhaitons évaluer ce type d'approche pour la reconnaissance d'écriture et d'entités nommées.

### 5.3.1 Nos propositions

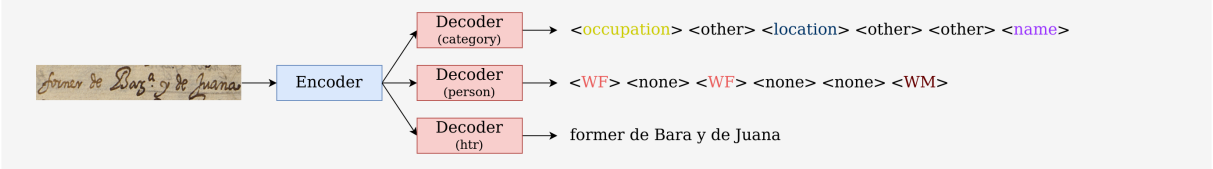
Pour cela, nous proposons trois nouvelles approches qui tiennent compte des atouts de modularité du seq2seq. Ces trois approches sont illustrées dans la figure 5.5.

#### 5.3.1.1 Stratégie combinée multi-tâches sans tags

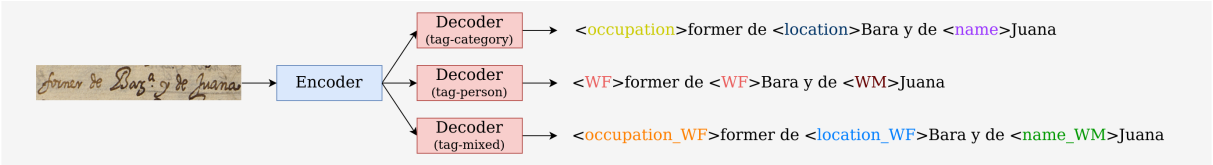
La première configuration, détaillée dans la figure 5.5a, consiste à partager un encodeur avec trois décodeurs. Chaque décodeur est spécialisé dans une tâche :

— Le décodeur *catégorie* réalise la tâche de reconnaissance de catégories à partir de l'image originale. Les éléments à prédire correspondent aux 6 classes correspondant aux catégories (en incluant la classe *other*).
— Le décodeur *personne* réalise la tâche de reconnaissance de personnes à partir de l'image originale. Les éléments à prédire correspondent aux 8 classes correspondant aux catégories (en incluant la classe *none*).
— Le décodeur *caractère* réalise la tâche de reconnaissance de caractères à partir de l'image originale. Les éléments à prédire correspondent aux lettres de l'alphabet.

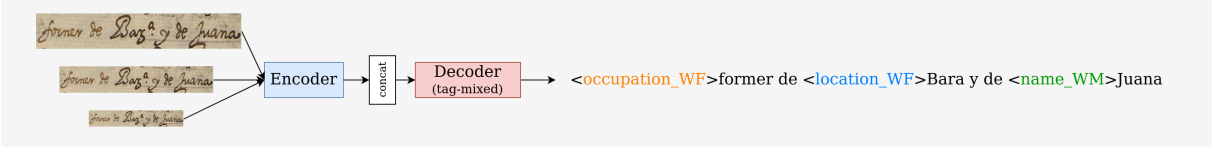
L'apprentissage conjoint de ces trois décodeurs permet d'apprendre à l'encodeur à extraire des caractéristiques communes pertinentes pour ces trois tâches. Chaque décodeur dispose d'un mécanisme d'attention propre qui se concentre sur la zone de l'image la




**(a)** *Stratégie combinée multi-tâches sans tags* : trois décodeurs permettent de générer la séquence des catégories sémantiques, des personnes, et des caractères.




**(b)** *Stratégie combinée multi-tâches avec tags* : trois décodeurs permettent de générer des séquences avec des tags de catégorie, des tags de personnes, et des tags mixtes.




**(c)** *Stratégie combinée multi-échelles avec tags* : l'image entre dans l'encodeur à trois échelles différentes afin d'avoir une représentation vectorielle plus riche. Le décodeur prédit une séquence avec tags à partir de cette représentation.

**Figure 5.5** – Les différentes stratégies multi-tâches et multi-échelles proposées.

plus pertinente. L'architecture est entraînée avec une fonction de coût unique, qui correspond à la moyenne des fonctions de coût individuelles. Lorsque le réseau n'apprend plus, l'apprentissage s'arrête et le modèle est évalué. Pour chaque image d'entrée, le réseau produit trois séquences de sortie. Ces séquences sont recalées afin d'aligner les catégories sémantiques et les personnes sur les mots de la transcription.

Cette configuration présente deux problèmes majeurs. D'une part, cet alignement est problématique : les deux décodeurs *catégorie* et *personne* prédisent une séquence à l'échelle des mots tandis que le décodeur *caractère* prédit une séquence à l'échelle des caractères. Une erreur sur la catégorie ou la personne, plus particulièrement un ajout ou un oubli, mène à un décalage du reste de la séquence. D'autre part, nous observons que les décodeurs *catégorie* et *personne* convergent plus rapidement que le décodeur *caractère*, ce qui déséquilibre l'apprentissage. Concrètement, l'apprentissage global s'arrête alors que le décodeur *caractère* n'a pas fini d'apprendre.

#### 5.3.1.2 Stratégie combinée multi-tâches avec tags

La seconde configuration, illustrée sur la figure 5.5b, est semblable à la précédente, mais utilise des tags afin d'éviter les problèmes de convergence et d'alignement. Chaque décodeur est entraîné avec un type de tag différent.

— Le décodeur *tag-catégorie* réalise la tâche d'extraction d'information avec des tags qui correspondent uniquement aux 5 classes correspondant aux catégories (en excluant la classe *other*).
— Le décodeur *tag-personne* réalise la tâche d'extraction d'information avec des tags qui correspondent uniquement aux 7 classes correspondant aux personnes (en excluant la classe *none*).
— Le décodeur *tag-mixte* réalise la tâche d'extraction d'information avec des tags mixte.

L'intuition derrière cette approche est de forcer l'encodeur à extraire des caractéristiques contextuelles riches. L'ajout des branches *tag-catégorie* et *tag-personne* peut également favoriser la reconnaissance des classes mixtes peu représentées. Chaque décodeur dispose d'un mécanisme d'attention propre qui se concentre sur la zone de l'image la plus pertinente.

L'architecture est également entraînée avec une fonction de coût unique, qui correspond à la moyenne des fonctions de coût individuelles. En revanche, les trois décodeurs effectuent des tâches de difficultés similaires, ce qui garantit une meilleure convergence globale.

A la fin de l'apprentissage, seule la branche *tag-mixte* est évaluée, ce qui permet d'évaluer l'apport des autres branches sur cette tâche.

#### 5.3.1.3 Stratégie combinée multi-échelle avec tags

La dernière stratégie, illustrée dans la figure 5.5c étudiée consiste à extraire des caractéristiques locales et globales à partir de l'image à différentes échelles. L'encodeur extrait trois vecteurs de caractéristiques, qui correspondent à la même image à différentes échelles. Les vecteurs de caractéristiques sont fusionnés, et servent d'entrée au décodeur.

Le décodeur prédit une transcription enrichie de tags mixtes.

### 5.3.2 Résultats

Nous évaluons dans cette section les performances de chaque configuration multi-tâche ou multi-échelle.

**Table 5.6** – Comparaison des quatre stratégies multi-tâches et multi-échelles.

| Modèle | CER (%) ⬇ | WER (%) ⬇ | $S_{basique}$ ⬆ | $S_{complete}$ ⬆ |
|---|---|---|---|---|
| *Séquentiel* | 2.82 | 8.33 | 91.2 | 86.7 |
| *Combiné* | 1.81 | 6.10 | 94.7 | 94.0 |
| *Combiné multi-tâches sans tags* | 7.75 | 17.38 | 61.8 | 48.1 |
| *Combiné multi-tâches avec tags* | **1.74** | **5.38** | **95.2** | **94.4** |
| *Combiné multi-échelles avec tags* | 5.61 | 15.13 | 83.0 | 80.3 |

| Modèle | *Name* | *Surname* | *Location* | *Occupation* | *State* |
|---|---|---|---|---|---|
| *Séquentiel* | 88.2 | 83.1 | 87.1 | 91.5 | 84.7 |
| *Combiné* | 96.1 | 91.0 | 93.7 | **95.3** | **97.8** |
| *Combiné multi-tâches sans tags* | 63.2 | 41.0 | 48.2 | 69.8 | 86.7 |
| *Combiné multi-tâches avec tags* | **97.0** | **92.6** | **94.5** | **95.3** | 96.7 |
| *Combiné multi-échelles* | 87.4 | 61.1 | 84.2 | 87.5 | 96.9 |

#### 5.3.2.1 Comparaison des stratégies d'apprentissage

Les expériences montrent que l'approche multi-tâches avec tags permet de surpasser l'approche mono-tâche. En effet, l'ajout des branches spécialisées dans la reconnaissance des catégories et personnes individuelles améliore les résultats de la branche mixte. En revanche, le gain est faible au regard de la complexité de l'architecture multi-tâches.

En revanche, les autres stratégies d'apprentissage ne permettent pas d'amélioration. L'approche multi-tâches sans tags obtient un taux d'erreur caractère élevé, ce qui confirme notre intuition sur la convergence asymétrique des trois branches : les deux branches (NER) convergent rapidement, ce qui ne permet pas à la troisième branche (HTR) d'apprendre suffisamment. De plus, les difficultés liées à l'alignement des trois séquences prédites peuvent expliquer les faibles scores sur la compétition IEHHR. L'approche multi-échelles n'est pas non plus convaincante. Les résultats montrent que l'ajout de caractéristiques multi-échelles ne permet pas au réseau d'améliorer ses performances. Pourtant, la visualisation de l'attention montre que le réseau se concentre sur les caractéristiques

correspondant à l'image à petite échelle pour extraire les tags, et à grande échelle pour extraire les caractères.

#### 5.3.2.2 Comparaison avec la compétition IEHHR

Nous comparons à présent, dans le tableau 5.7 nos contributions aux méthodes soumises à la compétition IEHHR. Nous notons que seules les méthodes qui effectuent la reconnaissance à partir des lignes de texte comme entrée sont prises en compte. Nos méthodes *combinée* and *combinée multi-tâches avec tags* sont compétitives. Notons qu'elles ne bénéficient d'aucun post-processing ni modèle de langue.

La comparaison entre le travail de Carbonell et al. [Carbonell 2018] et notre approche combinée est intéressante. En effet, la même méthodologie est utilisée pour ces deux contributions, seule l'architecture diffère. Nous utilisons un réseau à base d'attention, alors qu'ils utilisent un CRNN-CTC. Grâce à notre architecture, le score complet gagne près de 5%, passant de 89.40% à 94.0%. Ce résultat illustre l'intérêt de notre architecture avec mécanisme d'attention pour l'extraction d'information.

Le modèle d'InstaDeep [Rouhou 2021], publié très récemment, reprend également l'approche combinée en utilisant un modèle Transformer. Les performances de notre modèle seq2seq *multi-tâches avec tags* obtient des performances comparables à leur approche, avec un score de reconnaissance complet légèrement supérieur. Les résultats obtenus par les modèles seq2seq et Transformer démontrent l'intérêt du mécanisme d'attention pour la tâche d'extraction d'informations.

Enfin, le modèle de NaverLabs [Prasad 2018] est plus performant, bien que les auteurs reportent un CER plus élevé (environ 5%). La force de leur proposition réside probablement dans des caractéristiques basées sur des n-grams de caractères pour la phase de reconnaissance d'entités nommées. En revanche, l'article ne précise pas si la phase de reconnaissance des entités nommées se fait sur le texte des lignes ou des actes. Il est à noter que cet article a été distribué en accès libre et n'a pas bénéficié d'une évaluation par les pairs.

### 5.3.3 Discussion

Dans cette étude, nous avons mesuré l'apport des approches combinées par rapport aux approches séquentielles. Ces stratégies bénéficient d'un contexte global qui permet de fiabiliser la transcription des mots.

**Table 5.7** – Comparaison des méthodes soumises à la compétition IEHHR. Les méthodes situées au dessus de la ligne horizontale correspondes à celles soumises pendant la compétition, les autres correspondent à des méthodes proposées après la fin de la compétition. Les deux méthodes en italiques correspondent à nos propositions.

| Nom | Modèle | Stratégie | $S_{basique}$ ↑ | $S_{complet}$ ↑ |
|---|---|---|---|---|
| Baseline HMM | HMM | Séquentielle | 80.2 | 63.1 |
| CITlab-ARGUS-1 | CRNN-CTC | Séquentielle | 89.5 | 89.2 |
| CITlab-ARGUS-2 | CRNN-CTC | Séquentielle | 91.9 | 91.6 |
| CITlab-ARGUS-3 | CRNN-CTC | Séquentielle | 91.6 | 91.2 |
| CVC-tags [Carbonell 2018] | CRNN-CTC | Combinée | 90.6 | 89.4 |
| Naver Lab [Prasad 2018] | CRNN-CTC | Séquentielle | **95.5** | **95.0** |
| InstaDeep [Rouhou 2021] | Transformer | Combinée | 95.2 | 93.3 |
| *Séquentiel* | seq2seq | Séquentielle | 91.2 | 86.7 |
| *Combiné* | seq2seq | Combinée | 94.7 | 94.0 |
| *Combiné multi-tâches avec tags* | seq2seq | Combinée | 95.2 | 94.4 |

Nous avons également proposé différentes stratégies d'apprentissage en multi-tâches avec les réseaux basés sur un mécanisme d'attention. La configuration multi-tâches avec tags est plus performante que la configuration classique, ce qui souligne le fait que le réseau bénéficie de l'apprentissage de différentes représentations sémantiques. Ces résultats sont en accord avec l'observation de Luong et al. [Luong 2015a] : l'apprentissage multi-tâches améliore, sous certaines conditions, les performances des réseaux de neurones avec mécanisme d'attention.

Enfin, nous avons montré l'intérêt des réseaux à mécanisme d'attention pour cette tâche.

## 5.4 Applicabilité aux registres paroissiaux

Nous discutons ici de l'applicabilité des stratégies présentées dans ce chapitre à l'extraction d'information dans des registres paroissiaux.

Les registres catalans de la base de données Esposalles et les registres paroissiaux ont de nombreux points communs. Ce sont des registres semi-structurés en actes. Ce sont les mêmes types de documents. Ainsi, les stratégies mises au point sur la base Esposalles doivent pouvoir s'appliquer aux registres paroissiaux. Cependant, certains points de différence posent problème.

Le premier point de différence entre les registres paroissiaux et les documents de la

base de données Esposalles est la difficulté des corpus. En effet, les documents de la base de données Esposalles sont homogènes : ils ne contiennent qu'un seul style d'écriture. De plus, les documents sont tous des actes de mariages et ont été numérisés en haute qualité. Au contraire, les registres paroissiaux sont hétérogènes, avec de nombreux styles d'écriture issus de différentes périodes temporelles. De nombreuses dégradations sont présentes sur le papier et l'encre. De plus, les documents présentent de nombreuses abréviations, annotations interligne et ratures. Ils ont également été numérisés avec une densité plus faible.

Le second point à souligner est la différence de taille de l'ensemble d'apprentissage. Seulement 700 lignes sont disponibles pour les registres paroissiaux. Ces 700 lignes doivent composer les ensembles d'apprentissage, de validation et de test. De son côté, la base Esposalles dispose de 2600 images de lignes exclusives pour l'apprentissage sur Esposalles. Ce manque de données forme un point de blocage crucial. Dans ces conditions, l'extraction d'informations n'est pas envisageable, car les réseaux de neurones ne sont pas capables de modéliser des documents aussi complexes à partir de si peu d'exemples. Cette affirmation est étayée par les résultats du chapitre précédent : le manque de documents ne permet pas d'apprendre un système de reconnaissance d'écriture. Il semble donc difficilement envisageable d'effectuer une tâche encore plus difficile dans ces conditions.

Dans le chapitre précédent, nous avons estimé qu'environ 10000 lignes d'apprentissage seraient nécessaires pour envisager la reconnaissance de ces documents. Or, l'annotation d'une telle quantité de données n'est pas envisageable à court terme. La génération de documents synthétiques fait l'objet du dernier chapitre. Elle pourrait permettre de réduire considérablement le besoin en annotations.

## 5.5 Conclusion du chapitre

Dans ce chapitre, nous avons étudié la tâche d'extraction d'information dans des documents historiques semi-structurés. Cette tâche relativement nouvelle a connu un essor avec la publication de la base Esposalles [Romero 2013] et de la compétition IEEHR associée [Fornés 2017].

Nous avons comparé les types de stratégies principales qui ont été proposées par la communauté scientifique : la stratégie séquentielle, pour laquelle la phase de reconnaissance d'écriture intervient avant la phase de classification en entités nommées, et la *stratégie combinée*, pour laquelle ces deux tâches sont effectuées en même temps. Les résultats de

notre étude montrent que, dans des conditions expérimentales similaires, la combinaison de ces deux tâches permet une hausse de 8% du score de reconnaissance complet. Les taux d'erreurs caractère et mot diminuent avec cette approche.

Cette étude met également en évidence l'intérêt des réseaux de neurones avec mécanisme d'attention pour l'approche combinée. La comparaison de notre travail avec l'approche proposée par [Carbonell 2018] montre que l'utilisation d'un modèle d'attention permet une hausse d'environ 5%, par rapport à une architecture CRNN-CTC. La visualisation des cartes d'attention montre que le mécanisme d'attention permet au réseau de se concentrer sur des caractéristiques locales pour la génération de caractères, et sur des caractéristiques contextuelles plus globales pour la génération de tags. De plus, les couches de récurrence du décodeur sont en mesure d'apprendre un modèle de langue implicite par catégorie sémantique, ce qui augmente le taux de reconnaissance.

Les conclusions des travaux de Luong et al. [Luong 2015a] nous ont poussé à explorer de nouvelles stratégies d'apprentissage multi-tâches et multi-échelles avec notre réseau basé sur un mécanisme d'attention. Si certaines stratégies d'apprentissage présentent des performances limitées, l'une d'entre-elles sort du lot. Nous observons qu'un apprentissage multi-tâche avec différentes représentations sémantiques aide le réseau à extraire des caractéristiques communes plus pertinentes. Cette configuration permet au réseau d'obtenir des performances proches de l'état de l'art à l'échelle des lignes sur la compétition IEHHR, atteignant un score complet de 94.4%, sans post-processing, ni modèle de langue explicite.

Ce travail met en lumière certaines perspectives intéressantes. D'une part, un point critique est d'effectuer l'analyse à l'échelle des actes. Les informations citées dans les lignes précédentes doivent pouvoir aider le réseau, en particulier pour la reconnaissance de personnes. En effet, si la ligne précédente présentait les informations sur la mariée, il est probable que la ligne suivante présente des informations sur ses parents. Certaines approches ont été proposées pour la reconnaissance de caractères à partir de paragraphes [Coquenet 2021]. Il est par exemple possible de concaténer les vecteurs caractéristiques des lignes successives dans un même acte. L'inconvénient de cette approche est qu'elle nécessite de nombreuses images d'apprentissage. Or, la base Esposalles n'est pas très grande, et il n'existe pas d'autre base avec des annotations aussi complètes. Une autre possibilité consiste à initialiser l'état caché du décodeur avec l'état caché final de la ligne précédente. Cette approche permet d'effectuer l'analyse à l'échelle des lignes, en transférer l'information de chaque ligne avant de traiter la suivante. Cette idée pose certaines difficultés d'implémentation pour le traitement par batch, car il faut s'assurer que les lignes sont

traitées dans l'ordre. De plus, il n'est pas garanti que l'état caché du décodeur fournisse un contexte suffisamment global sur le contenu de la ligne.

Enfin, nous avons discuté des conditions d'applicabilité de ces méthodes pour la reconnaissance de registres paroissiaux. Le manque de données reste un problème majeur pour l'apprentissage de tels modèles. Le prochain chapitre adresse la génération de documents synthétiques, qui permet de réduire le besoin en données annotées.

Chapitre 6

# Génération de données synthétiques et augmentation

Dans ce chapitre, nous proposons d'appliquer les méthodes de reconnaissance de documents présentées dans les chapitres 4 et 5 sur des registres paroissiaux. Pour pallier le manque de transcriptions d'actes, nous introduisons une méthodologie de génération d'actes synthétiques réalistes et étudions les stratégies d'apprentissage statistique des modèles de reconnaissance à l'aide de ces données synthétiques. Enfin, nous présentons les résultats obtenus sur des actes réels et montrons l'intérêt des documents synthétiques pour la reconnaissance de texte manuscrit et l'extraction d'information, en l'absence de documents réels dans l'ensemble d'apprentissage.

## 6.1 Introduction

Dans les chapitres 4 et 5, nous avons proposé une méthodologie de reconnaissance d'écriture et d'extraction d'information qui permet d'obtenir une performance à l'état de l'art sur la base de données Esposalles [Romero 2013], dans le cadre de la compétition internationale IEHHR [Fornés 2017]. Nous souhaitons à présent appliquer cette méthodologie pour la reconnaissance des registres paroissiaux.

Cependant, nous ne disposons pas de suffisamment d'actes transcrits pour effectuer l'apprentissage d'un tel modèle de reconnaissance. En effet, environ 700 lignes de texte issues de registres paroissiaux ont été annotées, et celles-ci doivent être utilisées pour la validation et l'évaluation du modèle. Il n'a pas été possible d'annoter plus de documents en raison du temps considérable que représente le processus d'annotation. Ce faible nombre d'échantillons ne permet pas d'apprendre correctement un modèle de reconnaissance stable. Il n'est pas non plus envisageable d'effectuer de la validation croisée, qui nécessiterait un temps d'apprentissage colossal. Or, nous avons montré qu'il est nécessaire de spécialiser le modèle de reconnaissance sur des registres paroissiaux. En effet, les

résultats du chapitre 4 confirment qu'un modèle appris sur des bases de données variées et génériques ne permet pas d'obtenir des résultats satisfaisants sur des images de registres paroissiaux.

Une solution rapide et peu coûteuse pour dépasser cette contrainte consiste à générer des images d'actes synthétiques. En effet, la génération d'un grand nombre d'images synthétiques permet de spécialiser un modèle générique sur des documents ayant des caractéristiques similaires à la collection à traiter. Ceci nous permet d'envisager la mise en place de l'approche de reconnaissance *séquentielle*, décrite dans le chapitre précédent, qui associe à chaque mot reconnu une ou plusieurs catégories sémantiques.

Dans ce chapitre, nous proposons donc une méthodologie pour générer des images d'actes synthétiques. Nous étudions également l'impact de ces données synthétiques lors de l'apprentissage du modèle de reconnaissance d'écriture. Enfin, nous présentons quelques résultats obtenus avec cette approche et démontrons que l'utilisation de documents synthétiques rend possible la reconnaissance de registres paroissiaux.

## 6.2 Production d'actes synthétique réalistes

La génération d'images d'actes synthétiques passe par trois étapes majeures que nous illustrons sur la figure 6.1. La première étape consiste à générer un contenu textuel semblable aux textes issus d'actes réels. La seconde étape permet de générer l'image correspondant au texte simulé à l'aide de polices manuscrites. Enfin, la dernière étape vise à accroître le réalisme et la variabilité des actes générés en appliquant des transformations aux images.

### 6.2.1 Génération du texte

Les registres paroissiaux sont composés d'actes dont le texte est structuré, récurrent et basé sur un vocabulaire spécifique. Par exemple, deux actes de baptême partagent de nombreuses similarités, car ils contiennent les mêmes informations, souvent écrites dans le même ordre. Nous souhaitons identifier ces structures d'actes afin de générer des textes qui ressemblent à ceux des actes des registres paroissiaux. La similarité entre les textes générés et réels doit permettre au modèle de reconnaissance d'écriture d'apprendre des caractéristiques linguistiques propres au vocabulaire et au style de l'époque.

Pour générer ces textes, nous modélisons les structures de phrase récurrentes dans


**Figure 6.1** – Illustration de la méthodologie mise au point pour générer des actes synthétiques.

les actes. Les informations variables (noms, prénoms, âge, métier, date...) sont ensuite modifiées pour créer des actes variés à partir d'une même structure. Cette approche est particulièrement intéressante, car elle permet de produire des transcriptions réalistes dont la structure sémantique est connue. Ainsi, la catégorie sémantique et le rôle de la personne relatifs à chaque mot sont connus.

#### 6.2.1.1 Modélisation des structures de phrases

Nous souhaitons identifier les structures de phrase récurrentes pour chaque type d'acte afin de générer des textes dans ce style. Pour cela, les transcriptions de la base HTR-BMS sont analysées afin de repérer les informations qui apparaissent fréquemment dans les actes, ainsi que leur ordre d'apparition. Les actes partageant la même structure sont ensuite mis en commun afin de définir une structure globale permettant de simuler ce type de texte.

Dans la figure 6.2, nous présentons la méthodologie employée pour modéliser un style d'acte de sépulture. L'analyse des quatre actes sur la figure 6.2a permet d'isoler une structure commune présentée sur la figure 6.2b. L'acte commence par le prénom de la personne, parfois suivi du nom de la personne. Si la personne est mineure, le prénom et

nom de ses parents sont souvent écrits. Il est ensuite précisé l'âge de la personne au moment du décès, ainsi que la date et le lieu du décès ou de l'inhumation. D'autres informations facultatives peuvent apparaître, comme les noms et prénoms de ses parents ou la liste des personnes présentes à la cérémonie. Cette méthodologie a été utilisée pour modéliser des actes de baptême, mariage, bans, fiançailles et sépulture. La table 6.1 dénombre les structures extraites pour chaque type d'acte.

Ces structures servent de squelette à la génération des actes synthétiques. Pour chaque acte généré, une structure est sélectionnée aléatoirement. Certains morceaux de phrase facultatifs sont ensuite supprimés afin d'augmenter la variabilité des transcriptions générées tout en garantissant un texte cohérent. Par exemple, pour un acte de sépulture, la date et le lieu du décès n'apparaissent pas systématiquement. La casse des noms propres est également transformée aléatoirement afin de correspondre à l'usage des majuscules dans les actes. Enfin, certains mots sont remplacés par des synonymes. Par exemple, l'expression « *décéda* » peut être remplacée par « *est décédé.e* », « *a été retrouvé mort.e* », ou encore « *est mort.e* ».

**Table 6.1** – Dénombrement des structures extraites par type d'acte.

| Type d'acte | Transcriptions | Structures extraites |
|---|---|---|
| Baptême | 42 | 5 |
| Bans | 13 | 6 |
| Fiançailles/promesses | 6 | 5 |
| Mariage | 23 | 3 |
| Sépulture | 56 | 7 |
| Autre | 6 | 0 |
| Total | 146 | 26 |

#### 6.2.1.2 Complétion des structures avec des informations variables

Les structures extraites doivent ensuite être complétées avec des informations variables, afin de créer des actes différents basés sur une même structure. Il est donc nécessaire de générer des prénoms, noms, métiers ou encore des dates.

**Prénom et nom** Les noms et prénoms sont tirés aléatoirement à partir de 5 447 noms et 712 prénoms relevés. Les prénoms ont préalablement été genrés à l'aide d'un modèle de classification : 377 prénoms masculins et 335 prénoms féminins ont été extraits.

**(a)** Quatre actes de sépulture issus de différentes paroisses et périodes.

*{prenom_principal} <{nom_principal}>*
*<fil.s.le de {prenom_pere} {nom_principal} <et {prenom_mere} {nom_mere}> >*
*âgé de <environ> {age_deces}*
*<est mort {date_deces} {lieu_deces}>*
*<a été inhumé {date_inhumation} {lieu_inhumation}>*
*<en présence de {témoins}> <{signatures}*

**(b)** Une structure générique basée sur ces quatre actes.

**Figure 6.2** – Modélisation du texte issu des actes. Les transcriptions des actes partageant le même squelette sont regroupées. Une structure d'acte est synthétisée et peut ensuite être utilisée afin de générer des actes similaires. Légende : {champ à peupler}, <facultatif>, synonymes possibles.

**Date** La date est générée aléatoirement entre le XVI$^{e}$ et le XVIII$^{e}$ siècle. Le jour, le mois et l'année sont ensuite formatés aléatoirement en lettres ou en format numérique. Les mois peuvent être écrits de différentes manières. En particulier, les mots de septembre, octobre, novembre, décembre s'écrivent fréquemment « *7bre* », « *8bre* », « *9bre* », « *10bre* » ou « *Xbre* ». Enfin, dans certains cas, la date exacte n'est pas précisée, et est remplacée par des expressions relatives : « *le même jour* », « *le jour suivant* », ou « *aujourd'hui* ». Ainsi, la date du 11/12/1642 peut être écrite de différentes façons, toutes équiprobables :

— « *le onze décembre mil six cent quarante-deux* »
— « *le onzième jour de décembre 1642* »
— « *le onze 10bre 1642* »
— « *le 11 Xbre 1642* »
— « *ledit jour et an ci-dessus* »
— ...

**Âge** L'âge des personnes est tiré aléatoirement entre 0 et 85 ans, et est formaté aléatoirement en lettres ou en nombres. S'il est inférieur à deux ans, le nombre de mois est également précisé.

**Lieu** Plusieurs types de lieux ont été identifiés : 418 lieux-dits ont été extraits des relevés auxquels nous avons ajouté les 333 communes d'Ille-Et-Vilaine. Des lieux spécifiques ont été créés pour les lieux de décès : « *chez lui* », « *chez son oncle* », « *à son domicile* ». Certains lieux religieux ont également été définis pour certaines cérémonies : « *dans l'église de {ville}* » ou « *dans le cimetière de {lieu_dit}* »

**Titre et métiers** Certaines personnes sont caractérisées aléatoirement par des adjectifs, comme « *feu* », « *défunt* », ou « *honorable* ». Nous avons également identifié 22 métiers à partir des relevés.

**Témoins** Pour un baptême, la liste des témoins inclut un ou les deux parents et éventuellement d'autres membres aléatoires de la famille (tante, oncle, grands-parents). Pour un mariage, la liste des témoins inclut les parents des deux mariés, ainsi que d'autres personnes. Enfin, pour une sépulture, elle est composée du conjoint, des parents ou des enfants, ainsi que potentiellement d'autres personnes.

### 6.2.2 Génération d'images d'actes synthétiques

L'image associée au texte doit ensuite être générée. Pour cela, nous utilisons des polices manuscrites de style historique. Des transformations sont ensuite appliquées aux images afin de rendre les images plus réalistes et d'ajouter de la variabilité.

#### 6.2.2.1 Génération de l'image initiale

Nous avons sélectionné 22 polices manuscrites historiques issues des fonderies typographiques P22 [1] et OlfFonts [2]. Ces polices ont été sélectionnées pour leur apparence proche de celles des prêtres. L'image associée au texte est générée avec l'une police sélectionnée aléatoirement, grâce à la bibliothèque Python Imaging Library (PIL).

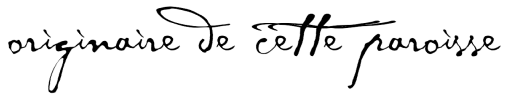
Certaines caractéristiques sont également sélectionnées aléatoirement, en particulier la distance interligne, l'inclinaison du texte, ou encore l'épaisseur du trait. Certaines polices permettent un niveau supplémentaire de personnalisation, comme la possibilité de sélectionner différentes variantes pour certains caractères, de modifier des ligatures, ou d'utiliser des styles historiques pour certaines lettres. Quelques-unes de ces caractéristiques sont illustrées sur la figure 6.3. La police présentée permet, par exemple, des variations stylistiques ou contextuelles, principalement visibles sur le double « `t` » et le « `d` ». Elle permet également d'utiliser des variations historiques de certaines lettres, comme le « `s` » long, que l'on retrouve sur de nombreux actes. En plus de ces variations, certaines polices proposent plusieurs styles pour chaque lettre de l'alphabet, ce qui accroit la variabilité des styles d'écriture.

Des retours à la ligne sont ensuite ajoutés afin que les nombre de caractères par ligne soit réaliste et homogène. L'analyse statistique des transcriptions réelles montre que les lignes contiennent généralement entre 20 et 120 caractères par lignes. Le nombre de caractères par ligne est donc tiré aléatoirement entre ces bornes et les retours à la ligne sont effectués de sorte que les mots ne soient pas coupés.

#### 6.2.2.2 Simulation d'exemples réalistes

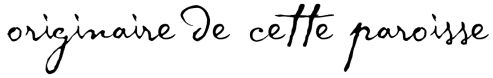
Certaines transformations ont été implémentées afin de rendre les images plus réalistes, dont certaines sont directement tirées du logiciel DocCreator [Journet 2017].

1. `https://p22.com/`
2. `https://www.oldfonts.com/`

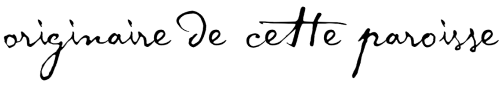


**(a)** Police originale

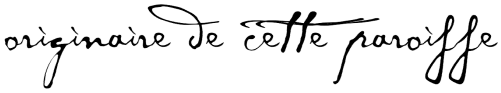
**(b)** Inclinaison aléatoire du texte



**(c)** Variations stylistiques (option `salt`). Le style des lettres `d` et `t` varie.



**(d)** Variations contextuelles (option `calt`)



**(e)** Variations historiques (option `hist`)

**Figure 6.3** – Illustration des différentes variations de la police OpenType P22Cezanne.

**Courbures des lignes de texte** Nous souhaitons reproduire les courbures du texte présent sur les actes des registres paroissiaux. Pour cela, nous avons extrait les lignes polygonales représentant les courbures des 21 000 lignes de texte de la base de données DLA-BMS-1. Ces courbures sont ensuite appliquées de manière aléatoire à chaque acte synthétique, en déplaçant les pixels colonne par colonne.

**Ajout de tâches** Les actes peuvent présenter des taches d'encre ou des ratures. Nous simulons ces artefacts en utilisant les motifs de tâches du logiciel DocCreator. Ces motifs sont redimensionnés et collés aléatoirement dans l'image. Le nombre d'artefacts par image synthétique varie aléatoirement entre 1 et 5.

**Coloration de l'encre** L'encre des registres paroissiaux est rarement complétement noire. Nous colorons le texte des actes synthétiques avec une couleur sombre aléatoire pour laquelle chaque valeur RGB est tirée aléatoirement entre 0 et 80.

**Effacement de l'encre** De la même façon, l'encre des actes n'est jamais uniforme. Nous simulons un effet d'effacement de l'encre à l'aide d'un bruit de Perlin [Perlin 2002] à l'image synthétique. Le coefficient du bruit de Perlin est tiré aléatoirement pour chaque image entre 0.2 et 0.8.

**Ajout d'un fond** La dernière transformation consiste à ajouter un fond de papier à l'acte synthétique. Pour cela, nous avons sélectionné certaines pages de registres paroissiaux qui ne contiennent pas de texte : 30 images de registres paroissiaux et 20 images de documents variés issus du logiciel DocCreator. Pour chaque image synthétique, un fond est tiré aléatoirement et l'image synthétique est collée par transparence à un endroit aléatoire de la page.

#### 6.2.2.3 Simulation d'exemples variés

Nous avons également implémenté d'autres transformations afin d'accroître la variabilité des images synthétiques. Chacune de ces transformations est appliquée avec une probabilité $p = 0.3$. La figure 6.4 présente quelques exemples d'actes synthétiques obtenus après l'application des différentes transformations.

**Apparence du texte** Les opérations morphologiques de dilatation et d'érosion sont appliquées à ces images synthétiques. Elles permettent d'impacter la taille du trait : l'érosion diminue le trait de la plume, alors que la dilatation grossit le trait. Le cisaillement est également utilisé pour modifier l'angle des caractères, ce qui impacte l'inclinaison du texte vers la gauche ou la droite.

**Prise de vue** Des déformations élastiques sont également appliquées au texte à l'aide d'une grille, afin de déformer localement les lettres et accroitre leur variabilité. Une rotation ou un changement de perspective aléatoire peuvent être appliqués au texte.

**Couleurs de l'image** Des transformations sont appliquées pour modifier l'intensité de l'image. Le contraste, la netteté, les couleurs et la luminosité de l'image sont ainsi ajustés de façon aléatoire, et certaines images sont également transformées en négatif.

**Qualité de l'image** Les images sont également bruitées en utilisant un bruit gaussien et un bruit de Poisson. Un flou de mouvement est appliqué afin de simuler un effet de mouvement de l'appareil photo lors de la prise de vue. L'algorithme de compression JPG est également appliqué. Enfin, la netteté et la résolution de l'image sont modifiées aléatoirement.

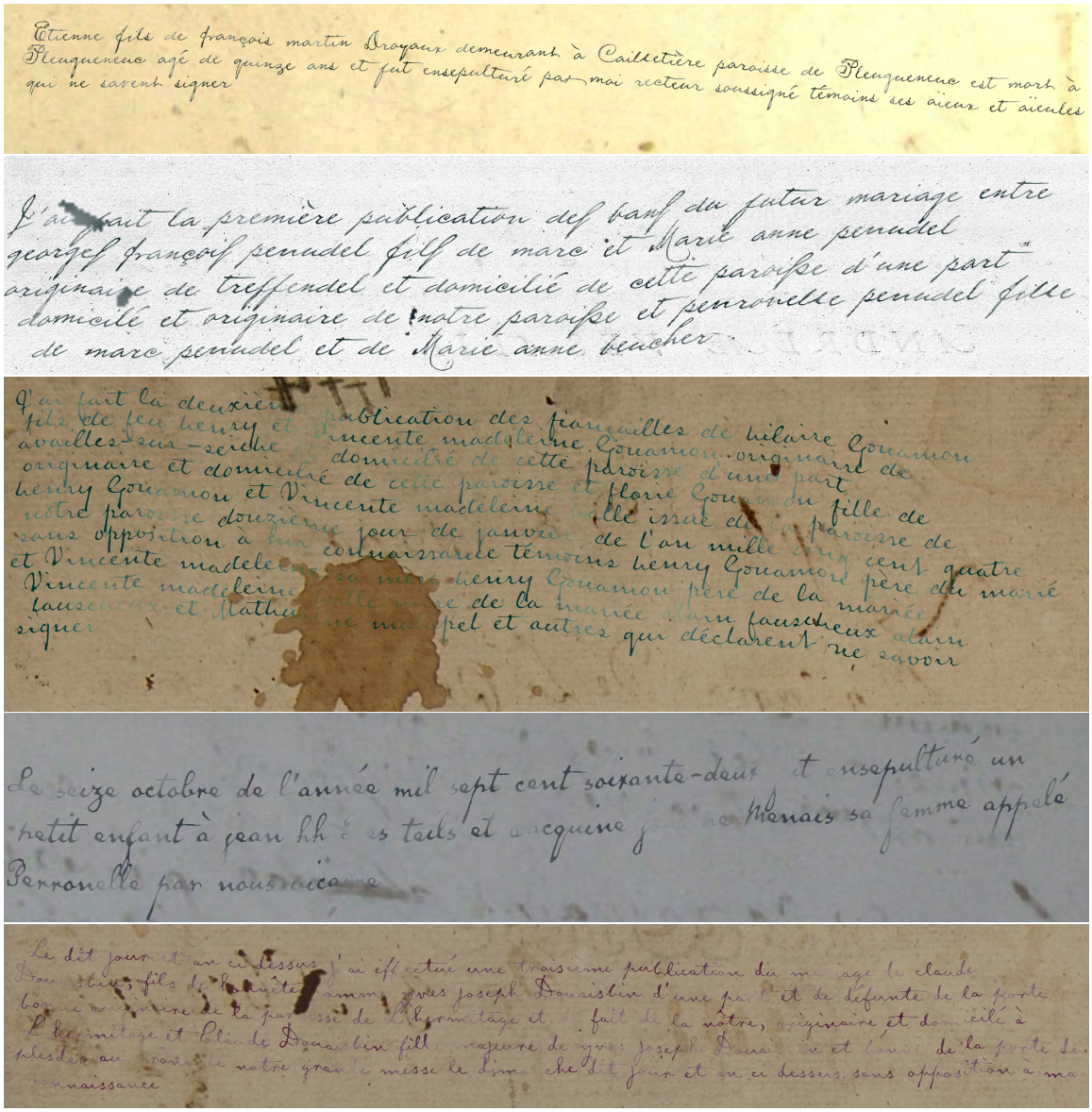


**Figure 6.4** – Exemples d'actes synthétiques générés avec notre méthodologie.

### 6.2.3 Génération d'images de lignes de texte synthétiques

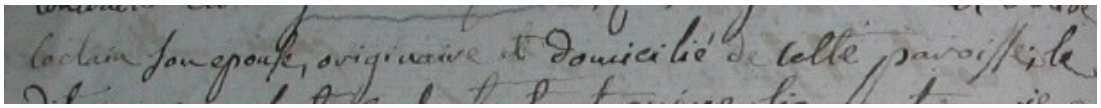
Notre méthodologie permet de générer des images d'actes synthétiques. Cependant, notre modèle de reconnaissance d'écriture prend en entrée des images de lignes de texte. Nous proposons donc trois stratégies pour générer des lignes de texte synthétiques. Celles-ci sont illustrées et comparées dans la figure 6.5.

**Génération de lignes de texte isolées** La première méthode consiste à segmenter le texte de l'acte en lignes, afin de générer une image par ligne de texte. Cette méthode est simple, efficace et rapide, mais le rendu n'est pas très réaliste, car les lignes de texte sont isolées : les ascendants et descendants des lignes voisines n'apparaissent pas sur l'image.

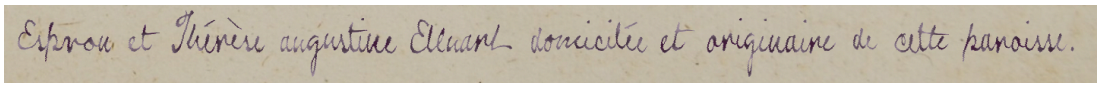
**Génération d'actes puis segmentation basée sur les caractéristiques de la police** Cette méthodologie consiste à effectuer la segmentation en lignes à partir de l'image de l'acte. La boite englobante des lignes de texte est calculée à partir des caractéristiques typographiques de la police. Ainsi, selon la présence de hampes ou jambages, les imagettes de lignes ne sont pas toutes de la même hauteur. Les coordonnées sont mises à jour lors de l'application des transformations de dégradation et d'augmentation, ce qui permet de connaître les boites englobantes de l'image transformée. Avec cette méthode, les boites englobantes sont suffisamment grandes pour englober tous les glyphes de la ligne. En conséquence, les descendants de la ligne précédente et les ascendants de la ligne suivante peuvent apparaître dans l'image segmentée.

**Génération d'actes puis segmentation basée sur le calcul de l'interligne** Cette méthodologie consiste également à effectuer la segmentation en lignes à partir de l'image de l'acte. La boite englobante des lignes de texte est calculée à partir de la connaissance des lignes de base et du calcul de l'interligne. Cette méthode est celle que nous avons utilisée pour créer les images de lignes à partir de registres paroissiaux réels. Là encore, les ascendants et descendants des lignes voisines peuvent apparaître dans l'image segmentée, comme pour les lignes réelles.

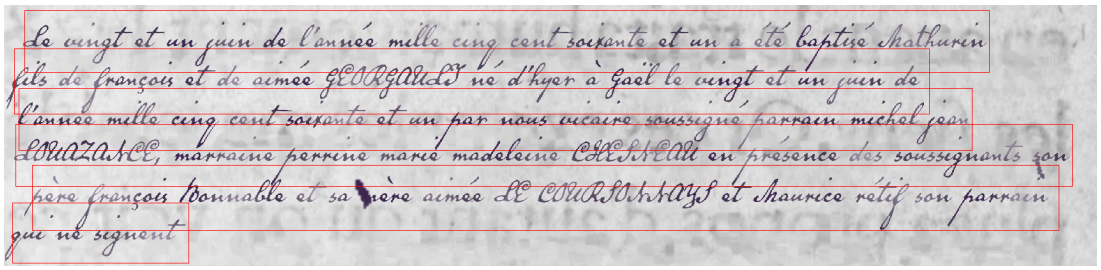
**Résultats** Nous comparons ces différentes méthodes de génération de lignes en évaluant leur impact sur les performances du modèle de reconnaissance d'écriture. L'apprentissage se fait en utilisant 16377 lignes de texte réelles de la base HTR-générique et 16377 lignes synthétiques. La validation et l'évaluation se fait sur la base de données HTR-BMS.



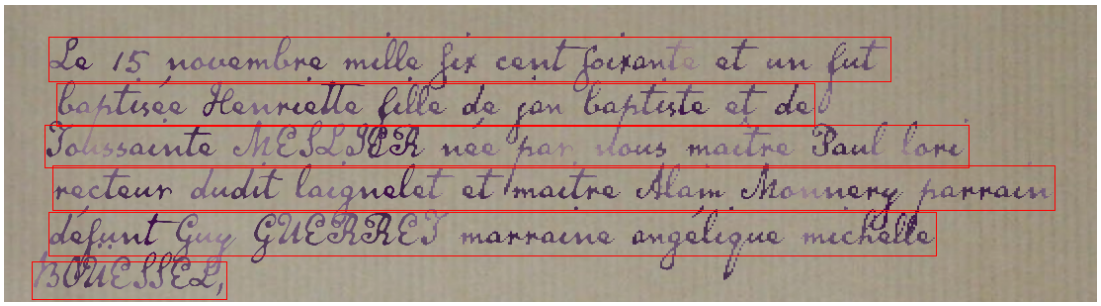
**(a)** Exemple de ligne de texte issue d'un acte réel après redressement de la courbure.



**(b)** Génération d'une ligne isolée. Le texte initial de l'acte est découpé en lignes, puis une image est générée pour chaque ligne de texte.



**(c)** Méthode basée sur les caractéristiques de la police. L'image de l'acte est générée, puis la segmentation est effectuée grâce aux caractéristiques de la police. La hauteur de la boite englobante est calculée de façon à ce que les ascendants et descendants soient complètement inclus dans l'image. En conséquence, elle peut inclure du texte des lignes adjacentes.



**(d)** Méthode basée sur l'interligne. L'image de l'acte est générée, puis la segmentation est effectuée grâce à la connaissance des lignes de base et de l'interligne. Avec cette méthode, les ascendants et descendants peuvent être coupés.

**Figure 6.5** – Un exemple de ligne réelle comparée à trois méthodes de génération de lignes de texte synthétiques.

Les résultats présentés dans la table 6.2 montrent que la génération de lignes synthétiques isolées est la méthode la moins convaincante. Elle mène à une augmentation des taux d'erreur caractères et mots d'environ 7% par rapport aux deux autres méthodes basées sur une segmentation de l'acte synthétique. Ces deux autres méthodes obtiennent des résultats comparables, avec une légère amélioration des performances pour la méthode basée sur les caractéristiques de la police. Nous rappelons que les images de lignes de la base HTR-BMS ont été créées à partir d'une méthode basée sur l'interligne. Ainsi, la méthode de segmentation basée sur la police ne bénéficie pas de ce biais. Nous retenons cette méthode de segmentation pour la suite.

**Table 6.2** – Évaluation des différentes méthodes de génération de lignes synthétiques sur la base BMS-HTR. Le modèle est appris sur 16 377 lignes synthétiques et 16 377 lignes réelles issues de la base HTR-générique.

| Méthode de génération des lignes | CER (%) ⬇ | WER (%) ⬇ |
|---|---|---|
| Lignes isolées | 30.44 | 75.34 |
| Segmentation basée sur la police | **28.44** | **72.68** |
| Segmentation basée sur l'interligne | 28.49 | 74.53 |

## 6.3 Apprentissage combinant données réelles et synthétiques

Nous souhaitons à présent optimiser l'utilisation des données synthétiques pour l'apprentissage d'un système de reconnaissance d'écriture. Pour cela, nous comparons différents scénarios d'apprentissage en combinant des données réelles et synthétiques. Nous étudions également l'impact de la sélection des bases de données réelles sur les performances du système.

### 6.3.1 Remplacer les lignes de texte réelles par des lignes de texte synthétiques

Dans un premier temps, nous souhaitons évaluer différents scénarios de mélange de données réelles et synthétiques avec un ensemble d'apprentissage de taille constant. Pour cela, nous sélectionnons aléatoirement des données réelles de la base de données HTR-générique, que nous remplaçons par des données synthétiques. La taille de l'ensemble

d'apprentissage reste donc constante, égale à la taille de la base de données HTR-générique (32 755 lignes). Nous rappelons que la base d'apprentissage ne contient aucun registre paroissial réel. Nous comparons différents ratios de données réelles remplacées par des données synthétiques.

**Table 6.3** – Scores sur BMS-HTR (test) pour la reconnaissance d'écriture. Pour ces expérimentations, la taille de l'ensemble d'apprentissage est constante à 32 755 : un pourcentage aléatoire de données réelles est remplacé par des données synthétiques. La base d'apprentissage ne contient aucun registre paroissial réel.

| Lignes synthétiques en remplacement (%) | CER (%) ⬇ | WER (%) ⬇ |
|---|---|---|
| 0 | 33.87 | 82.84 |
| 25 | **25.79** | **70.70** |
| 50 | 28.44 | 72.68 |
| 75 | 28.74 | 73.24 |
| 100 | 37.03 | 80.29 |

Les résultats de la table 6.3 démontrent l'intérêt des données synthétiques. En effet, remplacer 25% de données réelles par des données synthétiques se traduit par une baisse d'environ 25% du taux d'erreur caractère, à taille de l'ensemble d'apprentissage constant. L'amélioration est également visible sur le taux d'erreur mot, qui diminue de 82.8% à 70.70%. À l'inverse, les résultats montrent également l'intérêt d'utiliser des données réelles pendant l'apprentissage, car les performances du système appris uniquement sur des données synthétiques sont très faibles. Ces résultats montrent donc l'intérêt de combiner des données réelles issues d'une base de données générique à des données synthétiques spécialisées pour l'apprentissage d'un modèle de reconnaissance de registres paroissiaux. Le meilleur compromis semble de combiner 25% de données synthétiques et 75% de données réelles.

Cependant, dans cette configuration, les lignes réelles sont retirées de l'ensemble d'apprentissage de façon aléatoire. Or, certaines d'entre elles semblent plus adaptées pour l'apprentissage d'un système de reconnaissance de registres paroissiaux. Par exemple, les documents issus des bases de données READ ou HTR-Paléo ressemblent aux registres paroissiaux : leur retrait de l'ensemble d'apprentissage pourrait impacter négativement les performances du système. Par conséquence, remplacer des données réelles peut être dangereux.

### 6.3.2 Compléter les lignes de texte réelles avec des lignes de texte synthétiques

Pour dépasser cette limite, nous effectuons la même expérimentation en gardant tous les documents réels dans l'ensemble d'apprentissage. Ainsi, nous conservons le même nombre de lignes synthétiques dans l'ensemble d'apprentissage, mais combinées à toutes les lignes issues de la base de données HTR-générique, soit 32 755 lignes réelles et 8 188 lignes synthétiques.

Le tableau 6.4 compare les résultats lorsque ces lignes synthétiques remplacent ou complètent les documents réels de la base de données HTR-générique. Les résultats montrent une légère amélioration des performances lorsque la base HTR-générique est utilisée dans son intégralité, avec 25% de données synthétiques supplémentaires. En revanche, nous observons que l'ajout d'un pourcentage plus élevé de données synthétiques dégrade les performances du système. Une surreprésentation des images synthétiques dans l'ensemble d'apprentissage n'est donc pas recommandée.

**Table 6.4** – Scores sur BMS-HTR (test) pour la reconnaissance d'écriture. Pour ces expérimentations, la taille de l'ensemble d'apprentissage augmente : les données synthétiques sont ajoutées aux 30 000 lignes de la base HTR-générique. La base d'apprentissage ne contient aucun registre paroissial réel.

| Lignes synthétiques (%) | N_train | CER (%) ⬇ | WER (%) ⬇ |
|---|---|---|---|
| 25 (remplacement) | 32755 | 25.79 | 70.70 |
| 25 (complément) | 40943 | **25.21** | **67.97** |
| 50 (complément) | 49132 | 28.03 | 73.94 |

### 6.3.3 Sélection des documents réels

Nous avons démontré que l'ajout de données synthétiques spécialisées à des données réelles génériques permet d'améliorer la reconnaissance des registres paroissiaux. Nous souhaitons à présent évaluer l'influence des données réelles utilisées pour l'apprentissage. Par exemple, la base de données IAM est très différente des registres paroissiaux : les documents sont modernes et rédigés en anglais. Ainsi, il est légitime de se demander si l'apprentissage sur cette base de données a un intérêt pour la reconnaissance de registres paroissiaux. Au contraire, les documents de la base de données HTR-Paleo, détaillée dans le chapitre 1, sont très proches des registres paroissiaux, car elle est composée de

documents français datant de la même époque. De la même façon, les documents de READ et Esposalles sont historiques et présentent certaines caractéristiques visuelles communes aux registres BMS, bien que la langue soit différente. Enfin, la base de données RIMES présente du texte moderne, mais rédigé en français.

Pour répondre à cette question, nous entraînons notre architecture sur différentes combinaisons de bases de données réelles, avec 25% de données synthétiques supplémentaires :

— *Générique* aussi appelée HTR-générique, qui contient toutes ces bases de données, (32 755 lignes pour l'apprentissage) ;

— *Hist-1* qui combine les bases READ (historique) et HTR-Paleo (historique et français)- soit les deux bases de données qui ressemblent le plus au style d'écriture des registres paroissiaux (10 717 lignes pour l'apprentissage) ;

— *Hist-2* qui contient READ et HTR-Paleo, mais également la base de données Esposalles qui contient des registres historiques catalans ressemblant aux registres paroissiaux, mais très homogène (13 346 lignes pour l'apprentissage) ;

— *Hist-Fr* qui contient READ et HTR-Paleo, mais également la base de données RIMES qui contient des documents français moderne, qui peut permettre d'apprendre un modèle de langue implicite (20 664 lignes pour l'apprentissage).

Les résultats du tableau 6.5 montrent que la sélection de bases de données cohérentes n'a pas d'intérêt, et qu'il est préférable d'utiliser toutes les données disponibles.

**Table 6.5** – Influence des données réelles sur la performance du système de reconnaissance d'écriture sur BMS-HTR. Différentes combinaisons de bases de données réelles sont comparées, avec un pourcentage fixe (25%) de données synthétiques ajoutées lors de l'apprentissage.

| | Base de données réelles | | | | | CER (%) ⬇ | WER (%) ⬇ |
|---|---|---|---|---|---|---|---|
| | IAM | RIMES | Esposalles | READ | HTR-Paleo | | |
| Générique | ✔ | ✔ | ✔ | ✔ | ✔ | **25.21** | **67.97** |
| Hist-1 | | | | ✔ | ✔ | 28.44 | 72.68 |
| Hist-2 | | | ✔ | ✔ | ✔ | 37.03 | 80.29 |
| Hist-Fr | | ✔ | | ✔ | ✔ | 28.74 | 73.24 |

## 6.4 Résultats finaux et discussion

Nous avons montré que la combinaison de données synthétiques spécialisées et de données réelles génériques permet de diminuer le taux d'erreur caractères de 25.6%. Dans

cette dernière section, nous présentons des exemples de transcriptions effectuées par notre modèle et commentons les résultats obtenus. Nous proposons également un exemple de reconnaissance complète du texte et des entités nommées.

La stratégie globale d'apprentissage *séquentielle* mise en place pour la reconnaissance d'écriture et l'extraction d'informations dans les registres paroissiaux est illustrée dans la figure 6.6. Il n'a pas été possible d'appliquer l'approche *combinée*, pourtant plus performante, car les bases de données réelles génériques ne contiennent pas d'annotations sur les entités nommées.

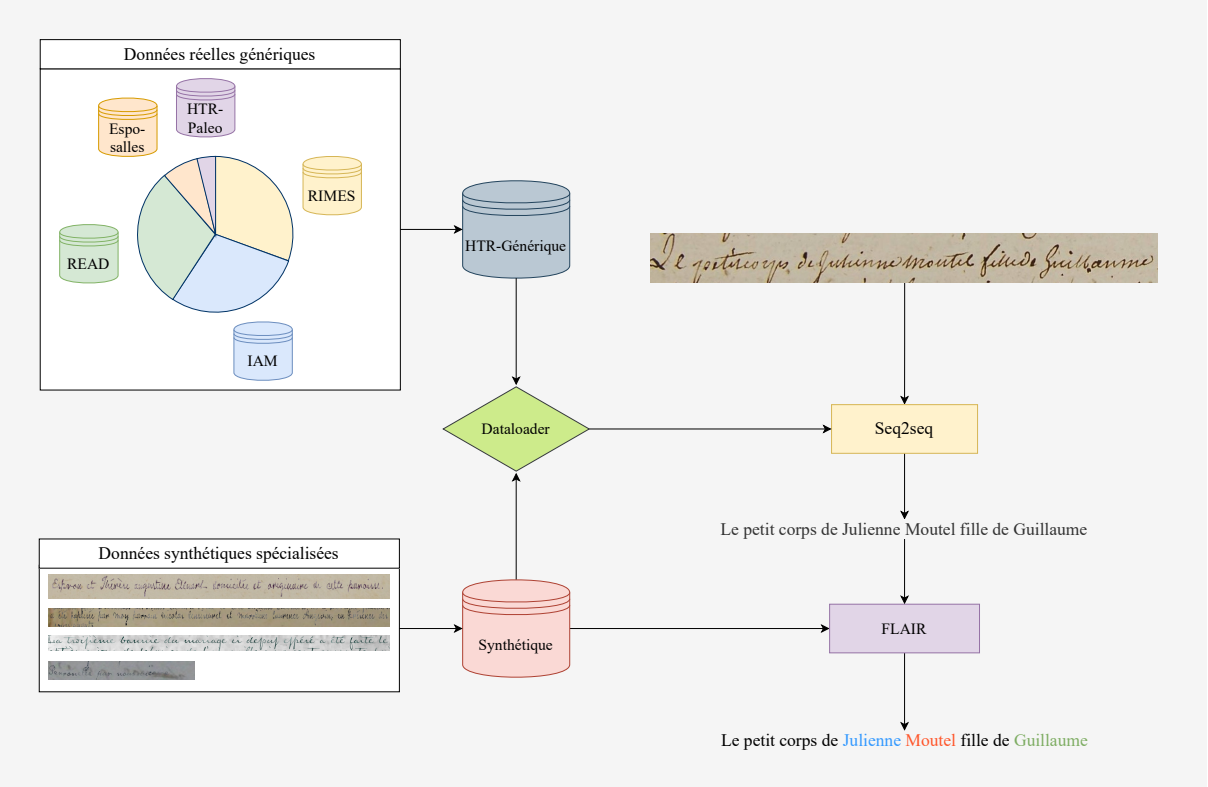
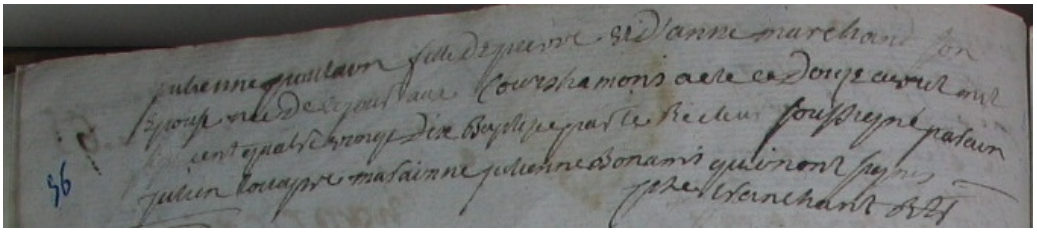



**Figure 6.6** – Illustration du mécanisme d'apprentissage avec des données réelles et synthétiques. Le modèle de reconnaissance d'écriture (Seq2seq) est appris sur des données réelles génériques et des données synthétiques spécialisées. Puis, la transcription prédite est analysée par le modèle de reconnaissance des entités nommées (FLAIR), appris uniquement sur des données synthétiques, afin d'associer chaque mot prédit à une catégorie sémantique et un rôle de personne.
Légende : prénom de la personne principale, nom de la personne principale, prénom du père.

### 6.4.1 Reconnaissance d'écriture

Nous avons montré que l'utilisation de données synthétiques permet d'envisager une reconnaissance automatique de l'écriture des registres paroissiaux lorsque aucun exemple réel n'est disponible pour l'apprentissage.

Nous présentons ici quatre exemples de transcriptions effectuées par notre modèle sur les figures 6.7, 6.8, 6.9, et 6.10. Sur ces images, le taux d'erreur caractères varie de 40% à 15%. Nous proposons une analyse qualitative des erreurs effectuées par le réseau afin d'expliquer ces variations et d'envisager des pistes d'améliorations.



**(a)** Image

*julienne poulain fille de pierre et d'anne marchand son*
*epouse nee de ce jour aux cours hamons a ete ce douze avril mil*
*sept cent quatre vingt dix Baptisee par le Recteur soussigné parain*
*julien louapre marainne julienne Bonami qui n'ont signes*

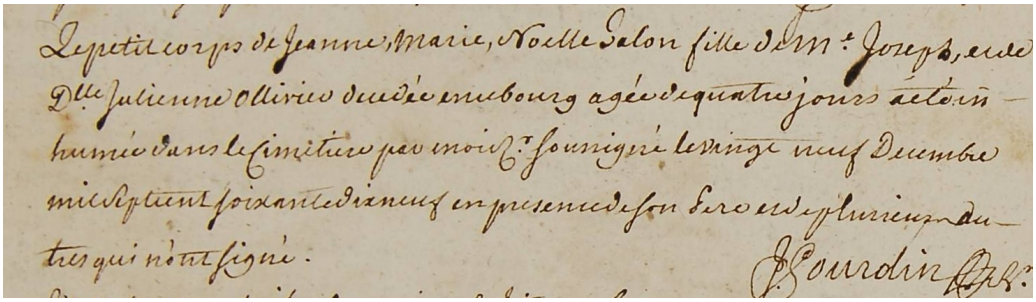
**(b)** Vérité

*entrenne quantaire fille de parres et janses marchen*
*le pour mon de monsieur lourstamons aire ce doupeures en*
*je vous quatre rouge des bayelle parle recedur soussigne partur*
*quille louape marainne pertenire sonant querront pages*

**(c)** Prédiction

**Figure 6.7** – Un exemple de transcription sur une image de faible résolution (180 dpi) pour laquelle le taux d'erreur caractères s'élève à 41.4%.

**Alphabet limité** Nous avons choisi de ne pas prendre en compte les caractères spéciaux lors de l'apprentissage du modèle : les majuscules sont transformées en minuscules, les caractères accentués sont remplacés par des caractères non accentués et caractères spéciaux sont supprimés. Ce choix est motivé par deux raisons. D'une part, ces caractères n'apparaissent que ponctuellement dans l'ensemble d'apprentissage, ce qui produit un déséquilibre des classes lors de l'apprentissage. D'autre part, les majuscules et les accents



**(a)** Image

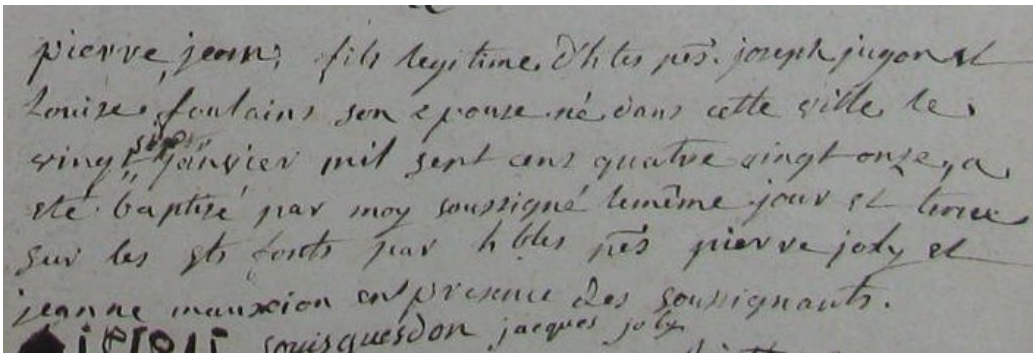
*Le petit corps de Jeanne Marie Noelle Salon fille de Me Joseph et de Dlle Julienne decedée en ce bourg agée de quatre jours a eté in- Humée dans le cimetière par moi Cr soussigné le vingt neuf decembre Mil sept cent soixante dix neuf en presence de son Pere et de plusieurs autres qui n'ont signé*

**(b)** Vérité

*lepetit corps de jenne marie noulle salon fille derne joreps ecde du suivente ollivier duder eubourg ager de quatrejours actoin trumie dans le fimitier par moi tr soussignes levringe nus demembre minseptent soixante dieneus en presencedesson serverseplusieurs de tre qui nont figne*

**(c)** Prédiction

**Figure 6.8** – Un exemple de transcription sur une image de bonne résolution (300dpi) pour laquelle le taux d'erreur caractères s'élève à 26.05%



**(a)** Image

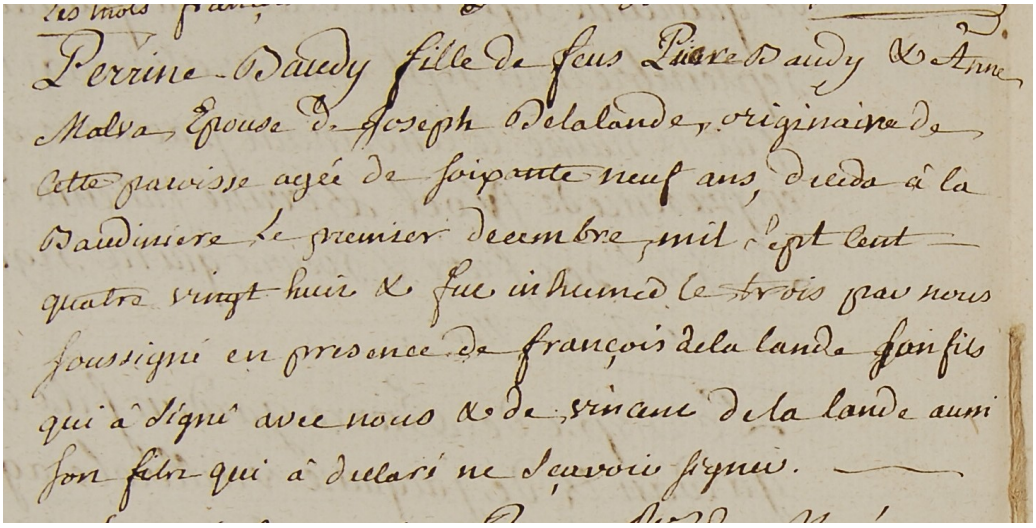
*pierre jean fils legitime d'htes pes joseph jugon et*
*louise foulains son epouse né dans cette ville le*
*vingt sept janvier mil sept cent quatre vingt onze a*
*ete baptisé par moy soussigne le même jour et tenu*
*sur les sts fonts par hbles pes pierre joly et*
*jeanne mauxion en presence des soussignants*

**(b)** Vérité

*pienve jeans fils rey time dh les res joreple jugonst*
*lonire foutains son eponse ne dians cette ville de*
*eing de janvier mit sent ans quintre singt onsega*
*ete baptise par moy soussigne rememe jour st line*
*sur les sts louts jur 10 blis nes pierre jotz et*
*reanne mauxion en presence des sonsignants*

**(c)** Prédiction

**Figure 6.9** – Un exemple de transcription sur une image de faible résolution (180 dpi) pour laquelle le taux d'erreur caractères s'élève à 19.87%.



**(a)** Image

*Perrine Baudy fille de feus Pierre Baudy & Anne Malva epouse de Joseph Delalande originaire de cette paroisse agée de soixante neuf ans deceda à la Baudiniere le premier decembre mil sept cent quatre vingt huit & fut inhumee le trois par nous soussigné en presence de françois de la lande son fils qui a signé avec nous & de vincent de la lande aussi son fils qui a declaré ne scavoir signer*

**(b)** Vérité

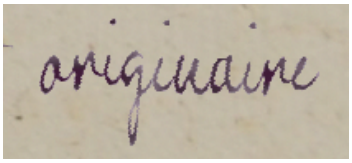
*perrine bandy fille de feus piere bandy de alve malva pouse de hosept betalande originairre de cette paroisse agee de hoiprente neuf ans decede a la baudimere te premier decembre mit ept lent quatre vingt huis x feus in numed le trois pran nous houssigne en presence de francois de la lande fan fils qui a signe avec nous de de rivant de la lande ausse hon fils qui a declari ne sauvoir signer*

**(c)** Prédiction

**Figure 6.10** – Un exemple de transcription sur une image de bonne résolution (300 dpi) pour laquelle le taux d'erreur caractères s'élève à 15.01%

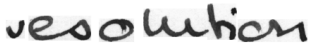
sont utilisés de manière hétérogène par les prêtres sur les registres paroissiaux. Cette homogénéisation des caractères lors de l'apprentissage permet de simplifier le modèle, mais conduit à des erreurs de transcription.
Par exemple, dans la figure 6.8, le nom « Marie » est reconnu « marie » : la différence de casse compte pour une erreur de transcription, même si la bonne lettre est reconnue. La même problématique se pose pour les accents, par exemple sur le mot « né » de la figure 6.9. Enfin, le caractère spécial « `&` » n'est pas reconnu dans la figure 6.10, car il ne fait pas partie de l'ensemble d'apprentissage. À la place, le modèle reconnaît la lettre « `x` » ou les caractères « de », ce qui conduit à des erreurs.
À terme, il sera nécessaire d'identifier les caractères spéciaux ayant une importance ou une fréquence d'apparition élevée dans les transcriptions des actes. La reconnaissance de ces caractères pourra contribuer à la diminution des taux d'erreur.

**Faible résolution des images** La résolution des images de registres paroissiaux peut également être une source d'erreur pour la transcription automatique. En effet, les registres paroissiaux sur lesquelles le modèle est évalué sont de faible résolution (180 dpi), alors que le modèle a été appris sur des documents génériques en bonne résolution (300 dpi). La figure 6.11 met en évidence cette différence de résolution.
Il n'est pas possible de mener une étude statistique sur l'impact de la résolution, car trop peu de registres paroissiaux en bonne qualité (300 dpi) ont été transcrits. En revanche, il est possible de comparer les transcriptions obtenues pour une même image à différentes résolutions. Ainsi, la figure 6.12 présente les transcriptions obtenues sur une image de bonne résolution (300 dpi), qui a été ré-échantillonné à des résolutions plus faibles. Nous observons que le taux d'erreur caractère augmente de 13.1% à 16.9% lorsque l'image est ré-échantillonnée à 180 dpi, et à 23.7% lorsque l'image est ré-échantillonnée à 72 dpi. Cette observation démontre que la faible qualité de numérisation peut avoir mené à des erreurs de transcription.
Dans l'idéal, nous souhaitons que le système de reconnaissance s'adapte à des documents de résolution variable (capture d'écran, image compressée). Pour cela, une première piste consisterait à apprendre le modèle avec des images de résolution variable, en ré-échantillonnant certaines images à 180 dpi. Une autre piste consisterait à s'intéresser aux algorithmes de super-résolution, déjà utilisés pour améliorer la reconnaissance du texte imprimé [Dong 2015] ou des caractères manuscrits isolés [Qian 2020].



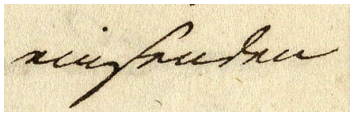
**(a)** Synthétique

**(b)** Esposalles (300 dpi)

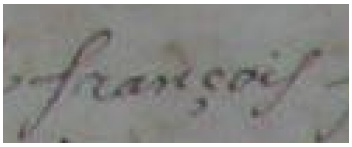


**(c)** IAM (300 dpi)

**(d)** RIMES (300 dpi)



**(e)** READ (300 dpi)

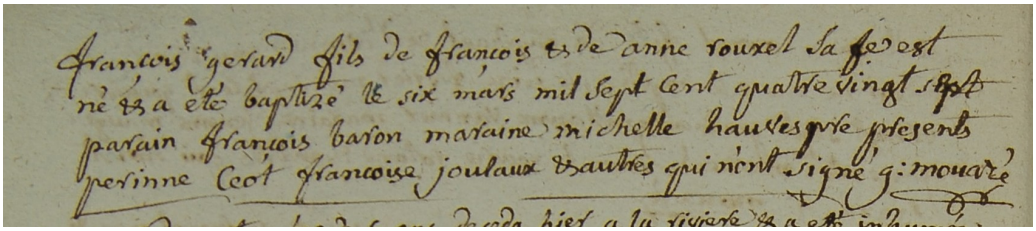


**(f)** HTR-BMS (180 dpi)

**Figure 6.11** – Zoom sur un mot dans des documents issus de différentes bases de données afin d'illustrer les différences de qualité des images. Tous les registres paroissiaux de la base HTR-BMS sont numérisés à 180 dpi.

**Confusion entre les lettres** D'autres erreurs courantes sont dues à la confusion entre plusieurs lettres. Cette confusion peut être induite par différents facteurs.
Certaines erreurs sont liées aux styles d'écriture de certains prêtres. Par exemple, sur la figure 6.9, les lettres « n », « r », et « v » se ressemblent, ce qui provoque des erreurs de reconnaissance (« pierre » reconnu « pienve »). De la même façon, sur l'acte de la figure 6.10, le prêtre trace la lettre « e » avec une largeur très faible : le modèle rencontre donc des difficultés à différencier les lettres « `i` » et « `e` » (« déclaré » reconnu « declari » et « aussi » reconnu « ausse »).
D'autres erreurs sont liées aux ligatures et décorations tracées par certains prêtres. Sur ce même acte, la dernière lettre du mot « inhumé » a été stylisée avec une ligature : le « `e` » a été reconnu comme un « `d` ». De la même façon, sur la figure 6.8, le `c` de « cimetière » est stylisé, il est donc mal reconnu (« fimetier »). Enfin, les erreurs les plus fréquentes viennent de la confusion entre le `s` long et le `f`. Cette confusion est visible sur les figures 6.8 (« signe » reconnu « figne ») et 6.10 (« son » reconnu « fon »).
La spécialisation du modèle sur des registres paroissiaux réels pourrait permettre de corriger certaines de ces erreurs. En effet, celui-ci pourrait s'adapter aux caractéristiques des écritures de l'époque et apprendre des caractéristiques linguistiques spécifiques aux actes.

**Espacement irrégulier entre les mots** Un espacement faible ou inexistant entre les mots peut également mener à des erreurs de transcription automatique.
La figure 6.8 présente, par exemple, un style d'écriture pour lequel les espaces inter-mots sont très faibles (« présence de son » reconnu « presencedesson ») La même problématique



| Transcription vérité |
|---|
| françois gerard fils de françois & de anne rouxel sa fe est<br>né & a été baptizé le six mars mil sept cent quatre vingt sept<br>parain françois baron maraine michelle hauvespre presents<br>perinne ceot francoise joulaux & autres qui n'ont signé |

| Transcription prédite | dpi | CER |
|---|---|---|
| clanteis queard fils de llancois et de anne souxe la feroit<br>ne sta ete baptere le six mais mil sept cent quatre vingt este<br>parrain glamois bason maraine in echelle hauvespele presents<br>resinne lect plancoir joulaux exautres qui nent vigne | 72 | 23.7.1 |
| franceis egeair fils de francois tred anne souxel sa ferest<br>ne es a ete baptise le six mars mil sept cent quatre vingt ent<br>parain francois bavon maraine michelle hauvespre presents<br>perinne lect prancoixe joulaux deaulles qui nent signe | 180 | 16.9 |
| franceis genard fils de francois tede anne souxel sa ferest<br>ne es a ete baptise le six mars mil sept cent quatre vingt enst<br>parain francois bavon maraine michelle hauvespre presents<br>perinne leot prancoixe joulaux davulles qui nent signe | 300 | 13.1 |

**Figure 6.12** – Influence de la résolution sur les performances du système de reconnaissance.

apparaît sur l'image de la figure 6.7. Au contraire, dans certains cas, l'espacement entre des lettres d'un même mot conduit à des erreurs de reconnaissance. Par exemple, sur la figure 6.10, le mot « inhumé » est coupé en deux à cause de l'absence de ligature entre le « n » et le « h ».
L'utilisation d'un modèle de langue implicite ou explicite pourrait permettre au modèle d'inférer la position des espaces dans les transcriptions prédites.

**Ratures et rajouts** Certaines erreurs sont également liées à des taches d'encre, des ratures ou des annotations présentes sur le papier.
Par exemple, sur la figure 6.9, un trait à la fin du mot « jean » fait que le modèle prédit « jeans ». De la même manière, le mot « sept », ajouté par le prêtre entre la deuxième et la troisième ligne, n'est pas correctement reconnu.
Enfin, sur la figure 6.12, le mot sept est raturé et est mal reconnu par le modèle. Ces erreurs pourraient être partiellement corrigées en apprenant le modèles sur des documents contenant ce type de ratures, taches ou annotations. Ces artefacts pourraient également être simulés lors de la génération de documents synthétiques.

**Incertitudes sur la transcription vérité** La transcription manuelle des actes est susceptible de contenir des erreurs ou des interprétations de lecture.
Par exemple, sur la figure 6.12, le nom de la marraine peut être lu de différentes façons : « hauvespre », « hauvesgure » ou « hauvesque ». Sur la figure 6.7, les derniers mots sont écrits sans espace ni apostrophe. Ils ont pourtant été transcrits « qui n'ont signes » par l'annotateur qui a interprété le texte lors de sa lecture.
Une perspective pour améliorer la fiabilité des transcriptions manuelle serait de s'appuyer sur des méthodes de transcription collaborative.

### 6.4.2 Extraction d'information

Nous souhaitons à présent effectuer une analyse qualitative de l'extraction d'information afin d'en tirer des pistes de réflexion pour améliorer le modèle de reconnaissance complet. Nous présentons un exemple de reconnaissance complète réalisée par le modèle, sur la figure 6.13.

Nous avons entraîné un modèle FLAIR [Akbik 2019] sur des actes synthétiques, comme illustré sur la figure 6.6. Une fois appris, le modèle est ensuite utilisé pour prédire les catégories sémantiques et les rôles des personnes à partir de la transcription prédite par le

modèle de reconnaissance d'écriture. Le modèle FLAIR obtient un taux de classification de 83% lorsqu'il est évalué sur la transcription vérité. À titre de comparaison, le même réseau appris sur la base Esposalles atteignait un taux de classification de 97%. Cette observation montre que la tâche de classification est plus difficile sur les registres paroissiaux. En effet, la tâche est plus complète, car nous souhaitons reconnaître des textes plus variés, avec trois types d'actes contre un seul pour la base Esposalles. De plus, le modèle FLAIR apprend à partir d'actes synthétiques qui sont trop propres (pas ou peu de fautes et d'abréviations, peu de variabilité).

La tâche d'extraction est difficile à effectuer sur les registres paroissiaux, car elle consiste à classer chaque mot reconnu par le modèle de reconnaissance d'écriture dans une catégorie sémantique. Or, les taux d'erreur à l'échelle des caractères et des mots sont élevés, ce qui signifie que les mots prédits sont incorrects, et donc difficiles à classer. En conséquence, les données extraites à ce jour sont pour la plupart inexploitables, soit parce que les mots sont incorrectement reconnus, soit parce qu'ils sont mal classés.

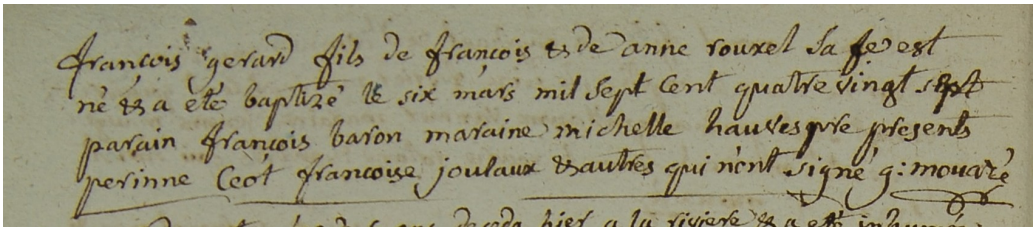
Cependant, sur certaines images lisibles et de bonne qualité, l'extraction automatique d'informations est encourageante. Par exemple, la table 6.13 présente les informations extraites par le système pour un document transcrit avec un taux d'erreur caractères de 15%. Ici, les catégories sémantiques sont bien reconnues, avec un score de reconnaissance basique [3] à 69%. En revanche, la reconnaissance du rôle des personnes est plus problématique : les mots se rapportant à la mère ont été reconnues comme appartenant au père, ce qui fait chuter le taux de reconnaissance.

Au-delà du score, il est légitime de se poser la question de la pertinence de cette métrique pour des applications généalogiques. En effet, dans ce cadre, la connaissance précise du patronyme reste primordial. De même, si la date est importante, il semble plus important de reconnaître correctement l'année que le jour précis. Ainsi, la métrique pourrait être améliorée en pondérant les champs les plus importants pour les généalogistes.

### 6.4.3 Intégration dans le système de reconnaissance complet

Nous avons montré que la génération de données synthétiques a permis de mettre au point ce système de reconnaissance, en l'absence de documents réels dans l'ensemble d'apprentissage. Notre méthode de génération d'actes est utilisée dans un `dataloader` Pytorch, ce qui permet de générer des documents synthétiques à la volée pour l'apprentissage

3. Voir la section 5.2.2 pour la définition de la métrique IEHHR



| Catégorie | Personne | Vérité | Prédiction |
|---|---|---|---|
| Date | Aucune | six mars mil sept cent quatre vingt | six mars mil sept cent quatre vingt |
| Prénom | Principal (M) | françois gerard | franceis egeraid |
| Prénom | Père | françois | francois tres anne |
| Nom | Père | | souxel |
| Prénom | Mère | anne | |
| Nom | Mère | rouxel | |
| Prénom | Parrain | françois | francois |
| Nom | Parrain | baron | bavon |
| Prénom | Marraine | michelle | michelle |
| Nom | Marraine | hauvesque | hauvespre |

**Figure 6.13** – Une image de bonne qualité sur laquelle de l'information peut être extraite. Le score basique de cette image est de 69% et le score complet est de 41%

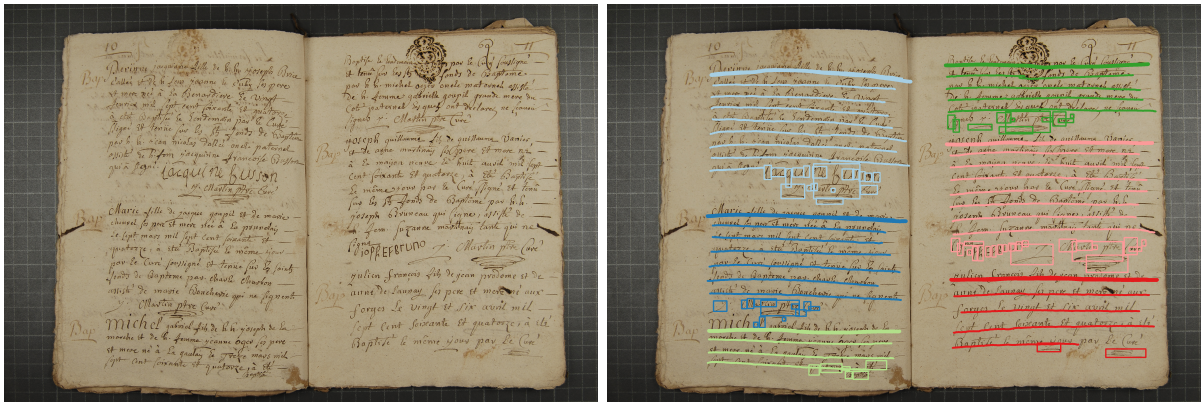
de modèles de reconnaissance d'écriture ou de reconnaissance d'entités nommées.

La chaîne de traitement complète permet de reconnaitre la mise en page des registres paroissiaux, avec notamment la localisation des actes et des lignes de texte, ainsi que la reconnaissance du texte manuscrit et les différentes entités nommées. La chaîne a été compartimentée grâce à des conteneurs Docker afin de faciliter l'appel aux différentes étapes du traitement.

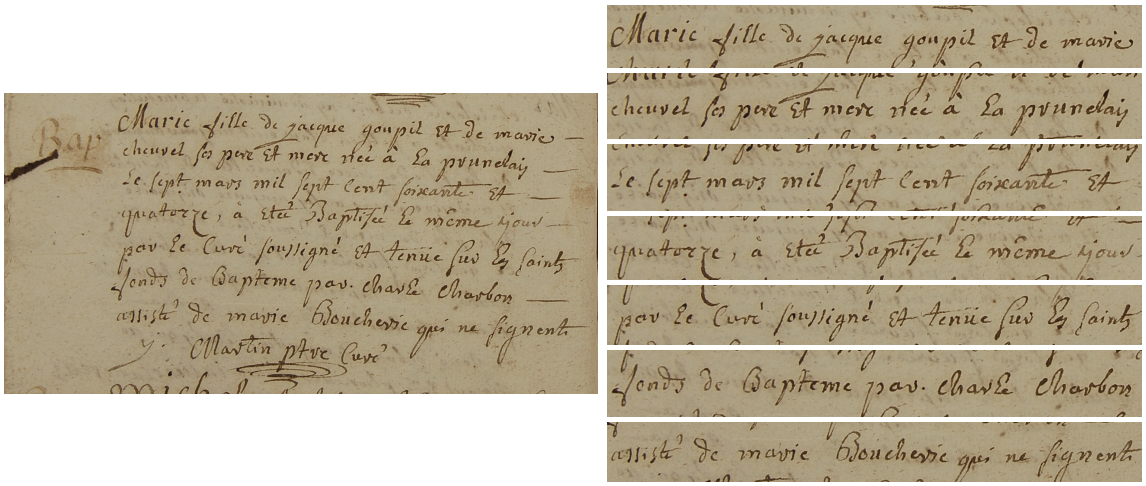
La figure 6.14 présente un exemple d'image de registre entièrement traitée grâce à notre système de reconnaissance. Les systèmes de reconnaissance d'écriture et d'entités nommées pourront être spécialisées sur des actes réels lorsque suffisamment de documents auront été transcrits.

## 6.5 Conclusion du chapitre

Dans ce chapitre, nous avons montré que l'utilisation combinée de données réelles génériques et de données synthétiques spécialisées permet d'envisager une reconnaissance



**(a)** Reconnaissance de mise en page



**(b)** Extraction et redressement des lignes de texte

| | |
|---|---|
| *Marie fille de jacque goupil et de marie* | *marie fille de jacque goupil et de marrie* |
| *chauvel ses pere et mere née à la prunelais* | *chavel ses pere et mer nee a la prundais* |
| *le sept mars mil sept cent soixante et* | *le sept marz mil sept cent soixantes et* |
| *quatorze, a éteé baptiseé le même jour* | *quatorce a ette baptisee et meme sjour* |
| *par le curé soussigné et tenue sur les saints* | *par et cure soussigne et tenie seer en saints* |
| *fonds de bapteme par charles charbon* | *fonds de bapteme par chavle charbon* |
| *assisté de marie boucherie qui ne signent* | *assiste de marie boucherie qui ne signents* |

**(c)** Reconnaissance de texte (vérité à gauche, prédiction à droite). Légende : Date, Personne principale (F) (prénom, nom), Père (prénom, nom), Mère (prénom, nom), Parrain (prénom, nom), Marraine (prénom, nom)

**Figure 6.14** – Un registre de bonne qualité (numérisé à 300 dpi) entièrement traité

du texte des registres paroissiaux, même en l'absence de ces registres dans l'ensemble d'apprentissage.

Dans un premier temps, nous avons présenté une méthodologie pour la génération d'actes synthétiques, qui consiste à modéliser le texte présent sur les actes puis à générer les images correspondantes. Le texte est généré en suivant des modèles d'actes de baptême, mariage ou sépulture, identifiés par l'analyse structurelle d'actes réels. L'image associée est ensuite générée en utilisant des polices d'écriture manuscrite, et de multiples transformations sont ensuite appliquées à l'image pour la rendre réaliste. Cette méthodologie permet d'obtenir des actes synthétiques intégralement transcrits, et dont les entités nommées sont connues. Nous avons également étudié différentes stratégies pour générer des images de lignes réalistes, afin de construire une base d'apprentissage synthétique.

Nous avons ensuite évalué l'impact des documents synthétiques sur la reconnaissance d'écriture et l'extraction d'information dans les registres paroissiaux. Pour cela, nous avons étudié différents scénarios d'apprentissage en combinant documents réels génériques et documents synthétiques spécialisés. Nous avons démontré qu'un apprentissage basé sur une combinaison de données réelles et synthétiques améliore considérablement les scores de reconnaissance par rapport à un apprentissage uniquement sur des données réelles ou synthétiques. Nous avons également étudié l'impact du dosage de données synthétiques, et montrons que la combinaison optimale consiste à utiliser un quart de données synthétiques et trois quarts de données réelles. Enfin, nous avons mis en évidence l'intérêt d'utiliser une base de données génériques la plus grande possible, incluant des documents d'autres langues et d'autres périodes.

Finalement, la méthodologie de génération d'actes synthétiques permet d'apprendre un système de reconnaissance d'écriture et d'entités nommées complet, même en l'absence de registres paroissiaux dans l'ensemble d'apprentissage. Nous avons présenté plusieurs exemples d'actes traités par notre système de reconnaissance, et montré qu'il est possible d'effectuer une reconnaissance complète sous certaines conditions. Nous avons également identifié les difficultés principales des registres, notamment la faible résolution et le style d'écriture spécifique des prêtres. Les résultats présentés dans ce chapitre sont encourageants, avec une marge de progression qui nous parait très importante. L'annotation de registres paroissiaux permettra sans doute de fiabiliser le système de reconnaissance. En combinant ces données réelles et synthétiques, il sera également possible d'utiliser la stratégie de reconnaissance *combinée* de caractères et d'entités nommées. Or, cette méthodologie de reconnaissance a été validée sur la base de données publique Esposalles

[Romero 2013], avec des performances à l'état de l'art. Il nous semble raisonnable d'espérer atteindre des performances proches de celles obtenues pour la base de données READ [Strauß 2018], soit environ 15% de taux d'erreur caractères. Or, nous avons montré dans ce chapitre que l'extraction d'information est envisageable avec ce taux d'erreur caractères. Une piste pour dépasser certaines limites présentées dans la section 6.4.1 serait d'adapter encore plus les documents synthétiques aux difficultés réellement rencontrées dans les registres paroissiaux : ajout de ratures, adaptation de l'espace inter-caractère et inter-mot, caractères spéciaux.

La génération de documents synthétiques ouvre également de nombreuses perspectives. D'une part, les documents générés sont associés à des annotations complètes : localisation des lignes de bases, des boites englobantes des lignes et des mots, localisation des pixels correspondant au texte. Ils peuvent donc être utilisés pour fiabiliser les modèles de reconnaissance de mise en page de documents. Il est également envisageable d'étendre le système pour générer des documents plus complets à l'échelle des pages.
Le système pourrait également être amélioré en ajoutant de polices d'écriture variées, ou en diversifiant les augmentations et dégradations réalisées. Les images de texte pourraient aussi être générées grâce à des réseaux génératifs Generative Adversarial Networks (GAN). En effet, de premiers travaux sur ces architectures parviennent à générer des mots complètement synthétiques [Kang 2020b ; Mattick 2021], même si les résultats sont encore peu réalistes pour de l'écriture ancienne. Enfin, la génération d'actes pourrait être améliorée en complétant la liste des structures de phrases possibles, ainsi que les dictionnaires de noms, prénoms, lieux et métiers qui gagneraient à être enrichis.

# Conclusion générale

Nous concluons ce manuscrit en proposant une synthèse des contributions de cette thèse. Nous identifions également les limites à dépasser, ainsi que certaines perspectives d'amélioration.

## Synthèse des travaux

Dans cette thèse, nous avons présenté une chaine complète pour la reconnaissance automatique de registres paroissiaux, contenant à la fois la segmentation des images en actes, la reconnaissance d'écriture et l'extraction d'information. Nos contributions se sont articulés autour de la problématique de l'apprentissage avec peu de données annotées.

Nos premières contributions se sont concentrées sur la reconnaissance de structure des registres. Dans le chapitre 3, nous avons proposé une approche hybride qui s'appuie sur l'apprentissage par réseaux de neurones de motifs structurels et sur la mise en place de règles logiques pour reconnaître la mise en page des registres. Les résultats mettent en évidence la robustesse de notre approche hybride par rapport à des approches complètement neuronales, même avec un faible nombre d'images d'apprentissage. En effet, cette approche permet une localisation de 90% des actes, soit une hausse de 7% comparé aux approches basées uniquement sur des réseaux de neurones de détection d'objets. Nous avons également montré que l'approche hybride surpasse les approches complètement neuronales, même lorsqu'elle est apprise avec trois fois moins de données d'apprentissage. Notre méthode hybride a été validée sur trois bases de données, ce qui démontre sa capacité de généralisation pour traiter d'autres types de documents. Cette capacité est très importante dans notre contexte industriel, car le système doit être en mesure de traiter des images de registres hétérogènes, qui n'ont pas forcément été représentés dans l'ensemble d'apprentissage. Enfin, l'approche hybridee permet également une reconnaissance complète et descriptive de la mise en page grâce à la localisation des lignes de texte, des signatures ainsi que des annotations marginales, avec leur appartenance à chaque acte.

Nous avons également adressé la question de la reconnaissance d'écriture manuscrite dans le chapitre 4. Nous avons démontré qu'une architecture basée sur un mécanisme

d'attention, déjà largement utilisée pour des tâches de tâches de traduction automatique ou de sous-titrage d'image, a un intérêt pour la reconnaissance d'écriture. Une étude approfondie sur les paramètres du modèle a été réalisée afin d'aboutir à l'architecture la plus adaptée pour cette tâche. Nous avons mis en évidence l'intérêt d'utiliser un encodeur convolutif et récurrent dont le squelette bénéficie d'un pré-entraînement sur ImageNet. Nous avons également démontré l'intérêt de la fonction de coût hybride, combinant la classification temporelle connectioniste (CTC) et l'entropie croisée (CE). Enfin, nous avons comparé différents mécanismes d'attention et mis en évidence la supériorité de l'attention hybride pour la reconnaissance d'écriture manuscrite. Notre architecture obtient des performances compétitives sur la plupart des bases de données publiques. Nous avons ensuite étudié différentes stratégies de transfert de connaissance pour la reconnaissance de texte des registres paroissiaux, dans un contexte où peu de données d'apprentissage sont disponibles. La stratégie la plus efficace consiste à apprendre un modèle générique en fusionnant plusieurs bases de données publiques. Cependant, les résultats mettent en lumière le manque de données d'apprentissage et donc, la nécessité de transcrire des registres paroissiaux.

Dans le chapitre 5, nous avons étudié les stratégies pour l'extraction d'information dans des documents structurés. Nous avons comparé deux approches : l'approche séquentielle, pour laquelle la reconnaissance d'écriture est effectuée en amont de la classification des mots, et l'approche conjointe, pour laquelle les deux tâches sont effectuées simultanément. Les résultats de notre étude montrent que, pour une même architecture, l'utilisation d'une approche combinée permet une hausse de 8% du taux de reconnaissance complet, démontrant ainsi l'intérêt de combiner les tâches de reconnaissance d'écriture et des entités nommées. Nous démontrons également l'intérêt de notre architecture basée sur un mécanisme d'attention pour la reconnaissance conjointe d'écriture et d'entités nommées, car elle mène à une hausse de 5% du score de reconnaissance par rapport à un réseau CRNN-CTC classique. Nous explorons également différentes stratégies d'apprentissage, et montrons l'intérêt d'une approche conjointe multi-tâches. Cette configuration multi-tâches établit un nouvel état de l'art à l'échelle des lignes sur la compétition IEHHR, atteignant un score complet de 94.4%, sans post traitement, ni modèle de langue explicite. Cependant, l'application de cette approche sur les registres paroissiaux est inenvisageable sans annotations supplémentaires sans annotations supplémentaires, car trop peu d'images contiennent des annotations sur les entités nommées.

Enfin, dans le chapitre 6, nous avons introduit une méthode permettant de générer

des actes synthétiques réalistes. Cette contribution est cruciale pour l'amélioration des performances des systèmes de reconnaissance d'écriture et d'extraction d'information sur les registres paroissiaux, en l'absence de documents annotés. Nous avons exploré différents scénarios d'apprentissage en combinant des données réelles génériques et des données synthétiques spécialisées. Les résultats montrent que l'ajout de données synthétiques mène à une diminution de 27% du taux d'erreur caractères sur les registres paroissiaux. Cependant, nos expérimentations mettent également en évidence la nécessité de combiner ces documents synthétiques à des documents réels lors de l'apprentissage d'un système de reconnaissance. Grâce à la génération d'actes synthétiques, nous avons montré que la reconnaissance d'écriture et l'extraction d'information est envisageable sur certains des registres paroissiaux, même sans aucunes données annotées issues de registres paroissiaux dans l'ensemble d'apprentissage. Les performances obtenues par notre système dans ces conditions difficiles nous laissent envisager une marge de progression importante.

## Limites et perspectives d'amélioration

La limite principale de ce travail vient du manque de transcriptions de registres paroissiaux. En effet, la transcription complète d'un acte nécessite près de 15 minutes, en incluant l'annotation des différentes catégories sémantiques et rôles des personnes se rapportant à chaque par mot. Si l'utilisation de données synthétiques permet de réduire le nombre de transcriptions manuelles à réaliser, l'annotation des registres paroissiaux nécessite tout de même un temps considérable. Pour dépasser cette contrainte, Doptim a prévu de dé · velopper un outil de transcription collaboratif intégré à Geneafinder, afin de recueillir un grand nombre de transcriptions d'actes réalisés par des généalogistes amateurs. Cependant, le déploiement d'un tel outil est complexe et n'a pas pu être réalisé durant ce travail de thèse. En effet, il nécessite le développement d'un outil de transcription robuste, combiné à une interface graphique intuitive. Il requiert également la définition d'un protocole d'annotation unifié (mots illisibles, abréviations, annotations inter-lignes, mots raturés...), et la création de ressources visant à former les utilisateurs à la lecture et à la transcription de ces documents historiques. Une phase de validation et de vérification systématique des annotations doit également être mise en place. Mais surtout, le projet nécessite une communauté d'utilisateurs désireux de s'engager dans ce projet d'annotation collaborative.

La constitution d'une base de données de registres paroissiaux est donc la première

perspective d'amélioration : elle permettra d'améliorer considérablement les performances du système de reconnaissance *séquentiel* d'écriture et d'entités nommées. En effet, la mise en place d'une telle base d'apprentissage facilitera la mise en place de la stratégie de reconnaissance *combinée*, à l'état de l'art sur la base de données Esposalles [Romero 2013]. La combinaison de ces données réelles avec des actes synthétiques permet d'envisager des performances similaires à celles obtenues sur la base de données READ [Strauß 2018], soit environ 15% de taux d'erreur caractères. Or, ce taux d'erreur permet une extraction complète des informations présentes sur les actes.

Au-delà de cette problématique de manque de données, nous proposons quelques améliorations qui nous semblent intéressantes pour la suite du projet.

Une première amélioration consisterait à combiner la reconnaissance de la mise à page et du contenu textuel. En effet, la tendance actuelle consiste à développer des systèmes *end-to-end*, c'est-à-dire capables de reconnaître le contenu textuel à partir d'images de pages. Or, la connaissance du texte doit permettre de segmenter les documents de façon efficace. Certains travaux ont notamment envisagé de fiabiliser la segmentation en actes en intégrant des indices visuels (localisation des lignes de texte) à des indices textuels (mots clés récurrents) [Boillet 2021b]. D'autre part, la reconnaissance du texte et des entités nommées au niveau des actes ou des pages permettrait au système de bénéficier d'informations contextuelles plus importantes. Ces informations contextuelles doivent permettre de fiabiliser la reconnaissance de contenu, et en particulier la reconnaissance des entités nommées, comme démontré par [Rouhou 2021] et observé pour le modèle FLAIR dans le chapitre 5. Notre méthodologie de génération d'actes synthétiques pourrait également être développée pour générer des images de pages, avec des annotations sur la localisation des lignes, des actes ainsi que leur contenu textuel.

Une autre amélioration consisterait à appliquer un modèle de langue ou un post-processing par catégorie sémantique, afin de fiabiliser la reconnaissance des mots importants. En effet, certaines catégories, comme les métiers ou les lieux, peuvent être réduites à des dictionnaires. D'autres, comme les prénoms et les noms, suivent des modes selon les périodes temporelles et les lieux. Il serait donc intéressant de combiner les informations statistiques à notre disposition pour fiabiliser la reconnaissance de ces mots.

Enfin, le traitement séquentiel de documents pourrait également permettre d'envisager une fiabilisation des informations extraites. En effet, les actes successifs sur une même page partagent certaines caractéristiques. D'une part, ils sont écrits de manière chronologique, ce qui donne une indication fondamentale sur la date. D'autre part, ils sont

généralement écrits par le même prêtre. Ainsi, les similarités de style d'écriture pourraient être exploitées afin de fiabiliser la reconnaissance. Il est également envisageable d'utiliser des connaissances logiques sur le contenu des actes relatifs à une même famille. En effet, on retrouve généralement sur ces actes les mêmes noms, prénoms et lieux, et parfois les mêmes témoins. Le croisement d'informations sur un même arbre généalogique pourrait donc permettre de fiabiliser la reconnaissance de noms, de prénoms, de métiers, et de lieux.

Afin de répondre à une problématique industrielle forte, avons utilisé des stratégies d'apprentissage permettant de garantir la capacité de généralisation du système, même avec très peu, voire aucune, données d'apprentissage issues des registres paroissiaux. D'une part, nous nous basons sur des motifs stables et des règles logiques afin de rendre notre modèle de reconnaissance de structure plus robuste. D'autre part, nous avons développé une méthodologie de génération d'actes synthétiques afin de fiabiliser la reconnaissance de caractères et d'entités nommées. Le système proposé pourra ainsi être déployé et amélioré graduellement lorsque des annotations réalisées par les généalogistes seront disponibles. Enfin, toutes les perspectives présentées dans cette section permettent d'envisager une amélioration importante du système de reconnaissance global.

# PUBLICATIONS PERSONNELLES

## Publication dans une revue internationale

[Tarride 2021c] Solène TARRIDE, Aurélie LEMAITRE, Bertrand B. COÜASNON et Sophie TARDIVEL, « Combination of deep neural networks and logical rules for record segmentation in historical handwritten registers using few examples », in : *International Journal on Document Analysis and Recognition* (jan. 2021), DOI : 10.1007/s10032-021-00362-8 (cf. p. 81).

## Publications dans un workshop international avec comité de lecture

[Tarride 2019] Solène TARRIDE, Aurélie LEMAITRE, Bertrand B. COUASNON et Sophie TARDIVEL, « Signature detection as a way to recognise historical parish register structure », in : *HIP 2019*, Sydney, Australia : ACM Press, sept. 2019, p. 54-59, DOI : 10.1145/3352631.3352636 (cf. p. 81, 104).

[Tarride 2021b] Solène TARRIDE, Aurélie LEMAITRE, Bertrand B. COÜASNON et Sophie TARDIVEL, « A comparative study of information extraction strategies using an attention-based neural network », DAS 2022 : 15th IAPR International Workshop on Document Analysis Systems, mai 2021, **En cours de soumission** (cf. p. 139).

## Publication dans un Doctoral Consortium sans comité de lecture

[Tarride 2021a] Solène TARRIDE, « Automatic recognition of historical handwritten parish records », ICDAR 2021 Doctoral Consortium : 16th Interna-

tional Conference on Document Analysis and Recognition, sept. 2021 (cf. p. 139).

# BIBLIOGRAPHIE


[Akbik 2018] Alan Akbik, Duncan Blythe et Roland Vollgraf, « Contextual String Embeddings for Sequence Labeling », in : *COLING 2018, 27th International Conference on Computational Linguistics*, 2018, p. 1638-1649 (cf. p. 146).

[Akbik 2019] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter et Roland Vollgraf, « FLAIR : An easy-to-use framework for state-of-the-art NLP », in : *NAACL Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 2019, p. 54-59 (cf. p. 73, 142, 187).

[Alaasam 2019] R. Alaasam, B. Kurar et J. El-Sana, « Layout Analysis on Challenging Historical Arabic Manuscripts using Siamese Network », in : *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 2019, p. 738-742 (cf. p. 57).

[Alberti 2018] Michele Alberti, Vinaychandran Pondenkandath, Marcel Würsch, Rolf Ingold et Marcus Liwicki, « DeepDIVA : A Highly-Functional Python Framework for Reproducible Experiments », in : *CoRR* abs/1805.00329 (2018), arXiv : 1805.00329 (cf. p. 57).

[Alberti 2019] Michele Alberti, Lars Vögtlin, Vinaychandran Pondenkandath, Mathias Seuret, Rolf Ingold et Marcus Liwicki, « Labeling, Cutting, Grouping : an Efficient Text Line Segmentation Method for Medieval Manuscripts », in : *CoRR* abs/1906.11894 (2019), arXiv : 1906.11894 (cf. p. 57).

[Antonacopoulos 2009] A. Antonacopoulos, D. Bridson, C. Papadopoulos et S. Pletschacher, « A Realistic Dataset for Performance Evaluation of Document Layout Analysis », in : *2009 10th International Conference on Document Analysis and Recognition* (2009), p. 296-300 (cf. p. 76).

[Antonacopoulos 2015] A. Antonacopoulos, C. Clausner, C. Papadopoulos et S. Pletschacher, « ICDAR2015 competition on recognition of documents with complex layouts - RDCL2015 », in : *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 2015, p. 1151-1155, doi : 10.1109/ICDAR.2015.7333941 (cf. p. 51).

[Asi 2015] A. Asi, R. Cohen, K. Kedem et J. El-Sana, « Simplifying the reading of historical manuscripts », in : *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 2015, p. 826-830 (cf. p. 56).

[Atienza 2021] Rowel Atienza, « Data Augmentation for Scene Text Recognition », in : *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021, published (cf. p. 77).

[Augustin 2006] Emmanuel Augustin, Jean-marie Brodin, Matthieu Carré, Edouard Geoffrois, Emmanuèle Grosicki et Françoise Prêteux, « RIMES evaluation campaign for handwritten mail processing », in : *Proc. of the Workshop on Frontiers in Handwriting Recognition*, 1, 2006 (cf. p. 47, 76, 77, 133).

[Ayyaz 2016] Muhammad Naeem Ayyaz, Imran Javed et Waqar Mahmood, « Handwritten Character Recognition Using Multiclass SVM Classification with Hybrid Feature Extraction », in : *Pakistan Journal of Engineering and Applied Sciences* (2016) (cf. p. 63).

[Baechler 2013] M. Baechler, M. Liwicki et R. Ingold, « Text Line Extraction Using DMLP Classifiers for Historical Manuscripts », in : *2013 12th International Conference on Document Analysis and Recognition*, 2013, p. 1029-1033 (cf. p. 54).

[Bahdanau 2014] Dzmitry Bahdanau, Kyunghyun Cho et Yoshua Bengio, *Neural Machine Translation by Jointly Learning to Align and Translate*, 2014, arXiv : 1409.0473 [cs.CL] (cf. p. 120, 123, 128, 131).

[Bahdanau 2015] Dzmitry Bahdanau, Kyunghyun Cho et Yoshua Bengio, « Neural Machine Translation by Jointly Learning to Align and Translate », in : *CoRR* abs/1409.0473 (2015) (cf. p. 66).

[Barlas 2014] P. Barlas, S. Adam, C. Chatelain et T. Paquet, « A Typed and Handwritten Text Block Segmentation System for Heterogeneous and Complex Documents », in : *2014 11th IAPR International Workshop on Document Analysis Systems*, 2014, p. 46-50 (cf. p. 56).

[Bazzi 1999] I. Bazzi, R. Schwartz et J. Makhoul, « An omnifont open-vocabulary OCR system for English and Arabic », in : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 21.*6* (1999), p. 495-504, doi : 10.1109/34.771314 (cf. p. 65).

[Benjlaiel 2014] M. Benjlaiel, R. Mullot et A. M. Alimi, « Multi-oriented Handwritten Annotations Extraction from Scanned Documents », in : *2014 11th IAPR International Workshop on Document Analysis Systems*, 2014, p. 126-130 (cf. p. 54).

[Binmakhashen 2019] Galal M. Binmakhashen et Sabri A. Mahmoud, « Document Layout Analysis : A Comprehensive Survey », in : *ACM Comput. Surv.* 52.*6* (oct. 2019), issn : 0360-0300, doi : 10.1145/3355610 (cf. p. 50).

[Biswas 2021a] Sanket Biswas, Pau Riba, Josep Lladós et Umapada Pal, « Beyond document object detection : instance-level segmentation of complex layouts », in : *International Journal on Document Analysis and Recognition (IJDAR)* 24 (sept. 2021), p. 1-13, doi : 10.1007/s10032-021-00380-6 (cf. p. 57, 59, 85).

[Biswas 2021b] Sanket Biswas, Pau Riba, Josep Lladós et Umapada Pal, « DocSynth : A Layout Guided Approach for Controllable Document Image Synthesis », in : *CoRR* abs/2107.02638 (2021), arXiv : 2107.02638 (cf. p. 77).

[Bluche 2013] Théodore Bluche, Hermann Ney et Christopher Kermorvant, « Tandem HMM with convolutional neural network for handwritten word recognition », in : *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, p. 2390-2394, doi : 10.1109/ICASSP.2013.6638083 (cf. p. 61, 62, 65).

[Bluche 2016] Théodore Bluche, *Joint Line Segmentation and Transcription for End-to-End Handwritten Paragraph Recognition*, 2016, arXiv : 1604.08352 [cs.CV] (cf. p. 121, 133).

[Bluche 2017] Théodore Bluche et Ronaldo Messina, « Gated Convolutional Recurrent Neural Networks for Multilingual Handwriting Recognition », in : *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, t. 01, 2017, p. 646-651, doi : 10.1109/ICDAR.2017.111 (cf. p. 65, 120).

[Bockholt 2011] Tiago Bockholt, George Cavalcanti et Carlos Mello, « Document Image Retrieval with Morphology-based Segmentation and Features Combination », in : t. 7874, jan. 2011, p. 1-10, doi : 10.1117/12.876727 (cf. p. 53).

[Boillet 2021a] Mélodie Boillet, Christopher Kermorvant et Thierry Paquet, « Multiple Document Datasets Pre-training Improves Text Line Detection With Deep Neural Networks », in : *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, p. 2134-2141, doi : 10.1109/ICPR48806.2021.9412447 (cf. p. 57, 118).

[Boillet 2021b] Mélodie Boillet, Martin Maarand, Thierry Paquet et Christopher Kermorvant, *Including Keyword Position in Image-based Models for Act Segmentation of Historical Registers*, 2021 (cf. p. 78, 196).

[Britto 2001] A. Britto, R. Sabourin, F. Bortolozzi et C.Y. Suen, « A two-stage HMM-based system for recognizing handwritten numeral strings », in : *Proceedings of Sixth International Conference on Document Analysis and Recognition*, 2001, p. 396-400, doi : 10.1109/ICDAR.2001.953820 (cf. p. 62, 63).

[Brunessaux 2014] Sylvie Brunessaux et al., « The Maurdor Project : Improving Automatic Processing of Digital Documents », in : *2014 11th IAPR International Workshop on Document Analysis Systems*, 2014, p. 349-354, doi : 10.1109/DAS.2014.58 (cf. p. 47, 56, 76, 89).

[Bukhari 2011a] Syed Bukhari, Faisal Shafait et Thomas Breuel, « Coupled snakelets for curled text-line segmentation from warped document images », in : *International Journal on Document Analysis and Recognition (IJDAR)* 16 (mars 2011), p. 1-21, doi : 10.1007/s10032-011-0176-2 (cf. p. 55).

[Bukhari 2011b] Syed Bukhari, Faisal Shafait et Thomas Breuel, « Improved Document Image Segmentation Algorithm using Multiresolution Morphology », in : t. 7874, jan. 2011, p. 1-10, doi : 10.1117/12.873461 (cf. p. 53).

[Bukhari 2011c] Syed Saqib Bukhari, Faisal Shafait et Thomas M. Breuel, « High Performance Layout Analysis of Arabic and Urdu Document Images », in : *2011 International Conference on Document Analysis and Recognition*, 2011, p. 1275-1279, doi : 10.1109/ICDAR.2011.257 (cf. p. 53).

[Bulacu 2007] M. Bulacu, R. Koert et Lambert Schomaker, « Layout analysis of handwritten historical documents for searching the archive of the cabinet of the Dutch queen », in : (jan. 2007) (cf. p. 56).

[Camastra 2007] Francesco Camastra, « A SVM-based cursive character recognizer », in : *Pattern Recognition* 40.*12* (2007), p. 3721-3727, issn : 0031-3203, doi : https://doi.org/10.1016/j.patcog.2007.03.014 (cf. p. 62, 63).

[Carbonell 2018] Manuel Carbonell, Mauricio Villegas, Alicia Fornés et Josep Lladós, « Joint Recognition of Handwritten Text and Named Entities with a Neural End-to-end Model », in : *CoRR* abs/1803.06252 (2018) (cf. p. 71, 72, 142, 157, 158, 160).

[Carbonell 2020] Manuel Carbonell, Alicia Fornés, Mauricio Villegas et Josep Lladós, « A Neural Model for Text Localization, Transcription and Named Entity Recognition in Full Pages », in : *Pattern Recognition Letters* 136 (mai 2020), doi : 10.1016/j.patrec.2020.05.001 (cf. p. 71, 73).

[Carel 2015] Elodie Carel, Jean-Christophe Burie, Vincent Courboulay, Jean-Marc Ogier et Vincent Poulain d'Andecy, « Multiresolution approach based on adaptive superpixels for administrative documents segmentation into color layers », in : *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 2015, p. 566-570, doi : 10.1109/ICDAR.2015.7333825 (cf. p. 53).

[Caruana 2006] Rich Caruana et Alexandru Niculescu-Mizil, « An Empirical Comparison of Supervised Learning Algorithms », in : *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06,

Pittsburgh, Pennsylvania, USA : Association for Computing Machinery, 2006, p. 161-168, ISBN : 1595933832, DOI : 10.1145/1143844.1143865 (cf. p. 63).

[Casey 1996] R.G. Casey et E. Lecolinet, « A survey of methods and strategies in character segmentation », in : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 18.7 (1996), p. 690-706, DOI : 10.1109/34.506792 (cf. p. 61).

[Chen 2013] Kai Chen, Fei Yin et Cheng-Lin Liu, « Hybrid Page Segmentation with Efficient Whitespace Rectangles Extraction and Grouping », in : *2013 12th International Conference on Document Analysis and Recognition*, 2013, p. 958-962, DOI : 10.1109/ICDAR.2013.194 (cf. p. 55).

[Chen 2014] K. Chen, H. Wei, M. Liwicki, J. Hennebert et R. Ingold, « Robust Text Line Segmentation for Historical Manuscript Images Using Color and Texture », in : *2014 22nd International Conference on Pattern Recognition*, 2014, p. 2978-2983 (cf. p. 54).

[Chen 2015a] K. Chen, M. Seuret, M. Liwicki, J. Hennebert et R. Ingold, « Page segmentation of historical document images with convolutional autoencoders », in : *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 2015, p. 1011-1015 (cf. p. 54).

[Chen 2015b] Kai Chen, Mathias Seuret, Marcus Liwicki, Jean Hennebert et Rolf Ingold, « Page segmentation of historical document images with convolutional autoencoders », in : *2015 13th International Conference on Document Analysis and Recognition (ICDAR)* (2015), p. 1011-1015 (cf. p. 57).

[Chen 2017a] Kai Chen et Mathias Seuret, « Convolutional Neural Networks for Page Segmentation of Historical Document Images », in : *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* 01 (2017), p. 965-970 (cf. p. 57).

[Chen 2017b] Liren Chen, Ruijie Yan, Liangrui Peng, Akio Furuhata et Xiaoqing Ding, « Multi-layer recurrent neural network based offline Arabic handwriting recognition », in : *2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR)*, 2017, p. 6-10, DOI : 10.1109/ASAR.2017.8067749 (cf. p. 66).

[Chorowski 2015a] Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, KyungHyun Cho et Yoshua Bengio, « Attention-Based Models for Speech Recognition », in : *CoRR* abs/1506.07503 (2015), arXiv : 1506.07503 (cf. p. 66).

[Chorowski 2015b] Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho et Yoshua Bengio, *Attention-Based Models for Speech Recognition*, 2015, arXiv : 1506.07503 [cs.CL] (cf. p. 120, 128, 129, 131).

[Chowdhury 2018] Arindam Chowdhury et Lovekesh Vig, « An Efficient End-to-End Neural Model for Handwritten Text Recognition », in : *CoRR* abs/1807.07965 (2018) (cf. p. 66, 133).

[Chung 2014] Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho et Yoshua Bengio, « Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling », in : *CoRR* abs/1412.3555 (2014), arXiv : 1412.3555 (cf. p. 128).

[Clausner 2012] C. Clausner, A. Antonacopoulos et S. Pletschacher, « A robust hybrid approach for text line segmentation in historical documents », in : *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, 2012, p. 335-338 (cf. p. 56).

[Clausner 2017] C. Clausner, A. Antonacopoulos et S. Pletschacher, « ICDAR2017 Competition on Recognition of Documents with Complex Layouts - RDCL2017 », in : *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* 01 (2017), p. 1404-1410 (cf. p. 51).

[Clausner 2018] C. Clausner, A. Antonacopoulos, Nora McGregor et Daniel Wilson-Nunn, « ICFHR 2018 Competition on Recognition of Historical Arabic Scientific Manuscripts – RASM2018 », in : *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)* (2018), p. 471-476 (cf. p. 52).

[Clausner 2019] Christian Clausner, Apostolos Antonacopoulos et Stefan Pletschacher, « ICDAR2019 Competition on Recognition of Documents with Complex Layouts - RDCL2019 », in : *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 2019, p. 1521-1526, doi : 10.1109/ICDAR.2019.00245 (cf. p. 51).

[Coquenet 2020] Denis Coquenet, Clément Chatelain et Thierry Paquet, « Recurrence-free unconstrained handwritten text recognition using gated fully convolutional network », in : *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 2020, p. 19-24, doi : 10.1109/ICFHR2020.2020.00015 (cf. p. 66, 120).

[Coquenet 2021] Denis Coquenet, Clément Chatelain et Thierry Paquet, « SPAN : a Simple Predict & Align Network for Handwritten Paragraph Recognition », in : *CoRR* abs/2102.08742 (2021), arXiv : 2102.08742 (cf. p. 66, 68, 78, 160).

[Corbetta 2002] Maurizio Corbetta et Gordon L Shulman, « Control of goal-directed and stimulus-driven attention in the brain », en, in : *Nat. Rev. Neurosci.* 3.*3* (mars 2002), p. 201-215 (cf. p. 119).

[Coüasnon 2006a] Bertrand Coüasnon, « DMOS, a generic document recognition method : Application to table structure analysis in a general and in a specific way », in : *IJDAR* 8 (juin 2006), p. 111-122, doi : 10.1007/s10032-005-0148-5 (cf. p. 54).

[Coüasnon 2006b] Bertrand Coüasnon, « DMOS, a generic document recognition method : Application to table structure analysis in a general and in a specific way », in : *IJDAR* 8 (juin 2006), p. 111-122, doi : 10.1007/s10032-005-0148-5 (cf. p. 93).

[Coüasnon 2017] Bertrand B. Coüasnon et Aurélie Lemaitre, « DMOS, It's your turn ! », in : *1st International Workshop on Open Services and Tools for Document Analysis (ICDAR-OST)*, Kyoto, Japan, nov. 2017 (cf. p. 106).

[Cruz 2014] F. Cruz et O. R. Terrades, « EM-Based Layout Analysis Method for Structured Documents », in : *2014 22nd International Conference on Pattern Recognition*, 2014, p. 315-320 (cf. p. 55).

[Das 2012] Rajib Das, Binod Kumar Prasad et Goutam Sanyal, « HMM based Offline Handwritten Writer Independent English Character Recognition using Global and Local Feature Extraction », in : *International Journal of Computer Applications* 46 (2012), p. 45-50 (cf. p. 62, 63).

[Deng 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li et Li Fei-Fei, « Imagenet : A large-scale hierarchical image database », in : *2009 IEEE conference on computer vision and pattern recognition*, Ieee, 2009, p. 248-255 (cf. p. 63, 75, 126, 127).

[Devlin 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee et Kristina Toutanova, *BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding*, 2019, arXiv : 1810.04805 [cs.CL] (cf. p. 73).

[Dhaka 2015] Vijaypal Singh Dhaka, « Character Recognition of Offline Handwritten English Scripts : A Review », in : 2015 (cf. p. 62).

[Diaz 2021] Daniel Hernandez Diaz, Siyang Qin, Reeve Ingle, Yasuhisa Fujii et Alessandro Bissacco, *Rethinking Text Line Recognition Models*, 2021, arXiv : 2104.07787 [cs.CV] (cf. p. 66, 67).

[Diem 2011] M. Diem, F. Kleber et R. Sablatnig, « Text Classification and Document Layout Analysis of Paper Fragments », in : *2011 International Conference on Document Analysis and Recognition*, 2011, p. 854-858 (cf. p. 54).

[Diem 2013] Markus Diem, Florian Kleber et Robert Sablatnig, « Text Line Detection for Heterogeneous Documents », in : août 2013, p. 743-747, doi : 10.1109/ICDAR.2013.152 (cf. p. 53).

[Diem 2017] Markus Diem, Florian Kleber, S. Fiel, Tobias Grüning et B. Gatos, « cBAD : ICDAR2017 Competition on Baseline Detection », in : *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* 01 (2017), p. 1355-1360 (cf. p. 51, 75, 76, 95, 101, 104).

[Diem 2019a] M. Diem, F. Kleber, R. Sablatnig et B. Gatos, « cBAD : ICDAR2019 Competition on Baseline Detection », in : *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 2019, p. 1494-1498 (cf. p. 76, 95).

[Diem 2019b] Markus Diem, Florian Kleber, Robert Sablatnig et Basilis Gatos, « cBAD : ICDAR2019 Competition on Baseline Detection », in : *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 2019, p. 1494-1498, doi : 10.1109/ICDAR.2019.00240 (cf. p. 51, 56, 57).

[Dong 2015] Chao Dong, Ximei Zhu, Yubin Deng, Chen Change Loy et Yu Qiao, « Boosting Optical Character Recognition : A Super-Resolution Approach », in : *ArXiv* abs/1506.02211 (2015) (cf. p. 184).

[Dutta 2018] Kartik Dutta, Praveen Krishnan, Minesh Mathew et C.V. Jawahar, « Improving CNN-RNN Hybrid Networks for Handwriting Recognition », in : *16th International Conference on Frontiers in Handwriting Recognition*, 2018, p. 80-85, doi : 10.1109/ICFHR-2018.2018.00023 (cf. p. 65, 133).

[Dutta 2019] Abhishek Dutta et Andrew Zisserman, « The VIA Annotation Software for Images, Audio and Video », in : *Proceedings of the 27th ACM International Conference on Multimedia*, Nice, France : ACM, 2019, doi : 10.1145/3343031.3350535 (cf. p. 39).

[Eskenazi 2016] Sébastien Eskenazi, Petra Gomez-Krämer et Jean-Marc Ogier, « A comprehensive survey of mostly textual document segmentation algorithms since 2008 », in : *Pattern Recognition* 64 (oct. 2016), doi : 10.1016/j.patcog.2016.10.023 (cf. p. 50).

[España-Boquera 2011] S. España-Boquera, M.J. Castro-Bleda, J. Gorbe-Moya et F. Zamora-Martinez, « Improving Offline Handwritten Text Recognition with Hybrid HMM/ANN Models », in : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33.*4* (2011), p. 767-779, doi : 10.1109/TPAMI.2010.141 (cf. p. 61, 65).

[Ferilli 2009] Stefano Ferilli, Marenglen Biba, Floriana Esposito et Teresa Basile, « A Distance-Based Technique for Non-Manhattan Layout Analysis », in : août 2009, p. 231-235, doi : 10.1109/ICDAR.2009.37 (cf. p. 53).

[Fernández 2012] F. C. Fernández et O. R. Terrades, « Document segmentation using Relative Location Features », in : *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, 2012, p. 1562-1565 (cf. p. 54).

[Fink 2007] Gernot Fink, *Markov Models for Pattern Recognition : From Theory to Applications*, jan. 2007, isbn : 978-1-4471-6307-7, doi : 10.1007/978-1-4471-6308-4 (cf. p. 65).

[Fischer 2014] A. Fischer, M. Baechler, A. Garz, M. Liwicki et R. Ingold, « A Combined System for Text Line Extraction and Handwriting Recognition in Historical Documents », in : *2014 11th IAPR International Workshop on Document Analysis Systems*, 2014, p. 71-75 (cf. p. 54).

[Fornés 2017] A. Fornés et al., « ICDAR2017 Competition on Information Extraction in Historical Handwritten Records », in : *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* 01 (2017), p. 1389-1394 (cf. p. 70, 71, 76, 140, 141, 144, 149, 159, 163).

[France 2008] Archives de France, *Écrire un cahier des charges de numérisation du patrimoine*, 2008 (cf. p. 26, 35).

[Fu 2019] Jun Fu et al., « Dual Attention Network for Scene Segmentation », in : *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, juin 2019 (cf. p. 120).

[Gaceb 2008] Djamel Gaceb, Véronique Eglin, Frank Le BOURGEOIS et Hubert Emptoz, « Application of graph coloring in physical layout segmentation », in : *International Conference on Pattern Recognition (ICPR 2008)*, sous la dir. d'IEEE, Tampa, Floride, USA, United States, nov. 2008, p. 1-4, doi : 10.1109/ICPR.2008.4761641 (cf. p. 53).

[Galibert 2015] Olivier Galibert, Juliette Kahn et Ilya Oparin, « The zonemap metric for page segmentation and area classification in scanned documents », in : *2014 IEEE International Conference on Image Processing, ICIP 2014* (jan. 2015), p. 2594-2598, doi : 10.1109/ICIP.2014.7025525 (cf. p. 89).

[Garz 2011] A. Garz, R. Sablatnig et M. Diem, « Layout Analysis for Historical Manuscripts Using Sift Features », in : *2011 International Conference on Document Analysis and Recognition*, 2011, p. 508-512 (cf. p. 54).

[Gildas 1988] Bernard Gildas, *Guide des recherches sur l'histoire des familles*, 1988 (cf. p. 17).

[Girshick 2013] Ross B. Girshick, Jeff Donahue, Trevor Darrell et Jitendra Malik, « Rich feature hierarchies for accurate object detection and semantic segmentation », in : *CoRR* abs/1311.2524 (2013), arXiv : 1311.2524 (cf. p. 87).

[Girshick 2015] Ross B. Girshick, « Fast R-CNN », in : *CoRR* abs/1504.08083 (2015), arXiv : 1504.08083 (cf. p. 86, 87).

[Goubert 1954] P. Goubert, « Une richesse historique en cours d'exploitation. Les registres paroissiaux », in : *Annales. Histoire, Sciences Sociales* 9 (1954), p. 83-93 (cf. p. 17, 37).

[Gras 1939] P. Gras, *Le registre paroissial de Givry (1334-1357) et la peste noire en Bourgogne*, t. 100, 1939, p. 295-308 (cf. p. 17).

[Graves 2006] Alex Graves, Santiago Fernández, Faustino Gomez et Jürgen Schmidhuber, « Connectionist Temporal Classification : Labelling Unsegmented Sequence Data with Recurrent Neural Networks », in : *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, Pittsburgh, Pennsylvania, USA : Association for Computing Machinery, 2006, p. 369-376, isbn : 1595933832, doi : 10.1145/1143844.1143891 (cf. p. 65).

[Graves 2009a] Alex Graves, Marcus Liwicki, Santiago Fernández, Roman Bertolami, Horst Bunke et Jürgen Schmidhuber, « A Novel Connectionist System for Unconstrained Handwriting Recognition », in : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31.*5* (2009), p. 855-868, doi : 10.1109/TPAMI.2008.137 (cf. p. 65).

[Graves 2009b] Alex Graves et Jürgen Schmidhuber, « Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks », in : *Advances in Neural Information Processing Systems*, sous la dir. de D. Koller, D. Schuurmans, Y. Bengio et L. Bottou, t. 21, Curran Associates, Inc., 2009 (cf. p. 65).

[Grüning 2017] Tobias Grüning, Roger Labahn, Markus Diem, Florian Kleber et Stefan Fiel, « READ-BAD : A New Dataset and Evaluation Scheme for Baseline Detection in Archival Documents », in : *CoRR* abs/1705.03311 (2017), arXiv : 1705.03311 (cf. p. 76, 102).

[Grüning 2018] Tobias Grüning, Gundram Leifert, Tobias Strauss et Roger Labahn, « A Two-Stage Method for Text Line Detection in Historical Documents », in : *CoRR* abs/1802.03345 (2018), arXiv : 1802.03345 (cf. p. 43, 44, 57-59, 77, 92, 102).

[Guerry 2019] Camille Guerry, Bertrand B. Coüasnon et Aurélie Lemaitre, « Combination of deep learning and syntactical approaches for the interpretation of interactions between text-lines and tabular structures in handwritten documents », in : *15th International Conference on Document Analysis and Recognition (ICDAR)*, Sydney, Australia, sept. 2019 (cf. p. 93).

[Hazem 2020] Amir Hazem et al., « Books of Hours : the First Liturgical Corpus for Text Segmentation », in : *12th Language Resources and Evaluation Conference*, Proceedings of the 12th Language Resources and Evaluation Conference, Marseille (Virtual), France : European Language Resources Association (ELRA), mai 2020, p. 776-784 (cf. p. 76).

[He 2017] Kaiming He, Georgia Gkioxari, Piotr Dollár et Ross B. Girshick, « Mask R-CNN », in : *CoRR* abs/1703.06870 (2017), arXiv : 1703.06870 (cf. p. 83, 86, 87, 90).

[Hebert 2011] D. Hebert, T. Paquet et S. Nicolas, « Continuous CRF with Multi-scale Quantization Feature Functions Application to Structure Extraction in Old Newspaper », in : *2011 International Conference on Document Analysis and Recognition*, 2011, p. 493-497 (cf. p. 54).

[JADLI 2020] Aissam JADLI, Mustapha HAIN, Adil CHERGUI et Abderrahman JAIZE, « DCGAN-Based Data Augmentation for Document Classification », in : *2020 IEEE 2nd International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS)*, 2020, p. 1-5, doi : 10.1109/ICECOCS50124.2020.9314379 (cf. p. 77).

[Journet 2008] Nicholas Journet, Jean-Yves Ramel, Véronique Eglin et Rémy Mullot, « Document Image Characterization Using a Multiresolution Analysis of the Texture : Application to Old Documents. », in : *International Journal on Document Analysis and Recognition* Volume 11.*Number 1* (oct. 2008), p. 9-18, doi : 10.1007/s10032-008-0064-6 (cf. p. 53).

[Journet 2017] Nicholas Journet, Muriel Visani, Boris Mansencal, Kieu Van-Cuong et Antoine Billy, « DocCreator : A New Software for Creating Synthetic Ground-Truthed Document Images », in : *Journal of imaging* 3.*4* (2017), p. 62 (cf. p. 77, 78, 169).

[Kaltenmeier 1993] A. Kaltenmeier, T. Caesar, J.M. Gloger et E. Mandler, « Sophisticated topology of hidden Markov models for cursive script recognition », in : *Proceedings of 2nd International Conference on Document Analysis and Recognition (ICDAR '93)*, 1993, p. 139-142, doi : 10.1109/ICDAR.1993.395764 (cf. p. 65).

[Kang 2019a] Lei Kang, Pau Riba, Mauricio Villegas, Alicia Fornés et Marçal Rusiñol, « Candidate Fusion : Integrating Language Modelling into a Sequence-to-Sequence Handwritten Word Recognition Architecture », in : *CoRR* abs/1912.10308 (2019), arXiv : 1912.10308 (cf. p. 132).

[Kang 2019b] Lei Kang, J. Toledo, Pau Riba, Mauricio Villegas, Alicia Fornés et Marçal Rusiñol, « Convolve, Attend and Spell : An Attention-based Sequence-to-Sequence Model for Handwritten Word Recognition : 40th German Conference, GCPR 2018, Stuttgart, Germany, October 9-12, 2018, Proceedings », in : jan. 2019, p. 459-472, isbn : 978-3-030-12938-5, doi : 10.1007/978-3-030-12939-2_32 (cf. p. 121).

[Kang 2020a] Lei Kang, Pau Riba, Marçal Rusiñol, Alicia Fornés et Mauricio Villegas, « Pay Attention to What You Read : Non-recurrent Handwritten Text-Line Recognition », in : *CoRR* abs/2005.13044 (2020) (cf. p. 67, 78, 133).

[Kang 2020b] Lei Kang, Pau Riba, Yaxing Wang, Marçal Rusiñol, Alicia Fornés et Mauricio Villegas, « GANwriting : Content-Conditioned Generation of Styled Handwritten Word Images », in : nov. 2020, p. 273-289, isbn : 978-3-030-58591-4, doi : 10.1007/978-3-030-58592-1_17 (cf. p. 78, 192).

[Kim 2016] Suyoun Kim, Takaaki Hori et Shinji Watanabe, « Joint CTC-Attention based End-to-End Speech Recognition using Multi-task Learning », in : *CoRR* abs/1609.06773 (2016), arXiv : 1609.06773 (cf. p. 125).

[Kumar 2010] Jayant KUMAR, Wael ABD-ALMAGEED, Le KANG et David DOERMANN, « Handwritten Arabic Text Line Segmentation Using Affinity Propagation », in : *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, DAS '10, Boston, Massachusetts, USA : Association for Computing Machinery, 2010, p. 135-142, ISBN : 9781605587738, DOI : 10.1145/1815330.1815348 (cf. p. 53).

[Kumar 2014] Gaurav KUMAR et Pradeep Kumar BHATIA, « A Detailed Review of Feature Extraction in Image Processing Systems », in : *2014 Fourth International Conference on Advanced Computing Communication Technologies*, 2014, p. 5-12, DOI : 10.1109/ACCT.2014.74 (cf. p. 62, 63).

[LeCun 1995] Yann André LECUN et al., « Comparison of learning algorithms for handwritten digit recognition », in : 1995 (cf. p. 61, 63).

[LeCun 2010] Yann LECUN et Corinna CORTES, « MNIST handwritten digit database », in : (2010) (cf. p. 63).

[Lemaitre 2011] Aurélie LEMAITRE, Jean CAMILLERAPP et Bertrand COÜASNON, « A perceptive method for handwritten text segmentation », in : *Document recognition and retrieval XVIII* 7874 (jan. 2011), DOI : 10.1117/12.873037 (cf. p. 53, 56).

[Lemaitre 2014] Aurélie LEMAITRE, Jean CAMILLERAPP et Bertrand COÜASNON, « Handwritten text segmentation using blurred image », in : *DRR - Document Recognition and Retrieval XXI*, DRR - Document Recognition and Retrieval XXI, San Francisco, United States, jan. 2014 (cf. p. 102).

[Lemaitre 2021] Aurélie LEMAITRE et Jean CAMILLERAPP, « Segmentation of historical maps without annotated data », in : *6th International Workshop on Historical Document Imaging and Processing (HIP'21)*, Lausanne, France, sept. 2021, DOI : 10.1145/3476887.3476909 (cf. p. 93).

[Li 2020a] Minghao LI, Lei CUI, Shaohan HUANG, Furu WEI, Ming ZHOU et Zhoujun LI, « TableBank : Table Benchmark for Image-based Table Detection and Recognition », in : *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France : European Language Resources Association, mai 2020, p. 1918-1925 (cf. p. 76).

[Li 2020b] Minghao Li et al., « DocBank : A Benchmark Dataset for Document Layout Analysis », in : *CoRR* abs/2006.01038 (2020) (cf. p. 76).

[Lin 2014] Tsung-Yi Lin et al., « Microsoft COCO : Common Objects in Context », in : *CoRR* abs/1405.0312 (2014) (cf. p. 63, 75, 86, 88).

[Lin 2018] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He et Piotr Dollár, *Focal Loss for Dense Object Detection*, 2018, arXiv : 1708.02002 [cs.CV] (cf. p. 83, 86, 87, 90).

[Liu 2015a] Qi Liu, Lijuan Wang et Qiang Huo, « A study on effects of implicit and explicit language model information for DBLSTM-CTC based handwriting recognition », in : *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 2015, p. 461-465, doi : 10.1109/ICDAR.2015.7333804 (cf. p. 66).

[Liu 2015b] Wei Liu et al., « SSD : Single Shot MultiBox Detector », in : *CoRR* abs/1512.02325 (2015), arXiv : 1512.02325 (cf. p. 86).

[Liu 2017] Hairong Liu, Zhenyao Zhu, Xiangang Li et Sanjeev Satheesh, « Gram-CTC : Automatic Unit Selection and Target Decomposition for Sequence Labelling », in : *CoRR* abs/1703.00096 (2017), arXiv : 1703.00096 (cf. p. 66).

[Liu 2019] Yinhan Liu et al., « RoBERTa : A Robustly Optimized BERT Pretraining Approach », in : *CoRR* abs/1907.11692 (2019), arXiv : 1907.11692 (cf. p. 73).

[Louloudis 2008] G. Louloudis, B. Gatos, I. Pratikakis et C. Halatsis, « Text line detection in handwritten documents », in : *Pattern Recognition* 41.*12* (2008), p. 3758-3772, issn : 0031-3203, doi : https://doi.org/10.1016/j.patcog.2008.05.011 (cf. p. 55).

[Luong 2015a] Minh-Thang Luong, Quoc Le, Ilya Sutskever, Oriol Vinyals et Lukasz Kaiser, « Multi-task Sequence to Sequence Learning », in : *Proceedings of ICLR, San Juan, Puerto Rico* (nov. 2015) (cf. p. 66, 74, 153, 158, 160).

[Luong 2015b] Minh-Thang Luong, Hieu Pham et Christopher D. Manning, *Effective Approaches to Attention-based Neural Machine Translation*, 2015, arXiv : 1508.04025 [cs.CL] (cf. p. 128, 131).

[Marti 2001] Urs-Viktor MARTI et Horst BUNKE, « Using a Statistical Language Model to Improve the Performance of an HMM-Based Cursive Handwriting Recognition System », in : *Int. J. Pattern Recognit. Artif. Intell.* 15 (2001), p. 65-90 (cf. p. 65).

[Marti 2002] Urs-Viktor MARTI et H. BUNKE, « The IAM-database : An English sentence database for offline handwriting recognition », in : *International Journal on Document Analysis and Recognition* 5 (nov. 2002), p. 39-46, DOI : 10.1007/s100320200071 (cf. p. 47, 76, 77, 126, 127, 133).

[Martin 2020] Louis MARTIN et al., « CamemBERT : a Tasty French Language Model », in : *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online : Association for Computational Linguistics, juill. 2020, p. 7203-7219 (cf. p. 73).

[Masci 2011] Jonathan MASCI, Ueli MEIER, Dan CIREŞAN et Jürgen SCHMIDHUBER, « Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction », in : *Artificial Neural Networks and Machine Learning – ICANN 2011*, sous la dir. de Timo HONKELA, Włodzisław DUCH, Mark GIROLAMI et Samuel KASKI, Berlin, Heidelberg : Springer Berlin Heidelberg, 2011, p. 52-59 (cf. p. 63).

[Mattick 2021] Alexander MATTICK, Martin MAYR, Mathias SEURET, Andreas MAIER et Vincent CHRISTLEIN, « SmartPatch : Improving Handwritten Word Imitation with Patch Discriminators », in : *CoRR* abs/2105.10528 (2021), arXiv : 2105.10528 (cf. p. 78, 192).

[Mehri 2013a] Maroua MEHRI, Petra GOMEZ-KRÄMER, Pierre HÉROUX, Alain BOUCHER et Rémy MULLOT, « Texture Feature Evaluation for Segmentation of Historical Document Images », in : *Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing*, HIP '13, 2013, p. 102-109, ISBN : 9781450321150, DOI : 10.1145/2501115.2501121 (cf. p. 53).

[Mehri 2013b] Maroua MEHRI, Pierre HEROUX, Petra GOMEZ-KRÄMER, Alain BOUCHER et Remy MULLOT, « A Pixel Labeling Approach for Historical Digitized Books », in : août 2013, p. 817-821, DOI : 10.1109/ICDAR.2013.167 (cf. p. 53).

[Mehri 2015] Maroua MEHRI, « Historical document image analysis : a structural approach based on texture », thèse de doct., mai 2015 (cf. p. 50).

[Mehri 2019] Maroua Mehri, Pierre Héroux, Rémy Mullot, Jean-Philippe Moreux, Bertrand Coüasnon et Bill Barrett, « ICDAR2019 Competition on Historical Book Analysis - HBA2019 », in : *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 2019, p. 1488-1493, doi : 10.1109/ICDAR.2019.00239 (cf. p. 51, 52, 56, 57).

[Menasri 2012] Fares Menasri, Jérôme Louradour, Anne-Laure Bianne-Bernard et Christopher Kermorvant, « The A2iA French handwriting recognition system at the Rimes-ICDAR2011 competition », in : *Proceedings of SPIE - The International Society for Optical Engineering* 8297 (jan. 2012), p. 51-, doi : 10.1117/12.911981 (cf. p. 65).

[Michael 2019] Johannes Michael, Roger Labahn, Tobias Grüning et Jochen Zöllner, « Evaluating Sequence to Sequence Models for Handwritten Text Recognition », in : *CoRR* abs/1903.07377 (2019) (cf. p. 66, 121, 125, 131, 133, 138).

[Michael 2021] Johannes Michael, Max Weidemann, Bastian Laasch et Roger Labahn, « ICPR 2020 Competition on Text Block Segmentation on a NewsEye Dataset », in : *Lecture Notes in Computer Science, (LNCS, volume 12668)*, Springer, fév. 2021, doi : 10.5281/zenodo.4555751 (cf. p. 76).

[Moysset 2014] Bastien Moysset et al., « The A2iA Multi-lingual Text Recognition System at the Second Maurdor Evaluation », in : *2014 14th International Conference on Frontiers in Handwriting Recognition* (2014), p. 297-302 (cf. p. 65).

[Moysset 2015] B. Moysset, C. Kermorvant, C. Wolf et J. Louradour, « Paragraph text segmentation into lines with Recurrent Neural Networks », in : *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 2015, p. 456-460 (cf. p. 56).

[Murdock 2015] Michael Murdock, Shawn Reid, Blaine Hamilton et Jackson Reese, « ICDAR 2015 competition on text line detection in historical documents », in : *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 2015, p. 1171-1175, doi : 10.1109/ICDAR.2015.7333945 (cf. p. 51, 57, 76).

[Nasien 2010] Dewi Nasien, Habibollah Haron et Siti Sophiayati Yuhaniz, « Support Vector Machine (SVM) for English Handwritten Character Recognition », in : *2010 Second International Conference on Computer Engineering and Applications*, t. 1, 2010, p. 249-252, doi : 10.1109/ICCEA.2010.56 (cf. p. 63).

[Nopsuwanchai 2003] Roongroj Nopsuwanchai et D. Povey, « Discriminative training for HMM-based offline handwritten character recognition », in : *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.* 2003, 114-118 vol.1, doi : 10.1109/ICDAR.2003.1227643 (cf. p. 62, 63).

[Oliveira 2017] Dário Oliveira et Matheus Viana, « Fast CNN-Based Document Layout Analysis », in : oct. 2017, p. 1173-1180, doi : 10.1109/ICCVW.2017.142 (cf. p. 57, 59, 85).

[Oliveira 2018] Sofia Ares Oliveira, Benoit Seguin et Frédéric Kaplan, « dhSegment : A generic deep-learning approach for document segmentation », in : *CoRR* abs/1804.10371 (2018), arXiv : 1804.10371 (cf. p. 57-59, 95, 96, 102).

[Ouwayed 2011] Nazih Ouwayed et Abdel Belaïd, « A general approach for multi-oriented text line extraction of handwritten document », in : *International Journal on Document Analysis and Recognition* 14.*4* (sept. 2011), doi : 10.1007/s10032-011-0172-6 (cf. p. 55).

[Palm 2019] Rasmus Berg Palm, Florian Laws et Ole Winther, « Attend, Copy, Parse End-to-end Information Extraction from Documents », in : *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 2019, p. 329-336, doi : 10.1109/ICDAR.2019.00060 (cf. p. 70).

[Papavassiliou 2010] Vassilis Papavassiliou, Themos Stafylakis, Vassilis Katsouros et George Carayannis, « Handwritten document image segmentation into text lines and words », in : *Pattern Recognition* 43.*1* (2010), p. 369-377, issn : 0031-3203, doi : https://doi.org/10.1016/j.patcog.2009.05.007 (cf. p. 55).

[Peng 2011] Xujun Peng, Srirangaraj Setlur, Venu Govindaraju et Ramachandrula Sitaram, « Handwritten text separation from annotated machine printed documents using Markov Random Fields », in : *In-*

*ternational Journal on Document Analysis and Recognition (IJDAR)* 16 (2011), p. 1-16 (cf. p. 54).

[Perlin 2002] Ken Perlin, « Improving Noise », in : *ACM Trans. Graph.* 21.*3* (juill. 2002), p. 681-682, issn : 0730-0301, doi : 10.1145/566654.566636 (cf. p. 170).

[Pham 2013] Vu Pham, Christopher Kermorvant et Jérôme Louradour, « Dropout improves Recurrent Neural Networks for Handwriting Recognition », in : *CoRR* abs/1312.4569 (2013), arXiv : 1312.4569 (cf. p. 66).

[Pinson 2011] S. J. Pinson et W. A. Barrett, « Connected Component Level Discrimination of Handwritten and Machine-Printed Text Using Eigenfaces », in : *2011 International Conference on Document Analysis and Recognition*, 2011, p. 1394-1398 (cf. p. 54).

[Plamondon 2000] R. Plamondon et S.N. Srihari, « Online and off-line handwriting recognition : a comprehensive survey », in : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.*1* (2000), p. 63-84, doi : 10.1109/34.824821 (cf. p. 60, 61).

[Plötz 2009] Thomas Plötz et Gernot A. Fink, « Markov models for offline handwriting recognition : a survey », in : *International Journal on Document Analysis and Recognition (IJDAR)* 12 (2009), p. 269-298 (cf. p. 60, 62, 64, 65).

[Poulos 2017] Jason Poulos et Rafael Valle, « Character-Based Handwritten Text Transcription with Attention Networks », in : *CoRR* abs/1712.04046 (2017), arXiv : 1712.04046 (cf. p. 121, 133).

[Prasad 2018] Animesh Prasad, Hervé Déjean, Jean-Luc Meunier, Max Weidemann, Johannes Michael et Gundram Leifert, « Bench-Marking Information Extraction in Semi-Structured Historical Handwritten Records », in : *ArXiv* abs/1807.06270 (2018) (cf. p. 72, 157, 158).

[Prusty 2019] Abhishek Prusty, Sowmya Aitha, Abhishek Trivedi et Ravi Kiran Sarvadevabhatla, *Indiscapes : Instance Segmentation Networks for Layout Parsing of Historical Indic Manuscripts*, 2019, arXiv : 1912.07025 [cs.CV] (cf. p. 57, 59, 85).

[Ptucha 2019] Raymond Ptucha, Felipe Petroski Such, Suhas Pillai, Frank Brockler, Vatsala Singh et Paul Hutkowski, « Intelligent character recognition using fully convolutional neural networks », in : *Pattern Recognition* 88 (2019), p. 604-613, issn : 0031-3203, doi : https://doi.org/10.1016/j.patcog.2018.12.017 (cf. p. 66).

[Puigcerver 2017] Joan Puigcerver, « Are Multidimensional Recurrent Layers Really Necessary for Handwritten Text Recognition ? », in : *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, t. 01, 2017, p. 67-72, doi : 10.1109/ICDAR.2017.20 (cf. p. 66, 133).

[Qian 2020] Zhuang Qian, Kaizhu Huang, Qiu-Feng Wang, Jimin Xiao et Rui Zhang, « Generative adversarial classifier for handwriting characters super-resolution », in : *Pattern Recognition* 107 (2020), p. 107453, issn : 0031-3203, doi : https://doi.org/10.1016/j.patcog.2020.107453 (cf. p. 184).

[Quirós 2018] Lorenzo Quirós, « Multi-Task Handwritten Document Layout Analysis », in : *ArXiv* abs/1806.08852 (2018) (cf. p. 57).

[Redmon 2015] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick et Ali Farhadi, « You Only Look Once : Unified, Real-Time Object Detection », in : *CoRR* abs/1506.02640 (2015), arXiv : 1506.02640 (cf. p. 86).

[Redmon 2016] Joseph Redmon et Ali Farhadi, « YOLO9000 : Better, Faster, Stronger », in : *CoRR* abs/1612.08242 (2016), arXiv : 1612.08242 (cf. p. 86).

[Redmon 2018] Joseph Redmon et Ali Farhadi, « YOLOv3 : An Incremental Improvement », in : *CoRR* abs/1804.02767 (2018), arXiv : 1804.02767 (cf. p. 83, 86, 90).

[Ren 2015] Shaoqing Ren, Kaiming He, Ross B. Girshick et Jian Sun, « Faster R-CNN : Towards Real-Time Object Detection with Region Proposal Networks », in : *CoRR* abs/1506.01497 (2015) (cf. p. 86, 87).

[Rensink 2000] Ronald Rensink, « The Dynamic Representation of Scenes », in : *Visual Cognition* 7 (jan. 2000), p. 17-42, doi : 10.1080/135062800394667 (cf. p. 119).

[Renton 2017] Guillaume Renton, Clément Chatelain, Sébastien Adam, Christopher Kermorvant et Thierry Paquet, « Handwritten Text Line Segmentation Using Fully Convolutional Network », in : *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* 05 (2017), p. 5-9 (cf. p. 57).

[Renton 2018] Guillaume Renton, Yann Soullard, Clement Chatelain, Sébastien Adam, Christopher Kermorvant et Thierry Paquet, « Fully convolutional network with dilated convolutions for handwritten text line segmentation », in : *International Journal on Document Analysis and Recognition (IJDAR)* (mai 2018), doi : 10.1007/s10032-018-0304-3 (cf. p. 57).

[Romero 2013] Verónica Romero et al., « The ESPOSALLES database : An ancient marriage license corpus for off-line handwriting recognition », in : *Pattern Recognit.* 46 (2013), p. 1658-1669 (cf. p. 60, 76, 77, 111, 114, 116, 124, 139, 144, 159, 163, 192, 196).

[Rouhou 2021] Ahmed Cheikh Rouhou, Marwa Dhiaf, Yousri Kessentini et Sinda Ben Salem, « Transformer-based approach for joint handwriting and named entity recognition in historical document », in : *Pattern Recognition Letters* (2021), issn : 0167-8655, doi : https://doi.org/10.1016/j.patrec.2021.11.010 (cf. p. 142, 157, 158, 196).

[Ryu 2014] J. Ryu, H. I. Koo et N. I. Cho, « Language-Independent Text-Line Extraction Algorithm for Handwritten Documents », in : *IEEE Signal Processing Letters* 21.*9* (2014), p. 1115-1119 (cf. p. 55).

[Saha 2019] Ranajit Saha, Ajoy Mondal et C. V. Jawahar, « Graphical Object Detection in Document Images », in : *2019 International Conference on Document Analysis and Recognition (ICDAR)* (2019), p. 51-58 (cf. p. 57, 59).

[Schwartz 1996] R. Schwartz, Christopher LaPre, J. Makhoul, C. Raphael et Ying Zhao, « Language-independent OCR using a continuous speech recognition system », in : *Proceedings of 13th International Conference on Pattern Recognition* 3 (1996), 99-103 vol.3 (cf. p. 65).

[Shafait 2008] F. Shafait, J. v. Beusekom, D. Keysers et T. M. Breuel, « Structural Mixtures for Statistical Layout Analysis », in : *2008 The Eighth IAPR International Workshop on Document Analysis Systems*, 2008, p. 415-422 (cf. p. 54).

[Simistira 2016] Foteini Simistira, Mathias Seuret, Nicole Eichenberger, Angelika Garz, Marcus Liwicki et Rolf Ingold, « DIVA-HisDB : A Precisely Annotated Large Dataset of Challenging Medieval Manuscripts », in : *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 2016, p. 471-476, doi : 10.1109/ICFHR.2016.0093 (cf. p. 76).

[Simistira 2017] Fotini Simistira et al., « ICDAR2017 Competition on Layout Analysis for Challenging Medieval Manuscripts », in : *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, t. 01, 2017, p. 1361-1370, doi : 10.1109/ICDAR.2017.223 (cf. p. 52).

[Simonyan 2015] Karen Simonyan et Andrew Zisserman, « Very Deep Convolutional Networks for Large-Scale Image Recognition », in : *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, sous la dir. d'Yoshua Bengio et Yann LeCun, 2015 (cf. p. 122).

[Soullard 2020] Yann Soullard, Pierrick Tranouez, Clément Chatelain, Stéphane Nicolas et Thierry Paquet, « Multi-scale Gated Fully Convolutional DenseNets for semantic labeling of historical newspaper images », in : *Pattern Recognit. Lett.* 131 (2020), p. 435-441 (cf. p. 120).

[Strauß 2018] Tobias Strauss, Gundram Leifert, Roger Labahn, Tobias Hodel et Günter Mühlberger, « ICFHR2018 Competition on Automated Text Recognition on a READ Dataset », in : *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 2018, p. 477-482, doi : 10.1109/ICFHR-2018.2018.00089 (cf. p. 76, 77, 137, 192, 196).

[Tang 2014] Youbao Tang, Xiangqian Wu et Wei Bu, « Text Line Segmentation Based on Matched Filtering and Top-Down Grouping for Handwritten Documents », in : *2014 11th IAPR International Workshop on Document Analysis Systems*, 2014, p. 365-369, doi : 10.1109/DAS.2014.14 (cf. p. 53).

[Tarn 2013] Archives départementales du Tarn, *Petit guide d'initiation : la paléographie*, 2013 (cf. p. 33).

[Toledo 2019] J. I. Toledo, Manuel Carbonell, A. Fornés et J. Lladós, « Information extraction from historical handwritten document images with a context-aware neural model », in : *Pattern Recognit.* 86 (2019), p. 27-36 (cf. p. 71, 72).

[Toselli 2004] Alejandro Héctor Toselli et al., « Integrated Handwriting Recognition And Interpretation Using Finite-State Models », in : *Int. J. Pattern Recognit. Artif. Intell.* 18 (2004), p. 519-539 (cf. p. 65).

[Trivedi 2021] Abhishek Trivedi et Ravi Kiran Sarvadevabhatla, « BoundaryNet : An Attentive Deep Network with Fast Marching Distance Maps for Semi-automatic Layout Annotation », in : *CoRR* abs/2108.09433 (2021) (cf. p. 59).

[UNESCO 1972] UNESCO, *Convention concernant la protection du patrimoine mondial culturel et naturel*, `https://whc.unesco.org/archive/convention-fr.pdf`, Accessed : 2021-08-17, 1972 (cf. p. 17).

[Vaswani 2017] Ashish Vaswani et al., « Attention is All you Need », in : *Advances in Neural Information Processing Systems*, sous la dir. d'I. Guyon et al., t. 30, Curran Associates, Inc., 2017 (cf. p. 66, 128).

[Vinciarelli 2002] Alessandro Vinciarelli, « A survey on off-line Cursive Word Recognition », in : *Pattern Recognition* 35.7 (2002), p. 1433-1446, issn : 0031-3203, doi : `https://doi.org/10.1016/S0031-3203(01)00129-7` (cf. p. 61, 65).

[Vinciarelli 2004] Alessandro Vinciarelli, Samy Bengio et Horst Bunke, « Offline recognition of unconstrained handwritten texts using HMMs and statistical language models », in : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26 (2004), p. 709-720 (cf. p. 65).

[Wan 2019] Zhaoyi Wan, Fengming Xie, Yibo Liu, Xiang Bai et Cong Yao, « 2D-CTC for Scene Text Recognition », in : *CoRR* abs/1907.09705 (2019), arXiv : `1907.09705` (cf. p. 66).

[Wang 1994] Jin Wang et Jack Jean, « Segmentation of merged characters by neural networks and shortest path », in : *Pattern Recognition* 27.5 (1994), p. 649-658, issn : 0031-3203, doi : `https://doi.org/10.1016/0031-3203(94)90044-2` (cf. p. 62).

[Wei 2013] H. Wei, M. Baechler, F. Slimane et R. Ingold, « Evaluation of SVM, MLP and GMM Classifiers for Layout Analysis of Historical Documents », in : *2013 12th International Conference on Document Analysis and Recognition*, 2013, p. 1220-1224 (cf. p. 54).

[Wei 2014] H. Wei, K. Chen, R. Ingold et M. Liwicki, « Hybrid Feature Selection for Historical Document Layout Analysis », in : *2014 14th International Conference on Frontiers in Handwriting Recognition*, 2014, p. 87-92 (cf. p. 56).

[Wigington 2017] Curtis Wigington, Seth Stewart, Brian Davis, Bill Barrett, Brian Price et Scott Cohen, « Data Augmentation for Recognition of Handwritten Words and Lines Using a CNN-LSTM Network », in : *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, t. 01, 2017, p. 639-645, doi : 10.1109/ICDAR.2017.110 (cf. p. 77).

[Wigington 2018a] Curtis Wigington, Chris Tensmeyer, Brian Davis, William Barrett, Brian Price et Scott Cohen, « Start, Follow, Read : End-to-End Full-Page Handwriting Recognition », in : *Computer Vision – ECCV 2018*, sous la dir. de Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu et Yair Weiss, Springer International Publishing, 2018, p. 372-388 (cf. p. 66, 68, 78).

[Wigington 2018b] Curtis Wigington, Chris Tensmeyer, Brian L. Davis, W. Barrett, Brian L. Price et Scott Cohen, « Start, Follow, Read : End-to-End Full-Page Handwriting Recognition », in : *ECCV*, 2018 (cf. p. 133).

[Xu 2016] Kelvin Xu et al., *Show, Attend and Tell : Neural Image Caption Generation with Visual Attention*, 2016, arXiv : 1502.03044 [cs.LG] (cf. p. 66, 119-121, 125).

[Xu 2017] Yue Xu, Wenhao He, Fei Yin et Cheng-Lin Liu, « Page Segmentation for Historical Handwritten Documents Using Fully Convolutional Networks », in : *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* 01 (2017), p. 541-546 (cf. p. 57).

[Yi 2017] Xiaohan Yi, Liangcai Gao, Yuan Liao, Xiaode Zhang, Runtao Liu et Zhuoren Jiang, « CNN Based Page Object Detection in Document Images », in : *2017 14th IAPR International Conference on*

*Document Analysis and Recognition (ICDAR)*, t. 01, 2017, p. 230-235, DOI : 10.1109/ICDAR.2017.46 (cf. p. 57, 59).

[Yin 2009] Fei YIN et Cheng-Lin LIU, « Handwritten Chinese text line segmentation by clustering with distance metric learning », in : *Pattern Recognition* 42.*12* (2009), New Frontiers in Handwriting Recognition, p. 3146-3157, ISSN : 0031-3203, DOI : https://doi.org/10.1016/j.patcog.2008.12.013 (cf. p. 53, 55).

[Yousef 2018] Mohamed YOUSEF, Khaled F. HUSSAIN et Usama S. MOHAMMED, « Accurate, Data-Efficient, Unconstrained Text Recognition with Convolutional Neural Networks », in : *CoRR* abs/1812.11894 (2018), arXiv : 1812.11894 (cf. p. 66, 120).

[Yu 2021] W. YU, Ning LU, X. QI, Ping GONG et Rong XIAO, « PICK : Processing Key Information Extraction from Documents using Improved Graph Learning-Convolutional Networks », in : *2020 25th International Conference on Pattern Recognition (ICPR)* (2021), p. 4363-4370 (cf. p. 70, 78).

[Zhang 2021] Peng ZHANG et al., « VSR : A Unified Framework for Document Layout Analysis combining Vision, Semantics and Relations », in : *CoRR* abs/2105.06220 (2021), arXiv : 2105.06220 (cf. p. 78).

[Zhong 2019] Xu ZHONG, Jianbin TANG et Antonio JIMENO-YEPES, « PubLayNet : largest dataset ever for document layout analysis », in : *CoRR* abs/1908.07836 (2019), arXiv : 1908.07836 (cf. p. 76, 85).

[Ziaratban 2010] Majid ZIARATBAN et Karim FAEZ, « An Adaptive Script-Independent Block-Based Text Line Extraction », in : *2010 20th International Conference on Pattern Recognition*, 2010, p. 249-252, DOI : 10.1109/ICPR.2010.70 (cf. p. 53).

DOCTORAT BRETAGNE LOIRE / MATHSTIC

INSA RENNES

**Titre :** Combinaison de connaissances physiques et textuelles pour la reconnaissance d'images de registres anciens

**Mot clés :** Documents historiques, reconnaissance de structure, reconnaissance d'écriture manuscrite, extraction d'information, génération de documents synthétiques

**Résumé :** Nos travaux portent sur la reconnaissance automatique de documents historiques ayant un intérêt fondamental en généalogie : les registres paroissiaux. Nos contributions s'articulent en trois axes. Le premier axe porte sur la reconnaissance de la structure de ces registres. Nous proposons une méthode hybride capable de localiser les actes en s'appuyant sur l'apprentissage de motifs structurels et leur groupement logique à l'aide de règles. Le deuxième axe adresse la reconnaissance du contenu textuel de ces documents. Nous adaptons une architecture neuronale avec mécanisme d'attention pour la reconnaissance d'écriture manuscrite, et montrons l'intérêt d'une reconnaissance conjointe d'écriture manuscrite et entités nommées. Nous étudions également plusieurs stratégies d'apprentissage conjointes. Le troisième axe aborde la génération de documents synthétiques. Nous proposons une méthode pour générer des actes synthétiques, en modélisant les structures de phrases récurrentes des actes et en implémentant des transformations permettant de déformer et dégrader les images synthétisées grâce à des polices manuscrites. Nous montrons l'intérêt de cette stratégie pour réduire le besoin en transcriptions manuelles, qui est un enjeu majeur de ce travail.

**Title:** Combining physical and textual approaches for parish register recognition

**Keywords:** Historical documents, document layout analysis, handwriting recognition, information extraction, synthetic document generation

**Abstract:** This thesis focuses on automatic recognition of historical French registers. These documents contain a series of records, and contain valuable information for genealogists. We present three main contributions. Firstly, we introduce a hybrid methodology for document layout recognition, combining neural networks and logical rules. We demonstrate the strength of this approach, especially when few training documents are available. Secondly, we focus on automatic handwritten text understanding . We adapt an attention-based neural network for this task and demonstrate that combining handwriting recognition and named entity recognition increases performance. We also study various training strategies for multi-task and multi-scale analysis. Finally, we adress synthetic document generation, modeling the textual content and the visual appareance of real records. This approach is crucial as it allows to reduce the dependancy on annotated documents, which is a key issue in this work.