HAL
open science

Books of Hours: the First Liturgical Corpus for Text Segmentation

Amir Hazem, Béatrice Daille, Marie-Laurence Bonhomme, Martin Maarand, Mélodie Boillet, Christopher Kermorvant, Dominique Stutzmann

To cite this version:

Amir Hazem, Béatrice Daille, Marie-Laurence Bonhomme, Martin Maarand, Mélodie Boillet, et al.. Books of Hours: the First Liturgical Corpus for Text Segmentation. 12th Language Resources and Evaluation Conference, May 2020, Marseille (Virtual), France. pp.776-784. hal-02931294

HAL Id: hal-02931294

https://hal.science/hal-02931294

Submitted on 5 Sep 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 776–784
Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

Books of Hours: the First Liturgical Corpus for Text Segmentation

Amir Hazem*, Béatrice Daille*, Marie-Laurence Bonhomme†, Martin Maarand†, Mélodie Boillet†‡, Christopher Kermorvant†‡ and Dominique Stutzmann§
* LS2N - Université de Nantes, Nantes
§ Institut de recherche et d'histoire des textes (IRHT), Paris
† TEKLIA, Paris
‡ LITIS - Université de Rouen-Normandie, Rouen
amir.hazem@ls2n.fr, beatrice.daille@univ-nantes.fr, Dominique.Stutzmann@irht.cnrs.fr, Kermorvant@teklia.com

Abstract

The Book of Hours was the bestseller of the late Middle Ages and Renaissance. It is a historical invaluable treasure, documenting the devotional practices of Christians in the late Middle Ages. Up to now, its textual content has been scarcely studied because of its manuscript nature, its length and its complex content. At first glance, it looks too standardized. However, the study of book of hours raises important challenges: (i) in image analysis, its often lavish ornamentation (illegible painted initials, line-fillers, etc.), abbreviated words, multilingualism are difficult to address in Handwritten Text Recognition (HTR); (ii) its hierarchical entangled structure offers a new field of investigation for text segmentation; (iii) in digital humanities, its textual content gives opportunities for historical analysis. In this paper, we provide the first corpus of books of hours, which consists of Latin transcriptions of 300 books of hours generated by Handwritten Text Recognition (HTR) - that is like Optical Character Recognition (OCR) but for handwritten and not printed texts. We designed a structural scheme of the book of hours and annotated manually two books of hours according to this scheme. Lastly, we performed a systematic evaluation of the main state of the art text segmentation approaches.
Keywords: text segmentation, books of hours, structural scheme, hierarchical segmentation

1. Introduction

In the late medieval Europe, the book of hours was a Christian devotional manuscript used by lay people as a guidance book in their daily prayers (Leroquais, 1927; Wieck et al., 1988). While this personal object of about 300 pages was mainly used by aristocracy and, later on, by the middle class, the emergence of the printing press made books of hours available for nearly every European family (Hindman and Marrow, 2013). With a massive production resulting in more than 10,000 extant manuscripts, this number one liturgical best seller is an invaluable source of information of the late medieval usage and practices and witness to the profound changes in the European society on cultural, religious, and industrial levels. Since they were illuminated with beautiful miniature paintings and decorations (De Hamel, 1992), books of hours have attracted great attention. They became very popular for art historians and many studies have been conducted for aesthetic purposes (Wieck, 2001). However, their textual content is still scarcely studied, often because of their length, the complexity of their liturgical content and the fact that they appear as being too standardized.

Despite the standardized character of books of hours in the way they were organized around the eight hours of the day, with a set of required and pre-selected readings, extra readings were often added. Each book of hours was a customized devotional object dedicated to a specific patron for whom the content was adapted according to their gender, geographical location, preferred saints and prayers (Clark, 2003), etc. As a consequence, each book of hours is unique with regard to its content and structure, which make their study highly complex since within the same structure the content may slightly or drastically differ, depending partially on the commissioner, on the crafts(wo)men, on the intended use, and on other phenomena yet to be discovered. Furthermore, within the large amount of devotional readings, some prayers or sequence of prayers were used in several hours of the day (either they were repeated, or different users read them at different times), and this increases segmentation complexity.

Many books of hours have been digitized worldwide, with some dedicated online corpora and exhibitions[1]. Conversely, available transcriptions and linguistically annotated books of hours are very rare as mentioned in De Hamel (1994). The very limited available texts reproduced here and there, illustrate few pages of a given book and are often fragmented or incomplete. A recent work studied calendars in medieval manuscripts, including books of hours (Heikkilä and Roos, 2018). In this paper, we present the first corpus release of books of hours in Latin issued from an automatic transcription via HTR. The corpus is composed of 300 automatic book transcriptions and two extra books manually compiled.

The present paper is the first attempt to process the whole content of books of hours. This is a great challenge in digital humanities that brings together researchers from the fields of document recognition, NLP and history. The study of the content of books of hours aims at providing opportunities for historical analysis to better understand the cultures and faiths from the 13th c. to the 16th c. In addition, its complex logical entangled structure offers a new type of resource for text segmentation. Since traditional segmentation data sets lie within the scope of expository texts, narrative or issued from spoken an written dialogues (Kozima, 1993; Hearst, 1994; Nomoto and Nitta, 1994; Utiyama and Isahara, 2001) or more recently from Wikipedia (Arnold et

[1] https://library.harvard.edu/collections/picturingprayer/exhibition.html

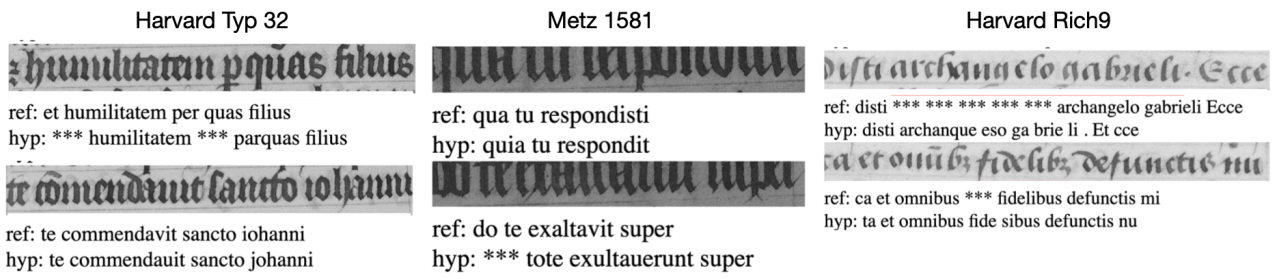

Figure 1: Examples of text recognition results (hyp) and their manual annotations (ref) on three books of hours manuscripts: Harvard Typ 32, Metz 1581 and Harvard Rich 9. The writings are of type "Textualis" on the left and center sides, and of type "Cursiva" on the right side.

al., 2019), books of hours arguably constitute a new genre of segmentation texts that exhibits difficulties with regard to length, structure and many ambiguities. Also, the vast majority of state of the art approaches dealing with text segmentation are based on topical shifts to build their segmentation models. On the contrary, sections and subsections of books of hours are much more correlated which makes the topical shift hypothesis inconsistent for this type of data as we will show in this paper. Our work has three main contributions. First, we release the BOH300 corpus of 300 transcriptions of manuscripts. Second, we define an annotation scheme of the complex structure of books of hours and release a manual annotation of three books at three levels of granularity. Finally, we conduct experiments on text segmentation of books of hours based on the main state of the art approaches and show their inappropriateness with regard to our data.

2. BOH300 Corpus

The BOH300 corpus contains transcriptions from two different sources: a majority of automatic raw transcriptions, with the addition of two manually transcribed books. The annotation of books of hours at three levels of granularity is presented in Section 3.

2.1. Automatic Transcriptions

We provide a set of 300 transcribed books. Each book is about 300 pages long on average. Transcriptions have been generated from images made available by the BVMM Bibliothèque Virtuelle des Manuscrits Médiévaux[2], the French national library Gallica[3], Harvard library[4], and from the Walters Art Museum and three other libraries. The document recognition system used to generate these transcriptions is based on a U-shaped fully convolutional neural network (Boillet et al., 2020) for line detection and layout analysis, and on the KALDI library (Arora et al., 2019) for text recognition. With an image such as the one in Figure 2 as input, the system outputs a line by line recognized text. The presence of decorative elements, such as miniatures, border or inline ornaments, as well as the gothic script, makes the task more difficult.

[2]https://bvmm.irht.cnrs.fr/
[3]https://gallica.bnf.fr
[4]https://library.harvard.edu/

Figure 2: Two pages from the so-called "Grandes Heures d'Anne de Bretagne", a book of hours in Latin, Paris, Bibliothèque nationale de France, Latin 9474.

Figure 1 presents examples of text recognition on three separate books of hours with two different writing styles. The manual transcription which represents the gold standard is labelled as 'ref', while the result of the automatic transcription is labelled as 'hyp'. We notice several recognition errors: on the left side, the abbreviated '*et*' at the beginning of the line is not recognized, while the abbreviated '*per*' is mistakenly expanded as '*par*'. In the second line, *i/j* and *u/v* are incorrectly disambiguated; the first line of the middle book (Metz 1581) is truncated; on the right side (Harvard Rich 9), if the abbreviation '*et*' is recognized, there is a confusion between the letters *c* and *t*. Overall, the word error rate (WER) is of about 30%.

2.2. Manual Transcriptions

Because of the length of books of hours (thousands of lines) as well as their special character writing (Gothic scripts, such as Textualis, Cursiva, etc.), their manual transcription requires huge efforts and reading skills to decipher these special characters. For this reason, manual transcriptions

are very rare. The only available transcriptions of an entire book are presented here. We provide manual transcriptions[5] of two books of hours. The first, to which we refer as Medievalist, is taken from the Latin/English Primer book of 1599[6]. It has been made available as html pages by Senior lecturer Glenn Gunhouse[7] in the Medievalist website[8]. We manually extracted each part of the book of hours from the html pages and arranged them in a raw document accompanied with annotations according to the structure of the book of hours. The second, i.e. Arsenal1194, is a manuscript digitized and available at the French digital library Gallica[9]. The provided text in the BOH300 corpus is hybrid: it has been partly transcribed by a historian on the basis of the digitized version of MS Paris, BnF, Arsenal 1194[10] and partly aligned with enhanced editions. Here also, we arranged the book in a raw document with its structure annotations.

3. Structure of Books of Hours

The nature and origin of books of hours make their structure quite complex. They appeared in the 13th century and evolved over time, with additional texts that were included as a means of enrichment. The whole was arranged in a particular structure that varied in its details depending on origins, commissioners, worshiped saints, etc. Hence, the structure of books of hours is not uniform. Also, if the main prayers, which correspond to the highest structure level, can often be identified thanks to their corresponding paintings[11] positioned at the beginning of each prayer, lower level indications are often omitted. Due to the lack of extensive comparative studies of the structure of books of hours, up to now, no official, unique standard has been established.
We conducted a detailed analysis of the content and structure of books of hours, and with the help of experts we propose a structural scheme at three levels of granularity. The first level is the core of the book of hours, it represents the major addressed prayers. The second level mainly represents the eight hours of the day and so, it indicates when prayers are theoretically to be read. Finally, the third level is more fine-grained: it represents prayer cycles and contains specific passages that might be excerpted from the Bible, the book of psalms, or from any non biblical texts. A summary of the categories of each level is shown in Table 1. Sections 3.1, 3.2 and 3.3 provide more details on the book of hours three levels structure.

Level1 1	Level 2	Level 3
Calendar	John	Hymn
Gospel Readings	Luke	Canticle
Obsecro Te	Mark	Capitule
O Intemerata	Matthew	Oratio
Hours of the Virgin	Matins	Psalms
Hours of the Cross	Lauds	Nocturns
Hours of the Holy Spirit	Prime	Lesson
Penitential Psalms / Litany	Terce	Invitatory
Hours of the Dead	Sext	
Suffrages	None	
Prayers	Vespers	
	Compline	
	Penitential	
	Litany	

Table 1: Books of Hours at three levels of granularity

3.1. First Level

For the first level of granularity, books of hours traditionally begin with a liturgical calendar followed by short extracts from each of the four Gospels, then by the Hours of the Virgin which is made up of eight sets of devotional prayers to Mary that define the book of hours. The Hours of the Virgin are followed by several other prayers including the Hours of the Cross and the Hours of the Holy Spirit. Sometimes it can be followed by the Hours of the Passion or a set of hours devoted to particular saints. Finally, books of hours include the Office of the Dead and are completed by the seven Penitential Psalms, Litanies and prayers to the Virgin and various saints. Hence, we propose 11 categories of level one. This short description is of course a generic theoretical definition. Actually, the content and books of hours section ordering vary a lot as shown in Table 2. The next Sections present each part of the first level categories.

3.1.1. Calendar

The Calendar[12] is placed at the beginning of the book of hours and is used to record the seasons, saints and days of the week. It generally includes Roman calendrical dates: Nones, Kalends and Ides. The usual Monday to Sunday names are referred to by dominical letters (a-g). Roman numerals (i-xix) refer to the dates depending on the Easter cycle. Major feast days are frequently distinguished by a gold or red color. Local saints are often included in calendars, which help identifying the geographical origin.

3.1.2. Gospel Readings

The Calendar is often followed by extracts from the four Gospels[13] (John, Luke, Matthew and Mark), representing the Church's major feasts (Christmas: John 1:1-14; Annunciation: Luke 1:26-38; Epiphany: Matthew 2:1-12; Ascension: Mark 16:14-20). An additional reading from John's gospel may be present, covering the Passion (John 18:1-19:42). This part may be illustrated by portraits of the four apostles with their symbols (an angel for Matthew, a lion for Mark, an ox for Luke, and an eagle for John).

[5] Manual transcription guarantees no error transcriptions.
[6] https://babel.hathitrust.org/cgi/pt?id=uc1.b3359820
[7] http://medievalist.net/bio.htm
[8] http://medievalist.net/hourstxt/home.htm
[9] https://gallica.bnf.fr/
[10] https://gallica.bnf.fr/ark:/12148/btv1b550103864
[11] There are some very commonly used iconographic cycles, with, e.g. the Annunciation to the Virgin Mary at the start of the hours of the Virgin, a Crucifixion painting for the hours of the cross, David for the penitential psalms, etc.
[12] An example of the calendar can be visualized here: https://gallica.bnf.fr/ark:/12148/btv1b550103864/f3
[13] Gospels extracts example: https://gallica.bnf.fr/ark:/12148/btv1b550103864/f27

Harvard251	Harvard253	Harvard32	Harvard1000	Harvard464	Poitiers1097	Poitiers43	Poitiers46
Calendar	Calendar	Calendar	Calendar	Suffrages	Gospel Lections	Calendar	Calendar
Gospel Lections	Gospel Lections	Gospel Lections	Virgin	Virgin	Prayers	Gospel Lections	Gospel Lections
Virgin	Obsecro Te	Obsecro Te	Litany of Mary	Cross	Virgin	Obsecro Te	Obsecro Te
Cross	Virgin	Virgin	Psalms Litany	Holy Spirit	Holy Spirit	Virgin	O Intemerata
Psalms Litany	Psalms Litany	O Intemerata	Dead	Psalms Litany	Mixed VCS	Cross	Virgin
Suffrages	Cross	Psalms Litany	Cross	Dead	Psalms Litany	Holy Spirit	Cross
Orationes	Holy Spirit	Cross	Holy Spirit	Gospel Lections	Dead	Psalms Litany	Holy Spirit
	Dead	Holy Spirit	Orationes	Obsecro Te	Suffrages	Dead	Psalms Litany
		Dead			Prayers	Suffrages	Dead
		Joys Virgin			Suffrages	7 req Lord	Verses
		7 req Lord			Gospels (Passio)	Suffrages Prayers	Suffrages

Table 2: A sample of books of hours first level segments. Each colour corresponds to a distinct first level annotation.

3.1.3. Obsecro Te and O Intemerata

Obsecro Te[14] (I beseech You...) and O Intemerata (O chaste...) are two prayers that may appear early in the book of hours (e.g. right after the Gospel readings) or within the prayers section at the end of the book. The prayers are named after their first words (*incipit*): *Obsecro Te domina sancta Maria...*, and *O Intemerata, et in aeternum...*, etc. Often, the beginnings of these prayers are illustrated with ornaments around the first letter of the prayer as illustrated in Figure 3[15]. This may lead to errors in transcriptions.

Figure 3: Example of a book of hours folio of Obsecro Te prayer that begins with ornaments.

[14]Obsecro Te example: `https://gallica.bnf.fr/ark:/12148/btv1b550103864/f37`

[15]`https://www.flickr.com/photos/medmss/13900826141`

3.1.4. Hours of the Virgin

The Hours of the Virgin[16] are considered as central and constitute the main texts of any book of hours. Each hour is composed of various couples of antiphons[17] and responses, psalms, hymns, canticles and prayers to the Virgin Mary. The basic decomposition of the prayers to the Virgin is related to the hours of day and very often illustrated with scenes depicting events connected with the infancy of the Christ. Also, different prayers are used depending on the period: during the greater part of the year or during the advent season, as well as on the feast of the annunciation, or during Christmas.

3.1.5. Hours of the Cross and the Holy Spirit

The texts of these prayers are respectively addressed to the Holy Cross and to the Holy Spirit. The Hours of the Cross[18] usually follow the Hours of the Virgin, as can be seen in Table 2 (Poitiers1097, Poitiers43 and Poitiers46). Also, the Hours of the Holy Spirit[19] often follow the Hours of the Virgin (Table 2: Poitiers 1097) or the Cross (Table 2: Harvard253, Harvard32, etc.). These prayers are much shorter than the Hours of the Virgin.

3.1.6. Penitential Psalms and Litany

Penitential Psalms and Litany[20] are composed of a sequence of seven Psalms (6, 31, 37, 50, 101, 129 and 142) and a series of invocations and prayers Litanies. The litany encompasses a call for the mercy of God, then lists of saints where each name is followed by the invocation: *Ora pro nobis*, then petitions and prayers. The order of saints reflects the heavenly hierarchy. Sometimes *Ora pro nobis* is not used. After the elicitation of several saints, the psalm 69 may follow and several verses and responses. Litany may ends with an oratio and a couple of verses and responses.

[16]Hours of the Virgin example: `https://gallica.bnf.fr/ark:/12148/btv1b550103864/f47`

[17]Underlined terms will be defined in Sections 3.2 and 3.3.

[18]Hours of the Cross example: `https://gallica.bnf.fr/ark:/12148/btv1b550103864/f225`

[19]Hours of the Holy Spirit example: `https://gallica.bnf.fr/ark:/12148/btv1b550103864/f233`

[20]Penitential Psalms and Litany example: `https://gallica.bnf.fr/ark:/12148/btv1b550103864/f187`

3.1.7. Office of the Dead

Office of the Dead[21] is repeats the prayer said on the day of the memory of all the dead. It contains only three hours of prayers (Vespers, Matins and Lauds, so that it is generally called "Office" rather than "Hours"). Each hour contains a series of psalms. Vespers usually contains psalms: 114, 119, 120, 129, 134 and 145, while Matins contains psalms: 5, 6, 7 or 22, 24, 26 or 39, 40, 41 followed by lessons (lectio). Finally, Lauds contains psalms: 50, 64, 62, 66, 148, 149, 150, 129 and ends with a prayer (Oratio).

3.1.8. Suffrages

Suffrages[22] are prayers to the saints presented as models to follow for Christians. This section starts with an antiphon followed by a versicle and a response. Each antiphon may contain the name of a saint (Sebastianus, Dionysius, Anthonius, etc.). Each suffrage ends with a prayer (Oratio). The number of suffrages and the saints can vary widely. Also, Suffrages are not always illustrated.

3.1.9. Prayers

Additional Prayers, also called with the Latin term Oratio, are optional. Sometimes they are added at the end of the book of hours. They are usually preceded by Suffrages. Their number is arbitrary and can be whether short or long.

3.2. Second Level

In the second level of granularity, we find Gospel readings of the four apostles: John, Luke, Matthew and Mark. We also find the eight hours of the day, from Matins to Compline that are composed of psalms, canticles and hymns. The content of the Hours of the Virgin is divided in eight hours of the day and the beginnings are often illustrated in manuscripts with the miniatures with an iconographic cycle of the Infancy of Christ

- Matins: at day break (most common miniature: Annunciation)
- Lauds: at day break (Visitation)
- Prime: at 6 am (Nativity)
- Terce: at 9 am (Annunciation to the shepherds)
- Sext: at noon (Adoration of the Magi)
- None: at 3 pm (Presentation in the Temple)
- Vespers: at sunset (Flight to Egypt)
- Compline: before sleep (Coronation of the Virgin)

Other iconographic cycles are implemented, such as Virgin's life or to the Passion of the Christ between the Betrayal and the Entombment events. Not all the eight hours are used in the book of hours. For instance, the Hours of the Cross and the Holy Spirit do not include Lauds, while the Hours of the Dead contains only Vespers, Matins and Lauds. Finally, we separate the Penitential Psalms and Litany section of level one into two sections of level two that is: (i) Penitential Psalms and (ii) Litany. Hence, we come up with 14 categories of level two.

3.3. Third Level

The categories belonging to the third level represent cycles of prayers and include several characteristics such as: length, text nature (whether they come from the Bible or not), etc. Hereafter, we briefly give some definitions.

- Hymn: is a non biblical chant.
- Canticle: is a chant or song extract from the Bible.
- Capitule: is a short extract from the Bible that is usually present in all the hours except Matins.
- Oratio: is a non biblical text that is composed of an invocation to God and/or specific saints. It is also present in all the hours except Matins.
- Psalms: are songs of praise present in the Bible and of a number of 150.
- Nocturns: this category indicates the moment in which prayers are pronounced, i.e., at night.
- Lesson: is usually an excerpt of the Bible, used in the nocturns of Matins.
- Invitatory: appears at the beginning of the hours and is composed of one or two verses of psalms 50 and 69, and the following psalm (usually psalm 94) is sometimes also called "invitatory".

Hence, we come up with eight categories of level three. At a very fine-grained level, additional information can be considered as: antiphons, verses and versicles as well as responses. These labels are liturgical functions that we do not address as another level of annotation for segmentation. They constitute a functional information at the sentence level. We let this annotation and identification for future work.

- Antiphons: biblical chant extracts of one or two lines.
- Verse: can be a verse of a psalm or a verse of an hymn.
- Response: is a short meditation chant usually preceded by a verse.
- Versicle: is composed of a verse and a response.

4. Data Set

The book of hours data set is composed of the automatic transcription of 300 books and two manually transcribed books. Also, we provide annotations at three levels of granularity of the manual transcription books (Arsenal 1194 and Medievalist) as well as one for the automatically transcribed book (Harvard 253).

[21] Example of the office of the Dead: `https://gallica.bnf.fr/ark:/12148/btv1b550103864/f239`

[22] Suffrages example: `https://gallica.bnf.fr/ark:/12148/btv1b550103864/f275`

4.1. Transcription Data Format

Each book of hours is a JSON file including two main objects: ”page” and ”lines”. Both of them contain three attributes: an identification string (”id”), the text of the page (”text”) and a confident transcription score (”score”). The ”text” attribute of ”page” contains the entire text of the page while the ”text” attribute of ”lines” contains one line of the page. Figure 4 illustrates the JSON format.

```
page: {
    id:   ”328af124b”,
    text: ”spectu tui sacratissim me et coronia
          super in caput tuum impositum...”
    score: null
},
lines: [
{
    id:   ”31feb02b”,
    text: ”spectu tui sacratissim me”,
    score: 0.4256
},
{
    id:   ”4fcbf610”,
    text: ”et coronia super in”,
    score: 0.4875
{
    id:   ”55411fee”,
    text: ”caput tuum impositum...”,
    score: 0.452
}
...]
```

Figure 4: JSON representation of a transcription page of a book of hours.

4.2. Annotations

Due to the characteristics and complexity of books of hours (content, length, writing type...), manual annotation requires strong skills and expertise which considerably limits the number of candidates. Annotations were performed by two historians of medieval period with strong skills on books of hours structure and liturgy. Each book is represented as a CSV file where the first column is an identification line number, the second column is a line text of the book of hours and the three levels of annotation are included from column three to five.

Num	Line Text	Level1	Level2	Level3
248	benedicite nomini ...	Virgin	Matins	Nocturn
249	annuntiate inter gen...	Virgin	Matins	Nocturn
250	quoniam magnus...	Virgin	Matins	Nocturn
251	quoniam omnes dii...	Virgin	Matins	Nocturn
252	confessio et pulchrit...	Virgin	Matins	Nocturn
253	afferte domino patri...	Virgin	Matins	Nocturn

Figure 5: CSV representation of annotations of a book of hours (Extracts from the Arsenal 1194).

Figure 5 illustrates an example of the provided annotations extracted from Arsenal 1194. The ”Virgin” label refers to the Hours of the Virgin (level 1), the ”Matins” label refers to the Matins of level 2 and ”Nocturn” label, refers to Nocturn of level 3. For text segmentation, we provide the gold standard files with section delimitation following Choi (2000) format. Hence, each of the annotated books has one Choi format per level. This allows to evaluate the segmentation of each level separately or all together.

5. Text Segmentation

Text segmentation is closely related to topic analysis and can be addressed following three axis: (i) syntagmatic axis where a text is delimited into homogeneous topics (ex: audio transcriptions); (ii) paradigmatic axis in which topics are identified and (iii) functional axis in which segment topics relations are linked for text structuring (ex: summarization) (Ferret et al., 1998).

Segmentation can be content-based where each topic is characterized by a specific vocabulary and each vocabulary change implies a topic change (Hearst, 1994). It can also use topic markers whether (i) oral: such as prosody, silence; (ii) written: using connectors, introductory expressions or (iii) visual: using line breaks, bullets, numbering, bold, etc. The addressed texts were mainly linear, in which case a sequential analysis of topical changes was applied (Hearst, 1994; Choi, 2000) (expository texts were usually used). Later on, hierarchical texts were addressed which required a more fine-grained subtopic structure analysis (Yaari, 1997; Eisenstein, 2009).

Most of the approaches dedicated to text segmentation perform a lexical level analysis to detect segments coherence, and use (i) patterns of lexical co-occurrence (Hearst, 1997) such as discourse structure (Nomoto and Nitta, 1994) or (ii) lexical cohesion (Morris and Hirst, 1991) based on term repetition (Hearst, 1994) and semantic relations extracted via a thesaurus (Morris and Hirst, 1991), a dictionary (Kozima, 1993) or automatically using a collocation network (Ferret et al., 1998). Lexical cohesion has inspired a broad range of unsupervised approaches including TextTiling (Hearst, 1994) (a tfxIdf Cosine-based approach), LSeg based on lexical chains (Galley et al., 2003), U00 (Utiyama and Isahara, 2001) a probabilistic dynamic programming approach, TopicTiling (Riedl and Biemann, 2012) a topic modeling approach based on Latent Dirichlet Analysis (LDA), etc. Two main types of texts were addressed by lexical cohesion based approaches, that is: technical and scientific documents (Hearst, 1997) in which term repetition is a strong indicator when a specific vocabulary is used; and narrative texts (Morris and Hirst, 1991; Kozima, 1993) where term repetition is not sufficient as concepts may be expressed in different ways and so, thesaurus and dictionaries may be required to extract semantic relations between terms. Ferret et al. (1998), introduced a mixed approach to deal with both types of texts.

Also, a bunch of supervised approaches was introduced mainly to deal with discourse (Joty et al., 2015), multiparty dialogue and chat forums segmentation (Hsueh et al., 2006; Hernault et al., 2010) and to perform segmentation at the sentence level to discover Elementary Discourse Units

(EDU) (Hernault et al., 2010; Joty et al., 2015). These approaches often combine lexical coherence information with dialogue features using for instance a decision tree classifier (Hsueh et al., 2006), Conditional Random Fields (CRF) (Hernault et al., 2010; Joty et al., 2015) or Neural network approaches such as TextTiling-like embedding approach for query-reply dialog segmentation (Song et al., 2016), multi-party dialog for EDU using sequential model (Shi and Huang, 2019), reinforcement learning (Takanobu et al., 2018). Recently, Li et al. (2018) proposed SegBot a bidirectional RNN coupled with a pointer network that addresses both topic segmentation and EDU. The aforementioned approaches mainly deal with linear texts, one of the first proposed approaches that tackled hierarchical text segmentation was introduced in (Yaari, 1997) using a supervised agglomerative bottom-up clustering method. Paragraph hierarchy information was used as key information during the segmentation process. A pioneer unsupervised approach for hierarchical text segmentation was introduced in (Eisenstein, 2009) using a bayesian generative model with dynamic programming. Finally, to infer the logical structure of a text, visual forms information (such as title, paragraph, item) was also used as additional features for classifiers such as CRF (Fauconnier et al., 2014). As previously mentioned, the segmented texts are often scientific expository texts, narrative or issued from spoken an written dialogues. More recently, Arnold et al. (2019) released a multi-label annotation corpus of wikipedia and proposed a deep neural network approach that segments documents into coherent sections and assigns topic labels to each section. It learns a latent topic embedding over the course of a document by combining topical (latent semantic content) and structural information.

6. Experiments and Results

We evaluate several state of the art approaches: (i) five unsupervised approaches: TextTiling (Hearst, 1994), clustering model (C99) (Choi, 2000), probabilistic dynamic programming model (U00) (Utiyama and Isahara, 2001), minimum cut model (MinCut) (Malioutov and Barzilay, 2006), the hierarchical bayesian model (HierBays) (Eisenstein, 2009); and (ii) one supervised approach: TopicTiling (Riedl and Biemann, 2012), a topic modeling-based approach. Due to the lack of large annotated training data, we do not evaluate other supervised (Hsueh et al., 2006; Joty et al., 2015) and deep learning (Song et al., 2016; Koshorek et al., 2018; Shi and Huang, 2019) approaches. The experiments were conducted on the manual transcribed book Medievalist and on the automatically transcribed book Harvard 253. The Arsenal 1194 book of hours was used as training data for the TopicTiling approach.

6.1. Experimental Data

Table 3 resumes the number of segments according to the first and second levels of segmentation. At the first level, we note that the number of shifts[23] is equal except (Hardvard 253 that does not contain the Suffrages section). However, for the second level, we see different number of shifts. If a change in shifts corresponds to a topical change in traditional segmentation data sets, this is not the case for books of hours where some segments can be highly correlated.

	Number of categories	
BoH	#Level1	#Level2
Arsenal 1194	8	34
Medievalist	8	55
Harvard 253	7	38

Table 3: Illustration of the number of boundaries or shifts per level for each book of hours.

6.2. Evaluation Metrics

The approaches are evaluated in terms of P_k (Beeferman et al., 1999) and Windowdiff (WD) (Pevzner and Hearst, 2002) metrics. P_k is an error metric which combines precision and recall to estimates the relative contributions of the different feature types. Nonetheless, it exhibits several drawbacks as mentioned in (Pevzner and Hearst, 2002). P_k is affected by segment size variation. It also penalizes more heavily false negatives than false positives and overpenalizes near misses. Hence, a second measure, WindowDiff, a variant of P_k, has been also used as it equally penalizes false positives and near misses.

6.3. Results

Table 4 reports the segmentation results of the Medievalist and the Harvard 253 books of hours on the first and second levels[24].

	Medievalist				**Harvard 253**			
	Level1		**Level2**		**Level1**		**Level2**	
	P_k	WD	P_k	WD	P_k	WD	P_k	WD
TextTiling	64.4	94.2	47.3	52.5	65.7	91.3	49.7	51.9
C99	64.4	100	56.2	71.5	67.4	97.9	50.9	62.7
U00	37.7	37.7	34.1	**34.6**	26.2	**29.4**	**36.8**	**37.1**
MinCut	58.4	61.8	48.8	52.1	52.5	58.1	44.1	51.6
HierBays	**24.3**	**30.3**	**33.1**	35.1	**14.1**	35.3	37.1	39.7
TopicTiling	50.2	58.2	43.6	49.1	60.3	87.1	44.5	51.9

Table 4: Performance analysis of different segmentation algorithms using P_k and WindowDiff (WD) on first and second levels of two Books of Hours (Medievalist and Harvard 353). P_k and WD are penalties, so lower scores are better.

Firstly, TextTiling, C99 and MinCut obtained very weak results for both levels of granularity. Secondly, the unsupervised U00 and HierBays approaches showed the best results. Based on P_k score, HierBays obtained the best results for Medievalist (P_k = 24.3% for level 1 and and P_k =

[23] A shift corresponds to a frontier between two segments or categories

[24] We do not report the results for the third level for sake of clarity. However, they showed the same tendency.

33.1% for level 2) and for Harvard 253 at the first level (P_k = 14.1%). Also, equivalent results as U00 are obtained for the second level of Harvard 253 (P_k = 37.1% whereas U00 obtained a slightly better score with P_k = 36.8%). Based on the WD measure however, HierBays and U00 are more equivalent. Nonetheless, for both books, HierBays showed a segmentation error between 30.3% to 35.1% for level 1 and 35.1% to 39.7% for level 2, while U00 showed a segmentation error between 29.4% to 37.7% for level 1 and 34.6% to 37.1% for level 2. Finally, even with supervision, TopicTiling approach has also shown weak results. This may be explained by the lack of training data. Indeed, the recommended number of topics for TopicTiling training is 100 (Riedl and Biemann, 2012). As shown in Table 3, book of hours number of "topics" is around 8 for level 1 and around 35 for level 2.

Overall, the state of the art approaches obtained weak results (except HierBays on level 1 which obtained a P_k score of 14.1% for Harvard 253). This confirms their inappropriateness for books of hours segmentation. Moreover, even with good segmentation from unsupervised approaches, it is necessary to identify each section of the book. This information is not provided by any of the evaluated approaches. Finally, the obtained results are equivalent for manual and automatic transcriptions. This may suggest no impact of errors in transcription during the segmentation process.

7. Conclusion

In this paper we presented the first corpus of books of hours which is made up of 300 automatic and 2 manual transcriptions. We also proposed an annotation scheme of the complex structure of books of hours and released three annotated books (Arsenal 1194, Harvard 253 and Medievalist). Finally, we have conducted experiments on a variety of text segmentation state of the art approaches and have shown their inappropriateness with regard to this complex texts. We hope that this work will serve as baseline for future work and leads to the proposition of new text segmentation methods dedicated to books of hours.

Acknowledgments

This work is part of the HORAE project (Hours - Recognition, Analysis, Editions) and is supported by the French National Research Agency under grant ANR-17-CE38-0008.

8. Bibliographical References

Arnold, S., Schneider, R., Cudré-Mauroux, P., Gers, F. A., and Löser, A. (2019). Sector: A neural model for coherent topic segmentation and classification. *Transactions of the Association for Computational Linguistics*, 7:169–184.

Arora, A., Chang, C. C., Rekabdar, B., Povey, D., Etter, D., Raj, D., Hadian, H., Trmal, J., Garcia, P., Watanabe, S., Manohar, V., Shao, Y., and Khudanpur, S. (2019). Using ASR methods for OCR. In *International Conference of Document Analysis and Recognition*.

Beeferman, D., Berger, A., and Lafferty, J. (1999). Statistical models for text segmentation. *Mach. Learn.*, 34(1-3):177–210, February.

Boillet, M., Kermorvant, C., and Paquet, T. (2020). Multiple document datasets pre-training improves text line detection with deep neural networks. In *submitted*.

Choi, F. Y. Y. (2000). Advances in domain independent linear text segmentation. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, NAACL 2000, pages 26–33, Stroudsburg, PA, USA. Association for Computational Linguistics.

Clark, G. T. (2003). *The Spitz Master. A Parisian Book of Hours*. Getty Museum Studies on Art publications, Los Angeles.

De Hamel, C. F. R. (1992). *A History of illuminated manuscripts*. Phaidon Press, London.

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

Eisenstein, J. (2009). Hierarchical text segmentation from multi-scale lexical cohesion. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, May 31 - June 5, 2009, Boulder, Colorado, USA*, pages 353–361.

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

Ferret, O., Grau, B., and Masson, N. (1998). Thematic segmentation of texts: Two methods for two kinds of texts. In *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, pages 392–396, Stroudsburg, PA, USA. Association for Computational Linguistics.

Galley, M., McKeown, K., Fosler-Lussier, E., and Jing, H. (2003). Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 562–569.

Hearst, M. A. (1994). Multi-paragraph segmentation expository text. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 9–16, Las Cruces, New Mexico, USA, June. Association for Computational Linguistics.

Hearst, M. A. (1997). Texttiling: Segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.*, 23(1):33–64, March.

Heikkilä, T. and Roos, T. (2018). Quantitative methods for the analysis of medieval calendars. *Digital Scholarship in the Humanities*, 33(4):766–787, 05.

Hernault, H., Bollegala, D., and Ishizuka, M. (2010). A sequential model for discourse segmentation. In *Computational Linguistics and Intelligent Text Processing, 11th International Conference, CICLing 2010, Iasi, Romania, March 21-27, 2010. Proceedings*, pages 315–326.

Sandra Hindman et al., editors. (2013). *Books of hours reconsidered*. Brepols, Turnhout.

Hsueh, P.-y., Moore, J. D., and Renals, S. (2006). Automatic segmentation of multiparty dialogue. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.

Joty, S., Carenini, G., and Ng, R. T. (2015). Codra: A novel discriminative framework for rhetorical analysis. *Computational Linguistics*, 41(3):385–435.

Koshorek, O., Cohen, A., Mor, N., Rotman, M., and Berant, J. (2018). Text segmentation as a supervised learning task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 469–473, New Orleans, Louisiana, June. Association for Computational Linguistics.

Kozima, H. (1993). Text segmentation based on similarity between words. In *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics*, ACL '93, pages 286–288, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Li, J., Sun, A., and Joty, S. (2018). Segbot: A generic neural text segmentation model with pointer network. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence and the 23rd European Conference on Artificial Intelligence*, IJCAI-ECAI-2018, pages xx – xx, Stockholm, Sweden, July.

Malioutov, I. and Barzilay, R. (2006). Minimum cut model for spoken lecture segmentation. In *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*.

Morris, J. and Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Comput. Linguist.*, 17(1):21–48, March.

Nomoto, T. and Nitta, Y. (1994). A grammatico-statistical approach to discourse partitioning. In *Proceedings of the 15th Conference on Computational Linguistics - Volume 2*, COLING '94, pages 1145–1150, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pevzner, L. and Hearst, M. A. (2002). A critique and improvement of an evaluation metric for text segmentation. *Comput. Linguist.*, 28(1):19–36, March.

Riedl, M. and Biemann, C. (2012). Topictiling: A text segmentation algorithm based on lda. In *Proceedings of ACL 2012 Student Research Workshop*, pages 37–42. Association for Computational Linguistics.

Shi, Z. and Huang, M. (2019). A deep sequential model for discourse parsing on multi-party dialogues. In *AAAI*.

Song, Y., Mou, L., Yan, R., Yi, L., Zhu, Z., Hu, X., and Zhang, M. (2016). Dialogue session segmentation by embedding-enhanced texttiling. In *Interspeech*, pages 2706–2710, 09.

Takanobu, R., Huang, M., Zhao, Z., Li, F., Chen, H., Zhu, X., and Nie, L. (2018). A weakly supervised method for topic segmentation and labeling in goal-oriented dialogues via reinforcement learning. In *IJCAI-ECAI*, pages 4403–4410.

Utiyama, M. and Isahara, H. (2001). A statistical model for domain-independent text segmentation. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, pages 499–506.

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

Wieck, R. (2001). *Time Sanctified: The Book of Hours in Medieval Art and Life*. G. Braziller.

Yaari, Y. (1997). Segmentation of expository texts by hierarchical agglomerative clustering. *CoRR*, cmp-lg/9709015.