Transcription automatique et segmentation thématique de livres d'heures manuscrits

Béatrice Daille[*] — **Amir Hazem**[*] — **Christopher Kermorvant**[†‡] — **Martin Maarand**[†] — **Marie-Laurence Bonhomme**[†] — **Dominique Stutzmann**[§] — **Jacob Currie**[§] — **Christine Jacquin**[*]

[*] *LS2N - Université de Nantes, Nantes* *prenom.nom@ls2n.fr*
[§] *Institut de recherche et d'histoire des textes (IRHT), Paris*
prenom.nom@irht.cnrs.fr
[†] *TEKLIA, Paris* *nom@teklia.com*
[‡] *LITIS, Université de Rouen-Normandie, Rouen*

RÉSUMÉ. *Les livres d'heures sont le plus grand best-seller de tout le Moyen Âge, avec plus de 10 000 témoins conservés. Incontournables pour comprendre l'univers mental médiéval, leurs textes ont été très peu étudiés. Ils sont très longs et ont une structure complexe correspondant à l'organisation liturgique médiévale et la prière quotidienne de l'office. Cet article décrit les méthodes et les traitements automatiques mis en œuvre sur les livres d'heures : la reconnaissance de l'écriture manuscrite et la segmentation adaptées à ces manuscrits. L'approche de segmentation semi-supervisée proposée tire profit de la constitution spécifique du manuscrit pour mieux retrouver leur structure malgré le bruit engendré par la reconnaissance de l'écriture.*

ABSTRACT. *Books of Hours are the number one best seller of the Middle Ages, with more than 10 000 copies preserved. They are a crucial witness to the medieval mindset, but their textual contents have been very scarcely studied. They are very long and offer a complex hierarchical entangled structure, with several characteristics specific to medieval daily Prières office. This paper presents the methods and processing applied to books of hours: handwritten text recognition and text segmentation adapted to medieval manuscripts. We propose a weak supervised approach, based on the overarching structure of the manuscripts, that provides the first state-of-the-art results on transcript texts and despite remaining errors for this new challenging task.*

MOTS-CLÉS : *reconnaissance de l'écriture manuscrite, segmentation thématique, livre d'heures.*

KEYWORDS: *handwritten text recognition, text segmentation, Book of hours.*

1. Introduction

Les livres d'heures sont un recueil de prières à l'usage des fidèles (Leroquais, 1927 ; Wieck *et al.*, 1988). Souvent richement enluminés, et répandus dès le XIII^e siècle en France, au sud des Pays-Bas, en Angleterre et plus tard en Italie et en Espagne, ils constituent une part importante de l'ensemble des manuscrits médiévaux préservés et sont une source d'information sur la vie et la chrétienté au Moyen Âge. Ils font partie des textes les plus lus au Moyen Âge et, par la richesse de leur décor, plusieurs livres d'heures font aussi partie des objets d'art les plus connus du Moyen Âge français, comme le livre d'heures d'Étienne Chevalier peint par Jean Fouquet et les *Très Riches Heures du duc de Berry*, peint par les frères de Limbourg. En empruntant leurs principaux éléments au bréviaire, l'un des types de livres liturgiques que la religion chrétienne utilise pour régler son culte, et en reproduisant ainsi partiellement le contenu de livres destinés aux prêtres et au clergé, ils permettent aux laïques de prier, comme ceux-ci, selon les heures canoniales tout au long de la journée (matines, laudes, prime, tierce, sexte, nones, vêpres, complies), d'où leur nom générique de « livres d'heures ». Leur noyau est en latin (Heures de la Vierge, Heures de la Croix et du Saint-Esprit, office des morts) et présente des additions en latin et dans des langues vernaculaires (souvent en français). Malgré leur succès à l'époque, leur contenu textuel reste actuellement très peu étudié, alors que la production d'un si grand nombre de manuscrits est un phénomène culturel et industriel capital qui manifeste les profonds changements du monde religieux du bas Moyen Âge, avec, à la fois, le développement d'une production livresque proto-industrielle et le passage de l'économie de la demande à celle de l'offre, mais aussi ce que J. Burckhart a nommé « l'éveil de l'individu » (Rosenwein, 2005) et surtout l'intériorisation de la foi, à une époque où l'encadrement ecclésial devient de plus en plus contraignant.

Malgré plus de 10 000 manuscrits témoins conservés, il existe très peu de livres d'heures transcrits en entier et annotés d'un point de vue linguistique. Comme l'affirme Christopher De Hamel (1994) : « *It sometimes seems surprising, therefore, that there is still no critical edition of the text [...]. Its cultural impact (if that is not too pompous a term for an illuminated prayer-book) was wider and deeper than that of many rare literary texts worked over and over again by modern editors. It reached people too with no other knowledge of literacy. Anyone who could be encouraged to edit the first proper printed edition of the Book of Hours since the sixteenth century would win the gratitude of all historians of manuscripts. The task, however, will be made immensely complicated by the number of surviving manuscripts and their endless subtle differences.* » L'une des rares ressources sur le texte des livres d'heures est la base *Beyond Use*, qui contient, en particulier, une section sur l'*Obsecro Te* (Plummer et Clark, 2015). Cette prière à la Vierge a été transcrite et annotée manuellement à partir de plus de 772 livres d'heures (Plummer et Clark, 2015)[1].

Les livres d'heures sont très longs, avec une moyenne de 300 pages. Ils ont une structure complexe correspondant à l'organisation liturgique médiévale et, en parti-

1. `http://www6.sewanee.edu/beyonduse/`

culier, à la prière quotidienne de l'office, avec plusieurs parties, sections et sous-sections et de nombreux textes dits « accessoires ». À l'heure actuelle, la majorité des livres d'heures sont faiblement catalogués. L'étude de l'usage liturgique qui en résulte repose en conséquence sur un faible nombre de points de repère textuels parmi les plus courts (antiennes, versets et répons) (Leroquais, 1927 ; Ottosen, 1993 ; Ottosen, 2008 ; Drigsdahl, 2013). Or, ceux-ci ne reflètent pas la structure globale et n'empêchent pas les ambiguïtés. Un même texte biblique peut apparaître dans des sections ou sous-sections différentes d'un livre d'heures à l'autre. Parmi les textes accessoires, les prières latines et vernaculaires ont, certes, fait l'objet de repérages, mais presque tout reste à faire. De très nombreux textes restent à découvrir : les textes latins sont surtout repérés pour les manuscrits les plus anciens ; les prières françaises vernaculaires font l'objet de recensements (Sonet, 1956 ; Sinclair, 1978 ; Sinclair, 1987 ; Sinclair, 1979 ; Sinclair, 1982 ; Sinclair, 1988 ; Rézeau, 1986), voire d'éditions (Rézeau, 1983), mais, latins comme vernaculaires, de nombreux textes sont inédits et la diffusion de ces textes par les livres d'heures reste à explorer.

Cet article présente nos premiers travaux pour identifier automatiquement la structure logique des livres d'heures, une étape nécessaire pour permettre une analyse textuelle complète par les historiens médiévistes. L'accès au texte à partir des images de livres d'heures numérisés nécessite une transcription automatique de l'écriture manuscrite. Dans un premier temps, une analyse automatique de la mise en page est réalisée pour identifier les différents éléments présents : iconographie, décoration et zones de texte. La transcription des lignes de texte est ensuite réalisée par un système automatique entraîné spécifiquement sur le type d'écriture manuscrite présent dans les livres d'heures.

2. Composition et structure du livre d'heures

Le livre d'heures, apparu au XIII^{e} siècle en se détachant du psautier dont il était un appendice, a évolué au cours du temps, et des textes non présents dans les premières versions ont été ajoutés.

2.1. *Composition du livre d'heures*

Le livre d'heures inclut un certain nombre de textes de référence, possédant les caractéristiques suivantes (Lebigue, 2007).

Antienne (*antiphona*) est une pièce chantée courte (une à deux lignes) dont le texte est généralement d'origine biblique. Dans le cadre des livres d'heures, ces textes apparaissent principalement autour d'un texte central qui peut être un psaume, un groupe de psaumes ou un cantique, généralement, pour les offices présents dans les livres d'heures, avec l'intonation (début du chant pour « imposer l'antienne ») avant le texte central, puis la pièce entière après le texte central. D'autres formes de l'antienne sont possibles : sous une forme complexe dans

l'invitatoire, où elle est dite en entier avant et après le psaume et la doxologie ; après les versets impairs, où seule la fin de l'antienne est chantée après les versets pairs. Chaque suffrage comporte aussi une antienne.

Absolution et bénédiction sont des textes prononcés avant les leçons de matines.

Cantique (*canticum*) est un chant tiré de la Bible, utilisé comme les psaumes, dont sept, dits « bibliques » sont tirés de l'Ancien Testament tels que le Cantique des trois enfants (*Benedicite*) et le Cantique d'Isaïe (*Confitebor tibi*) et sept sont tirés du Nouveau Testament, dont le Cantique de Zacharie (*Benedictus*), le Cantique de Vierge (*Magnificat*) et le Cantique de Siméon (*Nunc dimittis*).

Capitule (*Capitulum*) est une lecture brève tirée de la Bible, présente dans toutes les heures sauf les matines où contiennent des lectures longues.

Doxologie est une formule conclusive de prières. On distingue notamment la grande doxologie « *Gloria in excelsis...* » de la petite doxologie *Gloria Patri...*, qui est en particulier récitée à la fin des psaumes, cantiques et dans les répons ; dans l'office des morts, le verset *Requiem aeternam...* est utilisé comme doxologie.

Hymne (*hymnus*) est un chant métrique ou rythmique d'origine non biblique.

Invocation est la première partie des heures et comprend un ou deux versets de psaumes de la Bible qui invitent à la prière (Ps. 50,17 et Ps. 69,2 à matines, puis seulement Ps. 69,2 aux autres heures), puis la doxologie (*Gloria patri*).

Invitatoire (*invitatorium*) suit, à matines, l'invocation et comprend un psaume, lui-même dit « invitatoire », avec son antienne intercalée ; le psaume invitatoire le plus courant est le psaume 94 (*Venite exultemus*).

Leçon ou Lecture (*lectio*) est extraite de la Bible (dans certains offices, on trouve également des extraits d'œuvres patristiques ou hagiographiques) et lue au sein des nocturnes de matines.

Oraison (*oratio*) est une prière. Des oraisons forment la conclusion des offices (sauf matines), suffrages et litanies.

Preces est une partie de l'office et de la litanie rassemblant des formules de supplication et principalement constituées d'un ou de deux versicules et de leurs réponses, du *Kyrie eleison* et du *Pater noster*.

Psaume (*psalmus*) est un chant de louange. Au nombre de cent cinquante dans la tradition latine, ils sont regroupés, au sein de la Bible, dans le livre des Psaumes. Il s'agit de textes poétiques divisés en versets ; ils constituent le fondement de la liturgie chrétienne et de la prière continuelle de l'Église.

Répons (*responsorium*) est un court chant de méditation après une lecture. Il est composé de (1) un répons proprement dit (anglais « *respond* »), lui-même divisé en (1a) une première partie du répons et (1b) une « réclame » ou « reprise », en

anglais « *partial respond* »), puis (2) un « verset » (lat. *Versus* ou *Versus responsorii*).Il existe des répons de deux sortes : les « répons prolixes », longs, utilisés après les longues leçons de matines, et les « répons brefs » après les capitules. On dit le répons proprement dit (1) en entier (une fois pour le répons prolixe, deux pour le répons bref), le verset (2), la réclame (1b), puis la doxologie et, pour le répons bref, le verset (2).

Verset (*versus*) désigne soit (1) un vers d'un psaume ou d'une hymne, soit (2) la deuxième partie d'un répons (lat. *responsorium*, angl. « *verse* »). Le mot français est parfois utilisé à égalité avec « versicule » ou pour désigner l'ensemble formé par le versicule et la réponse.

Salutation conclut les offices et est constituée du verset *Benedicamus Domino*, de son acclamation *Deo gratias* et d'un verset (généralement *Fidelium animae*).

Versicule (*versiculus*, abrév. *Vers.*) est un vers suivi d'une « réponse » (lat. *responsio* ou *responsum versiculi*, abrév. *Resp.*, angl. *response*). Dans la liturgie collective, le versicule est chanté par les solistes et la réponse par le chœur. Il intervient au début d'un office, après une hymne et dans les *preces*.

Ainsi, la majeure partie des textes constituant les livres d'heures sont extraits de la Bible. D'autres textes, tels que le Notre Père (*Pater noster*), les doxologies, ou d'autres prières interviennent dans la composition des offices ou dans l'agencement des livres d'heures, en particulier la prière *Obsecro Te*, une supplication à la Vierge afin de recevoir son assistance au moment de la mort.

2.2. *Structure du livre d'heures*

Le livre d'heures comporte une structure complexe qui peut varier selon le lieu d'origine, la destination liturgique (« usage liturgique ») et les désirs du commanditaire.
Traditionnellement, le livre d'heures débute avec le calendrier liturgique. Celui-ci permet au fidèle d'avoir connaissance des fêtes religieuses et, le cas échéant, du ou des saints à célébrer selon le jour. Le calendrier peut être suivi d'extraits de chacun des quatre Évangiles (« péricopes évangéliques »), puis par les « Heures de la Vierge » ou petit office de la Vierge (*Officium parvum beatae Mariae Virginis*), office votif en l'honneur de la Vierge Marie. Ces Heures de la Vierge constituent la section la plus importante du livre d'heures. Divisées en huit sections selon chacune des heures de la journée, elles sont composées de psaumes, de cantiques et d'hymnes, tous thématiquement liés, au moins partiellement, à la Vierge. Ces textes sont eux-mêmes séparés par des antiennes, versets et répons. Généralement placées à la suite des Heures de la Vierge, parfois divisées par heures et intercalées à l'intérieur de celles-ci, surtout dans les manuscrits de l'Ouest de la France, se trouvent les Heures de la Croix et les Heures du Saint-Esprit. Deux autres parties sont presque systématiquement présentes : d'une

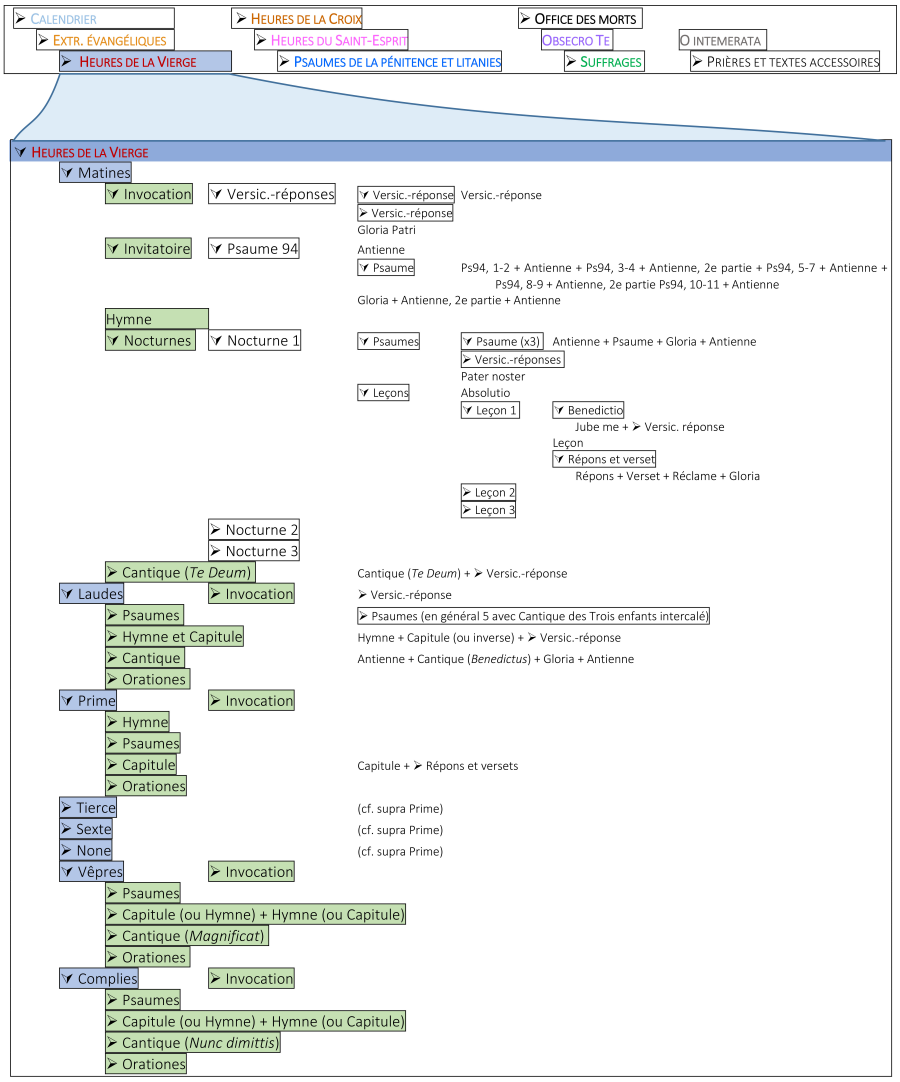

Figure 1. *Structure d'un livre d'heures : parties principales et subdivisions des Heures de la Vierge*

part, l'office des morts, traditionnellement appelé « office » et non « heures » car il ne se compose pas des huit heures, mais seulement de vêpres, matines et laudes, contenant les prières récitées par le clergé pour le salut de l'âme des défunts, et, d'autre part, les sept psaumes de la pénitence ou psaumes pénitentiels (psaumes 6, 31, 37, 50, 101, 129 et 142 dans la numérotation de la Vulgate) qui sont complétés par les

Niveau 1	Nombre de textes de niveau 3
Heures de la Vierge	228
Heures de la Croix	85
Heures du Saint-Esprit	68
Office des morts	156
Suffrages	101

Tableau 1. *Nombres de textes de niveau 3 pour chaque partie constitutive de niveau 1 du livre d'heures ms. Paris, bibliothèque de l'Arsenal, 1194.*

litanies et prières adressées à Dieu, aux anges, à la Vierge et aux saints appelés hiérarchiquement (apôtres, martyrs, confesseurs, etc.). De nombreux autres textes peuvent également être copiés : offices complets, comme les Heures de la Passion ; suffrages, c'est-à-dire des mémoires votifs, composés d'une antienne, d'un verset et d'une oraison ; prières additionnelles, en latin ou en vernaculaire, parfois liées à des indulgences, et dont les deux plus fréquentes sont les prières *Obsecro Te* et *O intemerata*. L'ordre de chacune des parties peut être interverti et tous les offices appelés « heures » se subdivisent en huit sections selon les heures de la journée.

La figure 1 récapitule la structure générique d'un livre d'heures. Plusieurs niveaux de structures sont distingués. Le premier niveau correspond aux grandes catégories de prières comme les péricopes évangéliques ou les Heures de la Vierge. Le second niveau décline les huit prières selon l'échelle temporelle, des matines à complies. Au troisième niveau apparaissent l'invitatoire, les hymnes, les nocturnes et cantiques.

Idéalement, une segmentation automatique doit pouvoir identifier ces trois niveaux. Le tableau 1 indique le nombre de textes présents dans le livre d'heures conservé à la bibliothèque de l'Arsenal (Ms-1194 réserve). Le tableau 2 indique la structuration au niveau 1 de huit livres d'heures. Cet examen comparatif montre que l'ordre type de succession des grandes prières n'est pas stable et qu'elles ne sont pas toutes présentes. Ce tableau illustre les difficultés qui vont être rencontrées pour la segmentation automatique.

3. Reconnaissance du texte manuscrit des livres d'heures

Nous décrivons dans cette section le système de reconnaissance d'écriture manuscrite utilisé pour obtenir une transcription automatique d'un corpus de livre d'heures numérisé sous forme d'images.

Harvard Lat 251	Harvard Lat 253	Harvard Typ 32	Harvard Typ 1000	Harvard Typ 464	Poitiers 1097	Poitiers 43	Poitiers 46
Calendrier	Calendrier	Calendrier	Calendrier	Suffrages	Extr. Évangiles	Calendrier	Calendrier
Extr. Évangiles	Extr. Évangiles	Extr. Évangiles	Vierge	Vierge	Obsecro Te	Extr. Évangiles	Extr. Évangiles
Vierge	Obsecro Te	Obsecro Te	Litanies de la Vierge	Croix	O Intemerata	Obsecro Te	Obsecro Te
Croix	Vierge	Vierge	Psaumes, litanies	Esprit	Prières	Vierge	O Intemerata
Psaumes, litanies	Psaumes, litanies	O Intemerata	Morts	Psaumes, litanies	Heures mêlées (Vierge	Croix	Vierge
Suffrages	Croix	Psaumes, litanies	Croix	Morts	+ Esprit	Esprit	Croix
Prières	Esprit	Croix	Esprit	Extr. Évangiles	+ Croix)	Psaumes, litanies	Esprit
	Morts	Esprit	Prières	Obsecro Te	Psaumes, litanies	Morts	Psaumes, litanies
		Morts			Morts	Suffrages	Morts
		Joies de la Vierge			Suffrages Prières	Sept requêtes à N.S.	Versets de s. Bernard
		Sept requêtes à N.S.			Extr. Évangiles (Passio)	Suffrages Prières	Suffrages

Tableau 2. *Exemples de la segmentation obtenue pour le niveau 1 pour huit livres d'heures (Harvard : Cambridge, Ma., Harvard University, Houghton Library, et Poitiers, médiathèque François-Mitterrand). Une couleur différente est attribuée à chaque élément de niveau 1.*

3.1. *Reconnaissance automatique de documents et d'écritures*

La reconnaissance de l'écriture imprimée (OCR), (*Optical Character Recognition*) sur des documents récents est considérée comme un problème résolu : des systèmes disponibles dans le commerce ou d'accès libre (*open source*) atteignent des taux d'erreurs bien inférieurs à 1 %. La situation est différente en ce qui concerne la reconnaissance de l'écriture manuscrite (HTR) (*Handwritten Text Recognition*) : il existe peu de systèmes commerciaux et les taux d'erreurs restent très variables et bien supérieurs aux taux obtenus par les OCR. La reconnaissance des écritures médiévales, dans une langue très éloignée de la langue actuelle et avec des spécificités sur les formes de lettres et l'usage des abréviations, est encore plus complexe. Il est dans ce cas nécessaire d'entraîner des systèmes spécifiques à la fois à la graphie et au contenu textuel. Les récentes avancées apportées par les techniques d'apprentissage statistique à base de réseaux de neurones profonds ont grandement amélioré les performances des systèmes qui sont maintenant capables de lire une grande diversité d'écritures anciennes après entraînement, comme l'ont montré de récentes applications (Bluche *et al.*, 2017a ; Lang *et al.*, 2018) et les compétitions internationales (Sánchez *et al.*, 2016 ; Sánchez *et al.*, 2017 ; Strauß *et al.*, 2018).

Avant d'appliquer un système de reconnaissance d'écriture, il est d'abord nécessaire d'analyser la structure de l'image des documents afin d'en extraire les zones de texte. Cette étape d'analyse de la mise en page (DLA) (*Document Layout Analysis*) est elle aussi beaucoup plus simple sur les documents imprimés que sur les documents manuscrits. L'extraction des lignes de texte dans un document manuscrit est généralement rendue plus complexe par les variations de taille d'écriture et l'inclinaison des lignes. Là encore, les modèles les plus performants actuellement sont les modèles par apprentissage automatique à base de réseaux de neurones profonds (Diem *et al.*, 2017 ; Renton *et al.*, 2018 ; Moysset *et al.*, 2018 ; Ares Oliveira *et al.*, 2018 ; Grüning *et al.*, 2018). Un exemple d'analyse d'une page de livre d'heures est présenté sur la figure 2.

3.2. *Description du système de reconnaissance d'écriture*

Un système complet de transcription automatique de document est composé d'un certain nombre d'étapes exécutées séquentiellement. Premièrement, les lignes de texte sont localisées dans chaque image de page du manuscrit numérisé. Ces lignes de texte sont ensuite extraites, et le système de reconnaissance d'écriture est appliqué sur chacune des imagettes de ligne. La reconnaissance d'écriture comprend elle-même deux étapes, l'application d'un modèle optique qui reconnaît des caractères, des fragments de caractères ou de mots, et l'application d'un modèle de langue qui détermine les séquences de caractères et de mots les plus vraisemblables.

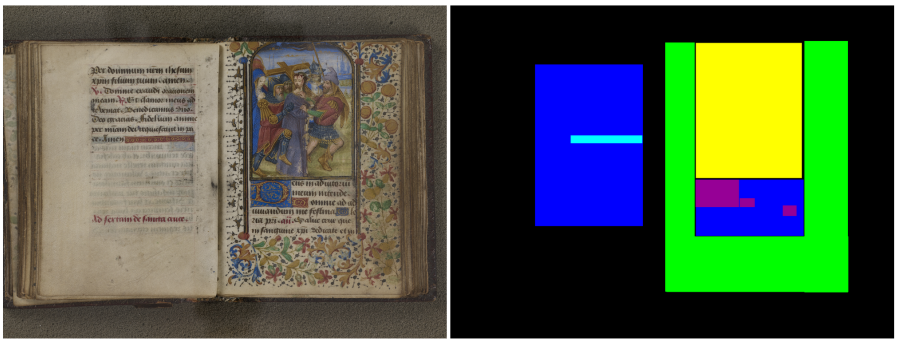

Figure 2. *Analyse d'une double page de livre d'heures : texte (bleu foncé), marge ornée (vert), miniature (jaune), lettrine (violet) et bout de ligne (bleu clair)*

3.2.1. *Détection des lignes de texte*

La détection des lignes de texte a été réalisée avec le logiciel Transkribus[2], une plate-forme de traitement de documents développée principalement à destination des chercheurs en humanités et pour le traitement des documents anciens. Transkribus permet à la fois de réaliser des opérations automatiques sur les images, comme la localisation des régions de texte et l'identification des lignes, mais aussi d'annoter manuellement les documents. La constitution d'un corpus de pages de livres d'heures annotées a ensuite permis d'entraîner un système spécifique de détection des zones et lignes de texte (Boillet *et al.*, 2019).

Un exemple d'extraction des lignes de texte est présenté sur la figure 3. Le texte est écrit de différentes couleurs, présente des lettrines (initiales décorées) et une marge ornée.

3.2.2. *Reconnaissance de l'écriture manuscrite*

Nous avons développé un système de reconnaissance de l'écriture manuscrite des livres d'heures fondé sur la librairie logicielle KALDI[3]. Bien qu'initialement développée pour la reconnaissance de la parole, elle peut être également utilisée pour la reconnaissance d'écriture car les deux applications partagent de nombreux points communs, surtout depuis la généralisation de l'utilisation des réseaux de neurones profonds pour ces deux applications (Bluche *et al.*, 2017b).

Le système développé repose sur une combinaison de réseaux de neurones profonds et de modèles de Markov cachés (HMM) (Peddinti *et al.*, 2015). Le modèle optique, en charge de modéliser la forme des lettres, est composé de plusieurs couches de réseaux de neurones à convolution suivies de couches TDNN (*Time Delay Neural*

2. `https://transkribus.eu`
3. `https://github.com/kaldi-asr/kaldi`

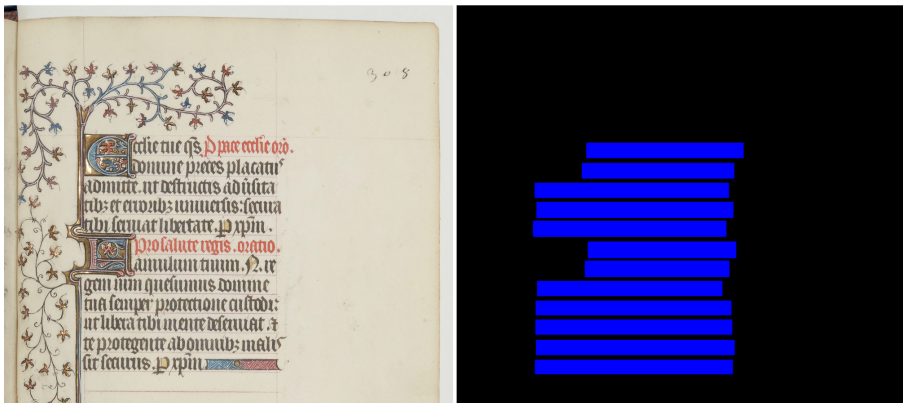

Figure 3. *Extraction des lignes de texte (en bleu) sur une page de livre d'heures présentant une marge ornée et des lettrines*

Networks) qui permettent de modéliser les caractères en contexte. Les prédictions sont ensuite utilisées par un modèle HMM qui modélise les mots comme des séquences de HMM de caractères. Un modèle de langue statistique de type n-gramme est ensuite utilisé pour modéliser les séquences de mots.

Le modèle a été entraîné sur des données provenant de trois corpus de documents médiévaux manuscrits, Psautiers, ECMEN et Fontenay, constitués dans le cadre de précédents projets de recherhche : ORIFLAMMS[4] (*Ontology Research, Image Features, Letterform Analysis on Multilingual Medieval Scripts*) et ECMEN (écriture médiévale et outils numériques). Les corpus Psautiers et Fontenay sont en latin, tandis que le corpus ECMEN est en ancien français. Dans un premier temps, il n'a pas été réalisé de détection de la langue des livres d'heures, ce qui aurait permis d'utiliser un modèle spécialisé soit en latin soit en français, car la quantité de données d'entraînement et de test est trop faible. Cette approche sera testée lorsque plus de documents annotés seront disponibles.

Le modèle a été évalué sur 247 lignes transcrites manuellement, issues des *Obsecro Te* de huit livres d'heures du corpus cible. Aucune transcription complète de livre d'heures n'étant disponible dans notre corpus, le système n'a été appris sur aucune donnée issue du corpus cible afin de les réserver pour l'évaluation des performances de la transcription automatique. Les données sont réparties en trois ensembles : les données d'entraînement, de validation et de test, comme présenté dans le tableau 3.

Ces documents sont assez hétérogènes, tant par la qualité des images, leur type (en couleur ou en niveaux de gris), la précision du découpage des lignes et les choix de transcriptions opérés. Cette hétérogénéité se manifeste dans les résultats de recon-

4. `https://oriflamms.hypotheses.org/`

	Entraînement	Validation	Test
Psautiers	4500	660	0
ECMEN	2000	542	0
Fontenay	723	0	0
Obsecro Te	0	0	247

Tableau 3. *Répartition des données et tailles en nombre de lignes des différents ensembles pour l'entraînement et l'évaluation du modèle de reconnaissance d'écriture manuscrite*

	Entraînement	Validation	Test
WER	14.51	24.46	34.19
WER (rescored)	–	32.36	26.32
CER	8.93	11.07	11.21
CER (rescored)	–	14.10	9.90

Tableau 4. *Évaluation des performances de la reconnaissance d'écriture manuscrite sur les différents ensembles, selon les taux d'erreurs mots (WER) et caractères (CER), avec (rescored) ou sans application du modèle de langue*

naissance obtenus et présentés dans le tableau 4. Ces résultats sont évalués selon deux métriques : le taux d'erreurs mots (WER) (*word error rate*) qui mesure le nombre de mots incorrects dans la transcription fournie par le système par rapport à la transcription humaine. Ce taux prend en compte les substitutions, insertions, suppressions et peut donc être supérieur à 100 %. Le taux d'erreurs caractères (CER) (*character error rate*) est son équivalent mesuré au niveau des caractères. Pour mesurer l'impact du modèle de langue, les taux WER et CER sont mesurés avant (*WER, CER*) et après réestimation des hypothèses de séquences de mots par le modèle de langue (*WER rescored, CER rescored*). Cette évaluation est réalisée en ignorant les confusions entre majuscules et minuscules, la ponctuation, et en assimilant les paires de lettres *u* et *v* ainsi que *j* et *i* qui sont des lettres identiques à cette époque.

Les taux d'erreurs sur le corpus de validation, qui contient des lignes de texte issues du même type de documents que l'ensemble d'entraînement (Psautiers et ECMEN), sont plus faibles que les taux d'erreurs sur l'ensemble de test, qui contient des lignes de texte issues de documents complètement disjoints (*Obsecro Te*). Les taux d'erreurs du modèle de reconnaissance d'écriture sont assez élevés sur le corpus de test, car aucune donnée d'entraînement n'est disponible pour le corpus cible. Des exemples d'erreurs sont présentés sur la figure 4. Cependant, ces taux élevés n'empêchent pas l'identification des textes comme il sera montré dans les sections suivantes.

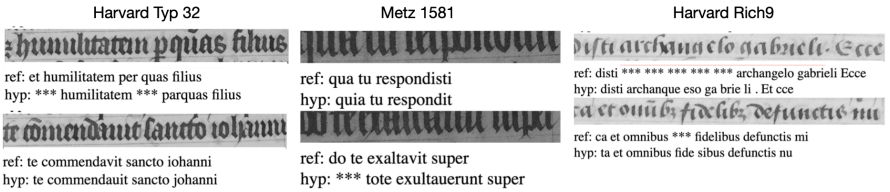

Figure 4. *Exemples de résultats de reconnaissance (*hyp*) et d'annotations manuelles (*ref*) sur les manuscrits Harvard Typ 32, Metz 1581, Harvard Rich 9. À gauche et au centre, écriture de type « Textualis », à droite de type « Cursiva ». À gauche, présence d'abréviations (*et *non reconnu,* per *reconnu comme 'par' qui est correct dans d'autres contextes,* m *correctement reconnu), confusion i/j et u/v sur la dernière ligne (les lettres ramistes ne sont pas distinguées au Moyen Âge et sont une restitution des annotateurs). Au centre, erreurs de segmentation, le haut de la ligne de texte est tronqué. À droite, erreurs de segmentation, confusion c/t, abréviations correctement reconnues.*

4. Segmentation de textes

La segmentation de textes est une problématique proche de l'analyse thématique et recouvre trois tâches distinctes (Ferret *et al.*, 1998) :

1) la segmentation thématique où le texte est découpé en segments thématiquement homogènes ;

2) l'identification thématique qui assigne aux segments de textes un thème. Les thèmes sont préalablement connus ;

3) le suivi thématique qui analyse les relations existant entre les thèmes des segments. Si les thèmes sont très différents, la segmentation thématique est conservée, si à l'inverse ils constituent une spécialisation ou une généralisation, des liens hiérarchiques peuvent être créés entre segments.

4.1. *Segmentation thématique*

La segmentation thématique vise à retrouver la structure sous-jacente d'un document. La segmentation thématique est considérée comme linéaire quand le texte est segmenté en sous-thèmes successifs (Skorochod'ko, 1972) ou comme hiérarchique quand il s'agit de distinguer thèmes et sous-thèmes (Grosz et Sidner, 1986). Les approches état de l'art font l'hypothèse d'une corrélation entre segments et thèmes. Deux segments adjacents seront réunis s'ils sont corrélés. Inversement, si la corrélation calculée apparaît comme faible, une frontière sera insérée entre les deux segments (Hearst, 1997 ; Choi, 2000 ; Riedl et Biemann, 2012). La segmentation thématique est nécessaire pour effectuer l'identification et le suivi thématique. Elle peut

s'appuyer sur le contenu textuel où chaque thème est caractérisé par un vocabulaire spécifique. La cohésion lexicale s'appuie sur la distribution des mots afin d'identifier les changements significatif de vocabulaire révélateurs de changements de thèmes (Hearst, 1994). Des marqueurs de rupture de thème peuvent être utilisés comme, à l'oral, l'intonation ou le silence et, à l'écrit, les caractères de mise en page (titres, espacements, séparateurs, liste d'éléments), les connecteurs de discours ou encore les expressions typiques fortement corrélées avec des frontières de segments thématiques. Les méthodes fondées sur la cohésion lexicale fonctionnent bien pour une segmentation linéaire (Hearst, 1994 ; Choi, 2000). Pour une segmentation hiérarchique, des méthodes plus élaborées requérant une analyse en sous-thèmes sont nécessaires (Yaari, 1997 ; Eisenstein, 2009).

Les principales approches utilisées pour segmenter un texte effectuent soit une analyse lexicale pour détecter les changements de thèmes à l'aide de patrons de cooccurrences (Hearst, 1997) comme les marqueurs de discours (Nomoto et Nitta, 1994), soit calculent la cohésion lexicale (Morris et Hirst, 1991) en exploitant des récurrences lexicales (Hearst, 1994) ou la présence de relations sémantiques fournies par un thésaurus (Morris et Hirst, 1991), un dictionnaire (Kozima, 1993) ou un réseau de collocations construit automatiquement (Ferret *et al.*, 1998).

La cohésion lexicale a inspiré un nombre important d'approches non supervisées. Ces méthodes sont les plus populaires, car elles sont indépendantes du document et ne nécessitent pas de phase d'apprentissage. Les principales approches sont TextTiling (Hearst, 1994), qui identifie localement les ruptures de la cohésion lexicale à l'aide de la mesure statistique du tf-idf et du cosinus, C99 (Choi, 2000), qui mesure globalement la cohésion lexicale au sein de chaque segment et cherche à en maximiser la cohésion lexicale, LSeg (Galley *et al.*, 2003), qui exploite les récurrences lexicales, U00 (Utiyama et Isahara, 2001) fondé sur les modèles probabilistes à facteurs latents, TopicTiling (Riedl et Biemann, 2012), exploitant l'allocation de Dirichlet latente *Latent Dirichlet Analysis (LDA)*, TOPICOLL (Ferret, 2002), exploitant conjointement la récurrence lexicale et des co-occurrences lexicales pour l'analyse thématique, etc. La cohésion lexicale a été appliquée sur deux genres principaux de textes : les documents scientifiques et techniques (Hearst, 1997) où la répétition de termes spécifiques du domaine constitue un indice fiable et les textes narratifs (Morris et Hirst, 1991 ; Kozima, 1993) où il est nécessaire d'utiliser des ressources lexicales pour identifier des relations sémantiques, la récurrence lexicale n'étant pas suffisante. Ferret *et al.* (1998) sont les premiers à proposer une approche mixte en combinant récurrence lexicale et identification de relations sémantiques.

Un ensemble de méthodes employant l'apprentissage supervisé a aussi été proposé pour traiter des discours (Joty *et al.*, 2015), des dialogues multiparties et des forums de chats (Hsueh *et al.*, 2006 ; Hernault *et al.*, 2010) ou pour segmenter des textes au niveau phrastique de manière à identifier les « unités élémentaires du discours » (Hernault *et al.*, 2010 ; Joty *et al.*, 2015). Ces approches combinent des traits de nature différente comme les indices de cohésion lexicale et les caractéristiques de dialogue dans différents classifieurs : arbre de décision (Hsueh *et al.*, 2006), champs

conditionnels aléatoires (CRF) (Hernault *et al.*, 2010 ; Joty *et al.*, 2015). Les travaux les plus récents utilisent des réseaux profonds : TextTiling intègre des plongements lexicaux pour la segmentation de dialogues de questions-réponses (Song *et al.*, 2016), des modèles séquentiels pour la segmentation de dialogues multiparties pour identifier les unités élémentaires de discours (Shi et Huang, 2019) ou encore des modèles d'apprentissage par renforcement (Takanobu *et al.*, 2018). Récemment, Li *et al.* (2018) ont proposé SegBot, un réseau neuronal récurrent (RNN) bidirectionnel couplé avec un mécanisme d'attention qui peut segmenter soit en unités élémentaires de discours, soit en unités thématiques.

Les méthodes ci-dessus s'appliquent principalement pour une segmentation linéaire. L'une des premières approches effectuant une segmentation hiérarchique a utilisé un algorithme de *clustering* hiérarchique (Yaari, 1997). Eisenstein (2009) a proposé un modèle génératif bayésien avec programmation dynamique. Enfin, pour inférer la structuration logique du texte, des traits additionnels de marques caractéristiques de changement de thèmes ont été inclus au sein du CRF (Fauconnier *et al.*, 2014).
Toutes ces méthodes ont été appliquées sur des textes scientifiques, narratifs ou des dialogues écrits ou retranscrits. Dans cet article, nous nous attelons à la segmentation automatique du livre d'heures après sa retranscription par les méthodes de reconnaissance de l'écriture manuscrite qui présentent un taux d'erreurs important.

4.2. *Approche semi-supervisée de segmentation de livres d'heures*

Nous proposons une approche semi-supervisée fondée sur une représentation par plongements de mots des parties du livre d'heures. Notre approche exploite l'idée de

Algorithm 1 Approche semi-supervisée

$Refs = ObsecroTe, Psalm6, Psalm50...$
$Blocks = block1, block2, block3...$
$bestblocks \leftarrow Empty$
for $doc \in Refs$ **do**
  $Max \leftarrow 0$
  for $block \in Blocks$ **do**
    if $sim(doc, block) < Max$ **then**
      $bestblocks[doc] \leftarrow block$
      $Max \leftarrow sim(doc, block)$
    end if
  end for
end for
print $bestblocks$

décomposition en blocs et de mesure de similarité comme utilisée dans (Choi, 2000 ; Utiyama et Isahara, 2001). Elle diffère, cependant, dans la manière de représenter les blocs d'un texte et dans la prise de décision quant à la sélection ou non d'un

bloc candidat, comme segment à part entière. En effet, les approches traditionnelles utilisent au sein d'un même document à segmenter, une similarité interbloc pour, soit les fusionner s'ils sont suffisamment similaires, soit, dans le cas contraire, décider d'une rupture de cohésion lexicale. Notre approche, en revanche, utilise une similarité, non pas entre blocs d'un même document, mais entre un bloc d'un document et une liste de parties du livre d'heures. Ainsi, la cohésion lexicale n'est plus détectée au niveau interne du texte à segmenter, mais au niveau externe, et ceci en s'appuyant sur une base de référence externe contenant des textes préalablement annotés. Notre approche peut être vue comme un alignement de textes de référence des livres d'heures et des textes transcrits découpés arbitrairement en blocs distincts. Nous illustrons notre démarche dans l'algorithme 1.

Premièrement, nous avons à disposition une liste de textes présents dans les livres d'heures qui vont nous servir de liste de référence ($Refs$). À partir de $Refs$, nous construisons une représentation par plongements de mots[5] de chaque texte. Le texte *Obsecro Te*, par exemple, sera représenté par un vecteur de plongements de mots. Ce vecteur est calculé à partir d'une combinaison linéaire des vecteurs de plongements des mots qui le composent (Arora *et al.*, 2017).

Deuxièmement, nous effectuons autant de segmentations de la transcription d'un livre d'heures que de textes présents dans $Refs$. Par exemple, pour le texte *Obsecro Te*, nous découpons le livre d'heures en blocs de taille égale à celle de l'*Obsecro Te*. Nous réitérons cette segmentation en blocs pour chaque texte de $Refs$. Chaque bloc (représenté dans $Blocks$) est représenté par un vecteur de plongements de mots de la même manière que pour $Refs$.

Enfin, pour chaque document référence, nous mesurons la similarité entre le vecteur de plongements de mots de celui-ci et les vecteurs de plongements de mots de tous les blocs de l'ensemble $Blocks$. Le bloc le plus similaire sera considéré comme étant la section correspondante du livre d'heures. Cette procédure est répétée pour chaque section et sous-section du livre d'heures.

La structure des livres d'heures n'étant pas identique d'un livre à l'autre comme l'illustre le tableau 2, un seuil appris sur un corpus d'entraînement est utilisé pour éliminer les documents de référence qui présentent une similarité faible au regard des blocs de la transcription. Une des limites de notre approche est l'hypothèse que la taille de la section de référence et de celle de la section présente dans la transcription sont similaires. Cependant, nous considérons que l'impact de la variabilité en termes de taille est faible vis-à-vis de la représentation par plongements de mots des blocs. Aussi, nous laissons une analyse fine de détection de ruptures entre les sections et sous-sections pour des travaux futurs.

5. Nous utilisons le modèle préentraîné latin de FastText qui s'appuie sur Wikipédia `https://github.com/facebookresearch/fastText/blob/master/docs/crawl-vectors.md`

4.3. *Expériences et résultats*

Dans une première expérience, nous évaluons plusieurs approches état de l'art ainsi que notre approche semi-supervisée sur un livre d'heures construit artificiellement et composé de plusieurs sous-sections d'un livre d'heures. La motivation de l'utilisation de données artificielles est de comparer les performances des méthodes sur des données propres, dépourvues d'erreurs de transcription, et sur les données bruitées proposées par la reconnaissance des caractères. Dans une seconde expérience, nous évaluons les différents systèmes sur la transcription du livre d'heures Harvard253[6], extrait de la collection digitale de Harvard (Harvard digital collections).

4.3.1. *Données expérimentales*

Le tableau 5 résume le nombre de segments de niveau 1, le nombre de segments de niveau 2 ainsi que la somme des nombres de segments des deux niveaux de segmentation notée « Niveaux 1 et 2 ». Le livre d'heures artificiel est construit aléatoirement à partir de 29 sous-sections de référence comportant entre autres les prières *Obsecro Te* et *O Intemerata* ainsi que plusieurs psaumes. La transcription du livre d'heures Harvard253 est composée de huit sections au premier niveau (comme l'illustre le tableau 1) et de 38 sections au niveau 2.

	Nombre de sections par niveau		
Livre d'heures	#Niveau 1	#Niveau 2	# Niveaux 1 et 2
Artificiel	-	29	29
Harvard253	8	38	46

Tableau 5. *Représentation du nombre de sections par niveau de la transcription Harvard253 ainsi que du livre d'heures artificiel*

4.3.2. *Mesures d'évaluation*

Les approches sont évaluées en termes de score d'erreurs par P_k (Beeferman *et al.*, 1999) et Windowdiff (WD) (Pevzner et Hearst, 2002). P_k est une mesure d'erreurs qui combine le rappel et la précision pour estimer la contribution relative de différents types de traits. Cependant, elle possède plusieurs désavantages et quelques points faibles épinglés par Pevzner et Hearst (2002). P_k est sensible à la variabilité de la taille des segments. Elle pénalise plus sévèrement les faux négatifs que les vrais positifs. Enfin, elle surpénalise les erreurs de segmentation qui sont proches des frontières (*near misses*) correctes. Pour pallier les manques de P_k, une seconde mesure, aussi état de l'art, WindowDiff (WD) est considérée. Cette mesure qui est une variante

6. `https://curiosity.lib.harvard.edu/medieval-renaissance-manuscripts/catalog/34-990094032810203941`

de P_k, pénalise de la même manière les faux positifs et les segmentations proches des vraies frontières.

4.3.3. *Approches état de l'art*

Nous évaluons, plusieurs approches état de l'art qui sont TextTiling (Hearst, 1994), le modèle par *clustering* (C99) (Choi, 2000), le modèle probabiliste par programmation dynamique (U00) (Utiyama et Isahara, 2001), le modèle de graphes partitionné par rupture minimale (MinCut) (Malioutov et Barzilay, 2006) et un modèle hiérarchique Bayésien (HierBays) (Eisenstein, 2009). Nous évaluons aussi une approche par modèle thématique TopicTiling (Riedl et Biemann, 2012). D'autres modèles par apprentissage comme dans (Koshorek *et al.*, 2018) auraient pu être considérés mais le manque de données d'apprentissage rend leur utilisation inefficace à ce stade.

4.4. *Résultats*

Le tableau 6 illustre les performances des différentes approches lors de la première expérience menée sur les données artificielles. Comme le montrent les résultats, aucune des approches état de l'art n'obtient des performances satisfaisantes. L'approche pionnière TextTiling obtient des résultats très faibles (P_k = 54 % et WD = 55,7 %). Une modélisation thématique par TopicTiling, bien que meilleure en termes de P_k avec un score de 42,5 %, obtient de moins bons résultats que TextTiling en termes de WD avec un score de 58,3 %. Ceci laisse à penser que le modèle thématique construit avec la LDA n'a pas permis une segmentation efficace. Un manque de données d'entraînement pourrait expliquer ces résultats. La meilleure approche état de l'art, MinCut, obtient un score P_k de 39 % et un score WD de 42,6 %. Enfin, notre approche semi-supervisée obtient les meilleurs résultats avec un score P_k de 27,5 % et un score WD de 29,2 %.

Le tableau 7 illustre les résultats des expériences menées sur la transcription du livre Harvard253 à la fois au premier niveau, au deuxième niveau ainsi que sur les deux niveaux de segmentation notés « Niveaux 1 et 2 ». La encore, nous observons les mêmes comportements que ceux constatés dans l'expérience précédente. Les résultats des approches état de l'art sont décevants. Cependant, une exception est à noter au niveau 1 concernant les approches U00, avec un score P_k de 23,6 % et HierBays, avec un score P_k de 14,2 % et un score WD de 25,3 %. De manière générale, notre approche obtient les meilleurs résultats au niveau 2 et au niveau d'une évaluation sur les deux niveaux (Niveaux 1 et 2) avec un score P_k de 29,9 % et un score WD de 31,4 % au niveau 1 et un score global de 31,4 % en termes de P_k et 32,6 % en termes de WD.

Ces résultats montrent d'une part, la pertinence d'utiliser une approche semi-supervisée dans ce type de textes que sont les livres d'heures et, d'autre part, que les erreurs de segmentation ont un impact faible sur la segmentation. Il est à noter que l'approche HierBays obtient les meilleurs résultats au niveau 1. Ainsi, une com-

	Data	
Approche	P_k	WD
TextTiling	54,0	55,7
C99	44,0	44,2
U00	47,6	51,7
MinCut	39.0	42.6
HierBays	41,3	50,3
TopicTiling	42,5	58,3
Approche proposée	**27,5**	**29,2**

Tableau 6. *Analyse de différentes méthodes de segmentation avec P_k et WindowDiff (WD) sur le livre d'heures artificiel. P_k et WD sont de taux d'erreurs, les scores les plus faibles caractérisent les meilleures méthodes.*

binaison de notre approche semi-supervisée avec l'approche HierBays pourrait être envisagée afin d'améliorer les résultats au niveau 1.

	Niveaux de segmentation					
	Niveau 1		**Niveau 2**		**Niveaux 1 et 2**	
Approche	P_k	WD	P_k	WD	P_k	WD
TextTiling	66,9	99,9	48, 1	60,4	46,0	57,5
C99	68,7	96,8	60,0	676	59,2	66,1
U00	23,6	39,5	38,0	39,4	35,6	38,7
MinCut	40,9	49,2	48,4	52,1	45,2	48,7
HierBays	**14,2**	**25,3**	36,7	39,9	32,9	38,5
TopicTiling	60,3	87,0	42,0	48,3	42,0	47,4
Approche proposée	27,2	33,5	**29,9**	**31,4**	**31,4**	**32,6**

Tableau 7. *Analyse de différentes méthodes de segmentation avec P_k et WindowDiff (WD) pour les deux premiers niveaux de segmentation du livre d'heures. P_k et WD étant des taux d'erreurs, les scores les plus faibles caractérisent les meilleures méthodes.*

5. Conclusion

Nous avons présenté dans cet article une chaîne globale de traitements automatiques du livre d'heures, allant de la reconnaissance de l'écriture manuscrite médiévale sur parchemin à sa segmentation en parties. Les traitements consistent en une première étape dédiée à la transcription du livre d'heures et en une seconde qui vise à fournir une segmentation hiérarchique de plusieurs niveaux. Le taux d'erreurs mots et caractères de la reconnaissance du document reste important, il est dû au manque de données d'entraînement. Nous avons évalué la segmentation au premier et second niveau en appliquant les principales approches état de l'art de segmentation thématique. Les résultats ne sont pas satisfaisants à l'exception de l'approche HierBays pour le seul premier niveau. Nous avons proposé une approche semi-supervisée entraînée sur les textes de référence connus, constitutifs du livre d'heures. Cette approche produit des résultats encourageants sur les deux niveaux de segmentation du livre d'heures. Ils laissent à penser que notre approche de segmentation n'est que peu sensible, voire insensible, aux erreurs de transcription puisqu'un même comportement à été observé à la fois sur un livre d'heures artificiel propre et sur le livre d'heures transcrit bruité. À visée globale, ce travail s'inscrit dans un continuum d'analyses et d'interprétations des textes liturgiques anciens que sont les livres d'heures. Si pour l'heure, nous ne sommes qu'au début de la constitution d'une chaîne de traitement robuste, nous envisageons, comme prochain travail, d'étendre l'évaluation à une base de données plus conséquente de livres pour ensuite permettre aux historiens d'interpréter nos résultats à des fins historiques et anthropologiques.

Remerciements

Ce travail qui s'inscrit dans le cadre du projet HORAE (Hours - Recognition, Analysis, Editions) a bénéficié d'une aide de l'Agence nationale de la recherche portant la référence ANR-17-CE38-0008.

6. Bibliographie

Ares Oliveira S., Seguin B., Kaplan F., « dhSegment : A generic deep-learning approach for document segmentation », *International Conference Frontiers in Handwriting Recognition*, 2018.

Arora S., Yingyu L., Tengyu M., « A Simple but Tough to Beat Baseline for Sentence Embeddings », *Proceedings of the 17th International Conference on Learning Representations (ICLR'17)*, p. 1-11, 2017.

Beeferman D., Berger A., Lafferty J., « Statistical Models for Text Segmentation », *Mach. Learn.*, vol. 34, n° 1-3, p. 177-210, February, 1999.

Bluche T., Hamel S., Kermorvant C., Puigcerver J., Stutzmann D., Toselli A. H., Vidal E., « Preparatory KWS Experiments for Large-Scale Indexing of a Vast Medieval Manuscript

Collection in the HIMANIS Project », *International Conference on Document Analysis and Recognition*, 2017a.

Bluche T., Kermorvant C., Ney H., *How to design deep neural networks for handwriting recognition*, 2017b.

Boillet M., Bonhomme M.-L., Stutzmann D., Kermorvant C., « HORAE : an annotated dataset of books of hours », *International Workshop on Historical Document Imaging and Processing*, 2019.

Choi F. Y. Y., « Advances in Domain Independent Linear Text Segmentation », *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, NAACL 2000, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 26-33, 2000.

De Hamel C., *A history of illuminated manuscripts*, 2nd ed. rev., enl. and with new ill edn, Phaidon P., London, 1994.

Diem M., Kleber F., Fiel S., Grüning T., Gatos B., « cBAD : ICDAR2017 Competition on Baseline Detection », *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2017.

Drigsdahl E., *Late Medieval and Renaissance Illuminated Manuscripts - Books of Hours 1300-1530*, 2013.

Eisenstein J., « Hierarchical Text Segmentation from Multi-Scale Lexical Cohesion », *Human Language Technologies : Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, May 31 - June 5, 2009, Boulder, Colorado, USA*, p. 353-361, 2009.

Fauconnier J.-P., Sorin L., Kamel M., Mojahid M., Aussenac-Gilles N., « Détection automatique de la structure organisationnelle de documents à partir de marqueurs visuels et lexicaux », *Actes de la 21e conférence sur le Traitement Automatique des Langues Naturelles*, Association pour le Traitement Automatique des Langues, Marseille, France, p. 340-351, July, 2014.

Ferret O., « Using Collocations for Topic Segmentation and Link Detection », *19th International Conference on Computational Linguistics, COLING*, 2002.

Ferret O., Grau B., Masson N., « t : Two Methods for Two Kinds of Texts », *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98/COLING '98, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 392-396, 1998.

Galley M., McKeown K., Fosler-Lussier E., Jing H., « Discourse Segmentation of Multi-party Conversation », *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, p. 562-569, 2003.

Grosz B. J., Sidner C. L., « Attention, Intentions, and the Structure of Discourse », *Computational Linguistics*, vol. 12, n° 3, p. 175-204, 1986.

Grüning T., Leifert G., Strauß T., Labahn R., « A Two-Stage Method for Text Line Detection in Historical Documents », 2018.

Hearst M. A., « MULTI-PARAGRAPH SEGMENTATION EXPOSITORY TEXT », *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Las Cruces, New Mexico, USA, p. 9-16, June, 1994.

Hearst M. A., « TextTiling : Segmenting Text into Multi-paragraph Subtopic Passages », *Comput. Linguist.*, vol. 23, n° 1, p. 33-64, March, 1997.

Hernault H., Bollegala D., Ishizuka M., « A Sequential Model for Discourse Segmentation », *Computational Linguistics and Intelligent Text Processing, 11th International Conference, CICLing 2010, Iasi, Romania, March 21-27, 2010. Proceedings*, p. 315-326, 2010.

Hsueh P.-y., Moore J. D., Renals S., « Automatic Segmentation of Multiparty Dialogue », *11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006.

Joty S., Carenini G., Ng R. T., « CODRA : A Novel Discriminative Framework for Rhetorical Analysis », *Computational Linguistics*, vol. 41, n° 3, p. 385-435, 2015.

Koshorek O., Cohen A., Mor N., Rotman M., Berant J., « Text Segmentation as a Supervised Learning Task », *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 2 (Short Papers)*, Association for Computational Linguistics, New Orleans, Louisiana, p. 469-473, June, 2018.

Kozima H., « Text Segmentation Based on Similarity Between Words », *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics*, ACL '93, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 286-288, 1993.

Lang E., Puigcerver J., Toselli A. H., Vidal E., « Probabilistic Indexing and Search for Information Extraction on Handwritten German Parish Records », *International Conference on Frontiers in Handwriting Recognition*, 2018.

Lebigue J.-B., *Initiation aux manuscrits liturgiques*, 2007.

Leroquais V., *Les Livres d'heures manuscrits de la Bibliothèque nationale*, [s. n.], Paris, 1927.

Li J., Sun A., Joty S., « SegBot : A Generic Neural Text Segmentation Model with Pointer Network », *Proceedings of the 27th International Joint Conference on Artificial Intelligence and the 23rd European Conference on Artificial Intelligence*, IJCAI-ECAI-2018, Stockholm, Sweden, p. xx - xx, July, 2018.

Malioutov I., Barzilay R., « Minimum Cut Model for Spoken Lecture Segmentation », *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006*, 2006.

Morris J., Hirst G., « Lexical Cohesion Computed by Thesaural Relations As an Indicator of the Structure of Text », *Comput. Linguist.*, vol. 17, n° 1, p. 21-48, March, 1991.

Moysset B., Kermorvant C., Wolf C., « Learning to detect, localize and recognize many text objects in document images from few examples », *International Journal on Document Analysis and Recognition*, vol. 21, p. 161-175, 2018.

Nomoto T., Nitta Y., « A Grammatico-statistical Approach to Discourse Partitioning », *Proceedings of the 15th Conference on Computational Linguistics - Volume 2*, COLING '94, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 1145-1150, 1994.

Ottosen K., *The responsories and versicles of the Latin office of the dead*, Aarhus university press, Aarhus, 1993.

Ottosen K., « Responsories of the Latin Office of the Dead », 2008.

Peddinti V., Povey D., Khudanpur S., « A time delay neural network architecture for efficient modeling of long temporal contexts », *INTERSPEECH*, 2015.

Pevzner L., Hearst M. A., « A Critique and Improvement of an Evaluation Metric for Text Segmentation », *Comput. Linguist.*, vol. 28, n° 1, p. 19-36, March, 2002.

Plummer J., Clark G. T., « Obsecro Te », *Beyond Use : A Digital Database of Variant Readings In Late Medieval Books of Hours*, 2015.

Renton G., Soullard Y., Chatelain C., Adam S., Kermorvant C., Paquet T., « Fully convolutional network with dilated convolutions for handwritten text line segmentation », *International Journal on Document Analysis and Recognition*, vol. 21, p. 177-186, 2018.

Riedl M., Biemann C., « TopicTiling : A Text Segmentation Algorithm based on LDA », *Proceedings of ACL 2012 Student Research Workshop*, Association for Computational Linguistics, p. 37-42, 2012.

Rosenwein B. H., « Y avait-il un « moi » au haut Moyen Âge ? », *Revue historique*, vol. n 633, n° 1, p. 31-52, 2005.

Rézeau P., *Les prières aux saints en français à la fin du Moyen Age*, Publications romanes et françaises, Droz, Genève, 1983.

Rézeau P., *Répertoire d'incipit des prières françaises à la fin du Moyen âge : addenda et corrigenda aux répertoires de Sonet et Sinclair, nouveaux incipit*, Droz, Genève, 1986.

Sánchez J. A., Romero V., Toselli A. H., Vidal E., « ICFHR2016 Competition on Handwritten Text Recognition on the READ Dataset », *International Conference on Frontiers in Handwriting Recognition*, 2016.

Shi Z., Huang M., « A Deep Sequential Model for Discourse Parsing on Multi-Party Dialogues », *AAAI*, 2019.

Sinclair K. V., *Prières en ancien français : nouvelles références, renseignements complémentaires, indications bibliographiques, corrections et tables des articles du "Répertoire" de Sonet*, Archon books, Hamden, 1978.

Sinclair K. V., *French devotional texts of the Middle Ages : a bibliographic manuscript guide*, Westport, 1979.

Sinclair K. V., *French devotional texts of the Middle Ages : a bibliographic manuscript guide. First supplement*, Westport, 1982.

Sinclair K. V., *Prières en ancien français : additions et corrections aux articles 1-2374 du "Répertoire" de Sonet. Supplément*, James Cook Univ. of North Queensland, Townsville, 1987.

Sinclair K. V., *French devotional texts of the Middle Ages : a bibliographic manuscript guide. Second supplement*, New York, 1988.

Skorochod'ko E. F., « Adaptive Method of Automatic Abstracting and Indexing », *Information Processing*, 1972.

Sonet J., *Répertoire d'incipit de prières en ancien français*, n° 54 in *Société de publications romanes et françaises*, Droz, Genève, 1956.

Song Y., Mou L., Yan R., Yi L., Zhu Z., Hu X., Zhang M., « Dialogue Session Segmentation by Embedding-Enhanced TextTiling », *Interspeech*, p. 2706-2710, 09, 2016.

Strauß T., Leifert G., Labahn R., Hodel T., Mühlberger G., « ICFHR2018 Competition on Automated Text Recognition on a READ Dataset », *International Conference on Frontiers in Handwriting Recognition*, 2018.

Sánchez J. A., Romero V., Toselli A. H., Villegas M., Vidal E., « ICDAR2017 Competition on Handwritten Text Recognition on the READ Dataset », 2017.

Takanobu R., Huang M., Zhao Z., Li F., Chen H., Zhu X., Nie L., « A Weakly Supervised Method for Topic Segmentation and Labeling in Goal-oriented Dialogues via Reinforcement Learning », *IJCAI-ECAI*, p. 4403-4410, 2018.

Utiyama M., Isahara H., « A Statistical Model for Domain-independent Text Segmentation », *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, p. 499-506, 2001.

Wieck R. S., Poos L. R., Reinburg V., Plummer J. H., Walters art museum, *Time sanctified : the Book of Hours in medieval art and life*, G. Braziller, New York, 1988.

Yaari Y., « Segmentation of Expository Texts by Hierarchical Agglomerative Clustering », *CoRR*, 1997.