# Apport du Text Mining pour l'exploration de relations dans les textes. Application à la découverte d'appariements entre objets d'intérêt et localisations dans la Tapisserie de Bayeux

David Condaminet, Antoine Widlöcher, Pierre-Yves Buard, Bruno Crémilleux, Julia Roger

# Apport du Text Mining pour l'exploration de relations dans les textes. Application à la découverte d'appariements entre objets d'intérêt et localisations dans la Tapisserie de Bayeux

David Condaminet*, Antoine Widlöcher*, Pierre-Yves Buard**
Bruno Crémilleux*, Julia Roger**

*Normandie Univ., UNICAEN, ENSICAEN, CNRS – UMR GREYC, Caen, France
prenom.nom@unicaen.fr
**MRSH, UNICAEN, CNRS – USR 3486, Caen, France
prenom.nom@unicaen.fr

**Résumé.** Nous présentons dans ce travail une approche fondée sur l'utilisation de méthodes de *Text Mining* pour l'exploration de relations dans des textes issus des Humanités, et plus précisément la découverte d'appariements entre descriptions textuelles d'objets d'intérêt (personnages, lieux, événements) et localisations dans la *Tapisserie de Bayeux*. Cette approche repose sur une phase d'amorçage permettant d'obtenir un premier appariement minimal mais fiable entre des objets d'intérêt et leurs localisations, le corpus contenant de tels appariements devant ensuite être exploré par une étape de fouille de données textuelles focalisée sur les mécanismes de mise en relation. Nous présentons ici notre démarche globale et détaillons le processus d'amorçage, qui repose notamment sur la mise en place d'un environnement d'observation mettant l'humain au centre du processus d'analyse et de contrôle.

## 1 Introduction

Les textes procèdent par nature massivement à des mises en relation, sur différents plans (syntaxique, sémantique...), à différentes échelles (propositions, phrases, discours, texte...) et selon différents paradigmes (relations de dépendance, causales, rhétoriques, temporelles, coréférentielles ...). L'accès aux informations véhiculées par ces relations constitue un enjeu majeur pour l'interprétation des données textuelles, et en particulier pour des tâches d'extraction d'information où il ne s'agit pas seulement d'identifier dans les textes des éléments de sens isolés (par exemple de entités nommées, des personnes, des lieux...), mais des dispositifs au sein desquels différents éléments interagissent (présence d'une personne en un certain lieu, à un certain moment... par exemple).

Dans la perspective d'exploiter automatiquement ces relations portées par les textes, la description manuelle des configurations linguistiques correspondantes, en vue de leur projection sur corpus, s'avère souvent si coûteuse qu'on souhaite au moins partiellement l'automatiser, et il n'est donc pas surprenant que différents travaux proposent des méthodes permettant l'extraction automatique des relations dans les textes, parmi lesquels on peut citer par exemple Mé-

tivier et al. (2015) dans le domaine de la recherche des relations entre gènes et maladies ou encore Tikk et al. (2013) et Cellier et al. (2015) pour la recherche d'interactions entre protéines ou entre gènes. Ces travaux sont fondés sur l'exploitation de méthodes de fouille de données ou d'apprentissage automatique.

Mais l'automatisation de l'extraction des règles suppose la disponibilité de données d'apprentissage en quantité suffisante, données au sein desquelles les objets liés sont pré-annotés. Faute d'en disposer, le recours à des ressources externes, telles que des terminologies ou ontologies de domaine porteuses de relations connues entre formes connues, peut permettre de sortir de l'impasse, par projection sur corpus de ces connaissances disponibles, en vue d'apprendre ensuite les modalités de leur mise en relation textuelle, modalités qui pourront ensuite être utilisées pour découvrir des configurations où s'articulent de manière similaire, au sein de relations inconnues, des formes connues ou elles-mêmes inconnues.

Dans le domaine des Humanités qui nous occupe ici, les mises en relation portées par les textes constituent aussi évidemment un enjeu d'importance. L'exploitation de méthodes automatisées du type de celles indiquées ci-dessus se heurte néanmoins à une double difficulté : (a) les données annotées sont encore globalement rares ; (b) les ressources externes projetables sur corpus pour procéder à une annotation automatique ne le sont pas moins.

Il s'avère donc nécessaire d'imaginer des méthodes pouvant être amorcées sur des données peu ou pas enrichies. C'est l'objet du travail que nous présentons ici.

Avant d'aborder la présentation de cette contribution, il convient de s'arrêter sur une difficulté inhérente aux phénomènes de mise en relation, difficulté qui n'est pas propre aux relations rencontrées dans les textes issus des Humanités, difficulté déterminante pour le choix de la méthode et suffisamment contre-intuitive pour être présentée ici, dès l'introduction. Une première approche des phénomènes de mise en relation pourrait en effet laisser penser qu'il s'agit là d'un phénomène pouvant être abordé de façon purement compositionnelle et ascendante, en considérant que l'analyse des relations suppose, dans cet ordre strict (1) la découverte des éléments isolés et (2) leur mise en relation. Il convient au contraire d'insister sur le fait que l'articulation entre éléments et relations s'avère en réalité beaucoup plus dialectique, l'élément isolé ne pouvant devenir significatif que parce qu'il entre dans une relation donnée. Si l'on vise par exemple les relations causales entre deux faits, il serait vain de penser que la cause ou la conséquence existent isolément et préalablement à cette relation.

Notre travail vise l'application de méthodes de *Text Mining* pour l'exploration de relations dans des textes issus des Humanités, et plus précisément la découverte d'appariements entre descriptions textuelles d'objets d'intérêt (personnages, lieux, événements) et localisations dans (c'est-à-dire positionnement sur) la Tapisserie de Bayeux, appariements décrits par des textes.

La méthode que nous présentons ci-après possède les propriétés suivantes, que nous souhaiterions mettre en emphase :

— elle est le fruit d'une étroite collaboration avec des chercheurs en SHS ;
— elle n'est pas limitée dans son principe au domaine spécifique dont elle relève ici ;
— elle met l'humain au centre du processus d'analyse et de contrôle ;
— elle permet de compenser partiellement la faible disponibilité des données enrichies.

La section 2 présente le contexte et les objectifs globaux de notre travail. La section 3 présente de manière détaillée la phase d'amorçage devant aboutir à la constitution d'un corpus d'apprentissage porteur d'appariements fiables entre localisations et objets d'intérêt. Nous y mettons notamment en évidence la nécessité d'interactions avec l'opérateur humain et présen-

tons, à la section 3.4, l'interface d'observation combinée des connaissances déjà acquises et des données textuelles.

# 2 Contexte et présentation schématique de notre méthode

Nous précisons dans cette section le contexte et les objectifs de notre travail. Nous commençons par donner quelques définitions.

## 2.1 Définitions

**La Tapisserie** est considérée, par simplification, comme une suite d'événements graphiquement décrits ;

**Un objet d'intérêt** (parfois noté OI ci-après) désigne la description textuelle d'un épisode, d'un personnage, d'un lieu ou de toute autre entité ou réalité prenant part à un événement décrit par la Tapisserie ;

**Une scène** est l'un des 58 segments conventionnellement délimités sur la Tapisserie ;

**Une figure** désigne une illustration de la Tapisserie présente dans notre corpus textuel et faisant référence, par sa légende, à une ou plusieurs scènes de la Tapisserie ;

**Une localisation** désigne la mention faite par le texte d'une séquence de la Tapisserie, celle-ci pouvant désigner une scène ou une figure – cette dernière étant indirectement et implicitement liée à une ou plusieurs scènes.

## 2.2 Contexte : événements et localisations dans la Tapisserie

Le travail que nous présentons ici s'inscrit dans un projet plus vaste visant à repenser un système documentaire donnant accès à l'abondante littérature relative à la Tapisserie de Bayeux. Dans ce cadre, notre travail doit plus particulièrement contribuer à l'effort de mise en relation des documents ou des parties de documents avec les zones de la Tapisserie qu'ils décrivent ou qu'ils commentent.

Notre corpus de travail est constitué des actes d'un colloque international (Bouet et al. (2004)) consacré à la Tapisserie de Bayeux, dont les contributions sont supposées représentatives des mécanismes par lesquels les chercheurs font référence aux différentes parties de l'œuvre brodée. Les événements rapportés par la Tapisserie peuvent ainsi faire l'objet, dans ce corpus, de descriptions textuelles désignant des objets d'intérêt (« l'envoi du messager », « l'arrivée de Guillaume en Angleterre », « la mort d'Edouard »...), ainsi que d'une localisation sur la Tapisserie (par exemple par l'indication d'un numéro de scène). L'ensemble documentaire est donc porteur d'une collection de références aux parties de l'ouvrage pouvant soit décrire textuellement les objets d'intérêt, soit donner des localisations, soit, dans certains cas, fournir à la fois la description textuelle et la localisation. Ainsi, on trouve par exemple : « Conan s'enfuit de Dol (scène 18) » (p. 191), « La scène 10 montre la rencontre des envoyés de Guillaume » (p. 191), « La scène 32, celle de la comète » (p. 193).

D'une manière générale, il est utile de disposer d'un appariement entre localisation et descriptions textuelles, pour que des localisations puissent être automatiquement rapportées

aux objets d'intérêt qui les concernent et que, inversement, les descriptions textuelles d'objets d'intérêt puissent être localisées sur la Tapisserie. Les avantages qui en résulteraient en matière d'aide à la navigation sont assez évidents : depuis la description textuelle d'un objet d'intérêt, on pourrait par exemple automatiquement rediriger le lecteur vers la ou les zones correspondantes sur la Tapisserie. Mais il deviendrait dès lors également possible d'enrichir significativement les modes d'indexation des contenus associés à la Tapisserie. Par exemple, les documents privilégiant le référencement par localisation pourraient bénéficier d'une indexation s'appuyant sur les descriptions textuelles issues du *mapping*, pour autoriser ainsi notamment une recherche par mots-clés, là où les seules localisations sont pourtant explicitement indiquées dans le texte.

## 2.3 Objectifs

L'objectif du travail présenté ici est la mise en place d'une solution permettant l'extraction semi-automatisée de ces appariements entre descriptions textuelles et localisations. Plus précisément, il s'agit d'extraire le *mapping* à partir de passages où coïncident une localisation et une description textuelle, passages pouvant ressembler aux exemples présentés ci-dessus. Pour y parvenir, il convient de localiser dans le corpus les passages porteurs de cette double identification. Pour cela, l'idée directrice de notre méthode est d'exploiter des techniques de fouille de texte pour la découverte de patrons linguistiques caractéristiques de la double identification, motifs généralisant des énoncés de mise en relation. La section suivante présente la méthode d'ensemble que nous proposons, méthode qui repose a) sur la construction d'un premier appariement minimal mais fiable (phase qui sera détaillée ensuite dans la section 3) puis b) son enrichissement grâce à la fouille de données séquentielles.

## 2.4 Présentation schématique de la méthode

La méthode proposée repose sur les deux phases complémentaires présentées ci-après, chacune pouvant être répétée itérativement.

### 2.4.1 Phase d'amorçage

L'objectif de la phase d'amorçage (cf. figure 1), phase qui pourra faire l'objet de plusieurs itérations, est d'obtenir un premier appariement minimal mais fiable, appariement associant des représentations textuelles d'OI et des localisations. La méthode retenue consiste à :

1. adopter un formalisme permettant l'expression de règles décrivant des patrons linguistiques susceptibles de capter un appariement entre description et localisation. Ce formalisme doit : a) être exploitable pour la définition manuelle de règles ; b) être utilisable lors des étapes ultérieures du projet pour encoder les motifs appris lors du processus de fouille (cf. phase suivante) ;
2. utiliser ce formalisme pour décrire manuellement les patrons correspondant à certains motifs triviaux immédiatement repérables (notamment dans les légendes des figures) ;
3. projeter ces patrons sur corpus pour extraire l'information utile des passages correspondants ;
4. capitaliser les informations d'appariement fiables ainsi extraites.

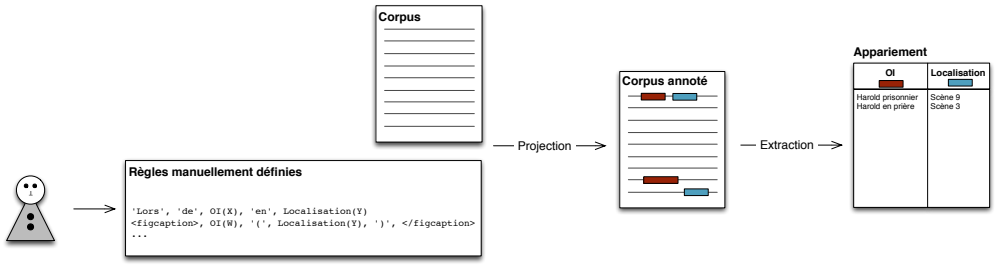


FIG. 1 – *Processus d'amorçage*

### 2.4.2 Phase d'enrichissement par exploration des relations

La phase d'enrichissement (cf. figure 2), qui pourra elle aussi faire l'objet de plusieurs itérations, consiste à exploiter l'appariement fiable pour isoler des énoncés où apparaissent OI et localisations que l'on sait fortement corrélés, pour découvrir automatiquement dans ces énoncés des structures textuelles caractéristiques de la mise en relation entre OI et localisations. La méthode retenue repose sur l'enchaînement suivant :

1. on dispose d'un corpus enrichi au sein duquel sont déjà annotées des informations relatives aux OI et aux localisations que l'on sait fortement corrélés ;
2. on sélectionne des énoncés comportant un OI et une localisation ;
3. on applique sur les passages sélectionnés une méthode de fouille séquentielle, pour apprendre des règles de mise en relation ;
4. ces règles sont reformulées selon le formalisme exploité à la phase 1 ;
5. ces règles reformulées sont projetées sur le reste du corpus, conformément au schéma présenté en figure 1, en vue d'enrichir nos connaissances sur l'appariement.

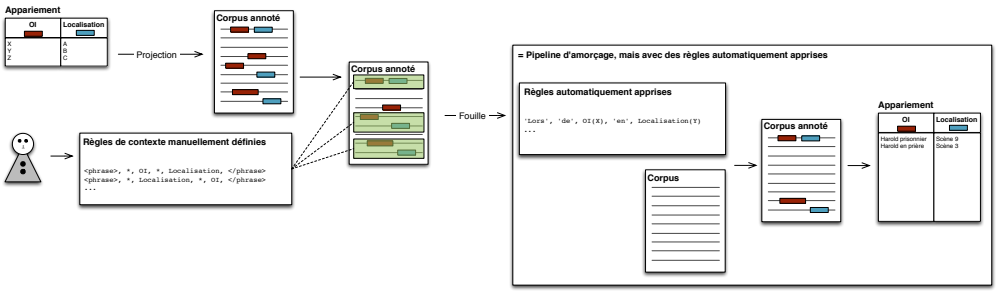


FIG. 2 – *Processus d'enrichissement*

Notons que l'application des règles apprises doit permettre non seulement de localiser de nouveaux appariements, mais aussi de déterminer les contours des OI (et dans une moindre mesure des localisations), ce qui fait écho au principe dialectique évoqué ci-dessus : un OI ne devient tel ici que dans la mesure où il intervient dans une relation. Intuitivement, l'OI est ce dont on parle et ce que l'on situe.

### 2.5 Avancement actuel de nos travaux

À l'heure où nous écrivons ces lignes, la phase 1, qui sera présentée en détail dans la section suivante, est toujours en cours. On le verra, l'amorçage, loin d'être réalisable naïvement, suppose la mise en place d'une interaction avec l'utilisateur, en vue d'obtenir un premier appariement suffisamment large pour permettre la constitution d'un corpus d'apprentissage suffisant pour la phase 2.

## 3 Extraction automatisée d'appariements entre descriptions textuelles et localisations

La première phase de notre travail consiste à mettre en place une méthode d'extraction semi-automatisée d'appariements fiables entre descriptions textuelles d'objets d'intérêt et leurs localisations à partir d'un corpus de textes d'étude, structuré au format XML, mais sans annotation préalable des localisations et objets d'intérêt. Ces appariements serviront de base pour la seconde phase du travail, phase présentée sommairement dans la section précédente mais qui ne sera pas détaillée dans cet article.

### 3.1 Initialisation de l'amorçage

Une lecture rapide du corpus permet d'identifier des énoncés d'appariement évidents dans les légendes de figures, ainsi que dans les titres de sections. Par exemple, on trouve dans Bouet et al. (2004) la légende "Fig. 3 : Tapisserie de Bayeux : Harold quitte l'Angleterre (scène 4)" (p. 200) ou le titre "La fuite de Conan (scène 18)" (p. 190).

On remarque également que, en présence d'illustrations, le numéro de figure est souvent utilisé dans le texte, pour mettre en correspondance indirectement les OI et les scènes décrites par les figures, sans mentionner explicitement les scènes. Il s'agit donc là également d'appariements fiables dont la détection est triviale et nous avons choisi d'utiliser les numéros de figure comme marqueurs de localisation : un énoncé mettant en relation un OI et une figure sera jugé aussi pertinent pour la suite qu'un énoncé établissant une relation entre un OI est un numéro de scène.

À partir de règles très simples exprimées avec des expressions régulières, nous avons extrait dans cette première phase d'amorçage 117 relations réputées certaines, i.e. 117 triplets distincts de type (*OI*, *scène*, *figure*) renvoyant globalement à 56 scènes – sur les 58 existantes –, 76 figures et 82 objets d'intérêt différents.

Sur la base de ces premiers appariements réputés fiables, une première projection sur le reste du corpus doit permettre de sélectionner un premier ensemble de phrases (ensemble devant être *in fine* utilisé comme corpus d'apprentissage pour la phase 2 d'exploration des relations) porteuses à la fois d'une localisation et d'un objet d'intérêt. Pour clarifier ce point, 3 ensembles de contextes doivent être distingués :

— $CS$ : phrases contenant au moins une expression de *scène* ;
— $CF$ : phrases contenant au moins une expression de *figure* ;
— $COI$ : phrases contenant au moins une expression d'un *objet d'intérêt*.

Les contextes dits de localisation correspondent à l'ensemble $(CS \cup CF)$. D'une manière générale, la sélection d'énoncés devant constituer le corpus d'apprentissage $\mathcal{C}$ sera un sous-ensemble de $C = (CS \cup CF) \cap COI$.

Si nous limitons notre corpus d'apprentissage aux phrases (ensembles de mots, $S$ ci-après) présentant une localisation ($l$ ci-après) et un OI ($o$ ci-après) ayant un lien fiable selon l'appariement $\mathcal{M}$ obtenu ci-dessus, le corpus d'apprentissage correspond alors à l'ensemble de phrases suivant :

$$\mathcal{C} = \{S \mid S \in C, \exists o \in S, \exists l \in S, (o, l) \in \mathcal{M}\}$$

Ainsi définie, la sélection d'énoncés devant constituer le corpus d'apprentissage pour la phase ultérieure serait de taille dramatiquement réduite. En effet, sur les données issues de Bouet et al. (2004) : $|\mathcal{C}| = 23$.

Notons que, même en relaxant la contrainte sur la fiabilité de l'appariement, le volume d'énoncés n'augmente pas de manière très importante, ce qui renvoie au fait que les OI et les localisations issues de cette toute première phase sont simplement en nombre trop réduit. En effet : $|C| = 56$.

## 3.2 Enrichissement de l'amorçage

Il est dès lors évidemment nécessaire de procéder à un enrichissement des données d'amorçage. Pour cela, il est notamment indispensable de pouvoir :

— contrôler l'état actuel des connaissances acquises, c'est-à-dire les appariements fiables ;
— prévisualiser l'état actuel du corpus $\mathcal{C}$ en cours de constitution ;
— identifier des configurations proches de celles qui sont actuellement sélectionnées et qui mériteraient une formalisation explicite dès la phase d'amorçage, en vertu de leur fiabilité.

De cette triple exigence ont résulté les étapes suivantes de notre travail, étapes ayant principalement consisté jusqu'ici à :

1. mettre en place un environnement permettant à l'opérateur informaticien et/ou humaniste une navigation efficace dans les connaissances acquises et les contextes d'acquisition (voir section 3.4) ;
2. proposer différentes méthodes d'élargissement des données visualisables via cette interface, pour permettre à l'opérateur d'identifier les configurations non encore prises en charge mais devant être traitées en priorité.

## 3.3 Élargissement des données observables

L'élargissement des données observables, qui vise à placer sous le regard de l'expert des configurations probablement remarquables, en vue d'améliorer l'amorçage, est réalisé selon deux axes complémentaires :

1. à un niveau local, un élargissement de la présentation des OI ;
2. à un niveau global, un élargissement de la présentation des contextes de mise en relation entre OI et localisation.

### 3.3.1 Élargissement au niveau des OI

Ce premier élargissement, au niveau local des OI, vise à permettre à l'opérateur de clarifier la nature des OI. Pour cela, la démarche retenue consiste à analyser les OI déjà sélectionnés, en quête de régularités exprimées par des motifs séquentiels, puis de projeter ces motifs sur le reste du corpus, pour voir s'ils correspondent à une formalisation acceptable de la notion d'OI.

Pour cela, nous nous appuyons sur :

— la bibliothèque Python *Natural Language Toolkit* (NLTK) Bird et al. (2009) pour la segmentation en phrases puis la projection des motifs ;

— *TreeTagger* de Schmid (1994) [1] pour la tokenisation, l'étiquetage morpho-syntaxique et la lemmatisation ;

— l'outil SDMC (*Sequential Data Mining under Constraints*) Béchet et al. (2015) pour la fouille de séquences.

Chacune des descriptions textuelles d'OI extraites précédemment est ici une transaction. Pour chaque transaction, chacun de ses mots, représenté par son étiquette, est un item. Nous extrayons alors les motifs séquentiels fréquents avec SDMC sur les OI préalablement étiquetés avec comme paramètres pour SDMC : $mingap = maxgap = 0$, $minsize = 3$, $maxsize = 5$ et $minsup = 10$. Après examen des motifs extraits, nous pouvons retenir 4 motifs, qui correspondent en fait, simplement, à des syntagmes nominaux :

```
{NOM_COMMUN} {PREPOSITION} {NOM_COMMUN}
{NOM_COMMUN} {PREPOSITION} {NOM_PROPRE}
{PREPOSITION} {DETERMINANT} {NOM_COMMUN}
{DETERMINANT} {NOM_COMMUN} {PREPOSITION} {NOM_PROPRE}
```

Nous pouvions évidemment largement subodorer ce fait, à la simple lecture des descriptions des OI. Toutefois, nous souhaitions mettre en place un *pipeline* réutilisable ensuite à chaque itération, en y incluant également la seconde phase non encore abordée (celle de l'exploration des relations) : à chaque nouvel élargissement de l'appariement, des nouveaux OI seront retenus, dont nous souhaitons pouvoir identifier les éventuelles structures communes, en permettant toujours à l'opérateur d'élargir aussi son horizon, en quête de nouvelles régularités significatives pouvant donner lieu à l'établissement manuel de règles jugées essentielles et fiables.

La projection des motifs sur le reste du corpus est réalisée à l'aide du module *RegexpParser chunker* de NLTK, dont l'expressivité est en adéquation avec notre représentation des données textuelles. Auparavant, l'ensemble du corpus est tokenisé à l'aide du TreeTagger et segmenté en phrases à l'aide du module *PunktSentenceTokenizer* de NLTK.

### 3.3.2 Élargissement au niveau des contextes de mise en relation entre OI et localisation

L'élargissement au niveau des contextes de mise en relation repose lui-même sur deux mesures complémentaires :

1. la relaxation de la contrainte d'appartenance du couple $(o, l)$ à l'appariement fiable $\mathcal{M}$, avec l'hypothèse, acceptable pour les besoins de l'observation, que la co-présence d'un OI et d'une localisation dans le même contexte (ici la phrase) résulte souvent d'une mise en relation. Cette relaxation revient, pour reprendre la notation utilisée ci-dessus, à passer de $\mathcal{C}$ à $C$ ;

1. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

2. la relaxation de la contrainte d'appartenance des OI à l'ensemble $\{o \mid \exists l, (o,l) \in \mathcal{M}\}$, c'est-à-dire l'élargissement des contextes à toute combinaison d'une localisation d'une part et, soit d'un OI, soit d'une instance quelconque de l'un des motifs extraits de la fouille d'autre part.

### 3.3.3 Effets du double élargissement

Sur la base de ce double élargissement, nous pouvons présenter à l'opérateur un ensemble de 151 contextes porteurs d'OI et de localisations, valeur à rapporter d'une part à $|\mathcal{C}| = 23$ et d'autre part à $|C| = 56$. Notons que, sans l'assimilation des mentions de figures à des localisations (assimilation scène/figure évoquée ci-dessus), le nombre de contextes observables tomberait de 151 à 77.

Reste alors à observer lesdits contextes, et à y repérer des structures fiables pouvant permettre d'enrichir le jeu de règles exploitables pour cette phase d'amorçage. Pour cela, nous plaçons entre les mains de l'opérateur une interface de navigation qui sera l'objet de la prochaine section.

## 3.4 Interface d'exploration des données

L'interface que nous présentons ici est un outil destiné à la fois au spécialiste de l'analyse des données, en charge du processus d'extraction automatisé, et au spécialiste de ces données spécifiques elles-mêmes, acteur des SHS.

Cette interface doit tout d'abord permettre l'exploration des données textuelles pour la découverte d'appariements fiables et la construction d'un corpus d'apprentissage utilisable lors de la phase d'analyse des relations. Elle sera également utile lors de cette seconde phase pour contrôler la progression de l'évolution des connaissances acquises.

Cet outil, dont le fonctionnement accorde une place privilégiée à l'utilisateur (cf. section 3.4.2), permet de guider le travail de construction incrémentale des connaissances en mettant l'humain au centre du processus d'analyse et de contrôle (cf. section 3.4.1).

### 3.4.1 Intérêt et rôle de l'interface

Cette interface web, dont la figure 3 fournit une première illustration, permet au spécialiste en sciences des données de *contrôler* l'évolution du processus d'extraction de connaissances, d'avoir une meilleure visualisation des contextes textuels dont les connaissances sont extraites et de prévisualiser le corpus d'apprentissage qui pourra résulter de l'exploitation des connaissances acquises. De surcroît, elle permet de confronter ce que le dispositif actuel capte déjà, et ce qu'il devrait idéalement capter, conformément aux mesures d'enrichissement présentées dans les sections précédentes. En conséquence de ces mesures, il pourra en effet observer des configurations proches de celles qui ont été déjà validées, y découvrant de nouvelles formes de mise en relation entre localisation et objets d'intérêt, dont certaines, suffisamment régulières et fiables, mériteront d'être explicitement formalisées et ajoutées au dispositif d'amorçage.

Pour un usager moins impliqué dans le processus d'extraction, mais spécialiste ou simple usager du corpus, par exemple documentaliste ou historien, cette interface offre un environnement efficace pour naviguer à la fois sur la *Tapisserie de Bayeux* et dans un corpus qui lui est dédié, au travers de différents critères de sélection : les scènes, les figures de l'ouvrage, ou

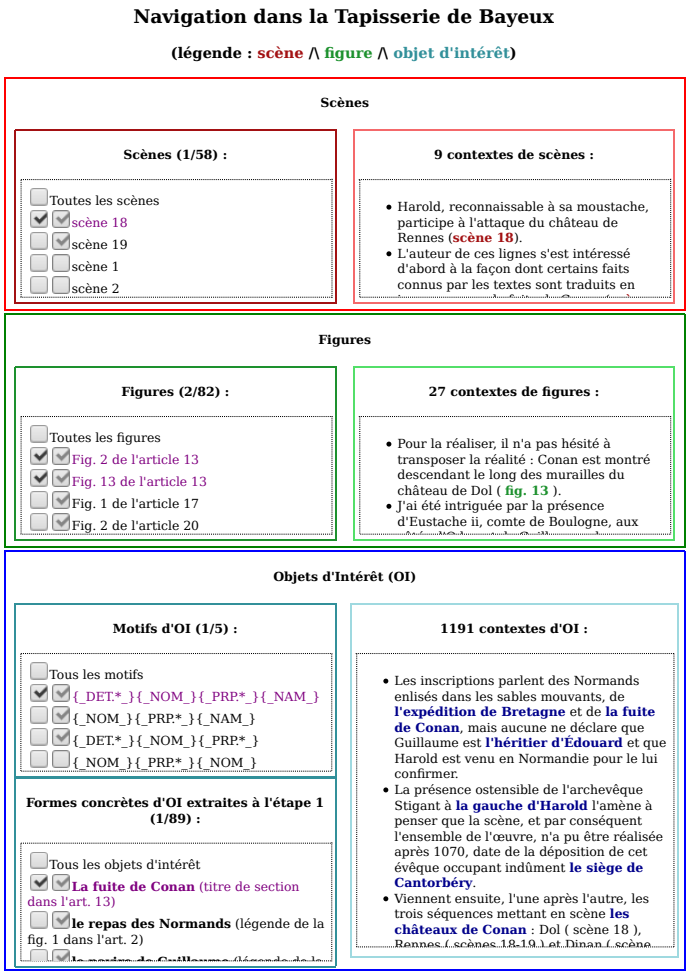


FIG. 3 – *Interface d'observation*

encore les objets d'intérêt, ces derniers pouvant être distingués par le degré de leur fiabilité, selon qu'ils sont extraits d'appariements jugés certains, ou de généralisations résultant de la projection de motifs.

### 3.4.2 Usage de l'interface

L'utilisateur est à l'initiative sur la construction des différents ensembles de contextes : $CS$, $CF$ et $COI$ (placés respectivement de haut en bas en colonne de droite), selon ce qu'il choisit de sélectionner respectivement comme scènes, figures, motifs et OI (de haut en bas en colonne de gauche). L'usage des quatre listes de sélection d'items se traduit par la mise à

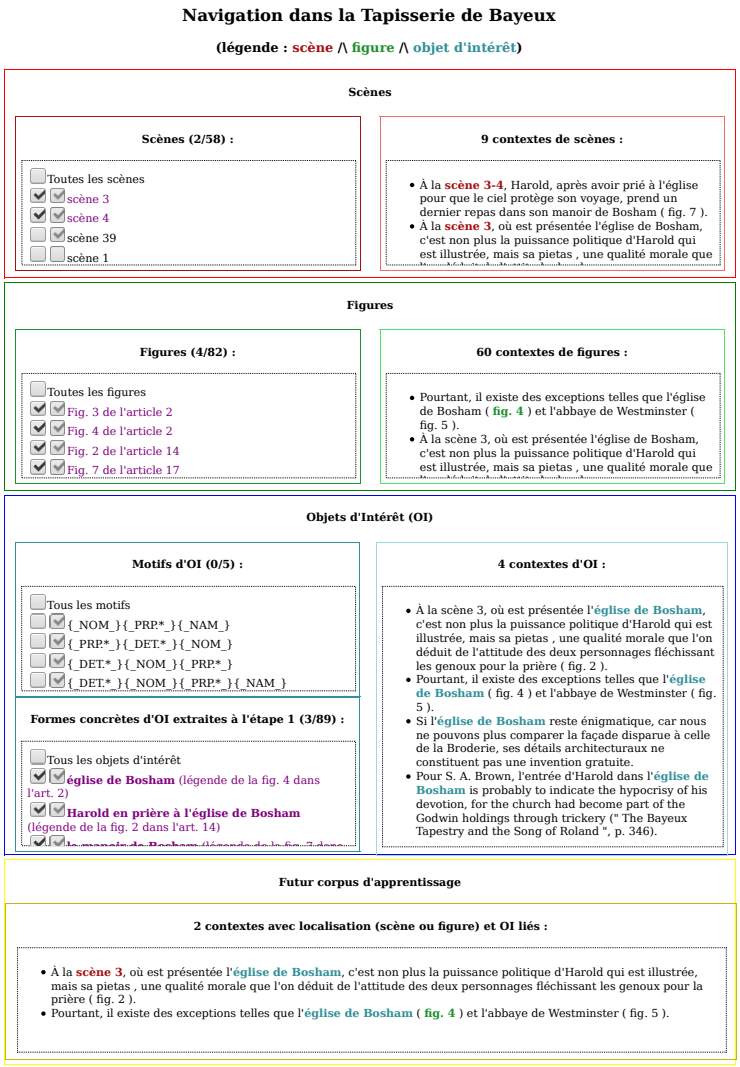


FIG. 4 – *Interface d'observation*

jour des contextes correspondants (à droite), et par la représentation conséquente du corpus d'apprentissage en cours de construction (en bas de la figure 4).

Lorsqu'il sélectionne un item d'une des catégories exposées ci-dessus, les items des 3 autres catégories étant liés à cet item, d'après l'état actuel des connaissances, sont pré-sélectionnés (via colonne des *checkboxes* située à droite). L'utilisateur peut alors, s'il le souhaite, poursuivre ce chemin de propagation des connaissances, en sélectionnant les items adéquats (via colonne des *checkboxes* située à gauche).

Le corpus d'apprentissage, pour sa part, est automatiquement déduit des ensembles de

contextes configurés par l'utilisateur, conformément aux mesures présentées en section 3.1. L'utilisateur contrôle ainsi, indirectement, la construction du corpus d'apprentissage.

La figure 4 illustre l'état de l'interface pour un cas d'utilisation lors duquel l'utilisateur recherche des événements qui ont eu lieu dans la ville de Bosham. On peut remarquer, au bas de cette figure, la construction résultante du corpus d'apprentissage $\mathcal{C}$.

# 4 Conclusion et perspectives

Dans cet article, nous avons présenté une démarche globale fondée sur l'utilisation de méthodes de fouille de données textuelles pour l'exploration de relations dans des textes issus des Humanités. Nous avons détaillé la première étape consistant à produire un premier appariement minimal mais fiable entre des objets d'intérêt et leurs localisations. Cette étape place l'humain au centre du processus, celui-ci intervenant via une interface de navigation et d'exploration des données. La poursuite de ce travail consistera désormais en l'exploitation de méthodes de fouille de données pour apprendre des règles de mise en relation.

De façon plus générale, la recherche de relations entre entités, par exemple entre personnes et localisations, est un sujet d'étude intéressant dans de nombreux autres corpus issus des Humanités. Nous projetons ainsi par exemple d'utiliser ce type de démarche pour l'étude du corpus des *Minutes du Procès Nuremberg* dont la MRSH de Caen a récemment mis une version numérisée à disposition de la communauté.

# Références

Béchet, N., P. Cellier, T. Charnois, et B. Crémilleux (2015). Sequence mining under multiple constraints. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing, Salamanca, Spain, April 13-17, 2015*, pp. 908–914.

Bird, S., E. Loper, et E. Klein (2009). *Natural Language Processing with Python*. O'Reilly Media Inc.

Bouet, P., B. Lévy, et F. Neveux (2004). *La Tapisserie de Bayeux : l'art de broder l'Histoire*. Office universitaire d'études normandes (ouen), Presses Universitaires de Caen.

Cellier, P., T. Charnois, M. Plantevit, C. Rigotti, B. Crémilleux, O. Gandrillon, J. Kléma, et J. Manguin (2015). Sequential pattern mining for discovering gene interactions and their contextual information from biomedical texts. *J. Biomedical Semantics 6*(27).

Métivier, J.-P., L. Serrano, T. Charnois, B. Cuissart, et A. Widlöcher (2015). Automatic Symptom Extraction from Texts to Enhance Knowledge Discovery on Rare Diseases. In J. H. Holmes, R. Bellazzi, L. Sacchi, et N. Peek (Eds.), *Proceeding of the 15th Conference on Artificial Intelligence in Medicine, AIME 2015*, Pavia, Italy, pp. 249–254.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Tikk, D., I. Solt, P. E. Thomas, et U. Leser (2013). A detailed error analysis of 13 kernel methods for protein-protein interaction extraction. *BMC Bioinformatics 14*(12).