Interprétation automatique d'itinéraires à partir d'un corpus de récits de voyages pilotée par un usage pédagogique.

Pierre Loustau*, Mauro Gaio*, Thierry Nodenot**

Laboratoire d'Informatique de l'Université de Pau et des Pays de l'Adour
http://liuppa.univ-pau.fr

*Université de Pau et des Pays de l'Adour
Avenue de l'Université, B.P. 1155
64013 Pau Cedex, France
prenom.nom@univ-pau.fr,
**IUT de Bayonne, Pays Basque
2 Allée du Parc Montaury
64600 Anglet, France
prenom.nom@iutbayonne.univ-pau.fr

Résumé. De larges corpus à fort ancrage territorial deviennent disponibles sous forme numérique dans les médiathèques et plus particulièrement dans les médiathèques de dimension régionale. Les défis qu'offrent ces gigas octets de documents bruts sont énormes en terme de traitement automatique des contenus. Nous proposons dans cet article deux modèles computationnels et une méthode complète permettant de réaliser un traitement automatique afin d'extraire des itinéraires dans des textes relatant des récits de voyage. Le premier modèle est un modèle des attendus. Il s'intéresse au concept d'itinéraire et adopte le point de vue du pédagogue et fait intervenir très tôt les usages envisagés. Le deuxième modèle est un modèle d'extraction, il permet de modéliser l'expression du déplacement dans des textes du genre *récit de voyage*. Nous proposons alors une méthode automatique pour : d'une part extraire et interpréter automatiquement les déplacements d'un récit et d'autre part passer des déplacements à l'itinéraire, c'est-à-dire alimenter de manière automatique le modèle des attendus à partir du modèle d'extraction. Nous montrons également comment les itinéraires extraits interviennent soit dans la phase de construction d'activités pédagogiques soit directement comme matériau dans une activité d'apprentissage. Nous présentons enfin $\pi\mathcal{R}$, un Prototype pour l'Interprétation d'Itinéraires dans des Récits de voyages, qui implémente notre approche. Il prend en entrée un texte brut et fournit l'interprétation de l'itinéraire décrit dans le texte. Il permet également de visualiser sur un fond cartographique l'itinéraire extrait.

1 Introduction

L'ensemble des propositions présentées dans cet article est élaboré dans le cadre du projet $\pi\mathcal{R}$[1], un des projets clefs d'une action de recherche appliquée initiée depuis trois ans [2].

Du point de vue approche adoptée, elle est résolument descendante (top-down). Par l'étude et la compréhension des usages cibles, elle débouche sur la formalisation du fonctionnement de l'information nécessaire à cet usage. Cette information étant contenue dans un corpus documentaire source. La formalisation obtenue permet alors de concevoir un modèle de l'information attendue, puis d'opérationaliser des outils de repérage et de représentations symboliques afin d'alimenter de manière automatique ce modèle.

Comme annoncé dans Casenave et al. (2004) et formalisé dans Nodenot et al. (2006) et Loustau et al. (2008), de plus en plus d'activités pédagogiques sont construites autour des documents patrimoniaux. C'est notamment le cas en géographie : localisation des principaux lieux constitutifs d'un itinéraire de voyage, positionnement sur une carte des lieux visités, lecture d'un itinéraire à différentes échelles, etc. L'enjeu ici est de se servir des documents analysés automatiquement pour ne concevoir que les activités que l'on va pouvoir encadrer de manière automatisée (ou assistée) par un tuteur informatique. Des expérimentations avec des enseignants et des apprenants sont menées depuis 2 ans pour définir les contextes d'utilisation opérationnels qui ont du sens pour de vrais enseignants.

Concernant le traitement automatique, la réponse à l'usage cible retenu nécessite une phase de marquage de l'information mettant en œuvre des méthodes s'apparentant à de l'extraction d'information (IE) et une phase de Recherche d'Information Géographique (GIR), concept proposé en 2004 et précisé en 2005 par Jones et Purves (2006) les co-fondateurs du premier workshop de ce nom. Selon le niveau de la représentation symbolique souhaité par la suite, l'interprétation à réaliser dans la phase de GIR peut s'appuyer sur une granularité plus ou moins locale de l'information géographique. Compte tenu de l'approche adoptée, cette granularité variera du niveau syntagmatique[3] (souvent considéré comme atomique pour l'information géographique) à des agglomérats de ce premier niveau, de plus en plus importants (portés par une proposition, une phrase, un paragraphe ou plus) et ceci selon les besoins exprimés par l'usage.

Dès que l'on veut obtenir une représentation symbolique d'une information, les traitements en GIR doivent se baser sur des méthodes de calcul du sens. Dans ce cas il faut alors se questionner sur le type de représentation sémantique le plus adapté pour capturer automatiquement le sens de telle ou telle granularité textuelle. Comme le précise Blackburn et Bos (2003), il n'y a actuellement pas de réponse unique à ce type de questionnement : cela dépend essentiellement du niveau de finesse attendu et du type de « phénomène »linguistique auquel on s'intéresse.

[1] $\pi\mathcal{R}$: Prototype pour l'Interprétation d'Itinéraires dans des Récits de voyage

[2] Cette action de recherche est portée par diverses formes de collaborations réalisées entre chercheurs du LIUPPA (équipe-projet DeSI), chercheurs de laboratoires tels que le LRI, le COGIT (IGN), L'IRIT (grâce au projet ANR-GEONTO) et différents partenaires d'organisations privées ou publiques : les entreprises DIS et Géocime, la médiathèque de Pau (MIDR), la communauté d'agglo. (CDAPP), le conseil général (CG64). Cette action a comme objectif de créer un ensemble de ressources et de traitements afin de raisonner, dans le cadre d'usages ciblés, sur de l'information textuelle à contenu géographique, après l'avoir recherchée, marquée et annotée.

[3] c'est-à-dire composé de syntagme(s). Le syntagme est un groupe de termes dont la succession a un sens qui forme une unité fonctionnelle.

Dans cette contribution, nous proposons un ensemble d'outils pour l'extraction et l'interprétation automatiques d'informations géographiques d'origine textuelle afin de permettre la conception d'activités pédagogiques.

Compte tenu de cette approche descendante, les attendus didactiques concernant les descriptions d'itinéraires composent l'usage cible et l'information nécessaire à cet usage est constituée par la description d'itinéraires contenue dans des documents patrimoniaux de la catégorie *récits de voyages*. Les itinéraires sont une forme d'information géographique d'une granularité résolument supérieure à la forme atomique de l'information géographique (du niveau du syntagme). Nous proposons deux modèles ayant pour origine d'un part les attendus et d'autre part l'observation du corpus documentaire. Ces deux modèles doivent être computationnels afin de permettre une mise en application informatique.

Dans la section suivante, après avoir mentionné des projets qui présentent des similarités, soit dans leur approche, soit dans les objectifs fixés, nous nous intéressons aux travaux sur lesquels nous nous sommes appuyés concernant le concept d'itinéraire du point de vue des sciences cognitives. Nous montrons ensuite les différentes approches concernant l'évocation de l'itinéraire dans la langue. La section 3 constitue notre proposition : nous y décrivons le modèle des attendus et ses origines (les attendus des pédagogues) ainsi que le modèle d'extraction et ses origines (l'observation d'un échantillon du corpus). Dans la section 4, nous montrons comment procéder pour instancier automatiquement le modèle d'extraction puis, toujours de manière automatique, quelles ressources, quels calculs sont nécessaires pour créer des instances du modèle des attendus. La section 5 montre l'état actuel des implémentations réalisées par la description de $\pi\mathcal{R}$, un prototype qui prend en entrée un texte du genre *récit de voyage* et qui donne en sortie une instance du modèle des attendus (visualisable par des outils de cartographie). Nous montrons en section 6 des exemples de requêtes qu'un pédagogue peut être amené à faire sur ces itinéraires extraits de manière automatique.

2 État de l'art

Différentes actions de recherche ont été menées autour de la problématique de l'itinéraire, elles ont donné naissance à des plateformes comme nous allons le constater au travers des projets comme Egges et al. (2001), Coyne et Sproat (2001) ou encore Maaßet al. (1993). Certains de ces projets s'intéressent comme nous au lien entre expression textuelle en langue naturelle et concepts géographiques liés aux itinéraires.

De nombreux travaux ont d'autre part été effectués autour de la problématique de modélisation du concept d'itinéraire du point de vue cognitif. Le modèle des attendus que nous proposerons, bien qu'émanant des attentes pour un usage pédagogique, s'inspire de ces différents travaux.

Enfin, dans le domaine du TALN et chez les linguistes, divers travaux vont nous permettre de proposer un modèle d'extraction (c'est-à-dire de repérage pour une interprétation) afin de rendre possible un traitement automatique. Les uns concentrent leurs efforts sur la caractérisation d'expressions dont la granularité est inférieure à la proposition. Les autres s'intéressent à la manière dont l'itinéraire est exprimé dans la langue (orale ou écrite) et enfin d'autres encore focalisent leurs recherches sur les verbes de déplacements. Ces verbes occupent en effet une place prépondérante dans une description d'itinéraire.

2.1 Des projets autour de l'itinéraire

Près de nos préoccupations, à savoir établir un lien entre le langage écrit et un concept géographique, de nombreux auteurs se sont intéressés à des systèmes permettant de passer automatiquement d'un texte à une scène (*text-to-scene conversion*). Coyne et Sproat (2001) avec WordsEye en est un exemple : ce système permet de visualiser en trois dimensions des descriptions du type : *The cat is on the large chair. A dog is facing the chair.* WordsEye utilise une base lexicale de plus de 10000 objets (et leur représentation associée), l'ontologie Wordnet, un analyseur de dépendances et des grammaires de cas pour interpréter les textes soumis. Cependant, WordsEye ne semble pas prendre en compte des textes relatant des histoires réelles et les exemples donnés ne sont que des cas d'école relativement courts dans lesquels le système cherche à positionner correctement les objets les uns par rapport aux autres de manière statique.

Carsim (Egges et al. (2001)) est un projet qui cherche à offrir une visualisation en 3D de descriptions d'accidents de la route à partir de constats textuels. L'approche est quelque peu semblable à celle que nous proposons : les auteurs cherchent à décrire formellement l'accident à partir d'une analyse linguistique puis, dans un deuxième temps à générer la scène virtuelle associée.

Maaßet al. (1993) avec le projet VITRA (VIsual TRAnslator) s'intéresse à la connexion entre langage et perception visuelle. L'élaboration d'un système de représentation des connaissances qui permette un accès en langage naturel aux données visuelles est un objectif de ce projet. VITRA vise plus particulièrement la génération automatique de descriptions d'itinéraires. Nous sommes là en présence de travaux qui adoptent une démarche exactement inverse à la nôtre : à partir d'un modèle de données de l'itinéraire, on cherche à produire une description d'itinéraire. Dans ce projet, l'accent est mis sur l'apport de la bi-modalité dans la description de l'itinéraire (description verbale et description graphique).

Les spécificités de notre approche par rapport à ces projets sont triples. Tout d'abord dans notre projet nous travaillons sur des documents d'un contenu bien plus volumineux que celui des constats d'accidents de la route ou des descriptions du type *the cat is on a large chair*. Cela a pour conséquence non seulement un nombre de données spatio-temporelles à manipuler plus important mais également des échelles variables dans les références spatiales. De plus, les formes d'expression de l'espace et du temps sont différentes : si dans les constats d'accidents les objets, leur trajectoire, les chocs sont importants, dans les documents patrimoniaux à composante géographique, les lieux, les époques, et les déplacements le sont plus. D'autre part, nous souhaitons intégrer dans notre approche les usages qui seront faits des informations extraites : dans le cas d'étude que nous proposons, celui de l'extraction des itinéraires, nous souhaitons aller plus loin que la visualisation des informations extraites. L'objectif est en effet à terme de fournir des textes interprétés à des applications à vocation pédagogique.

2.2 L'itinéraire du point de vue conceptuel et sciences cognitives

La représentation mentale de l'espace, la psychologie de l'espace sont des préoccupations relativement anciennes. Dès le début du XXème siècle, et même avant avec quelques idées de Goethe, Köhler (1929) s'intéresse à cette problématique dans la théorie de la Gestalt – ou théorie des formes – et établit entre autre que *le tout est perçu avant les parties le formant.* Transposé à notre problème, l'itinéraire global est perçu avant les éléments qui le composent.

Dans les années quarante, des théories concernant l'espace et son modèle cognitif font leur apparition. Les travaux de Tolman (1948) sur la notion de *carte cognitive* chez l'animal et les recherches de Piaget qui s'est intéressé en 1948 au développement de l'apprentissage de l'espace chez les enfants en sont les principaux exemples. B. Kuipers est également à l'origine de travaux largement cités dans ce domaine. Kuipers (1977) propose le modèle TOUR dans lequel il décrit les fonctions de la *carte cognitive* qui sont d'assimiler les informations concernant l'environnement, de représenter la position actuelle, et de répondre à des questions de localisation et de recherche d'itinéraires. Le modèle TOUR est un modèle computationnel psychologique du sens commun spatial pour la description d'itinéraire.

Wunderlich et Reinelt (1982) s'intéressent, dans un dialogue, au processus de la description d'un chemin répondant à la question *How to get there from here ?* Ils décomposent ce processus en quatre phases : initiation (question initale, confirmation, reconfirmation), description de la route (par l'informateur), confirmation (confirmation et répétition éventuelle) et fermeture (remerciements, etc.). La phase 2 nous intéresse ici plus particulièrement puisqu'il s'agit de décrire un itinéraire. Dans ces travaux, les auteurs décomposent l'itinéraire ainsi : un point de départ, des destinations intermédiaires et un point d'arrivée. Ces destinations intermédiaires sont des *landmarks*, un *landmark* a des caractéristiques particulières qui le rendent facilement reconnaissable. Les auteurs introduisent également les *extended landmarks*. Au contraire des *landmarks* classiques qui s'apparentent à des points (un monument, un carrefour, etc.), les *extended landmarks* sont des portions de route que l'on doit suivre (une route, un tunnel, le bord d'un cours d'eau, etc.).

Denis (1994) puis Przytula-Machrouh et al. (2004) s'intéressent aux connaissances utilisées par les personnes décrivant un itinéraire, aux modes de représentation de ces connaissances et à leur utilisation. Ces travaux sont à cheval entre le concept et son évocation de manière verbale. Concernant le concept d'itinéraire, les auteurs se placent dans un contexte de description *a priori* d'un itinéraire en ville, dans le but d'apporter une information, de la même manière que Wunderlich et Reinelt (1982). Dans ces travaux, le concept de *scène élémentaire* est utilisé comme unité de base de la description d'itinéraire. Une scène élémentaire est fondamentalement constituée de *repères* et d'*actions*. Les auteurs se sont également intéressés au moyen de verbaliser ces descriptions d'itinéraire et actent que les repères sont représentés par des noms ou des groupes nominaux alors que les actions sont verbalisées grâce à des verbes de déplacement.

Fraczak et Lapalme (1999) se placent également dans le contexte de la description d'itinéraire *a priori*, ils décrivent la structure conceptuelle de l'itinéraire comme une succession d'entités spatio-temporelles appelées *segments* et *relais*. Le *segment* est un fragment de l'itinéraire durant lequel une ou plusieurs caractéristique(s) reste(nt) constante(s) tandis qu'un *relais* marque un changement de caractéristique(s). Ces caractéristiques pouvant être *une orientation*, *une direction*, *un type de chemin*, etc.

Si la modélisation de l'itinéraire a donné lieu à de nombreux travaux, ces travaux se placent en amont de la réalisation de l'itinéraire. Il s'agit d'une description *a priori* qui a un objectif bien particulier : conduire celui qui va réaliser l'itinéraire d'un point de départ à un point d'arrivée. Le référentiel de la description est donc celui qui effectue l'itinéraire. Dans nos travaux, l'itinéraire considéré est un itinéraire *a posteriori*. Le narrateur le décrit une fois qu'il a été réalisé non pas pour qu'un autre le réalise à son tour mais dans un but narratif, pour que la majorité des personnes puisse comprendre son voyage. Cela implique une description parfois

moins fine de repères visuels d'orientation et l'utilisation d'un référentiel *intrinsèque*. Autrement dit un référentiel peu dépendant de l'objet de référence et faisant appel à des relations ternaires impliquant un observateur, l'objet à situer et l'objet de référence. Ce référentiel est le seul qui puisse être à la fois partagé par celui qui a fait l'itinéraire et celui qui le lit. En effet, le lecteur n'est pas physiquement sur les lieux racontés, et, sauf exception, il n'a pas non plus en tête la configuration exacte des lieux traversés et racontés dans le récit.

D'autre part, nous pensons qu'un modèle conceptuel ne peut s'élaborer de manière purement objective : il adopte forcément un point de vue. Nous verrons (cf section 3.1) que notre point de vue est celui du pédagogue qui souhaite faire intervenir ces descriptions d'itinéraires dans ses activités pédagogiques.

Enfin, si l'itinéraire global est certes perçu avant les éléments qui le composent, il n'en est pas moins indissociable. En effet, dans un texte, il n'y a pas d'itinéraire sans l'évocation des emplacements successifs et des déplacements du narrateur. C'est le sujet de la section suivante.

2.3 L'itinéraire : son évocation dans la langue par le déplacement

Dans le développement de la méthode que nous proposons pour interpréter les itinéraires, nous nous appuyons sur les travaux de linguistes et du TALN. Ces travaux montrent que le déplacement est primordial dans la compréhension de l'itinéraire, ceci sera confirmé par des observations faites sur un échantillon de notre corpus. Nous donnons ici un aperçu des principaux travaux linguistiques qui ont été menés autour de l'expression du déplacement dans la langue.

Boons (1987) constate que les critères pour déterminer les compléments locatifs de la phrase sont insuffisants. Le plus souvent, la question *où ?* que l'on nous apprend à nous poser à l'école primaire pour trouver le complément locatif ne suffit pas. Boons propose alors de classer les verbes de mouvement selon la phase spatio-temporelle sur laquelle ils sont focalisés. Les verbes de déplacement peuvent donc avoir une *polarité aspectuelle initiale* (comme pour le verbe *sortir*), *médiane* (comme pour le verbe *passer*) et *finale* (comme pour le verbe *arriver*).

Laur (1991) a fait une étude complète et détaillée des combinaisons des prépositions spatiales avec les verbes de mouvement en français. Elle reprend la *polarité aspectuelle* définie par Boons pour les verbes mais l'étend quelque peu. Les prépositions de lieu sont classées selon plusieurs critères. Le critère positionnel (à, dans, sur, en face de, etc.) ou le critère directionnel qui indique un déplacement (vers, depuis, jusqu'à, etc.). Dans le cas des prépositions présentant un critère directionnel, les prépositions possèdent en plus un *trait aspectuel* (comme pour les verbes de Boons) : par exemple *depuis* est initial, *par* est médian et *jusqu'à* est final. Pour Laur, les verbes de déplacement (VDP) expriment le passage d'un lieu ou d'une entité à un autre alors que les verbes de mouvement font état d'un changement de position ou d'état. Les VDP constituent un sous-ensemble des verbes de mouvement. La cible (C) est l'objet qui se déplace ou est déplacé. Enfin, le lieu de référence verbal (LRV) est le lieu auquel fait référence le verbe de déplacement par la sémantique qu'il transporte. Le lieu site (S) est le lieu indiqué dans la phrase par le syntagme prépositionnel. Laur propose alors de classer les verbes de déplacement selon trois critères :

- la polarité aspectuelle (cf Boons) ;
- la relation de localisation que le verbe implique. Cette relation est interne (lorsque le verbe décrit une inclusion ou un contact de la cible C par rapport au site S comme *quitter,*

entrer, passer) ou externe (lorsqu'il y a séparation entre C et S comme dans *s'écarter, dépasser, contourner*) ;
– le lien du verbe avec son LRV. Laur distingue les verbes qui décrivent un changement de LRV *entrer, sortir, etc.* des verbes qui décrivent simplement un déplacement dans l'espace sans impliquer un changement entre deux états (*marcher, graviter, s'éloigner, etc.)*.

Plus tard, Sablayrolles (1995) reprend ces travaux et introduit le changement d'emplacement. Le troisième critère de Laur montre une large sous-spécification du déplacement. En effet, s'il n'y a effectivement pas de changement de LRV dans la phrase *Paul s'approche du mur*, il y a tout de même un changement d'état. *Paul* est effectivement plus près de la cible *mur* au temps t+1 qu'au temps t. Les lieux sont donc insuffisants pour décrire le déplacement dans la langue. C'est pour palier à ce problème que Sablayrolles introduit les emplacements, en plus des lieux. A la différence d'un lieu, un emplacement est une portion de surface, sans aucune fonctionalité ni élément lexical associé. Il est uniquement défini géométriquement par l'enveloppe pragmatique associée à l'entité concernée. De notre point de vue, l'approche de Sablayrolles nous semble être relativement semblable à celle de Laur, si ce n'est qu'il introduit un grain plus fin que le lieu : l'emplacement.

Sarda (2000) s'est penchée sur le cas particulier des verbes de déplacement transitifs directs et en propose une typologie. Elle cherche à raffiner la catégorie des verbes médians qui selon elle est plutôt définie par défaut, c'est-à-dire rassemblant les verbes qui ne sont ni initiaux, ni finaux. Sarda propose de catégoriser les verbes en fonction de la nature des relations de localisation qu'ils impliquent. Elle distingue en premier lieu les verbes relationnels (qui dénotent un déplacement quel que soit l'objet) des verbes référentiels (qui dénotent un déplacement seulement lorsque l'objet est un lieu). Les verbes relationnels sont catégorisés selon des relations de distance (*approcher, fuir, etc.*), d'orientation (*monter, descendre, etc.*) ou de passage (*traverser, sauter, etc.*) et les verbes référentiels sont soit neutres initiaux (*quitter, déserter, etc.*) soit neutres finaux (*atteindre, regagner, etc.*) soit de contact (*heurter, taper, etc.*). Une classe de verbes moins clairement définie subsiste : celle des verbes médians (*arpenter, sillonner, parcourir, etc.*) qui sont référentiels par rapport à un domaine topologique (un intérieur par exemple).

La représentation de la sémantique des verbes de déplacement a donné lieu a de très nombreux travaux, notamment chez les linguistes et dans la communauté TALN. Elle pose encore de nombreux problèmes et est sans cesse remise en question comme dans la thèse de Mathet (2000). Cependant, le type de documents sur lesquels nous travaillons (des récits de voyage) ainsi que les objectifs que nous souhaitons atteindre quant à la finesse de l'interprétation du déplacement dans la langue nous permettent de simplifier le problème. Le sous-ensemble des verbes de déplacement auxquels nous nous intéressons ici sont ceux qui entrent dans une construction *verbe,préposition (facultative),entité_spatiale*, que nous noterons triplet (V,P ?,E). Cette construction lève une grande partie des problèmes d'ambiguïté que l'on peut trouver dans des propositions comme *quitter son mari, traverser une mauvaise période, etc.*. De plus, nous souhaitons simplement inférer qu'à un moment donné, le sujet est localisé sur l'objet entité spatiale. Les préoccupations plus fines sont pour le moment écartées de nos attendus. En effet, dans un but de traitement automatique de corpus volumineux à des fins d'interprétation pour le domaine pédagogique, une première approximation du déplacement par la polarité aspectuelle des verbes de déplacement nous semble suffisante. Pour prendre un exemple, dans la compré-

hension globale de l'itinéraire décrit dans un récit, que le narrateur *s'éloigne de Pau* ou *quitte Pau* est relativement semblable.

3 Modélisation du concept d'itinéraire et son expression dans des récits de voyage : deux modèles computationnels

Les enseignants proposent de nombreuses activités pédagogiques à partir de documents à contenu géographique. Les origines de ces documents peuvent être multiples : des extraits de publications universitaires, des sélections obtenues dans des publications plus grand public (par exemple des articles de quotidiens, d'hebdomadaires, de guides touristiques, ...) ou encore des passages choisis dans des œuvres littéraires. Pour l'ensemble de ces « documents sources » se posent des problèmes pédagogiques importants : l'adéquation du matériau aux objectifs d'apprentissage visés. En effet, souvent l'exploitation directe du contenu de ce type de document reste une tâche difficile pour les élèves (du secondaire en particulier). Il s'ensuit que le pédagogue, auteur de manuel ou professeur, adapte le document afin de le rendre plus accessible aux élèves. Il s'agit très fréquemment d'une adaptation dans le sens de la simplification, d'où la mention fréquente dans les manuels : « d'après tel auteur ».

Pour les travaux présentés ici, nous avons retenu comme activité cible : *la mobilisation chez l'élève des repères spatio-chronologiques*. Elle est une composante incontournable de plusieurs objectifs pédagogiques. Étant données les diverses formes qu'une telle activité peut revêtir et afin de modéliser un processus de complexité maîtrisable, nous nous sommes restreints à l'explicitation de trois tâches :

1. localisation des principaux lieux constitutifs d'un itinéraire,
2. interprétation des déplacements plausibles entre les différents lieux empruntés ou traversés au sein l'itinéraire,
3. cartographie du parcours interprété.

Ces tâches devront être appliquées à une sous catégorie de « documents sources » des documents de type œuvre littéraire et plus spécifiquement du genre littéraire *récit de voyage*.

Dans la démarche du pédagogue de mise en adéquation de la source documentaire aux finalités pédagogiques, il lui est nécessaire de passer par trois grandes étapes :

1. une étape de recherche d'information et d'extraction de passages qui peut lui retourner plusieurs documents,
2. une étape de comparaison du matériau retourné en étape 1 qui lui fait choisir un document,
3. enfin une étape de reconstruction du « document source pédagogique », à partir du document choisi en étape 2.

Le corpus de « documents sources » sur lequel nous travaillons est constitué de monographies datant d'une même époque (fin du XIXème siècle) et évoquant des récits de voyage réalisés sur un même territoire : les Pyrénées.

Remarquons que les textes du corpus sont spécifiques par rapport aux nombreux travaux du domaine de la description d'itinéraire car leur but n'est pas de décrire uniquement un itinéraire. L'objectif premier des auteurs de ces documents est de relater l'expérience personnelle de l'auteur. Néanmoins, on peut observer que le genre descriptif est d'avantage utilisé que le genre

narratif et que la description de l'itinéraire y occupe une place prépondérante. On peut constater, d'autre part, qu'il s'agit dans notre cas d'une description *a posteriori* alors que la plupart des travaux existants qui traitent de la description d'itinéraire se placent *a priori*. Le récit de voyage ne contient donc pas des énoncés tels que ceux qui apparaissent dans les descriptions *a priori* comme *continuer tout droit, tourner à droite, etc.*. On peut remarquer également une certaine régularité dans l'évocation du territoire et des déplacements au sein de l'itinéraire, comme évoqué par Boons (1987), Laur (1991), Mathet (2000). Il est le plus souvent évoqué par des verbes de déplacement (*j'ai quitté Bordeaux à 8h00, je suis arrivé à Pau à 11h, j'ai gravi le Pic du Midi d'Ossau le lendemain*, etc.), et ce quel que soit l'auteur. Enfin, le récit de voyage obéit à des règles particulières concernant la chronologie des évènements relatés : il y a dans ce genre de document une synchronisation quasi parfaite entre la chronologie du texte et la chronologie du voyage.

Compte tenu de l'état actuel de la formalisation des connaissances dans le domaine de la compréhension automatique de texte concernant les concepts spatiaux, temporels et spatio-temporels, il n'est pas encore possible de concevoir un système permettant d'assurer de manière autonome les trois grandes étapes de préparation d'un tel matériau pédagogique. Par contre, comme la suite de ce papier tente de le montrer, il est possible de concevoir un outil accompagnant le pédagogue dans ces tâches.

Dans un premier temps, nous présenterons une formalisation de l'information attendue, puis après avoir discuté les résultats d'une étude réalisée sur un échantillon du corpus, nous détaillerons le modèle d'extraction. Sera enfin relatée la phase d'interprétation, autrement dit la transformation de l'information exprimée dans le modèle d'extraction pour sa ré-expression dans le modèle des attendus.

3.1 Le modèle des attendus

Dans le cas particulier de l'exemple de récit de voyage donné en figure 1, les éléments constituant une description d'itinéraire sont :

1. des lieux et des « temps calendaires »[4],
2. des déplacements,
3. des faits (activités ou situations) qui modifient la dynamique par défaut du déplacement.

Les lieux sont mentionnés lorsque l'auteur les a simplement traversés ou qu'il y situe des faits marquants. Tous les lieux traversés lors d'un itinéraire ne sont pas relatés ce qui nous rappelle les propriétés de saillance de Przytula-Machrouh et al. (2004); Denis (1994); Wunderlich et Reinelt (1982).

D'autre part, seuls les lieux associés à une description de faits (arrêt pour visiter, pour manger, pour changer de moyen de transport, etc.) peuvent éventuellement donner naissance à un itinéraire d'échelle inférieure. Nous les apparentons à la notion d'étape intermédiaire de Fraczak et Lapalme (1999). Nous proposons également de reprendre les notions de *relais* et de *segments* (figure 2) telles que décrites par Fraczak et Lapalme (1999)

Pour donner une représentation semi-formelle du modèle obtenu, nous avons utilisé le langage MADS de Parent et al. (2006). MADS est un modèle de données qui permet d'ajouter des

[4]Nous entendons pour temps calendaires toutes expressions faisant référence explicitement à un jour du calendrier ou à une portion de celui-ci.

5 juillet. Dimanche. Je me suis baigné. Je suis parti à 11h1/4 avec un groupe de gens pour Rochefort dans une pinasse. Vent contraire. En louvoyant, en ramant et avec l'aide de la marée, nous arrivâmes à Rochefort vers trois heures; cependant la journée fut belle et la compagnie plutôt agréable. J'ai vu toutes les parties du phare; la base fut construite au temps de Louis XIV, la partie supérieure en 1789. Le dispositif lenticulaire actuel sert depuis douze ans. Le gardien se plaint seulement de la lampe qui lui cause des ennuis. Nous avons quitté Rochefort vers 5 heures. Presque calme; nous avons mis deux heures pour arriver à Royan. J'y ai passé la nuit.

6 juillet. Je suis parti par le bateau à vapeur à 6 heures du matin et j'ai atteint Bordeaux à midi après une brève traversée. Après–midi pluvieuse. J'ai dîné avec M. Guestierl. J'ai écrit à Elisa.

7 juillet. J'ai quitté Bordeaux à 7 heures en diligence pour Pau. J'ai été agréablement surpris par la beauté de la campagne [...] Nous avons traversé Langon.

FIG. 1 – *Extrait d'un récit de voyage.*

types de données propres aux données spatio-temporelles mais aussi d'ajouter des relations spécifiques entre ces données. MADS se veut également un modèle de données résolument tourné vers une modélisation des concepts lors de la phase de modélisation. Les concepteurs doivent en effet s'abstraire totalement des contraintes d'implémentation pour être le plus près possible du monde réel et de ses représentations. Cela n'était pas le cas dans la gestion des données géographiques où les concepteurs ont longtemps été influencés par des préoccupations d'implémentation internes aux outils qui gèrent ces données.

Avant de décrire un à un les objets qui composent le modèle que nous proposons, il est bon de préciser que les quatre principaux objets sont tous des Entités Géographiques (EG). De manière abstraite une EG est définie par trois composantes : une spatiale, une temporelle et une thématique ou phénomène. Cette définition explicitée dans Usery (2003) est largement admise dans de nombreux autres travaux. De ce fait on admettra qu'une EG est la composition de trois sous-entités : une Entité Spatiale (ES), une Entité Temporelle (ET) et une entité phénomène (EP). Ces différentes facettes de l'information géographique sont traduites dans MADS par les propriétés spatiales et temporelles que nous positionons sur les objets.

Itinéraire : il s'agit d'un objet complexe de haut niveau qui correspond à l'itinéraire raconté par le narrateur. Il est associé à deux étapes (de départ et d'arrivée) et est composé de segments, eux-même associés à des relais. De ce fait, il a un attribut spatial de type MADS *Complex Geometry* ( ) et un attribut temporel de type MADS *Interval* ( ). Le lien de composition est une association de type composant/composé ( ).

Le relais : il s'agit d'une portion de l'espace géographique investi par l'itinéraire à un moment donné. Le relais n'a pas de contenu, il s'agit simplement d'un espace traversé par le narrateur. Il a donc un attribut spatial de type MADS *Point* ( ). Il n'est cependant pas suffisant d'en considérer uniquement l'aspect spatial. Du point de vue temporel, c'est la date à laquelle le narrateur a traversé le relais qui nous intéresse. De ce fait, la temporalité de l'objet Relais est de type MADS *Instant* ( ). L'attribut thématique est implicite, il s'agit du fait que le lieu ait été investi par le narrateur. Ces trois notions combinées constituent bien le concept d'EG.

L'étape : On différencie une étape d'un relais lorsqu'un fait y est relaté comme par exemple une activité. De ce fait, l'attribut temporel de l'objet *Etape* est du type MADS *Interval* ( ) et l'attribut spatial du type MADS *Surface* ( ). Ces étapes permettent le changement d'échelle

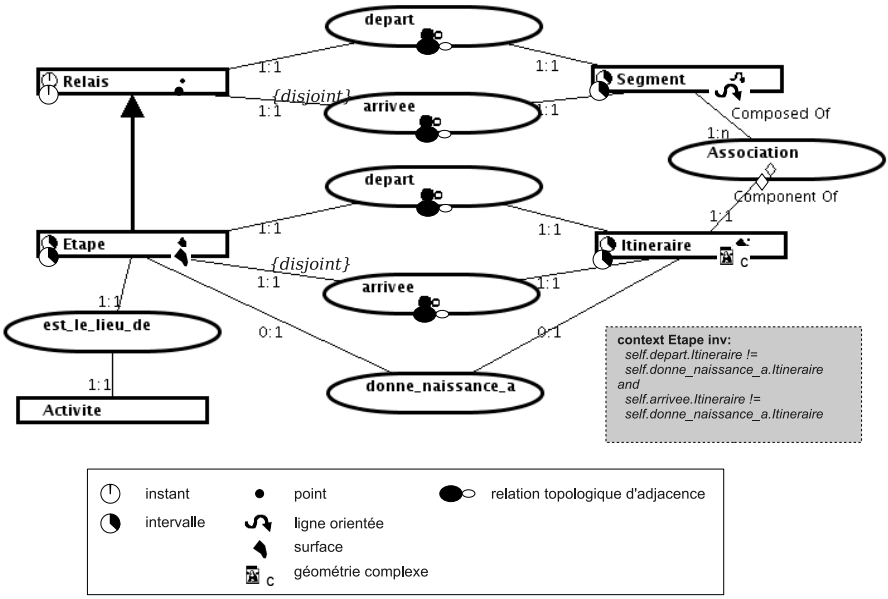


FIG. 2 – *Diagramme MADS simplifié du concept itinéraire. Il s'agit d'un diagramme purement conceptuel, il ne préjuge pas de la façon dont le modèle devrait être formalisé pour son exploitation informatique.*

par l'intermédiaire de la relation *donne_naissance_a*. Lors d'un changement d'échelle (par exemple lors de la description d'un itinéraire I' de plus petite portée à l'intérieur d'un itinéraire I de plus grande portée), un deuxième itinéraire est instancié (cf. schéma 3 de la figure 3). Nous verrons comment l'itinéraire I' peut être rattaché à l'itinéraire I (par déduction spatiale, temporelle ou spatio-temporelle par exemple).

Le segment (ou fragment d'itinéraire) : il est le chemin qui relie deux relais consécutifs. Il est du type spatial MADS *OrientedLine* ( ) et du type temporel MADS *Interval* ( ). Il est important de noter que ce chemin n'est qu'un chemin virtuel, c'est-à-dire une des représentations possibles de l'espace. Dans la carte cognitive - définie par Kuipers (1977) - que se construit le lecteur, le chemin qui relie deux relais n'est pas forcément celui qui a été véritablement emprunté par l'acteur de l'itinéraire. Ce chemin virtuel approche cependant le chemin véritable de manière plus ou moins juste selon les indications dont dispose le lecteur. Ces indications peuvent être aussi nombreuses que variées : la modalité du transport, la vitesse de parcours, la topologie du terrain qui sépare les deux relais, etc. Ces indices, qui aident le lecteur à approcher le chemin véritablement parcouru par l'acteur, peuvent cependant être classés en deux catégories. D'une part, les précisions données par l'acteur de l'itinéraire dans son récit (modalité du transport, vitesse de parcours, etc.), d'autre part, des connaissances dont dispose le lecteur sur la région traversée (topologie du terrain, difficulté de parcours, existence de telle voie de communication, etc.).

Tout segment est associé à un relais de départ et àun relais d'arrivée par les associations *depart* et *arrivee*. Elles sont du type MADS *TopoTouch* ((●○)), c'est-à-dire que le segment est adjacent du point de vue topologique aux relais (de départ comme d'arrivée).

L'activité : l'activité correspond aux occupations du narrateur lors de ses étapes (visites, repas, etc.).

Nous avons ajouté au modèle des attendus une contrainte OCL qui spécifie que dans le cas où un itinéraire I' naît à une étape de départ (respectivement d'arrivée), cet itinéraire I' est différent de l'itinéraire I pour lequel cette étape joue le rôle d'étape de départ (respectivement d'arrivée). Ceci pourrait être le cas dans des configurations telles que celles de la figure 3.

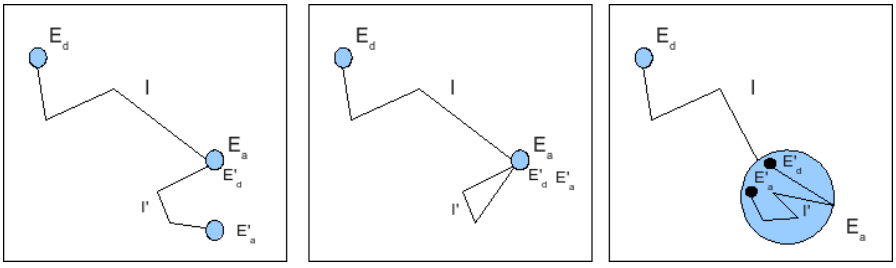


FIG. 3 – *Des exemples de configurations où plusieurs itinéraires émergent.*

Le modèle des attendus que nous proposons se veut compact de manière à pouvoir être manipulé dans une chaîne de traitement informatique permettant de reconstruire un itinéraire à partir de son évocation dans un texte. Le second modèle, le modèle d'entrée ou d'extraction, a pour but de modéliser la manière dont les auteurs expriment leur itinéraire dans leur récit de voyage. Pour sa mise au point, nous avons observé sur un échantillon du corpus, les différentes formes des expressions supportant l'information sur les itinéraires et leur organisation dans le discours.

3.2 Un modèle d'extraction

3.2.1 Observations sur un échantillon du corpus

Parmi les documents mis à notre disposition par la médiathèque, trois ont été retenus[5] afin d'y entreprendre plusieurs études pour valider plus finement les premières observations (description *a posteriori*, régularité dans l'évocation du territoire et des déplacements au sein de l'itinéraire, synchronisation entre la description et le déroulement du voyage).

Ces trois textes racontent le voyage d'un explorateur qui part d'une grande ville (Bordeaux, Paris) à la découverte des Pyrénées sur plusieurs jours.

Étude 1, importance des Entités Spatiales : cette première étude consiste à relever les propositions qui évoquent la position (PP) de l'auteur dans son voyage et à comptabiliser combien parmi elles utilisent des Entités Spatiales (ES). Ces PP qui utilisent des ES seront appelées

[5] *Le voyage de James David Forbes, l'excursion de J.-R. Bals et Le voyage d'Ann Lister*.

PPES. Le concept d'ES a fait l'objet de travaux antérieurs [6], succinctement, une ES est une zone géolocalisable. On définit une ES par rapport à un point d'ancrage dans l'espace : l'Entité Géographique Nommée (EGN). L'EGN est un objet dont on peut obtenir la géolocalisation à l'aide d'une ressource, grâce à son nom.
Les résultats sont donnés dans la figure 4. L'emploi des ES pour évoquer le déplacement y apparaît assez nettement. On atteint des taux de l'ordre de 80% selon les auteurs, mais aussi selon la nature du voyage qui est relaté. En effet, le déficit d'emploi d'ES dans l'évocation du déplacement apparaît le plus souvent lorsque l'auteur évoque des déplacements de plus petites tailles (*je suis allé au parc, j'ai fait une promenade sur le port, j'ai quitté l'hôtel*, etc.), le plus souvent lorsqu'il est à une étape intermédiaire. C'est le cas pour le texte de Ann Lister, dans lequel l'auteur évoque largement ses occupations lors des étapes le long de son voyage.

Étude 2, importance des formes verbales : cette seconde étude a pour but d'évaluer le poids des verbes. Elle consiste à comptabiliser dans les propositions (PP) qui évoquent le déplacement le nombre de propositions qui utilisent des formes verbales à cette fin (PPV). Avec une moyenne de 87,7% (les résultats par document sont donnés en figure 4) cette étude montre leur prédominance.

Si ces deux études doivent encore être approfondies en augmentant le nombre de récits étudiés, cela donne d'ores et déjà une bonne idée du poids qu'occupent les ES et les verbes de déplacement dans notre corpus. Ces observations corroborent également les travaux autour de l'évocation du déplacement dans la langue tels que ceux de Laur (1991); Sarda (1992); Muller et Sarda (1999).

		Étude 1			Étude 2		
Texte	Mots	PP	PP ES	%	PP	PP V	%
J.-D. Forbes	3662	75	63	80%	75	63	80%
A. Lister	3225	40	20	50%	40	35	87,5%
J.-R. Bals	19015	213	170	80%	213	190	89%

FIG. 4 – *Études 1 et 2 : le poids des propositions utilisant les ES (PPES) et des verbes (PPV) dans les propositions évoquant la position (PP)*

Étude 3, synchronisation entre le récit et le voyage : cette étude a pour but de montrer le parallèle qui existe entre la chronologie du voyage et celle du texte le décrivant. Nous avons relevé les dates et heures qui peuvent être attachées aux déplacements mentionnés dans les textes utilisés dans les études précédentes.

Les résultats sont donnés en figure 5. Ils montrent clairement la synchronisation entre le voyage et le récit qui en est fait. Les rares exceptions qui font que des déplacements ne sont pas évoqués dans l'ordre chronologique sont issus d'exemples du type : *avant d'arriver à* ES_x*, j'ai traversé* ES_y.

[6] Dans nos précédents travaux (Lesbegueries et Loustau (2006); Lesbegueries et al. (2006); Gaio et al. (2008)) nous avons distingué deux catégories d'ES ou ET (entité temporelle). On distingue : les *Entités Absolues* des *Entités Relatives*. Une ES Absolue (comme *Pau, le bois de Zouhoure, etc.*), on aura directement sa representation géométrique dans un répère terrestre. Pour une ES Relative (comme *près de Pau, à la lisière du bois de Zouhoure*), évoquée relativement à une ou plusieurs autres ES, sa géométrie est obtenue après inférence (donc généralement moins précise et soumise à caution).

		Étude 3		
Texte	Mots	PPV	PPV correctement ordonnés	pourcentage
J.-D. Forbes	3662	63	63	100%
A. Lister	3225	35	31	88%
J.-R. Bals	19015	58	57	98%

FIG. 5 – *Étude3 : la chronologie du récit*

3.2.2 Le modèle

Les études réalisées sur le corpus documentaire ont montré l'importance des ES et des verbes de déplacement dans l'évocation du déplacement des acteurs lors du récit de leur voyage (cf figure 4). Ces évocations de déplacements permettent au lecteur de se construire une représentation mentale de l'itinéraire parcouru au sein de sa carte cognitive, concept introduit par Kuipers (1977). Nous présentons dans ce paragraphe la manière dont les déplacements apparaissent dans la langue.

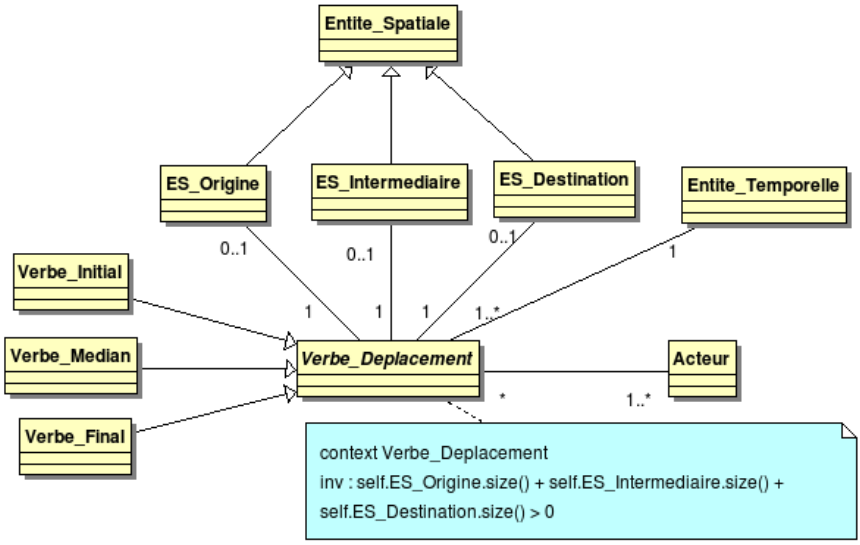


FIG. 6 – *Un modèle pour l'extraction des déplacements.*

Nous reprenons ici principalement le critère de polarité aspectuelle des verbes introduits par Boons (1987) et repris par Laur (1991). Nous modélisons donc les déplacements dans la langue en prenant en compte les verbes de déplacement obligatoirement associés à un acteur, des ES, et une ET (entité temporelle). De cette manière, nous résolvons en partie le difficile problème de la polysémie de certains verbes (cf. section 2.3 pour quelques exemples) qui renferment ou non, selon le contexte, un sens spatial.

Ainsi, conformément à la notion de polarité aspectuelle, les déplacements extraits seront *initiaux* (quitter), *médians* (traverser) ou *finaux* (arriver).

Les notions d'origine, de destination et de position intermédiaire interviennent également. Les constructions verbales du déplacement font en effet émerger une (*quitter Pau*), deux (*quitter Pau pour Bordeaux*) ou trois (*quitter Pau pour Bordeaux par la RN 134*) de ces propriétés. De même, les notions temporelles qui permettent de situer le déplacement dans le temps jouent également un rôle important. Cependant, le lien entre le verbe de déplacement et l'entité temporelle est moins fort syntaxiquement parlant que celui entre le verbe et la ou les entité(s) spatiale(s), notamment dans les textes que nous considérons.

Pour résumer (cf figure 6), un verbe de déplacement est spécialisé en verbe initial, médian ou final. Dans la langue, il est associé à un *Acteur* et à une *Entite_Spatiale* au moins (qu'elle soit d'origine, intermédiaire ou de destination). Il est également associé à une *Entite_Temporelle*, ce qui permet d'horodater le déplacement.

3.3 La phase d'interprétation : transformation du modèle d'extraction vers le modèle des attendus

Nous donnons ici, de manière abstraite, le parallèle qui peut être fait entre le modèle d'extraction et le modèle d'interprétation. Nous reviendrons par la suite sur la méthode, les outils et les ressources nécessaires au passage de l'un à l'autre. Attardons-nous donc exclusivement sur la manière dont les principaux objets du modèle des attendus naissent à partir du modèle d'extraction.

Naissance du *relais* : un *relais* naît lorsqu'un déplacement est évoqué par un verbe de polarité aspectuelle donnée associé à une entité spatiale compatible avec cette polarité (entité origine compatible avec verbe initial, entité indermédiaire avec verbe médian, entité destination avec verbe final).

Naissance du *segment* : un *segment* naît lorsque deux *relais* consécutifs sont identifiés (cela pré-suppose que les relais ont été ordonnés dans le temps selon leur propriété temporelle).

Naissance d'une *étape* : par défaut, le premier et le dernier *relais* deviennent *étapes*. D'autres relais peuvent devenir étapes : lorsqu'une activité peut être identifiée et rattachée à un *relais* (qui devient *étape* de ce fait).

Naissance d'une *activité* : les *activités* peuvent être variées (nuit dans un hôtel, repas, changement de modalité, etc.). Nous verrons plus loin comment procéder afin de détecter de manière automatique ces activités grâce à la notion de rupture.

Naissance d'un *itinéraire* : la naissance de l'*itinéraire* est directement liée à la naissance des *étapes*. Dès que deux étapes sont identifiées, un *itinéraire* est identifié.

Nous donnons dans la figure 7 un exemple de ce passage dans un cas simple : celui dans lequel un seul itinéraire est décrit dans un document. Les exemples d'instances du modèle d'extraction (du haut vers le bas) sont les résultats de l'interprétation des déplacements de l'extrait de document (*Le journal de James David Forbes*) donné en figure 1 :

– *je suis parti à 11h1/4 pour Rochefort*
– *nous arrivâmes à Rochefort vers trois heures*
– *nous avons quitté Rochefort vers 5 heures*
– *nous avons mis deux heures pour arriver à Royan*
– *j'ai atteint Bordeaux à midi*

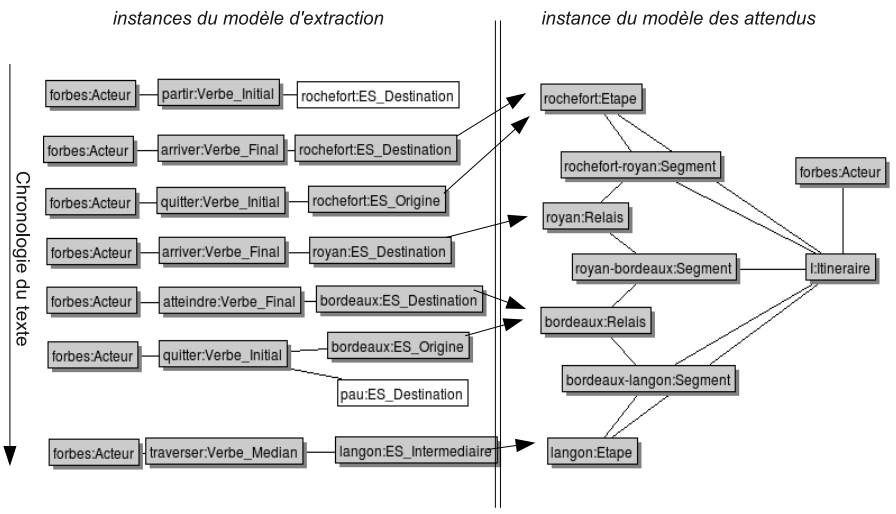


FIG. 7 – *Passage d'instances du modèle d'extraction à une instance du modèle des attendus.*

– *j'ai quitté Bordeaux à 7h en diligence pour Pau*
– *nous avons traversé Langon*

L'algorithme de passage d'instances du modèle d'extraction à une instance du modèle des attendus est alors le suivant :

1. les entités spatiales non compatibles dans les instances du modèle d'extraction (sur fond blanc) ne sont pas retenues comme *relais* dans le processus de découverte des *relais*. Par exemple dans le sixième déplacement (*j'ai quitté Bordeaux à 7h en diligence pour Pau*), seule l'ES *Bordeaux* est retenue car *partir est initial*, *Bordeaux* est l'entité origine, *Pau* l'entité destination. Ces déplacements donneront donc naissance à quatre relais : *Rochefort, Royan, Bordeaux et Langon* ;
2. trois segments seront déduits de la liste ordonnée de relais ;
3. les relais *rochefort* et *langon* seront transformés en étapes (application de la règle par défaut) ;
4. l'itinéraire I est alors créé et attaché à l'acteur $forbes$.

4 Interprétation et extraction d'itinéraire : mode opératoire

4.1 Extraction des déplacements : des méthodes issues du Traitement Automatique du Langage Naturel (TALN)

La problématique ici est la suivante : partant d'un texte brut, nous souhaitons en extraire automatiquement les déplacements. Nous faisons appel ici à des chaînes de traitements lin-

guistiques classiques dans le domaine du TALN[7]. Ces méthodes de traitements linguistiques seront largement utilisées dans les deux grandes phases que nous décrivons dans cette section (extraction des ES et extraction des déplacements).

Extraction des ES : l'extraction des ES est basée sur la reconnaissance d'EGN depuis lesquelles nous sommes capables d'interpréter des relations spatiales qui donnent naissances à des entités plus complexes du type *aux alentours de Pau, au Nord de Bordeaux, à 10km de Toulouse, etc.*. Cette méthode est détaillée dans Lesbegueries et al. (2006).

Extraction des déplacements : nous proposons ici de rendre opératoire la modélisation des déplacements donnée en section 3.2.2 grâce à des transducteurs. Nous rappelons qu'un transducteur est un dispositif qui transforme un langage donné en un autre. Les transducteurs sont basés sur des machines à états finis mettant en correspondance deux langages réguliers. Compte tenu des observations faites sur notre corpus documentaire, la construction des verbes de déplacements peut être apparentée à un langage régulier, modélisé par des machines à états finis, tel que présenté dans la figure 8. Cette modélisation du déplacement sous forme de transducteurs est générique aux documents du type *récits de voyage* et aux documents dans lesquels le déplacement est exprimé principalement sous forme verbale.

Les transducteurs sont traduits en règles de grammaire dans lesquelles nous retrouvons les principaux objets du modèle : le verbe, la préposition et l'ES. Cette analyse à base de règles s'appuie sur les résultats d'analyses plus en amont. La première est une analyse morphosyntaxique capable de retourner deux étiquettes concernant la forme (*@tag* et *@stag* dans la figure 8) et le lemme[8] de chaque unité lexicale (*@lemme* dans la figure 8). Elle permet de s'abstraire des formes fléchies des mots, notamment celles des verbes conjugués. La deuxième analyse est l'extraction des ES (*@sem* dans la figure 8).

Prenons par exemple l'interprétation du déplacement précédemment étudié : *nous arrivâmes à Rochefort*. Nous considérons à ce stade que des analyses en amont ont déjà été réalisées. De ce fait, nous avons accès à diverses informations pour chaque mot de cette phrase. Ces informations sont les suivantes.

- *nous* : @texte=nous / @lemme=nous / @tag=pro / @stag=null / @sem=null
- *arrivâmes* : @texte=arrivâmes / @lemme=arriver / @tag=ver / @stag=ppa / @sem=null
- *à* : @texte=suis / @lemme=être / @tag=pre / @stag=null / @sem=null
- *Rochefort* : @texte=Rochefort / @lemme=Rochefort / @tag=nom / @stag=prp / @sem=es

Nous pouvons alors analyser la phrase à l'aide du transducteur de la figure 8 comme ceci :

- en début d'analyse, le transducteur est en état 0 et le pointeur d'analyse positionné sur le mot *nous*.
- il n'y a pas de transition par *nous* (ou une de ses propriétés), on avance donc le pointeur sur le deuxième mot *arrivâmes*, le transducteur reste en état 0.
- il existe une transition par *arrivâmes* (concernant son lemme, *@lemme=arriver*), le transducteur passe en état 7 et nous avons détecté un déplacement *final*. Le pointeur est avancé sur le mot *à*.
- il existe une transition par *à* (concernant son texte, *@texte=à*), le transducteur passe en état 8. Le pointeur est avancé sur le mot *Rochefort*.

[7] De telles chaînes peuvent se décomposer ainsi : (i) découpage du texte brut en unités lexicales (pour simplifier, le mot), (ii) analyse morpho-syntaxique de chaque unité lexicale, (iii) analyse sémantique des séquences d'unités lexicales.

[8] Le lemme est la forme canonique de chaque unité lexicale (par exemple le verbe à l'infinitif pour un verbe conjugué, le nom masculin singulier pour un nom commun, etc.).

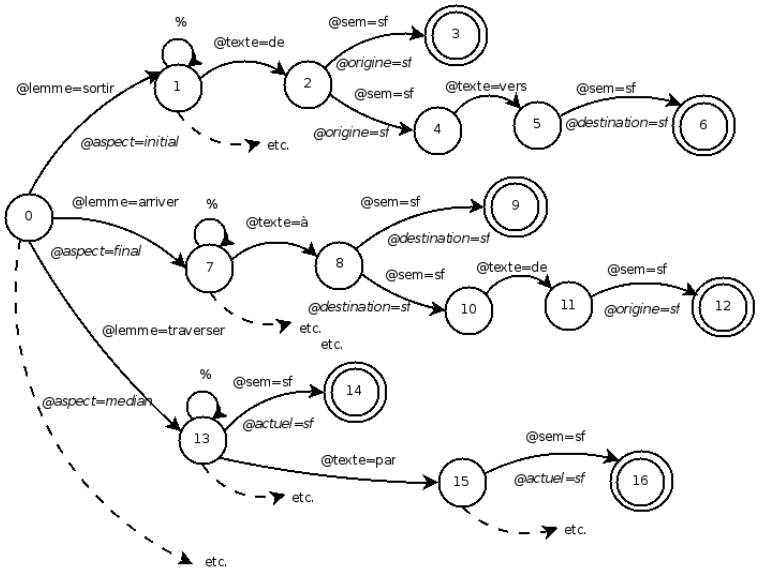


FIG. 8 – *Extrait simplifié de la base des transducteurs spatiaux. Le symbole % représente le « joker ». Les transitions en pointillé montrent que d'autres états existent.*

– il existe une transition par *Rochefort* (concernant sa propriété sem, *@sem=es*), le transducteur passe en état 9 et nous avons détecté un déplacement *final* qui a pour destination *Rochefort*. Le pointeur est en fin de phrase et le transducteur sur un état final, l'analyse est validée : nous avons reconnu un déplacement de polarité aspectuelle finale ayant pour destination *Rochefort*.

Dans la section suivante nous chercherons à répondre à cette question : comment s'élever au niveau du discours à partir de ces éléments extraits localement ?

4.2 Des déplacements à l'itinéraire

Tout comme l'humain doit avoir une connaissance géographique du monde pour pouvoir raisonner sur un itinéraire Kuipers (1977), un système doit également avoir des capacités équivalentes. Les principales ressources envisagées sont des ressources de type géographique et elles peuvent être classées en deux catégories : les données et les raisonnements.

Données factuelles : ce sont des données brutes, que l'on pourrait qualifier de données universelles. Elles s'apparentent aux connaissances du monde que peut avoir l'humain. Elles représentent des faits, au sens logique du terme, c'est-à-dire qu'elles sont considérées comme vraies. Les couches de données SIG[9] (comme celles des communes de France, du réseau rou-

[9]SIG : Système d'Information Géographique

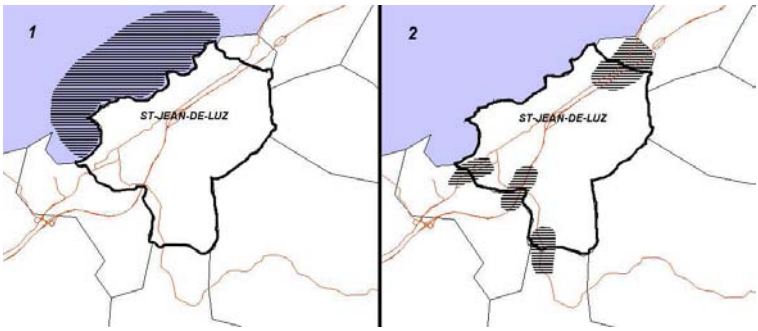


FIG. 9 – *Quitter Saint-Jean-de-Luz : les zones probables selon la modalité du transport. (1) en bateau,(2) en voiture*

tier, ou du réseau fluvial), les Gazeteers[10] , les bases toponymiques, sont des données factuelles envisageables.

Raisonnements : ils sont de deux types. Les plus génériques, s'effectuent directement sur les données factuelles et peuvent être ainsi qualifiés de raisonnement *bas niveau*. Ces raisonnements sont très souvent offerts par les outils traitant l'Information Géographique, ils permettent notamment d'utiliser des opérateurs et des fonctions afin de déduire de nouvelles informations à partir des données factuelles. Prenons un exemple simple et considérons que dans nos données factuelles nous ayons les coordonnées géographiques des villes de *Pau* et *Bordeaux*. Une fonction élémentaire d'un SIG nous permettra de connaître la distance entre *Pau* et *Bordeaux* ou de construire la ligne correspondant au segment reliant *Pau* et *Bordeaux*. Nous considérons ces fonctions comme prédéfinies dans notre système.
Des raisonnements plus spécifiques sur ces données tentent d'apporter au système des règles de bon sens ou règles de sens commun dans le domaine des itinéraires.

- Règles situant l'acteur par rapport à l'entité et mettant en relation la polarité aspectuelle du verbe avec les ES d'origine, intermédiaires et de destination. Un déplacement de polarité aspectuelle initiale P_i ayant pour origine ES_o situe l'acteur du déplacement en ES_o. Un déplacement de polarité aspectuelle médiane P_m ayant pour entité intermédiaire ES_i situe l'acteur en ES_i. Un déplacement de polarité aspectuelle finale P_f ayant pour destination ES_d situe l'acteur en ES_d.
- Règles concernant l'étendue des zones probables de localisation de l'acteur en fonction de la polarité aspectuelle des verbes et de la modalité de déplacement. Un déplacement évoqué avec un verbe initial dont l'origine est l'ES ES_o situe l'acteur dans une région limitrophe de la frontière de l'entité ES_o, la largeur de cette zone est différente selon la modalité de déplacement : elle sera plus large sur un déplacement en voiture que sur un déplacement à pied et cela est principalement dû à la vitesse du déplacement.
- Règles concernant les modalités de transport et leur localisation probable (ex : voiture sur route, bateau sur mer/océan/fleuve, vélo sur route mais pas autoroute, etc.).

[10]Gazeteers : une gazetteer est un dictionnaire géographique qui renferme des informations sur les lieux et les noms de lieux, leur localisation, etc.

Ces deux dernières règles concernant les modalités du transport sont illustrées sur la figure 9. L'utilisation de fonctions comme *Boundary*, *Intersection* et *Buffer* proposées dans les spécifications Open GIS[11] et implémentées dans la plupart des SIG permettent de construire ces zones. Par exemple, on obtient les zones hachurées de la figure 9.2 en appliquant une extension (*Buffer*) d'un facteur β, dépendant de la vitesse de déplacement sur l'intersection (*Intersection*) entre la frontière du polygone de St-Jean-De-Luz (*Boundary(Geom*$_{stjean}$*)*) et les géométries des routes du département des Pyrénées Atlantiques (*Geom*$_{routes64}$). Une dernière intersection du résultat obtenu avec les routes donne les zones hachurées. Ceci se résume par une requête de la forme :

Intersection(Buffer(Intersection(Boundary(Geom$_{stjean}$*),Geom*$_{routes64}$*),*β*),Geom*$_{routes64}$*)*

Enfin des règles de détection de ruptures permettent de différencier les étapes des relais, avec comme ruptures possibles (notons que c'est la combinaison de ces ruptures qui permet de déceler une activité) :

- rupture dans la modalité : détection simple dès lors que les modalités peuvent être repérées ;
- rupture dans l'amplitude dans le déplacement et/ou le temps (ie changement d'échelle) : détection possible par inclusion spatiale et/ou temporelle ;
- rupture dans la structure logique : dépend de la qualité de la phase ROC[12] (détection des changements de paragraphe possible avec une ROC basique, détection de titre de chapitre, de section voire plus avec des ROC de meilleure qualité) ;
- rupture dans la linéarité spatiale de l'itinéraire : apparition de boucles ;
- rupture dans la continuité spatiale de l'itinéraire : apparition de sauts spatiaux ;
- rupture dans la continuité temporelle de l'itinéraire : apparition de sauts temporels.

4.3 $\pi\mathcal{R}$: un Prototype pour l'Interprétation d'Itinéraires dans des Récits

La démarche générale consiste à construire une chaîne de traitement linguistico-géographique (figure 10), capable d'extraire les déplacements de manière locale au niveau phrastique dans les textes puis de reconstruire l'itinéraire en utilisant des ressources comme nous le décrivions dans la section précédente. Cette chaîne de traitement utilise le langage XML qui permet de facilement enchaîner différents traitements en ajoutant à l'information extraite dans une phase *n* l'information extraite à une phase *n+1*. Nous décrivons ici les outils utilisés par chaque phase du traitement.

Extraction des déplacements : correspond à la partie supérieure de la figure 10. Comme précédement évoqué, l'extraction des ES est effectuée par le prototype PIV de Lesbegueries et al. (2006).
L'extraction des déplacements est une chaîne de traitement linguistique à part entière. Elle est constituée des grandes phases que nous décrivions dans la section précédente et a été implémentée grâce à la plate-forme de traitement linguistique *Linguastream* de Widlöcher et Bilhaut

[11] Les spécifications Open GIS sont des documents techniques détaillant les objets permettant de représenter des données géographiques, les fonctions permettant de les manipuler, etc. Elles sont produites par l'Open Geospatial Consortium.

[12] ROC (ou OCR) : Reconnaissance Optique de Caractères (ou Optical Character Recognition).

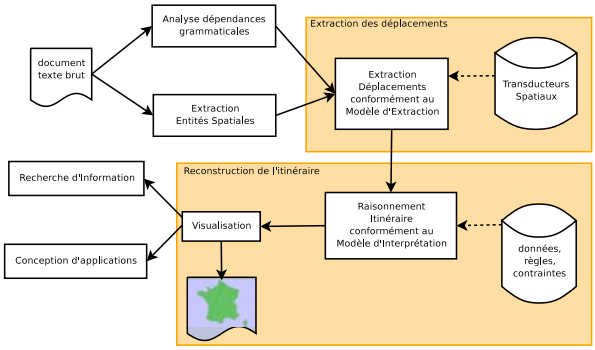


FIG. 10 – *La chaîne de traitement linguistico-spatiale envisagée : les zones encadrées sont les parties développées dans cet article.*

(2005). Au sein de celle-ci, l'analyse morpho-syntaxique est effectuée par Tree-Tagger, l'analyseur morphosyntaxique éprouvé de Schmid (1994). Les transducteurs des verbes de déplacements sont traduits en grammaires DCG[13] et l'analyse basée sur ces grammaires est confiée à Prolog afin de profiter des mécanismes de déduction et d'unification de ce langage.

On obtient ainsi la première partie de notre chaîne de traitement, celle qui est capable d'extraire les déplacements des textes (cf figure 10).

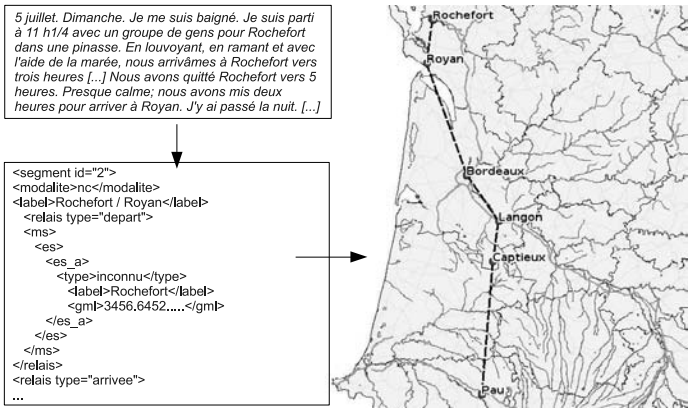
Reconstruction de l'itinéraire : ce module nous permet de passer des déplacements extraits à un itinéraire. Il prend en entrée un fichier XML dans lequel les déplacements sont représentés selon les critères que nous avons évoqués dans la section 3.2.2. Il utilise le SIG PostGIS[14] afin de mettre en application les règles de raisonnement spatial précédemment évoquées. Diverses ressources sont également nécessaires afin de reconstruire l'itinéraire. Dans $\pi\mathcal{R}$, nous utilisons différentes couches de données pour la géolocalisation des relais (communes de France, toponymes, etc.) et des segments (réseau autoroutier, routier, sentier, etc.). Ces couches de données sont également rendues disponibles dans PostGIS.

Pour chaque déplacement extrait, on cherche tout d'abord à produire une zone probable de localisation de l'acteur. Pour cela nous faisons appel aux règles concernant la polarité aspectuelle et la modalité du transport. Les fonctions *boundary*, *buffer*, *intersection* du SIG PostGIS ont été utilisées à ces fins. Considérons le déplacement de l'exemple : « Nous avons quitté Bordeaux pour Langon en voiture. ». Celui-ci a pour origine Bordeaux et pour destination Langon. Il fait donc émerger un segment reliant les deux relais que sont Bordeaux et Langon. La modalité du déplacement étant *voiture*, on va faire appel à un algorithme de calcul de trajet dans le réseau routier pour construire la route probablement empruntée par l'acteur du déplacement. Cette fonctionnalité est apportée par le module pgRouting[15]. Rappelons qu'il ne s'agit pas forcément de la réalité ; l'acteur a peut-être pris un autre chemin. Par cette méthode, on cherche à

[13] DCG : Definite Clause grammar

[14] PostGIS : extension GIS du système de gestion de base de donnée libre PostgreSQL

[15] pgRouting est un module qui donne au SIG PostGIS des fonctions de calcul de plus court chemin (algorithme de *Dijkstra*, *A-étoile*, etc.)


5 juillet. Dimanche. Je me suis baigné. Je suis parti à 11 h1/4 avec un groupe de gens pour Rochefort dans une pinasse. En louvoyant, en ramant et avec l'aide de la marée, nous arrivâmes à Rochefort vers trois heures [...] Nous avons quitté Rochefort vers 5 heures. Presque calme; nous avons mis deux heures pour arriver à Royan. J'y ai passé la nuit. [...]

<modalite>nc</modalite>
<label>Rochefort / Royan</label>
<relais type="depart">
<ms>
<es>
<es_a>
<type>inconnu</type>
<label>Rochefort</label>
<gml>3456.6452.....</gml>
</es_a>
</es>
</ms>
</relais>
<relais type="arrivee">
...
Rochefort
Royan
Bordeaux
Langon
Captieux
Pau

FIG. 11 – *Sortie simplifiée du prototype $\pi\mathcal{R}$: un fichier XML contenant l'interprétation de l'itinéraire. Visualisation effectuée avec MapServer*

approcher la représentation mentale de l'itinéraire que pourrait se faire le lecteur dans sa carte cognitive.

Une fois que chaque segment de l'itinéraire est ainsi construit, on stocke les segments dans le SIG PostGIS.

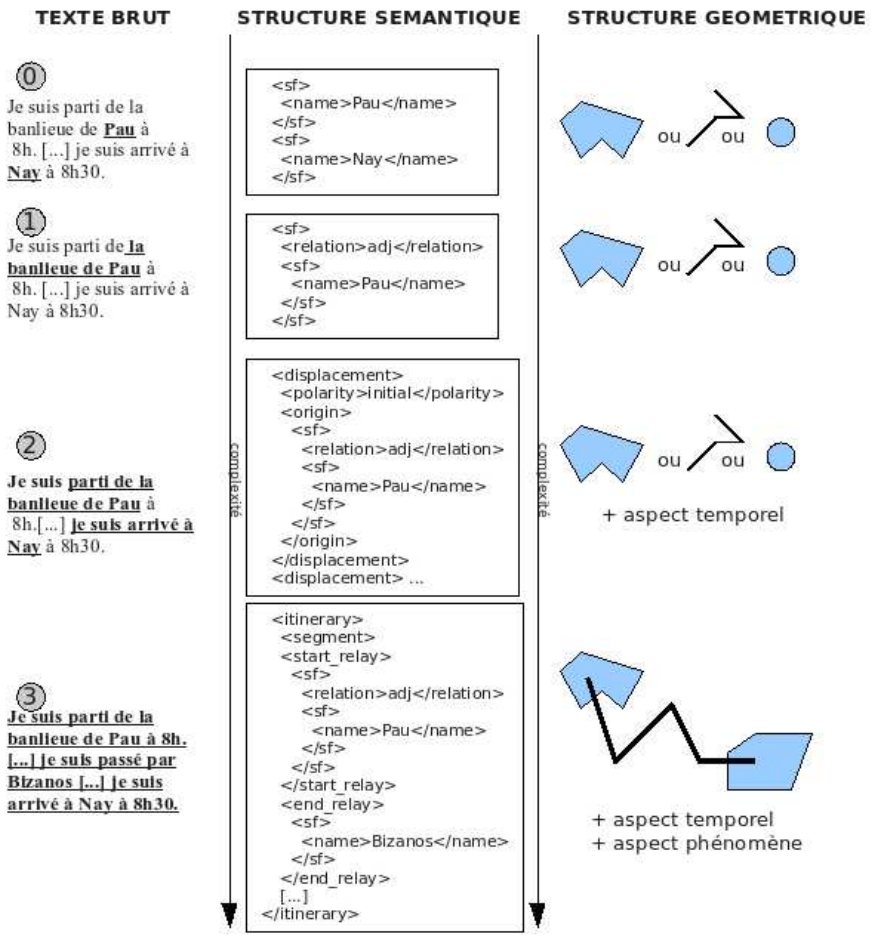
Les détections de ruptures qui donnent naissance à des activités ne sont pas encore implémentées. De ce fait, nous appliquons la règle par défaut pour la détection des étapes de départ et d'arrivée : le premier relais devient étape de départ, le dernier relais devient étape d'arrivée.

En sortie du prototype $\pi\mathcal{R}$, nous obtenons donc une instance du modèle des attendus au format XML. Ce fichier XML contient l'ensemble des objets qui permettent de décrire un itinéraire : des segments, des relais et des étapes. Tous ces objets sont des EG, ils contiennent donc un géocodage au format GML, et une marque temporelle. Ces interprétations d'itinéraires peuvent être alors visualisées par des outils de cartographie tels que MapServer[16] comme montré dans la figure 11. Ils peuvent également intervenir dans la phase de conception d'activités pédagogiques : c'est ce que nous montrons dans la section suivante par des exemples de requêtes multi-niveaux de l'information géographique.

5 Quelques exemples de requêtes multi-niveaux permettant de sélectionner des documents pertinents

Rappelons qu'il incombe aux enseignants de choisir dans le corpus de documents disponibles ceux qui sont les plus appropriés pour un usage pédagogique. Le processus de décision n'est pas totalement rigoureux et les enseignants sont donc amenés à formuler un ensemble de requêtes multi-niveaux qui leur permet, petit à petit, d'extraire les documents pertinents. La

[16]MapServer : environnement de développement Open Source permettant de construire des applications internet à contenu spatial (http ://mapserver.gis.umn.edu/)


TEXTE BRUT
STRUCTURE SEMANTIQUE
STRUCTURE GEOMETRIQUE
0
Je suis parti de la banlieue de Pau à 8h. [...] je suis arrivé à Nay à 8h30.
<sf>
<name>Pau</name>
</sf>
<sf>
<name>Nay</name>
</sf>
ou
ou
1
Je suis parti de la banlieue de Pau à 8h. [...] je suis arrivé à Nay à 8h30.
<sf>
<relation>adj</relation>
<sf>
<name>Pau</name>
</sf>
</sf>
ou
ou
2
Je suis parti de la banlieue de Pau à 8h.[...] je suis arrivé à Nay à 8h30.
<displacement>
<polarity>initial</polarity>
<origin>
<sf>
<relation>adj</relation>
<sf>
<name>Pau</name>
</sf>
</sf>
</origin>
</displacement>
<displacement> ...
ou
ou
+ aspect temporel
complexité
complexité
3
Je suis parti de la banlieue de Pau à 8h. [...] je suis passé par Bizanos [...] je suis arrivé à Nay à 8h30.
<itinerary>

<start_relay>
<sf>
<relation>adj</relation>
<sf>
<name>Pau</name>
</sf>
</sf>
</start_relay>
<end_relay>
<sf>
<name>Bizanos</name>
</sf>
</end_relay>
[...]
</itinerary>
+ aspect temporel
+ aspect phénomène

FIG. 12 – *Prise en compte de la complexité croissante des informations géographiques sur lesquelles portent les requêtes.*

figure 12 présente quatre exemples de requêtes de complexité croissante ; la première colonne montre un exemple de texte brut pour lequel l'information ciblée est soulignée. La deuxième colonne décrit la structure sémantique de l'information ciblée alors que la troisième colonne propose une projection de cette information afin de pouvoir la traiter avec un SIG par exemple.

De nombreux SIG permettent de résoudre la requête 0 (et même des systèmes d'information classiques se basant sur des appariements de type *fulltext*). Les autres requêtes sont beaucoup plus spécifiques à la problématique étudiée dans cet article puisqu'elles prennent en compte les entités géographiques associées aux verbes de déplacement qui apparaissent dans le texte (si nous avions proposé des modèles computationnels prenant en compte les descriptions de lieux associées à des verbes de perception, le système aurait également dû s'appuyer sur les requêtes de niveau 0).

Pour chacune des requêtes 0 à 3, nous présentons dans les paragraphes qui suivent les traitements effectués (niveau de granularité des informations recherchées) et nous indiquons quels modèles computationnels sont les plus appropriés pour répondre à ces requêtes.

Requête 0 : Trouver les documents qui parlent de Pau et de Nay
Ici, il s'agit de traiter une information géographique de base, donc de faire une simple comparaison entre les entités géographiques nommées (EG) du corpus documentaire et de la requête. Traitements sur le corpus : extraction des EG, géocodage des EG, indexation des EG. Traitements sur la requête : si elle est exprimée en langage naturel, des traitements similaires à ceux effectués sur le corpus sont nécessaires c'est-à-dire extraction des EG, géocodage des EG, indexation des EG. Si la requête est formulée via une interface-usager (via une carte par exemple), le géocodage est obtenu directement. Croisement de la requête et du corpus : il s'agit d'une comparaison géométrique au sens strict entre le géocodage de la requête et le géocodage des EG des documents.

Requête 1 : Trouver les documents qui parlent de la banlieue de Pau et de Nay
Dans cette requête, le niveau d'agrégation des informations géographiques à exploiter est plus élevé car une simple utilisation des EG ne permet pas de traiter la requête. Le système doit être en mesure d'interpréter les EG *Pau* et *Nay* mais aussi le concept de *banlieue* c'est-à-dire d'interpréter des Entités spatiales (ES). Traitements sur le corpus : extraction des ES, géocodage des ES, indexation des ES. Traitements sur la requête : si elle est exprimée en langage naturel, des traitements similaires à ceux effectués sur le corpus sont nécessaires c'est-à-dire extraction des ES, géocodage des ES, indexation des ES. Si la requête est formulée via une interface-usager (via une carte par exemple), le géocodage est obtenu directement. Croisement de la requête et du corpus : il s'agit d'une comparaison géométrique au sens strict entre le géocodage de la requête et le géocodage des ES issues des documents.

Requête 2 : Trouver les documents dans lesquels le narrateur part de Pau
Ici encore, le niveau d'agrégation de l'information géographique est augmenté. Le système doit non seulement interpréter les ES mais aussi les relations spatiales entretenues par le narrateur avec ces ES. En d'autres mots, le système doit être capable de déterminer si l'ES *Nay* apparaît dans un déplacement du narrateur de polarité finale. Traitements sur le corpus : extraction, interprétation, géocodage et indexation des ES, extraction, interprétation, géocodage et indexation des déplacements. Traitements sur la requête : si elle est exprimée en langage naturel, des traitements similaires à ceux effectués sur le corpus sont nécessaires. Si la requête est formulée via une interface-usager (via une carte par exemple), le géocodage est obtenu directement. Croisement de la requête et du corpus : il ne s'agit plus de faire une simple comparaison

géométrique comme dans les cas précédents. Le système doit également interroger la structure sémantique (codée dans un arbre XML) pour déterminer la sémantique du déplacement effectué pour une ES donnée. Dans notre exemple, l'ES *Pau* est associée à un déplacement de polarité initiale.

Requête 3 : Trouver les documents dans lesquels le narrateur part de la banlieue de Pau, traverse Bizanos et arrive à Nay

Dans cette requête, le niveau d'agrégation est encore augmenté puisque le système doit être en mesure de capter l'itinéraire complet effectué entre *la banlieue de Pau* et *Nay*. Traitements sur le corpus : extraction des ES, géocodage des ES, indexation des ES, extraction et interprétation des déplacements, reconstruction et indexation de l'itinéraire. Traitements sur la requête : si elle est exprimée en langage naturel, des traitements similaires à ceux effectués sur le corpus sont nécessaires. Si la requête est formulée via une interface-usager (via une carte par exemple), le géocodage est obtenu directement. Croisement de la requête et du corpus : il s'agit d'effectuer une comparaison géométrique entre le géocodage de la requête et le géocodage des itinéraires extraits des documents du corpus.

6 Bilan et perspectives

Dans cet article, nous avons présenté une approche computationnelle et un outillage pour capter et exploiter, à différents niveaux de granularité, la sémantique des informations géographiques contenues dans un corpus de récits de voyage. Les informations particulières visées par ces travaux sont (des plus simples aux plus agrégées) des entités géographiques nommées, des entités spatiales, des déplacements, des relais, des étapes et des segments constitutifs des itinéraires relatés par les auteurs.

Nous nous situons ici dans un courant très profond qui est celui du passage du web actuel (pour lequel on affiche des documents avec quelques fonctionnalités de recherche, y compris dans des projets comme Google Library) à un web de la connaissance et des services (le web 2.0).

Des niches très diverses se constituent des outillages qui sont capables d'exploiter les contenus textuels que l'on peut récupérer via les bibliothèques numériques en ligne. Cela justifie une démarche de traitements totalement automatisés en amont des usages (notre approche consiste à fournir des services d'interrogation / visualisation sur une clé d'accès au document très particulière qui est celle de l'itinéraire). Les traitements automatisés que nous proposons dans ce papier sont transparents pour les utilisateurs puisque les documents sont traités avant que l'utilisateur ne s'y intéresse via un traitement *back-office* (comme lors d'une indexation classique d'un moteur de recherche). Cela justifie également que l'on propose des services à valeur ajoutée utilisant le potentiel des documents ainsi sélectionnés dans des applications finalisées (la sélection de documents sur des critères géo-spatiaux n'étant pas une fin en soi). D'où la conception d'applications pédagogiques et les usages faits de ces applications par des apprenants.

L'intérêt des usages pédagogiques et l'utilisabilité des documents analysés par notre outillage est de se servir de leur sémantique pour guider l'activité de conception. D'une part pour définir les types d'interaction avec l'apprenant qui pourront être évalués par le système tuteur (diagnostic des connaissances de l'apprenant, diagnostic des erreurs de compréhension sur la base de la connaissance qu'a ce tuteur du texte mis à disposition de l'apprenant). D'autre part

pour éviter au concepteur de définir des activités que la machine ne pourrait pas interpréter parce que ce qui a été capté dans le texte de manière automatisée est trop pauvre par rapport à la compréhension que peut avoir l'apprenant de cette activité par sa lecture "humaine" du document. L'enjeu ici est de se servir des documents analysés automatiquement pour ne concevoir que les activités que l'on va pouvoir encadrer de manière automatisée (ou assistée) par un tuteur informatique.

Compte tenu de la taille du corpus documentaire mis à notre disposition, nous avons proposé une approche totalement automatisée d'extraction des itinéraires de ce corpus. Cette approche se base sur deux modèles computationnels et un processus automatisé de mise en relation de ces deux modèles. Le premier modèle est un modèle des attendus (cf. section 3.1). Il a pour but de donner une représentation au concept d'itinéraire et a été piloté par les usages que nous souhaitons en faire (à des fins de RI et de conception d'activités pédagogiques). Le modèle des attendus proposé s'inspire de travaux des nombreux auteurs (cf. section 2.2) ayant nourri les recherches dans ce domaine. Rappelons enfin que la formalisation MADS que nous avons proposée pour ce modèle des attendus est purement conceptuelle (cf figure 2) et ne préjuge pas de la façon dont le modèle est ensuite formalisé pour son exploitation par un Système d'Information de type Base de Données. Le second modèle (cf. section 3.2.2) est un modèle d'entrée ou d'extraction, il suit une logique du domaine de l'Extraction d'Information. Il permet par sa mise en application au sein de notre chaîne de traitement de capter environ 75% (cf figure 13) des déplacements importants d'un récit de voyage. Cette capacité varie en fonction de la précision avec laquelle l'auteur évoque ses déplacements : les déplacements de petite ampleur (le plus souvent ceux qui ne font pas intervenir d'ES) sont plus difficiles à capter. Comme dans la plupart des travaux, les capacités d'extraction dépendent fortement de la base lexicale utilisée. Ici, elles s'appuient sur l'existence de transducteurs de tous les verbes susceptibles d'évoquer un déplacement (cf. section 4.1). Des expérimentations en cours semblent toutefois montrer qu'il est possible d'augmenter automatiquement ces ressources lexicales par synonymie, par similarité de construction, etc. Le processus automatisé que nous avons proposé permet de reconstruire l'itinéraire (conformément au modèle des attendus) à partir des déplacements extraits (conformément au modèle d'entrée). La mise en oeuvre de ce processus repose sur une hypothèse relativement forte concernant la chronologie des déplacements. Cette hypothèse est acceptable car les récits de voyage constituent un domaine à part entière : ce qui est dit doit être fidèle à ce qui a été vu, le locuteur rendant compte de ses découvertes avec la plus grande exactitude (ce qui est visé, c'est la parfaite concordance des mots et des choses vues). Cependant, nous tentons actuellement de lever une partie de cette hypothèse en prenant en compte les aspects temporels et thématiques des relais mais aussi en tenant compte du temps (conjugaison) des verbes.

Du point de vue qualitatif, les résultats actuels sont très encourageants. Nous avons évalué les capacités de notre outillage en comparant les itinéraires annotés automatiquement à une annotation manuelle effectuée sur trois textes. Nous donnons en figure 13 les résultats de cette évaluation : Dans l'évaluation 1, nous avons comparé les déplacements captés automatiquement à tous les déplacements annotés manuellement. Les résultats sont mitigés car certains déplacements ne sont pas évoqués sous la forme attendu, c'est-à-dire avec le triplet (V,P ?,E). En comparant l'extraction automatique avec les déplacements évoqués grâce au triplet (V, P ?,E), nous avons obtenu de bien meilleurs résultats (évaluation 2). Dans une grande majorité des cas, c'est l'absence de l'ES qui met en échec notre processus d'extraction des déplacements.

		Evaluation 1		Evaluation 2	
Texte	Auto.	Manu. quelconque	%	Manu. avec (V,P ?,E)	%
J.-D. Forbes	44	75	58,6%	64	68,7%
A. Lister	22	40	55%	25	88%
J.-R. Bals	46	91	50,5%	58	79%

FIG. 13 – *Evaluation 1 par rapport au marquage manuel de tous les déplacements, évaluation 2 par rapport aux déplacements évoqués avec le triplet (V,P ?,E)*

Ce point est cependant à nuancer car les déplacements évoqués sans ES sont très souvent des déplacements de plus petite ampleur (je suis allé au port, j'ai quitté la gare, etc.). Ils sont donc de moins grande importance pour les usages que nous avons ciblés car notre outillage doit avant tout donner aux utilisateurs finaux une vue d'ensemble de l'itinéraire relaté dans chaque document textuel analysé.

Les améliorations envisagées portent principalement sur des traitements qui permettraient :

1. de capter automatiquement la modalité des déplacements en tenant compte d'une part, de la modalité implicite des verbes de déplacement utilisés (ex : *j'ai marché jusqu'au col des Pentes*, et d'autre part des indications de modalité attachées à des verbes qui ne permettent pas d'identifier une modalité précise (ex : *j'ai atteint le col des Tentes après 4 heures de marche*) ;
2. d'horodater automatiquement chacun des déplacements détectés en se basant sur la présence d'indications temporelles associés à des blocs de texte (section, paragraphe) constitutifs du récit de voyages analysé ;
3. de capter automatiquement les ruptures dans le récit (ex : passage du mode narratif au mode descriptif) afin de mieux identifier les activités effectuées par un acteur. Et par voie de conséquence mieux identifier les étapes constitutives d'un itinéraire (rappelons que dans la figure 2, une étape est un relais particulier auquel on associe une activité faite par un acteur) ;
4. d'attacher chacun des déplacements captés dans un récit à un acteur particulier, ce qui permettrait d'identifier les itinéraires particuliers faits par tel ou tel acteur (ce type d'analyse doit reposer sur des outils d'analyse des dépendances grammaticales de la phrase, notamment afin de détecter les relations acteur-déplacement, c'est-à-dire les relations sujet-verbe).

D'ores et déjà, l'outillage dont nous disposons est exploitable pour des usages finalisés, notamment pédagogiques. Deux types de travaux sont menés. Nous avons d'une part développé un prototype d'application Web[17] qui permet à des enfants de cycle 3 d'être assistés dans leur lecture et leur compréhension d'un récit de voyages : "Le voyage de James David Forbes dans les Pyrénées". L'interface est divisée en plusieurs zones d'interaction avec l'apprenant, chaque zone étant chargée de rendre compte d'un aspect du récit : nous avons ainsi pu intégrer le texte du récit balisé à partir du modèle computationnel présenté dans cet article, une carte géographique pour matérialiser la dimension spatiale du voyage, un calendrier pour rendre compte des éléments temporels du récit et un index d'activités pour mettre en évidence les activités décrites par le narrateur dans son récit. Par ailleurs, nous avons engagé un travail

[17] http ://erozate.iutbayonne.univ-pau.fr/forbes2007/exp/

de développement d'outils de conception pédagogique dont les briques de base sont constituées d'outils de création de cartes géographiques [18] que nous enrichissons de fonctionnalités permettant d'exploiter les éléments constitutifs d'un récit de voyages : entités géographiques nommées, entités spatiales, relais, étapes et itinéraires. Une fois mis au point, ce type d'outil de conception permettra à un concepteur de créer des scénarios pédagogiques du type : "Étant donnés un récit de voyage que lit l'apprenant et une carte interactive des lieux du récit (type carte GoogleMaps), permettre à l'apprenant de créer et placer des étiquettes des étapes de l'itinéraire en offrant des fonctionnalités d'assistance et de corrections automatisées sur la base de la connaissance de l'itinéraire réel emprunté (calculé automatiquement par l'outillage proposé dans cet article et validé ensuite par l'enseignant)".

Les perspectives des travaux présentés sont donc nombreuses. Nous pensons qu'envisager le développement d'applications finalisées à partir des travaux présentés dans cet article constitue un challenge intéressant mais accessible compte tenu des résultats actuels. Et il est clair qu'un tel objectif opérationnel nous obligera à conforter nos résultats actuels (cf. les 4 perspectives présentées précédemment pour améliorer la qualité des informations captées sur les itinéraires de voyage).

Des applications dans le domaine touristique sont également envisageables : avec la multiplication des équipements portables équipés de puces GPS, on peut facilement imaginer proposer à un promeneur de consulter des documents qui racontent des itinéraires qui sont passés par l'endroit où il se trouve ou des itinéraires semblables à celui qu'il est en train de faire.

Références

Blackburn, P. et J. Bos (2003). Computational semantics. *Theoria, Special Issue : Logic, Language and Information 18/1*(46), 27–45.

Boons, J.-P. (1987). La notion sémantique de déplacement dans une classification syntaxique des verbes locatifs. *Langue Française 76*, 5–40.

Casenave, J., C. Marquesuzaà, P. Dagorret, et M. Gaio (2004). La revitalisation numérique du patrimoine littéraire territorialisé. In *EBSI-ENSSIB, Montréal (CA)*.

Coyne, B. et R. Sproat (2001). Wordseye : An automatic text-to-scene conversion system. In E. Fiume (Ed.), *SIGGRAPH 2001, Computer Graphics Proceedings*, pp. 487–496. ACM Press / ACM SIGGRAPH.

Denis, M. (1994). La description d'itinéraires : des repères pour des actions. In *Notes et Documents LIMSI*.

Egges, A., A. Nijholt, et P. Nugues (2001). Generating a 3d simulation of a car accident from a formal description : the carsim system. In *Proceedings of The International Conference on Augmented, Virtual Environments and Three-Dimensional Imaging, ICAV3D*.

Fraczak, L. et G. Lapalme (1999). Utilisation de stratégies cognitives dans la génération automatique de descriptions d'itinéraires. In *TALN'99*, Cargese, pp. 10 pages.

Gaio, M., C. Sallaberry, P. Etcheverry, C. Marquesuzaa, et J. Lesbegueries (2008). A global process to access documents' contents from a geographical point of view. *J. Vis. Lang. Comput. 19*(1), 3–23.

[18] http ://erozate.iutbayonne.univ-pau.fr/plugin_carto

Jones, C. B. J. et R. Purves (2006). Gir05 2005 acm workshop on geographical information retrieval. In *2nd ACM workshop on Geographical information retrieval*, Volume 40, pp. 34–37.

Kuipers, B. (1977). Modeling spatial knowledge. In *IJCAI*, pp. 292–298.

Köhler, W. (1929). *Psychologie de la forme*. Paris : Gallimard.

Laur, D. (1991). *Sémantique du déplacement et de la localisation en français : une étude des verbes, des prépositions et de leur relation dans la phrase simple*. Ph. D. thesis, Université de Toulouse II.

Lesbegueries, J., M. Gaio, P. Loustau, et C. Sallaberry (2006). Geographical information access for non-structured data. In *21st ACM Symposium on Applied Computing - Advances in Spatial and Image based Information Systems track, SAC'06*, Dijon, pp. 83–89.

Lesbeguerries, J. et P. Loustau (2006). Extraction et interprétation d'information géographique dans des données non-structurées. In *Actes de la 3ème Conférence en Recherche d'Information et Applications (CORIA'06)*.

Loustau, P., T. Nodenot, et M. Gaio (2008). Spatial decision support in the pedagogical area : Processing travel stories to discover itineraries hidden beneath the surface. In Springer (Ed.), *Procedings of the 11th AGILE International Conference on Geographic Information Science. To be published*.

Maaß, W., P. Wazinski, et G. Herzog (1993). Vitra guide : Multimodal route descriptions for computer assisted vehicle navigation. In *Six Int. Conf. on Industrial and Engineer- ing Applications of Artificial Intelligence and Expert Systems*, Edinburgh, Scotland, pp. 144–147.

Mathet, Y. (2000). *Etude de l'expression en langue de l'espace et du déplacement : analyse linguistique, modélisation cognitive, et leur expérimentation informatique*. Ph. D. thesis, Université de Caen Basse Normandie.

Muller, P. et L. Sarda (1999). Représentation de la sémantique des verbes de déplacement transitifs du français. *T.A.L. 39*(2), 127–147.

Nodenot, T., P. Loustau, M. Gaio, C. Sallaberry, et P. Lopisteguy (2006). From electronic documents to problem-based learning environments : an ongoing challenge for educational modeling languages. In *7th International Conference on Information Technology Based Higher Education and Training*, pp. 75–86. IEEE.

Parent, C., S. Spaccapietra, et E. Zimányi (2006). *Conceptual Modeling for Traditional and Spatio-Temporal Applications - The MADS Approach*. Springer.

Przytula-Machrouh, E., G. Ligozat, et M. Denis (2004). Vers des ontologies transmodales pour la description d'itinéraires. *Revue Internationale de Géomatique : nř Spécial sur les ontologies spatiales 14*(2), 285–302.

Sablayrolles, P. (1995). *Sémantique formelle de l'expression du mouvement. De la sémantique lexicale au calcul de la structure du discours en français*. Thèse de doctorat, Université Paul Sabatier, Toulouse, France.

Sarda, L. (1992). *Syntaxe & Sémantique - Sémantique du lexique verbal*, Chapter L'expression du déplacement dans la construction transitive directe, pp. 121–137. 2. Presses Universitaires de Caen.

Sarda, L. (2000). L'expression du déplacement dans la construction transitive directe. *Syntaxe et Sémantique* (2), 121–137.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, Manchester, UK.

Tolman, E. (1948). Cognitive maps in rats and man. *Psychological Review 42*(55), 189–208.

Usery, E. L. (2003). Multidimensional representation of geographic features. Technical report, U.S Geological Survey (USGS).

Widlöcher, A. et F. Bilhaut (2005). La plate-forme LinguaStream : un outil d'exploration linguistique sur corpus. In *Actes de TALN 2005*, Dourdan, France, pp. 517–522.

Wunderlich, D. et R. Reinelt (1982). How to get there from here. In R. J. Jarvella et W. Klein (Eds.), *Speech, Place, and Action*, pp. 183–201. Chichester : Wiley.

Summary

Local cultural heritage documents are characterized by contents strongly attached to a territory (i.e. Geographical references). Numerous corpora of such local documents become available and a challenging task is to process them automatically in order to retrieve and to make explicit the geographical information that they contain. In this paper, we suggest two computational models and a complete implementable method to automaticaly extract geographic information at different level from text documents telling a trip. The first model is aimed at modeling the concept of an itinerary, it is an interpretation model. The second one is an extraction model, it can extract the displacements of the author from a document and it is based on the extraction of low-level geographical information (named entities). Then we suggest a complete method to go from the displacements to the itinerary : that is to say from the syntagm level to the discourse one. Finaly, we present $\pi\mathcal{R}$, a prototype that fully implements our approach. It can read a raw text route narrative and gives the interpretation of the itinerary described. It also displays a visual interpretation on a map.