École doctorale des sciences exactes et leurs applications
Escuela de Doctorado de la Universidad de Zaragoza

Automatic reconstruction of itineraries from descriptive texts

THÈSE

pour l'obtention du

Doctorat de l'Université de Pau de des Pays de l'Adour (France)
(mention Informatique)

et

Doctor por la Universidad de Zaragoza (España)
(Programa de Doctorado de Ingeniería de Sistemas e Informática)

par

Ludovic Moncla

Composition du jury

Rapporteurs :	Christophe Claramunt	Institut de Recherche de l'Ecole Navale (IRENav)
	Denis Maurel	LI, Université François Rabelais, Tours
	Ross Purves	University of Zurich
Examinateurs :	Philippe Muller	IRIT, Université Paul Sabatier, Toulouse
	Adeline Nazarenko	LIPN, Université Paris 13
Invité :	David Buscaldi	LIPN, Université Paris 13
Directeurs :	Mauro Gaio	LIUPPA, Université de Pau et des Pays de l'Adour
	Javier Nogueras Iso	DIIS, Universidad de Zaragoza
Co-Encadrant :	Sébastien Mustière	COGIT IGN, Université Paris-Est

Laboratoire d'Informatique de l'Université de Pau et des Pays de l'Adour - EA 3000
Departamento de Informática e Ingeniería de Sistemas, Universidad de Zaragoza
Laboratoire COGIT, IGN, Université Paris-Est

Acknowledgments

First of all, I wish to express my greatest thanks to my two supervisors Mauro Gaio and Javier Nogueras-Iso. Mauro Gaio for giving me the opportunity to do this PhD. His unconditional availability, encouragement and trust helped me a lot to accomplish this work. I thank him for all the interesting discussions we had, sharing ideas and talking about everything. Then, I thank Javier Nogueras-Iso for his availability, his patience, his hospitality and his precious advices. I also thank them for their great support during all stages of this work and for their help with administrative issues. My thanks go also to Sébastien Mustière for his support and his precious remarks. I thank all three of them for the time they have spent re-reading papers and documents including this dissertation.

I would thank the French National Mapping Agency (IGN) and the Communauté d'Agglomération Pau Pyrénées (CDAPP) for funding my PhD.

I am very grateful to Christophe Claramunt, Denis Maurel, Ross Purves, Adeline Nazarenko, Philippe Muller and David Buscaldi for accepting to be members of the jury. Also I would like to thank especially Christophe Claramunt, Denis Maurel and Ross Purves for accepting to review this dissertation, and to Philippe Muller for accepting to be present each year for the mid-term evaluations of my work, for his interest and his useful remarks.

I would like to thank all my colleagues from the University of Pau, researchers, teachers, staff members and more specifically current and former PhD students: Samson, Ehsan, Manzoor, Mamour and Tien.

I would also like to thank the members of the Advanced Information Systems Research Group (IAAA) of the Computer Science and Systems Engineering Department at the University of Zaragoza for their support and help during my stays in Zaragoza and especially to Walter Renteria-Agualimpia for our fruitful collaboration.

I would like to thank the members of the COGIT laboratory of IGN, for their support during my stays in Paris and especially Cécile Duchêne, Sidonie Christophe and Guillaume Touya for their encouragement during my talk at the GIScience conference in Vienna.

Also I want to thank my family for their encouragement and my friends who reminded me that there is more to life than academic research and for all the good times spent together.

Last but not least, special thanks to my fiancé, Camille, for her unconditional support, encouragement and understanding during these three years. Also thanks for lending a hand in making nice schemas for this dissertation and for my oral presentations. But the most important, I thank her for all those great moments we share together.

To my fiancé and family

Abstract

This PhD thesis is part of the research project 'PERDIDO', which aims at extracting and retrieving displacements from textual documents. This work was conducted in collaboration with the LIUPPA laboratory of the university of Pau (France), the Advanced Information Systems (IAAA) group of Universidad de Zaragoza (Spain) and the COGIT laboratory of IGN (France). The objective of this PhD is to propose a method for establishing a processing chain to support the geoparsing and geocoding of text documents describing events strongly linked with space.

We propose an approach for the automatic geocoding of itineraries described in natural language. Our proposal is divided into two main tasks. The first task aims at identifying and extracting information describing the itinerary in texts such as spatial named entities and expressions of displacement or perception. The second task deal with the reconstruction of the itinerary. Our proposal combines local information extracted using natural language processing and physical features extracted from external geographical sources such as gazetteers or datasets providing digital elevation models.

The geoparsing part is a Natural Language Processing approach which combines the use of part of speech and syntactico-semantic combined patterns (cascade of transducers) for the annotation of spatial named entities and expressions of displacement or perception. The main contribution in the first task of our approach is the toponym disambiguation which represents an important issue in Geographical Information Retrieval (GIR). We propose an unsupervised geocoding algorithm that takes profit of clustering techniques to provide a solution for disambiguating the toponyms found in gazetteers, and at the same time estimating the spatial footprint of those other fine-grain toponyms not found in gazetteers.

We propose a generic graph-based model for the automatic reconstruction of itineraries from texts, where each vertex represents a location and each edge represents a path between locations. Our model is original in that in addition to taking into account the classic elements (paths and waypoints), it allows to represent the other elements describing an itinerary, such as features seen or mentioned as landmarks. To build automatically this graph-based representation of the itinerary, our approach computes an informed spanning tree on a weighted graph. Each edge of the initial graph is weighted using a multi-criteria analysis approach combining qualitative and quantitative criteria. Criteria are based on information extracted from the text and information extracted from geographical sources. For instance, we compare information given in the text such as spatial relations describing orientation (e.g., going south) with the geographical coordinates of locations found in gazetteers.

Finally, according to the definition of an itinerary and the information used in natural language to describe itineraries, we propose a multi-scale markup langugage. This language relies on a core generic layer based on the Text Encoding and Interchange guidelines (TEI) which defines a standard for the representation of texts in digital form. We also define a second layer adding spatial semantics for encoding spatial and motion information.

Additionally, the rationale of the proposed approach has been verified with a set of experiments on a corpus of multilingual hiking descriptions (French, Spanish and Italian).

Keywords: Information Extraction, Automatic itinerary reconstruction, Natural Language Processing

Résumé

Cette thèse s'inscrit dans le cadre du projet PERDIDO dont les objectifs sont l'extraction et la reconstruction d'itinéraires à partir de documents textuels. Ces travaux ont été réalisés en collaboration entre le laboratoire LIUPPA de l'université de Pau et des Pays de l'Adour (France), l'équipe Systèmes d'Information Avancés (IAAA) de Universidad de Zaragoza (Espagne) et le laboratoire COGIT de l'IGN (France). Les objectifs de cette thèse sont de concevoir un système automatique permettant d'extraire, dans des récits de voyages ou des descriptions d'itinéraires, des déplacements, puis de les représenter sur une carte.

Nous proposons une approche automatique pour la représentation d'un itinéraire décrit en langage naturel. Notre approche est composée de deux tâches principales. La première tâche à pour rôle d'identifier et d'extraire les informations qui décrivent l'itinéraire dans le texte, comme par exemple les entités nommées de lieux et les expressions de déplacement ou de perception. La seconde tâche à pour objectif la reconstruction de l'itinéraire. Notre proposition combine l'utilisation d'informations extraites grâce au traitement automatique du langage ainsi que des données extraites de ressources géographiques externes (comme des gazetiers).

L'étape d'annotation d'informations spatiales est réalisée par une approche qui combine l'étiquetage morpho-syntaxique et des patrons lexico-syntaxiques (cascade de transducteurs) afin d'annoter des entités nommées spatiales et des expressions de déplacement ou de perception. Une première contribution au sein de la première tâche est la désambiguïsation des toponymes, qui est un problème encore mal résolu en NER et essentiel en recherche d'information géographique. Nous proposons un algorithme non-supervisé de géo-référencement basé sur une technique de clustering capable de proposer une solution pour désambiguïser les toponymes trouvés dans les ressources géographiques externes, et dans le même temps proposer une estimation de la localisation des toponymes non référencés.

Nous proposons un modèle de graphe générique pour la reconstruction automatique d'itinéraire, où chaque noeud représente un lieu et chaque segment représente un chemin reliant deux lieux. L'originalité de notre modèle est qu'en plus de tenir compte des éléments habituels (chemins et points de passage), il permet de représenter les autres éléments impliqués dans la description d'un itinéraire, comme par exemple les points de repères visuels. Un calcul d'arbre de recouvrement minimal à partir d'un graphe pondéré est utilisé pour obtenir automatiquement un itinéraire sous la forme d'un graphe. Chaque segment du graphe initial est pondéré en utilisant une méthode d'analyse multi-critère combinant des critères qualitatifs et des critères quantitatifs. La valeur des critères est déterminée à partir d'informations extraites du texte et d'informations provenant de ressources géographiques externes. Par exemple, nous combinons les informations issues du traitement automatique de la langue comme les relations spatiales décrivant une orientation (ex: se diriger vers le sud) avec les coordonnées géographiques des lieux trouvés dans les ressources pour déterminer la valeur du critère "relation spatiale".

De plus, à partir de la définition du concept d'itinéraire et des informations utilisées dans la langue pour décrire un itinéraire, nous avons modélisé un langage d'annotation d'information multi-couche. Ce language s'appuie sur une couche générique basée sur les recommendations du consortium TEI (Text Encoding and Interchange) et peut être adapté en plusieurs couches spécifiques ajoutant de la sémantique aux éléments et aux relations annotées.

Enfin, nous avons implémenté et évalué les différentes étapes de notre approche sur un corpus multilingue de descriptions de randonnées (Français, Espagnol et Italien).

Mots-clés: Extraction d'information, Reconstruction automatique d'itinéraire, Traitement Automatique du Language Naturel

Resumen

Esta tesis se inscribe dentro del marco del proyecto PERDIDO donde los objetivos son la extracción y reconstrucción de itinerarios a partir de documentos textuales. Este trabajo se ha realizado en colaboración entre el laboratorio LIUPPA de l'Université de Pau et des Pays de l'Adour (France), el grupo de Sistemas de Información Avanzados (IAAA) de la Universidad de Zaragoza y el laboratorio COGIT de l'IGN (France). El objetivo de esta tesis es concebir un sistema automático que permita extraer, a partir de guías de viaje o descripciones de itinerarios, los desplazamientos, además de representarlos sobre un mapa.

Se propone una aproximación para la representación automática de itinerarios descritos en lenguaje natural. Nuestra propuesta se divide en dos tareas principales. La primera pretende identificar y extraer de los textos describiendo itinerarios información como entidades espaciales y expresiones de desplazamiento o percepción. El objetivo de la segunda tarea es la reconstrucción del itinerario. Nuestra propuesta combina información local extraída gracias al procesamiento del lenguaje natural con datos extraídos de fuentes geográficas externas (por ejemplo, gazetteers).

La etapa de anotación de informaciones espaciales se realiza mediante una aproximación que combina el etiquetado morfo-sintáctico y los patrones léxico-sintácticos (cascada de transductores) con el fin de anotar entidades nombradas espaciales y expresiones de desplazamiento y percepción. Una primera contribución a la primera tarea es la desambiguación de topónimos, que es un problema todavía mal resuelto dentro del reconocimiento de entidades nombradas (Named Entity Recognition – NER) y esencial en la recuperación de información geográfica. Se plantea un algoritmo no supervisado de georrefenciación basado en una técnica de clustering capaz de proponer una solución para desambiguar los topónimos los topónimos encontrados en recursos geográficos externos, y al mismo tiempo, la localización de topónimos no referenciados.

Se propone un modelo de grafo genérico para la reconstrucción automática de itinerarios, donde cada nodo representa un lugar y cada arista representa un camino enlazando dos lugares. La originalidad de nuestro modelo es que además de tener en cuenta los elementos habituales (caminos y puntos del recorrido), permite representar otros elementos involucrados en la descripción de un itinerario, como por ejemplo los puntos de referencia visual. Se calcula de un árbol de recubrimiento mínimo a partir de un grafo ponderado para obtener automáticamente un itinerario bajo la forma de un grafo. Cada arista del grafo inicial se pondera mediante un método de análisis multicriterio que combina criterios cualitativos y cuantitativos. El valor de estos criterios se determina a partir de informaciones extraídas del texto e informaciones provenientes de recursos geográficos externos. Por ejemplo, se combinan las informaciones generadas por el procesamiento del lenguaje natural como las relaciones espaciales describiendo una orientación (ej: dirigirse hacia el sur) con las coordenadas geográficas de lugares encontrados dentro de los recursos para determinar el valor del criterio "relación espacial".

Además, a partir de la definición del concepto de itinerario y de las informaciones utilizadas en la lengua para describir un itinerario, se ha modelado un lenguaje de anotación de información espacial adaptado a la descripción de desplazamientos, apoyándonos en las recomendaciones del consorcio TEI (Text Encoding and Interchange).

Finalmente, se ha implementado y evaluado las diferentes etapas de nuestra aproximación sobre un corpus multilingüe de descripciones de senderos y excursiones (francés, español, italiano).

Palabras clave: Extraccion de Información, Reconstrucción automática de Itinerarios, Procesamiento del Lenguaje Natural

Contents

List of Figures

List of Tables

Chapter 1

Introduction

It's a dangerous business, Frodo, going out your door. You step onto the road, and if you don't keep your feet, there's no knowing where you might be swept off to.

— J.R.R. Tolkien, *The Lord of the Rings*

Contents

1.1 Motivation

Nowadays, with the rise of mobile mapping apps and navigation services which are available on phones, tablets, and smartwatches released with an embedded GPS, people get used to follow route instructions and record their route in their everyday life. For instance, a common behaviour is to ask for the nearest commodity or Point of Interest (POI) such as finding the nearest restaurant, hotel or museum. Furthermore, runners, bikers or hikers use apps[1] to plan, record and share their routes on social media. These apps map your route and may track your activity providing information such as geocoded route, distance covered, average pace and calories burned. However, considerable amounts of geographical data are still collected not in form of Geographic Information System (GIS) data but in natural language texts form, and are not adapted to new technologies and new behaviours. Moreover, a lot of documents describing travels or walks in different tourist sites are now available in digital form. For instance, just to provide an example in the French-Spanish transborder area, this kind of documents is very abundant thanks to different storage sites such as multimedia libraries. On the French side, the *Médiathèque Intercommunale à Dimension Régionale* (MIDR) of Pau[2] has digitalized more than 300,000 pages of documents describing mainly travel stories of the 19th century occurring in the Pyrenees. On the Spanish side, the *Sistema de Información del Patrimonio Cultural Aragonés* (SIPCA)[3] provides more than 500,000 documents from the Aragon's archives. Additionally, a lot of people describe and share their journeys (with daily descriptions, photos, etc.) on travel blogs, participative websites or social media. In the last few years, analysis of data coming from social media has become a challenge for researchers and data scientists (Sui and Goodchild, 2011) and particularly with the evolution of technologies and the increasing availability of geocoded data.

[1] Such as Runtastic, VTTrack, randogps, etc.
[2] http://mediatheques.agglo-pau.fr/
[3] http://www.sipca.es/

In the early nineties, Frank and Mark (1991) wrote "*It is conceivable that systems of the future might be able to assimilate and analyze explorer's journals such as Columbus' logs or the journals of Lewis and Clark, check them for consistency, and perhaps reach new inferences about the itineraries of their travels*". Since then, advances in automatic Natural Language Processing (NLP), processing and representation of geographical information, but also the explosion of open digital geographical resources, have made developing such systems now possible.

Although Artificial Intelligence (AI) and NLP have been widely studied for decades, they are still open fields of research which aim at the understanding of natural language by computer systems (e.g., question answering systems, machine translation, automatic summarization and natural language generation, etc.). Thus natural language understanding still remains a challenge with the objective to structure real-world information into computer-understandable data. With respect to our concern, this involves the understanding of human spatial cognition for automatic spatial representation and reasoning. Furthermore, the recent availability of textual descriptions associated with GIS data (e.g., GPS tracks) made now possible to design machine learning methods in order to build fully automatic systems for spatial representation and reasoning without introducing human interactions. A relevant example of such data is the case of hiking community websites in which volunteer hikers share a description of their hikes associated with real GPS tracks. This kind of data, considered as crowdsourcing, can be used to train and evaluate a fully automatic system dealing with NLP and GIS.

This PhD thesis has been developed under the framework of the research project called 'PERDIDO' (Project for Extracting and Retrieving DIsplacements from textual DOcuments). PERDIDO is a cross-disciplinary research project motivated by the concern of preservation and enhancement of the cultural heritage (e.g., old travelogues, explorer's journals) and also tied to digital preservation concerns. Digital preservation is defined as a set of processes oriented towards facilitating the conservation and access to information in digital format. According to Rothenberg (2000), the goal of preservation is to allow future users to retrieve, access, decipher, view, interpret, understand, and experience documents, data, and records in meaningful and valid (i.e., authentic) ways. There are different strategies for long term preservation such as backup policies, migration of formats, or emulation of technology. However, one of the most promising strategies for documental cultural heritage consists in the normalization or information extraction, whose objective is to extract the content of documents using markup languages in order to make the documents independent from proprietary formats or specific software for document edition.

The objective of this PhD is to propose a method for establishing an automatic processing chain to support the geoparsing and geocoding of text documents describing events strongly linked with space and motion. It involves three main fields of research: computer science, linguistics and geography (Figure 1.1).

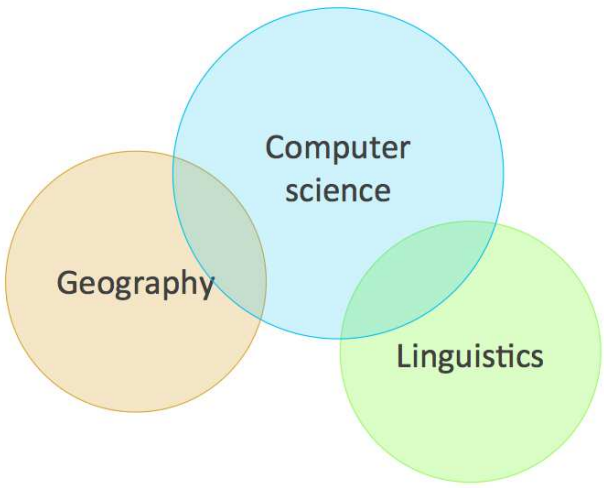


Figure 1.1: Interdisciplinary fields involved in this thesis

Linguistics and geography provide knowledge and methods for spatial cognition and spatial analysis, and computer science provides mechanisms for automatic computing, analysis, storage and representation.

Computer science also makes the connection between linguistics and geography thanks to NLP and GIS. The main goal of our work is to automatically extract and interpret information describing places and motion in textual documents in order to reconstruct and map the described itinerary. The main objective is thus to turn textual information into GIS data and to assign a geocoded representation to a textual document. This problem of connecting texts with geographic space is illustrated in Figure 1.2.

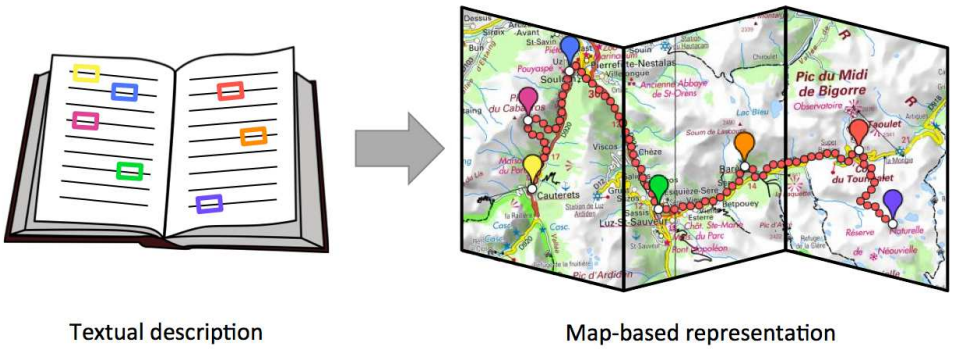


Figure 1.2: From text to map: geospatial information extraction and representation

1.2 Challenges

The main challenge of this PhD is to connect text with geographic space and to provide a map-based representation of itineraries described in textual documents. Our proposal is thus to combine the use of information expressed in texts and information found in external geographical resources to build a geocoded representation of an itinerary. Furthermore, an itinerary does not refer only to the route traveled but refers also to the context of the trip (description of landmarks and landscapes, purpose of the displacement, etc.). Vasardani et al. (2013) propose an approach for the reconstruction of the environment from a verbal description (translating spatial information into sketch maps). Although our work has some similarities with their proposal, our goal is different. Our approach is focused on the automatic reconstruction of routes and transcribed them in their geographical area of achievement (identifying waypoints and routes by interpreting spatial information in geographical context). The obtained geocoded representation of the itinerary may be used for geoindexing documents like old travelogues (Lesbegueries et al., 2006) or legal texts (Yahiaoui et al., 2014). It also paves the way to a further analysis of other information contained in the text like geocoding related events or landmarks (Li et al., 2014), or to an analysis of itineraries for searching spatial patterns (Laube et al., 2005). This kind of work may also have a great interest in digital humanities and in particular in spatial humanities (Gregory et al., 2015) such as displaying cultural heritage using textual analysis. Globally, one of the objectives of this project is to pave the way new interactions with textual data and add further knowledge (such as descriptions, demographic data, images, etc.) to improve the user experience and enhance the visibility and interest of textual documents.

One of the main tasks of NLP is text analytics (also known as text mining), which involves information retrieval and lexical analysis. The objective of this task is to turn unstructured text into machine-readable structured data. Moreover, geographical information text mining lies in extracting place names as mentioned in a first notable work that can be attributed to Woodruff and Plaunt (1994). Nowadays, most approaches focus on extracting explicit geographic data from text and associating extracted location references with other information resources (Jones et al., 2008). For that purpose, we need to identify how human conceptualize space and motion in the language in order to identify which elements in the texts are involved in the description of itinerary. The problem of the automatic reconstruction of itineraries from texts addresses several challenges. The first challenge is to identify relevant information expressed in texts (geospatial semantic information) in order identify waypoints and find the sequence that provides the order in which the waypoints are visited during the displacement. Furthermore, considering a fully

automatic process, another challenge is to turn text into structured data, i.e. to obtain an annotated text described with a formal markup language where tags could be a combination of part-of-speech and geo-semantic information. This challenge involves the annotation of spatial information (i.e., geoparsing) from natural language descriptions and the formalization of relationships between the elements used to express space and motion in the language. Additionally another challenge addresses the problem of geocoding (also known as toponym resolution), which aims to assign explicit georeferences to place names (i.e., geographical coordinates). Commonly this task uses external geographical resources to find referent locations for a given place name. Furthermore, this challenge may take profit of the increasing number of available geographical resources (e.g., gazetteers, any source of geo-referenceable open data, etc.). Finally, another challenge is to build a geocoded representation of the itinerary. This task is facilitated by the result of the first challenge of finding the sequence of waypoints. Then, to make links between all these tasks we also need a way to store and share information of text documents and additional data.

To summarize, the specific objectives of this thesis include:

- to identify how itineraries are described in natural language;
- to propose a method for the automatic reconstruction of itineraries using information expressed in text and additional data found in geographical resources;
- to propose an approach for the automatic annotation of geospatial information from texts;
- to propose a method for toponym disambiguation;
- to propose a solution for the storage of spatial and spatio-temporal information extracted from texts.

1.3 Contributions

We propose to divide the problem of the automatic reconstruction of itineraries from texts into three sub-problems: the annotation of geospatial information in texts (geoparsing), the toponym resolution (geocoding) and the reconstruction of the itinerary itself. Although we focus the main contribution of this dissertation on the automatic reconstruction of itinerary, this PhD also proposes solutions for the other sub-problems. Additionally, we propose a specialised markup language for the information storage. The relationships between the contributions are shown in Figure 1.3.

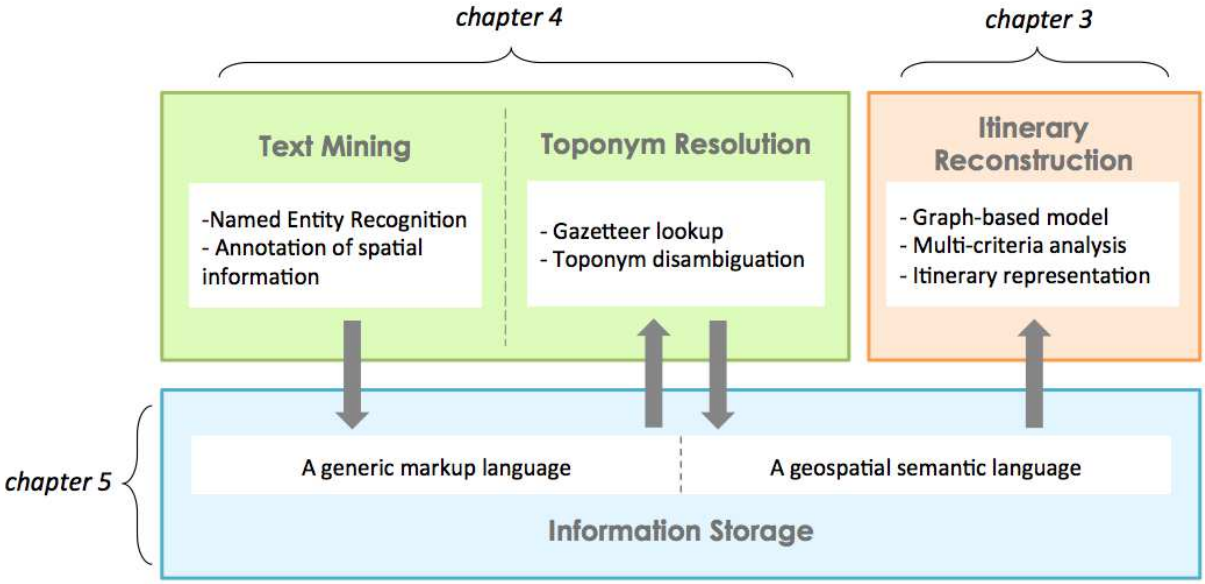


Figure 1.3: Relation between the contributions of this thesis

Our main contribution is described in Chapter 3 and refers to the reconstruction of itineraries using

information extracted from textual descriptions of the itineraries and additional data coming from external geographical resources. It addresses the problem of distinguishing waypoints from other types of locations and identifying the sequence order in which waypoints are visited during the displacement and build an approximation of the representation of the route of the itinerary. Our approach is based on a domain-specific corpus analysis. In order to solve the problem of itinerary reconstruction from text, we first define and analyse all the components of an itinerary and how they are expressed in natural language.

Our second contribution is described in Chapter 4 and addresses the problem of automatically annotating passages in the text that describe the various trips making up the itinerary. According to the analysis of elements used in description of itineraries described in Chapter 3, this problem involves the annotation, the resolution and the disambiguation of toponyms but also the annotation of spatial information such as spatial relations, expressions of motion and perception. Thus we propose a method for establishing a processing chain to support the geoparsing and geocoding of text documents describing itineraries.

Finally, we describe in Chapter 5 our proposal of a multi-scale markup language, which was applied for the annotation of geospatial information from textual descriptions of itineraries.

Each contribution described in this dissertation (Figure 1.3) can be seen as an independent task. For instance, the main contribution described in Chapter 3 may rely on manually annotated texts. Additionally, annotations provided by the approach described in Chapter 4 can be used for another purpose than reconstructing an itinerary. However, we have developed a fully automatic processing chain to show the feasibility of our proposal and to evaluate our approach on real data.

1.4 The Structure of the Thesis

The remaining chapters of this thesis are organised as follows.

- Chapter 2 provides an overview of works dealing with spatial cognition and spatial knowledge in language. This chapter also gives some background about natural language processing, markup languages and geographical resources.
- Chapter 3 describes our proposal for the automatic geocoding of itineraries. It addresses the problem of identifying waypoints and describes how we propose to interpret geospatial semantic information from annotated texts.
- Chapter 4 describes the proposed method for the automatic annotation of information in texts and describes also our proposal of a toponym disambiguation approach.
- Chapter 5 describes our proposal of a multi-scale markup language based on a TEI-compliant core layer and applied to geospatial semantic annotations, and introduces the notion of Expanded Named Entity.
- Chapter 6 demonstrates the feasibility of the contributions of this PhD with the integration of the proposed methods in a web-based architecture, from the automatic annotation of information to the automatic reconstruction of itineraries.
- Chapter 7 describes the PERDIDO gold-standard corpus and the evaluation results of the proposed methods. Each module of the proposed processing chain is evaluated and results are discussed for each language (French, Spanish and Italian).
- Finally, Chapter 8 summarises the contributions of this thesis and concludes with some suggestions for future work.

Chapter 2

Background and Related Work

An investment in knowledge always pays the best interest.

— Benjamin Franklin

Contents

2.1 Introduction

According to the problem of automatic reconstruction of itinerary from texts, this thesis involves connections between several disciplines of natural language processing: spatial cognition, discourse analysis (or text mining) and spatial analysis. Since the early 1990's few significant research, in the field of description of routes, have clarified relationships between these domains. The appropriate use of 'spatial language' thus depends on the addressee's capacity to translate linear linguistic information into multi-dimensional

internal representations that incorporate relations between the described objects (Denis, 1997). But many aspects still need clarifications, in particular dependencies between linguistic expressions, reasoning and visual information (Landau and Jackendoff, 1993; Jackendoff, 2012).

In this chapter, we describe the background and relevant works with our concern of automatic reconstruction of itinerary from texts. This problem involves several sub-problems. Firstly, we need to know how itineraries are conceptualized and described in texts. Then we need mechanisms for automatically extracting relevant information from texts, and finally, we need to interpret this information in order to reconstruct the itinerary.

The remainder of this chapter is structured as follows. Section 2.2 provides an overview of works dealing with spatial cognition and shows how humans communicate and conceptualize spatial knowledge. More specifically it shows how space and motion are represented in language and how it is possible to make links between cognition, language and Geographic Information System (GIS). Section 2.3 provides an overview of NLP methods for named entity recognition and classification and for toponym disambiguation. Then, Section 2.4 describes markup languages for encoding spatial and spatio-temporal information in texts. Section 2.5 describes the characteristics of geographical resources and describes some relevant gazetteers. Finally, Section 2.6 summarises and concludes this chapter.

2.2 Itinerary Descriptions

2.2.1 Overview

With the rise of new needs and behaviours (e.g., route planning and tourist applications), the democratisation of devices equipped with GPS and the wide availability of geographical information, the notion of itinerary is being studied more and more. For instance, Hao et al. (2010) put forward a probabilistic model to identify place elements taken from travelogues. The aim of this work being to improve the tourist experience, providing them with information or recommendations about the places they are visiting. Zhang et al. (2010) use learning methods to extract the three elements they consider to be the most important in an itinerary: origin, destination and the path taken (instructions). They work on a corpus of web pages where instructions giving directions can be found. Other studies (Breier, 2013) have focused on ancient documents with the aim of finding and modelling historical roads that no longer exist.

An itinerary description informs about the displacement sequence of a person along a path. It provides action plans (route instructions) and includes directive and descriptive statements. Directives provide information about the direction and distance of travel and descriptive statements provide information about the environment and the location of places (landmarks) (Allen, 1997). Furthermore, route directions describe not only places and the path of the displacement, but they also refer to landmarks located along the route (Michon and Denis, 2001) that are supposed to be seen during the displacement such as buildings (e.g., church, school, shop, etc.) and natural landmarks (e.g., mountain peak, lake, river, etc.).

Denis (1997) has proposed a general framework for the analysis of natural route descriptions, taking into account the spatial knowledge expressed in natural language. They consider three cognitive operations as being involved in route description discourses, i.e. an internal representation of the environment, the planning of a route and the formulation of the procedure that the user should execute. Several cognitive models of space, such as the one proposed by Lynch (1960), defined landmarks, nodes, and paths as the most important components in route descriptions. A landmark is defined as an environmental feature playing the role of a point of reference. Tom and Denis also show the important role of landmarks in the description of routes in comparison with the use of street names (Tom and Denis, 2004). A node refers to a place involved in the path of the displacement where actions and decisions are taken, it is also known as 'decision point' (e.g., turn left after the church), and paths refer to pathways such as streets or trails. Denis (1997) consider five classes of components that are used to construct abstract descriptions of routes, i.e. prescription of actions without referring to any landmark, prescription of actions with reference to landmarks, reference to landmarks without referring to any associated action, description of landmarks and commentaries.

Although studies about wayfinding and route descriptions are mainly focused on urban areas, few studies (Brosset et al., 2008; Sarjakoski et al., 2011) aim at providing knowledge on human verbal de-

scriptions of route and landmarks on natural environment and some others consider indoor navigation (MacMahon et al., 2006). Brosset et al. (2008) made a comparison study between wayfinding descriptions made in urban areas and those made in natural environments. They highlight the importance of landmarks, actions and the role of land features and topography for descriptions in natural environments. Their experiments are based on wayfinding descriptions made by experienced orienteers who were asked to remember and communicate their route at the end of a foot orienteering race. Their experiments confirm that landmarks are the most important component and that spatial orientation terms are often used while there are only few references to temporal aspects, which play a very minor role in wayfinding descriptions. Furthermore, they also show that 3D constructs, expressed by verbs (e.g., "to climb") and nouns (e.g., "valley") are very relevant but not always available. This can be explained by the fact that 3D descriptions are less precise and more complex to memorize and communicate. Finally, Brosset et al. (2008) also show the importance in navigation descriptions of land features (e.g., open land, forest, vegetation) associated with running performance (e.g., easy, slow, difficult). Sarjakoski et al. (2011) propose a study of verbal descriptions of hiking made by pedestrians in a national park. They classify expressions made by the participants according to the classification proposed by (Denis, 1997) (i.e., action, action and landmark, landmark, landmark description, and comment). They describe the distribution of classes of expressions in route descriptions and show that landmark-related expressions are the most frequently used and that action-related expressions are more frequent at decision points and are related to landmarks. Furthermore, they also show that comments depend on the season (winter or summer), are frequent and mention temporary phenomena such as snow, flowers, leaves on trees or birds.

Götze and Boye (2015) propose a study comparing landmark-based instructions with relative direction instructions (left, right, straight) in an urban area. It shows that relative direction instructions (e.g., 'turn left') and direction and landmark combined (e.g., 'go straight toward the school') work better in some situations such as when the decision point has a simple configuration (e.g., T-intersection or four-way intersection where all streets meet at right angles). Depending the cases relative direction instructions can imply more ambiguities than landmark-based instructions and vice versa.

Route descriptions may be very different from one author (or speaker) to another. The results of route description experiments (Denis, 1997; Brosset et al., 2008; Sarjakoski et al., 2011; Götze and Boye, 2015) highlight major differences for the same itinerary between the descriptions provided such as the length of descriptions and the number and type of landmarks. Sometimes these differences can be explained by the fact that people who already know the place where the itinerary occurs do not detect all the possible landmarks.

Furthermore, modelling and analysing itineraries lies in the general framework of Time-Geography (Winter and Raubal, 2006) and received much attention in the literature. In particular, Spaccapietra et al. (2008) propose a pattern for conceptually modelling itineraries and its implementation to store and query this model in a DBMS.

With respect to the problem of spatial knowledge representation Werner et al. (2000) introduced the concept of Route Graphs. This general concept refers to the union of routes and allows different routes to be integrated into a graph-like structure. Furthermore, Werner et al. (2000) defined a route as a concatenation of a sequence of route segments from one place to another, and a route segment consists of two places: an *entry* and an *exit*, which are connected by a *course*. Furthermore, Spatial Semantic Hierarchy (SSH) introduced by (Kuipers, 2000) is a model of cognitive spatial knowledge representing both qualitative and quantitative knowledge. It aims to model human cognitive maps, which can be used for robot exploration and map-building. It provides different levels of spatial knowledge representation based on ontology (i.e., sensory, control, causal, topological and metrical). The hierarchy of representations can be particularly useful for incomplete knowledge. Furthermore, according to Kuipers (2000), verbal route directions correspond to knowledge expressed in the SSH causal level.

A large number of studies have looked into trajectories (Kim et al., 2009; Yuan and Raubal, 2012), with a focus on the movement of the mobile objects (animal migrations, aeroplanes, ships, pedestrians, etc.). Although these works on trajectory are not based on verbal descriptions, the concepts explored in these studies can be considered similar to those applicable to itineraries.

Hiking descriptions can be considered to be a variant of route instructions, which is a composite of several forms of discourse, i.e., procedural discourse and descriptive discourse. A description of itinerary describes locations and motion events between locations. Then, as we have seen, locations may refer to

places reached during the displacement (waypoints) or to places used as visual landmarks. Furthermore, waypoints, route, landmarks, and actions are the main components used to described itineraries (Lynch, 1960; Denis, 1997). Descriptions of itineraries may refer to displacements in both urban and natural environments and with different land features within the routes. Moreover, most of the waypoints and landmarks are described in the text using place names, even if in natural environments these place names may be fine-grain.

2.2.2 Toponyms

A place is defined as a geographical area (i.e., a portion of space) with definite or indefinite boundaries and can be categorised (e.g., city, forest, building, etc.) (Borillo, 1998). Entities identifying places and landmarks in natural language can be named and may be represented with more or less accuracy on a map. For instance, a city may be represented as a point or as a polygon depending the scale of interest.

Spatial named entities are also called toponyms. The word toponym is derived from the Greek *tópos* (place) and *ónoma* (name) and is the name used by geographers to define geographical named entities.

The United Nations Group of Experts on Geographical Names (UNGEGN) provides Toponymic Guidelines[4] for several countries. The National Institute of Geographic and Forest Information (IGN[5]) classifies toponyms in two categories: official and non-official names (Lejeune et al., 2003). Official names, in France, refer to administrative entities such as country, state, region, city, or village. However, most toponyms are non-official names and are divided in several categories:

- non-administrative populated places such as hamlets (settlements);
- wooded areas and localities (*lieux-dits*), which refer to small geographic areas bearing a traditional name;
- oronyms, which refer to toponyms in mountain areas (e.g., le Mont Blanc, Pic du Midi);
- hydronyms, which refer to the names of water bodies (e.g., Lac de la Plagne, River Thames);
- odonyms, which refer to street names;
- and details of landscape or human activity (e.g., belvedere, cave).

Typography of toponyms must follow some specific rules. It is important to identify and specify the rules used for the construction of place names in order to propose an automatic recognition process of place names in written texts. These rules are not the same for official and non-offcial names. For instance, in France, official compound names must be written with hyphens (e.g., Saint-Etienne, Nord-Pas-de-Calais, Côte-d'Or), whereas non-official compound names should not be written with hyphens (e.g., le Mont Blanc, le Pic du Midi). Furthermore, concerning official names, substantives and adjectives must start with an uppercase (e.g., Noisy-le-Grand) and prepositions, conjunctions and articles beginning official compound names must start with an uppercase (e.g., La Rochelle, Mont-de-Marsan). On the contrary, concerning non-official names, articles, even located at the beginning of the name must start with a lowercase. Furthermore, prepositions, adverbs and conjunctions must start with an uppercase, if located at the beginning of non-official compound names and with a lowercase otherwise.

Toponyms are stored in geographical resources such as geographical databases, commonly gazetteers. These geographical resources are described in Section 2.5.

2.2.3 Spatial Relations in Language

Many linguistic and cognitive studies (Bloom, 1994; O'Keefe, 1996; Aurnague et al., 2010) deal with space and language and more specifically with spatial relations with a view to describe the object to be located and the point of reference used. Moreover, in many research studies on 'spatial language' Miller and Johnson-Laird (1976); Talmy (1985) and Vasardani et al. (2013) and many others, the linguistic representation of the place of an object requires three elements: the reference object, the object to be located and the spatial relation between them. In other words the cognitive model of space frequently lies on the one proposed by Lynch (1960). For instance, Talmy (1985) has proposed to describe two objects

[4] http://unstats.un.org/unsd/geoinfo/ungegn/toponymic.html
[5] Institut National de l'Information Géographique et Forestière: http://www.ign.fr/

related in space with the terms *figure* and *ground*. The *figure* represents the object of the description and the *ground* the reference. With the same idea, Vandeloise (1986) proposed the terms *target* and *site* and Borillo (1998) the terms *concrete entity* and *spatial reference* to describe the expression of spatial relations between two entities. Spatial relations are most of the time expressed by closed-class terms (i.e. functional categories of words) such as prepositions (e.g., 'at', 'in', 'near', etc.) in particular for topological constraints. Whereas metric relations such as distance are expressed in open-class terms (i.e. lexical categories of words) (Tversky and Lee, 1998).

The meaning of spatial relations such as 'near' or 'approach' depends on the nature of figures and grounds (Morrow and Clark, 1988). Asymmetry between *figure* and *ground* is based on different properties (Vandeloise, 1986), such as: size, visibility, salience, fixity or mobility, etc. These properties are close to the concepts of Gestalt's principles theory: continuity, similarity/complementarity and proximity (Gaio and Madelaine, 1996; Tversky and Lee, 1998). In sentences (1), (2) and (3), figures are 'statue', 'car', and 'book' and grounds are 'church', 'river' and 'table'.

(1) La statue est à l'avant de l'église.
The statue is in front of the church.

(2) La voiture est près de la rivièere.
The car is close to the river.

(3) Le livre est sur la table.
The book is on the table.

Furthermore, in natural languages the absolute location of an entity with an unknown location is defined using relations with the location of another entity with a known location. Levinson (1996, 2003) and Frank (1998) use the concept of 'frame of reference' to express the location of an object. They proposed three types of spatial frame of reference: intrinsic, absolute and relative. Intrinsic and absolute frame of references are binary spatial relations. Intrinsic relations define the location of the figure in relation to a part of the ground (see example (1)). Absolute relations use mostly cardinal directions to identify the relative location of an object (see example (4)). Relative frame of reference is a ternary relation in which the location of the object (figure) is in relation to both the viewpoint of the narrator and the location of another object (ground) (see example (5)). Frames of reference describe the geometric and linguistic principles that are used for the perspective representation of the environment described in natural language (Frank, 1998).

(4) Le refuge est au nord du lac.
The refuge is to the north of the lake.

(5) Le refuge est à gauche du lac.
The refuge is to the left of the lake.

Furthermore, (Aurnague and Vieu, 2015) propose a survey of approaches dealing with locative adpositions and make a comparison between geometrical approaches (such as the ones proposed by Talmy (2000) and Landau and Jackendoff (1993)) and functional aspects introduced by Vandeloise (1986). (Aurnague and Vieu, 2015) also show that function and geometry can be regarded as complementary tools through the study of several languages (French, Basque and Yuhup). They propose a three-level approach (geometrical, functional and pragmatic) with functional properties going together with geometrical constraints on entities and their relations. With regard to our concern, in this thesis we also consider both geometric and functional aspects of spatial expressions.

Lesbegueries (2007) proposes a cognitive model to define spatial entities with a geographical point of view. With the same idea as for temporal entities (Allen, 1983), Lesbegueries distinguishes two types of spatial entities: absolute and relative. Absolute spatial entities consist in a named entity allowing a geo-location (example (6)). These named entities may be associated with one or more terms (example (7)), called spatial indicators, contained within a geographic lexicon (e.g. river, lake, city, church, etc.). Relative spatial entities (examples (8) and (9)) are defined using spatial relations with at least one absolute spatial entity. Lesbegueries considered five spatial relation types: orientation (e.g. *south of*), distance (e.g. *at 20 kilometers of*), adjacency (e.g. *next to*), inclusion (e.g. *in*), and union or intersection relations between two spatial entities.

(6) à Pau
in Pau

(7) Château de Pau
Pau Castle

(8) proche de Pau
near Pau

(9) entre Pau et Saragosse
between Pau and Zaragoza

Kemmerer (2005) shows that languages such as English use the same prepositions to describe both spatial and temporal relationships (e.g., in the forest, in the morning). Furthermore, according to Kemmerer (2005): "With respect to space–time parallelisms, the Metaphoric Mapping Theory maintains that the three-dimensional domain of space is inherently more concrete and richly organized than the one-dimensional domain of time, so the relational structures of temporal concepts are given greater coherence by being aligned with the relational structures of spatial schemas through the application of a TIME IS SPACE metaphor".

Furthermore, as we have seen, in itinerary analysis not only spatial entities are important, but also their associated spatial relations. These enable spatial entities to be specified locally, as well as allowing the notion of movement between the different entities to be expressed. Spatial relations describe a relationship between the traveller and environmental features or between different environmental features.

In the current work we are mainly focused on spatial relationships and spatial reasoning. There are two types of approaches for spatial reasoning: qualitative and quantitative. Additionally, spatial relations are classified into three categories: topological, projective and metric relations.

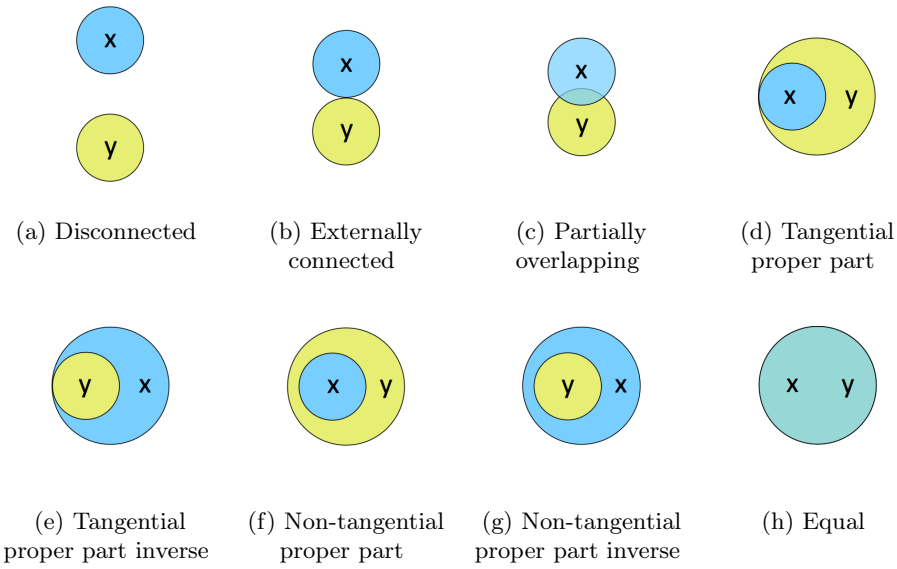
Qualitative approaches deal with spatial knowledge without using numerical computation, as opposed to quantitative approaches. Furthermore, qualitative approaches are considered to be closer to how space is defined in humans' cognition and representation (Renz, 2002). For instance, spatial relations expressed in natural language are most of the time qualitative, particularly in route descriptions (e.g., 'go straight', 'turn left after passing the church'). Moreover, spatial reasoning can be done by exploiting the combination of relations.

Topological Relations

Topological relations describe relationships between spatial regions rather than points and usually refer to relations of adjacency or inclusion between spatial regions (e.g. in, cross). Randell et al. (1992) proposed the well-known Region Connection Calculus (RCC). This approach is based on binary topological relations. They proposed a set of eight relations denoted *RCC-8* as *base relations*. Base relations are defined by Renz as a set of jointly exhaustive and pairwise disjoint relations (Renz, 2002). Relations of the *RCC-8* model may be extended to all other topological relations defined in the RCC theory. Examples for these relations are shown in Figure 2.1.

Egenhofer (1991) and Egenhofer and Franzosa (1991) have also defined a model for topological relations called the *9-intersection-model* which defines topological relations such as overlap or inclusion by comparing the intersection of the interior, the exterior, and the boundary of different spatial features of any dimensionality. Although the *RCC-8* and the *9-intersection model* are built using different approaches, they define exactly the same set of topological relations. Clementini and Cohn (2014) introduce the *RCC*-9* which is an extension of *RCC-8*, based on nine topological relations and capable of modelling topological relations between generic spatial features of dimension 2, 1 or 0, without forcing an interpretation in terms of regions, lines, or points. They introduce two new relations to express boundaries and *cross* relation.

Egenhofer and Shariff (1998) proposed a formal model to describe the semantics of spatial relations expressed in natural language. This proposal is a refinement of the *9-intersection-model* proposed for topological relations, and give details in the form of splitting and closeness ratios. This model was only developed for relations between a region and a line and was calibrated for 64 spatial English-language terms. More recently, Nedas et al. (2007) refine this model to specify the geometry of line-line relations.



(a) Disconnected (b) Externally connected (c) Partially overlapping (d) Tangential proper part

(e) Tangential proper part inverse (f) Non-tangential proper part (g) Non-tangential proper part inverse (h) Equal

Figure 2.1: The eight base relations of *RCC-8*

Projective Relations

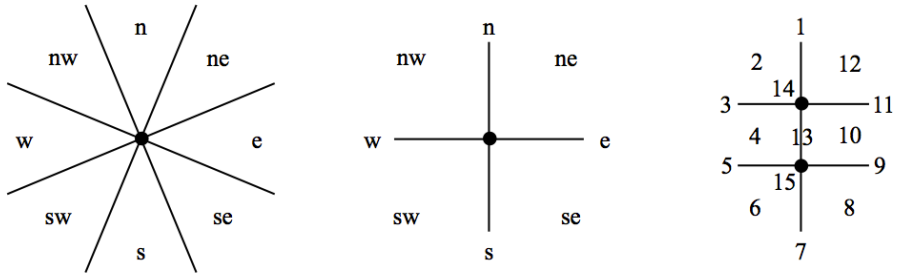
Projective relations may be qualitatively defined in projective geometry and is in between topology and metrics. They attempt to formalize relations expressed in natural language (Clementini, 2009) by orientation (e.g. *in the direction of*) and cardinal relations such as: *right of*, *in front of*, *between*, *along*, *in the suburbs of*, *north of* (e.g. phrases (10) and (11)). Projective relations are considered as qualitative, because they do not rely on an Euclidean representation when involved in a reasoning process. Unlike topological approaches, orientation is a ternary relationship depending on figure (located object), ground (reference object) and frame of reference (third spatial entity or a direction). There are three different kinds of frames of reference, extrinsic, intrinsic, and deictic, depending if the orientation is given by external factors, inherent properties or external point of view (Hernandez, 1994). Specific models have been proposed for orientation relations by Freksa (1992); Hernández (1993), and cardinal directions by Frank (1991); Ligozat (1998).

(10) Prendre à l'Ouest le chemin jusqu'à la rivière Mende.
Take to the West the way down to the River Mende.

(11) Desde la basílica nos dirigimos hacia el Puente de Piedra.
From the basilica we head towards the Stone Bridge.

Cardinal relations refer to the terms 'north', 'east', 'south', and 'west'. However, orientation relationship may also be expressed with the terms 'front', 'right', 'back', and 'left' in a local space. Furthermore, Frank (1991) distinguishes two methods for spatial reasoning with orientation relations, the 'projection-based' method (Fig. 2.2b) and the 'cone-based' method (Fig. 2.2a) . The projection-based method also called 'cardinal algebra' by Ligozat (1998), refers to a simple projection-based calculus of directions (Fig. 2.2b) with nine basic cardinal relations (n, ne, e, se, s, sw, w, nw, eq). Freksa (1992) proposed another approach of spatial reasoning based on relative orientation information about spatial environments called the Double Cross Calculus (DCC) (Fig. 2.2c). This approach used the notion of conceptual neighbourhood of spatial relations and path knowledge (Zimmermann, 1993; Zimmermann and Freksa, 1996) to define 15 base relations.

Distance relations may also be considered as qualitative (e.g., 'very close', 'close', 'far', etc.). For instance, Clementini et al. (1997) proposed a composition algorithm of a cone-based orientation approach and absolute distance relations for spatial reasoning.



(a) Cone-based (b) Projection-based (c) Double-cross

Figure 2.2: Orientation and cardinal relations between points (Figure 3.1 from (Renz, 2002).)

Metric Relations

Metric relations depend on measures, such as the expression of distances (e.g., phrases (12) and (13)) and are based on Euclidean representation. Furthermore, metric relations, such as the distance between two points, are generally considered to have a quantitative nature, such as in the approach proposed by Berretti et al. (2003). Unlike qualitative relations, quantitative relations give precise measurements and locations of spatial objects.

(12) Continuer tout droit sur 400 m.
Continue straight for 400 m.

(13) Ci si incammina sulla strada per circa 500 m.
Nous marchons sur la route pendant environ 500 m.

2.2.4 Motion Expression

Itineraries describe displacements between places using motion expressions or spatial relations. Syntactic parts of speech, in particular verbs, characterise a motion event. Many linguistic studies such as Boons (1987), Slobin (1996), Aurnague (2011) and Kokashvili (2012) have highlighted the importance of the use of motion verbs in language, especially in Romance languages (Talmy, 2000).

All aspects of a spatial scene are not represented by the language (Talmy, 1983), and according to Talmy (1985, 2000), a motion event is characterised by different conceptual components: a movement ('Motion'), a displaced object ('Figure'), a setting ('Ground'), a trajectory ('Path') and a 'Manner'. Talmy classified languages into two categories: *verb-framed language* and *satellite-framed language*. In verb-framed languages such as French or Spanish, the *Path* is expressed by verbs and the *Manner* by adverb phrases. In such languages, there are a lot of verbs referring to the direction such as *entrer* (go in), *sortir* (go out), *monter* (go up), *descendre* (go down). In satellite-framed languages such as English or German, the Path is expressed by satellites (in, out, up, down) and verbs refer usually to the mode of travel such as walking, running, swimming, etc.

However, languages are not fully part of one category or the other (Pourcel and Kopecka, 2005). For instance, in English language, which is mostly a satellite-framed language, there are motion verbs such as *enter*, *exit*, *ascend*, *descend* that refer both to Motion and Path. On the contrary, in verb-framed languages there are also some satellite-framed expressions such as *partir de* (to leave), *partir à* (to go) where the path is encoded in the French prepositions 'de' and 'à'. Furthermore, Pourcel and Kopecka (2005) give examples of a satellite-frame pattern (14) and a verb-frame pattern (15) to illustrate the dual typology for motion encoding proposed by Talmy and show the preferential lexicalisation of Path and Manner of motion in both satellite-framed and verb-framed languages.

(14)

Subject$_{(Figure)}$	Verb$_{(Manner)}$	Satellite$_{(Path)}$	Object$_{(Ground)}$
Julie	ran	across	the street.

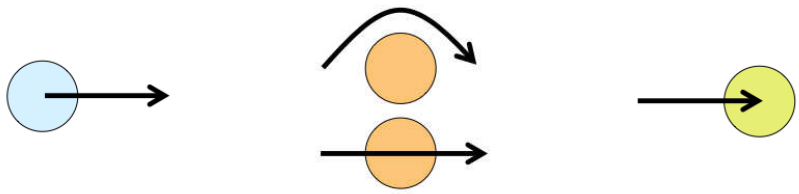
(15)

Subject$_{(Figure)}$	Verb$_{(Path)}$	Object$_{(Ground)}$	Gerund$_{(Manner)}$
Julie	traversa	la rue	en courant.
Julie	crossed	the street	running.

'Julie ran across the street.'

Egan (2015) proposes a comparative study of Path and Manner encoding according to English and French translations of Norwegian predications of motion events containing path prepositions (between, through and over). Berthele (2004) makes a cross-linguistic corpus analysis of Swiss German, German and French motion verbs based on a more fine-grained typology of motion verbs proposed by WäLchli (2001). Furthermore, Iacobini (2009) shows the important role of dialect analysis to explain the emergence of Italian phrasal verbs. They used the six cardinal kinds of displacement proposed by WäLchli (2001): *AD* (F displace to G), *IN* (F displace into G), *SUPER* (F displace up G), *AB* (F displace away from G), *EX* (F displace out to G), *DE* (F displace down G) . The results show that two typological categories are not sufficient and that a satellite-framed language does not express more Manner than a verb-framed language.

Then, with respect to the role played by verbs in the expression of motion, Boons (1987) proposed to categorise motion verbs according to the aspectual properties of movement, called hereafter 'aspectual polarity' (Fig. 2.3). The three polarities are initial (i.e., to leave), median (i.e. to cross) and final (i.e. to arrive). From the aspectual polarity classification, Laur (1991, 1993) shows the importance of the prepositions associated with motion verbs. Laur distinguishes two categories of spatial prepositions: prepositions of location (e.g., at, in, on) and preposition of direction (e.g., to, from, towards). Prepositions of direction have also an aspectual polarity like motion verbs. For instance *depuis* (from) is initial, *par* (through) is median and *jusqu'à* (up to) is final. Laur distinguishes also the motion verbs and the posture verbs. Motion verbs (e.g., go, walk, run) involve a change of location from a place to another, whereas posture verbs (e.g., stand, lie, sit) refer to static actions or a change of state.



(a) initial (b) median (c) final

Figure 2.3: Aspectual polarity of motion verbs

Without changing the intrinsic polarity of the verb, the preposition can change what would be called the *focus* of the displacement. More specifically, the association of a motion verb with a preposition of place (e.g., *from*, *in*, *at*, *to*, *by*, etc.) can change the focus of the displacement to take on the polarity of the preposition instead of the verb. Let us take the verb *to leave* for example. Alone or in association with the preposition 'from', the focus would be considered with initial polarity, but if used with the preposition 'for', the focus would then be considered as having final polarity. Undeniably, 'leaving from Paris' and 'leaving for Paris' are two expressions with a radical opposite meaning. If we consider the role played by the name, in one case, the place name is the origin of the displacement, and in the other case the place name is the destination. In the example 'leaving for Paris' it does not mean 'to arrive in Paris', because we do not know if the destination is reached or not, but we know that we are leaving a place to go to *Paris*. In terms of place name *Paris* is the focus, so the polarity of the whole expression may be considered as final.

In some specific cases, two prepositions with opposite polarities may be associated to a motion verb. For instance, in sentence (16), there is only one verb with the expression of two polarities, initial (going from) and final (going to). This 'mixed-polarity' is illustrated in Figure 2.4.

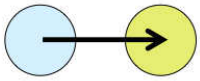


Figure 2.4: Motion verbs with mixed-polarity (initial+final)

(16) Je vais de Pau à Saragosse.
I am going from Pau to Zaragoza

Sarda (2001) has proposed to categorise French verbs according to the types of relations expressed. For instance, she distinguishes relational verbs from referential verbs. Relational verbs refer to displacements whatever the nature of the object, whereas referential verbs refer to displacements only when the object is a place. She categorised referential verbs according to the aspectual polarity proposed by Boons (1987) and relational verbs according to spatial relations: distance (e.g., approach, escape), orientation (e.g., ascend, descend) or transit (e.g. cross). Muller (1998) noticed that the final verb *approcher* (approach or come near) defines a change of distance between the figure and the ground and not a topological relation. For instance, this means that according to sentence (17) we do not know if the figure (*he*) will reach the ground (*Paris*).

(17) Il s'approche de Paris.
He approaches Paris

Muller proposes to categorise motion verbs according to the change of location relationship: he adds properties of telicity and transition. Telicity refers to the property of a verb or verb phrase that presents an action or event as being complete. And transition refers more to an activity (travel) than to an accomplishment (cross). This shows that motion verbs not only refer to spatial but also to temporal relations. For instance, Asher et al. (2008) argue that spatio-temporal primitives are crucial for the understanding of texts and show that motion verbs contribute to both temporal and spatial structure within discourse. Most of motion verbs indicate a spatial trajectory through time. Thus, analysing the types of motion helps to know the location of an object at a certain time.

In the next section, we describe NLP methods for the automatic annotation of information in texts and more specifically oriented to the annotation of spatial Named Entities (NEs).

2.3 Natural Language Processing for Information Extraction

2.3.1 Overview

NLP is a multidisciplinary field of computer science, artificial intelligence, and computational linguistics. The objective of NLP is to develop algorithms for processing texts by extracting and making information accessible to computer systems. Unlike formal languages that are artificially constructed following formally defined rules, the main challenge of NLP is related to the highly ambiguous nature of natural language.

Many research tasks are related to NLP such as Part-of-speech (POS) tagging (Schmid, 1994), automatic summarization (Mani, 1999; Hahn and Mani, 2000), question answering (Hirschman and Gaizauskas, 2001), Spatial Role Labeling (SpRL) (Kordjamshidi et al., 2011), Information Extraction (IE) and Information Retrieval (IR) (Lewis and Jones, 1996; Manning et al., 2008) and Named Entity Recognition (NER) (Nadeau and Sekine, 2007), which is considered as a subtask of IE and is used by most of NLP tasks. In recent years, with the rise of data coming from microblogging and social networks,

more and more studies deal with the task of sentiment analysis (Pang and Lee, 2008; Pak and Paroubek, 2010) also called opinion mining. They attempt to identify and extract subjective information such as 'good', 'bad', 'happy', or 'sad' about various topics such as movies, restaurants or politics.

Furthermore, some works deal with several NLP tasks. For instance, Collobert et al. (2011) propose a unified neural network architecture and learning algorithm that can be applied to several NLP tasks such as POS tagging, NER, and Semantic Role Labeling (Carreras and Màrquez, 2005).

Part-of-speech tagging is the process of assigning a grammatical category to each word in a sentence. A category indicates with a unique tag the syntactic role of each word (e.g. verb, noun, adjective). POS is usually used as a preprocessing step of parsing of unrestricted texts and it is very useful in several NLP tasks such as IR, Text to Speech (Dutoit, 1997) and Word Sense Disambiguation (WSD) (Navigli, 2009).

Carreras and Màrquez (2005) describe the CoNLL-2005 shared task on Semantic Role Labelling (SRL). SRL is a shallow semantic parsing of natural language, which aims at analysing the propositions expressed by verbs in a sentence. All the elements of the sentence that fill a semantic role of the verb must be identified. SRL aims at answering the question 'Who did What to Whom, and How, When and Where?' (Palmer et al., 2010). In a similar way, Kordjamshidi et al. (2011) describe the spatial role labelling task, which proposes machine learning methods to extract spatial roles and their relations.

In this thesis, we are focusing on the tasks of NER and IE and more specifically on Spatial Named Entity Recognition and Geographical Information Retrieval (GIR). GIR is considered as an extension of the field of Information Retrieval (IR) and includes two main tasks related to our concern: toponym recognition and toponym disambiguation (Jones and Purves, 2008).

2.3.2 Named Entity Recognition (NER)

Extraction and annotation of NE is an important task in NLP, particularly in the case of automatic information extraction (Poibeau, 2003). It is considered as a shallow analysis using light parsing methods. NER, also known as Named Entity Recognition and Classification (NERC), aims at identifying and classifying named entities into categories such as 'persons', 'organizations', 'time' or 'locations'. Although NE are considered as being 'proper names' since the Sixth Message Understanding Conference (MUC), named entity task distinguishes three types of NE: ENAMEX (person, organization and location), TIMEX (date and time), and NUMEX (money and percent). These three types of NE are defined by a markup language based on Standard Generalized Markup Language (SGML). Some examples of annotation of NE are shown on Figure 2.5.

```
<ENAMEX TYPE="LOCATION">Paris</ENAMEX>
<ENAMEX TYPE="PERSON">Mr Dupont</ENAMEX>
<TIMEX TYPE="DATE">2015</TIMEX>
<NUMEX TYPE="MONEY">15 euros</NUMEX>
```

Figure 2.5: MUC markup for NER

The NER is applied in several domains or topics such as newspapers, medical, biology or military datasets. Studies in the MUC NE task have proposed seven NE types (organization, location, person, date, time, money and percent expressions), and the Automatic Content Extraction (ACE) program has defined five NE types (organization, location, person, facility and geographical and political entity). Although the number of NE types is enough to cover general issues, many other categories or subcategories may be proposed. For instance, Sekine et al. (2002) proposed 150 types of NE organised in an extended NE hierarchy. Furthermore, Ehrmann (2008) and Tran (2006) propose an overview of the different typologies of Named Entities proposed by linguists or used in NLP. NE typologies may be classified in two categories: one mainly used for automatic named entity recognition and the second one used by linguists.

NER has two objectives: recognise the entity boundaries in the sentence and classify entities in the NE semantic categories. McDonald (1996) distinguishes *internal evidence* and *external evidence*. Internal evidence is derived from words within the name, whereas external evidence uses criteria provided by the context in which a name appears. Evidences may refer to definitive or heuristic criteria such as known

terms, abbreviations (e.g. *Mr.* for *Mister* and *Inc.* for *Incorporated*) or known names found in gazetteers. Evidences may also be defined by typography rules, such as words beginning with a capital letter. These rules may be language-dependant. For instance, in German common nouns may also begin with a capital letter.

Two types of NER approaches have been proposed: those that use learning techniques and those based on ad-hoc rules. In the particular case of annotation of spatial entities, approaches use external resources like gazetteers to search for and identify toponyms.

Data-driven approaches

Many current work address the NER problem using learning techniques. Different approaches exist, such as Hidden Markov Models (HMM) (Bikel et al., 1997; Zhou and Su, 2002), Decision Trees (Szarvas et al., 2006), Maximum Entropy Models (ME) (Borthwick, 1998), Conditional Random Fields (CRF) (McCallum and Li, 2003), and Support Vector Machines (SVM) (Takeuchi and Collier, 2002).

Learning systems are based on supervised, semi-supervised or unsupervised machine learning algorithms. Supervised learning method aims at automatically build rule-based systems using a large training corpus of annotated documents containing both positive and negative examples of NE. Supervised methods require large annotated corpora. The problem of availability of large amount of annotated resources led to semi-supervised or unsupervised learning methods.

Unsupervised learning methods rely on lexical resources, lexical patterns and statistics computed on a large unannotated corpus (Nadeau and Sekine, 2007). Some studies propose hybrid systems, such as Mikheev et al. (1998) that proposed a hybrid statistical-handcoded system for MUC-7.

Learning systems use statistic approaches or machine learning and require large amount of annotated training data. Although developers do not need linguistic or domain expertise, some changes may require re-annotation of the entire training corpus, which is very time consuming.

Knowledge-based approaches

The ad-hoc or linguistic approach (Poibeau, 2011) relies on syntactic-semantic patterns developed manually by experienced language engineers and with the help of experts. Among these hand-written rule-based algorithms, several use transducers with a finite number of states (Poibeau, 2003) to define linguistic patterns. Transducers are a type of finite-state machine that make insertions, replacements and deletions in a text, and may be also used in cascade (Friburger and Maurel, 2004). A transducer is a local grammar defined as an automaton with an input and output alphabet. A cascade consists of a sequence of levels and annotations are built on annotations done at the previous level. Although a transducer does not cover complete linguistic phenomena, a cascade of transducer participates in the coverage of a significant part. Furthermore, a cascade is robust because the decisions are taken locally (Abney, 1996). Defined patterns exploit morpho-syntactic structures and lexicons or gazetteers data.

For instance, for the recognition of person names, a simple rule may be expressed in natural language as : 'if a first name (knowledge found in lexicon or gazetteer) is followed by an unknown word beginning with a capital letter, then the phrase may be annotated as a person name.' This rule is defined using a transducer in Figure 2.6. Greyed boxes contain inclusions of other transducers, in this example it refers to a dictionary of first names. The tag <PRE> indicates that we are looking for a word beginning with an uppercase and <!DIC> refers to unknown words.

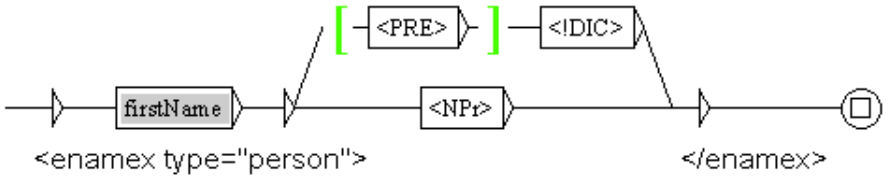


Figure 2.6: Example of transducer for person name recognition

However, a main drawback of this type of approach is that proper names belong to an open class of word and that most of them are unknown and are not stored in gazetteers (Mani and MacMillan, 1996). However, Mikheev et al. (1999) show that unlike persons or organizations, location names need gazetteers to be extracted. Moreover, according to Friburger and Maurel (2004), all named entities cannot be extracted in the same manner. For instance, whereas location names need gazetteers, person and organization names may be extracted using lexical contexts. Friburger and Maurel (2004) propose a tool (CasSys) to create a finite-state transducer cascade for the extraction of named entities in French journalistic texts. They are only interested in the ENAMEX type defined by the MUC Named Entity Task composed of organization, person, and location names. They obtain very good results on French newspapers especially on extracting names of persons, but they explain that place names or names of organizations are more difficult to extract because they have less contexts or internal evidences.

Tran and Maurel (2006) and Maurel et al. (2014) have proposed a multilingual relational database of proper names called Prolexbase, which is particularly dedicated for NLP and based on two main concepts: the *pivot* (i.e., language independent variation of a name) and the *prolexeme* (i.e., language-dependent projection of the pivot). They have described a four-level ontology composed of instances, linguistic, conceptual, and meta-conceptual hierarchical levels. They have proposed the Prolex typology classified in types (e.g., given name, surname, country, city, etc.) and supertypes (e.g., anthroponyms, toponyms, ergonyms and pragmonyms). They have also defined relations between instances, types and supertypes (such as synonym, meronym (*part-of*), hyponym, etc.) in order to build the Prolexbase relational dictionary.

Handcraft approaches could be time-consuming to develop. Each rule must take into account exceptions introduced by the ambiguity of the natural language. For instance, if a rule indicates that a name must begin with a capital letter, an exception would be the words beginning sentences. Patterns may depend on specific properties of each language or domain (i.e., specific corpus). And depending on the complexity and the number of patterns designed, some changes may be hard to accommodate. But handcraft approaches require only a small amount of training data in comparison with methods based on machine learning algorithms. Usually, ad-hoc systems make use of human intuition and achieve higher performance than learning systems.

Hybrid approaches

Knowledge-based and data-driven approaches can be used in a complementary manner in hybrid NER systems (Béchet et al., 2011). In most hybrid approaches, outputs of knowledge-based systems are considered as features by machine learning algorithms.

Nouvel et al. (2012) propose an approach called mXS that aims at combining knowledge-based and data-driven approaches in a modular way. The mXS system aims at locating NE boundaries by using sequential pattern mining and machine learning. A first module extracts knowledge and context patterns and another module annotates NE using the knowledge extracted by the first module and machine learning techniques. The knowledge-based system used by Nouvel et al. (2012) is called CasEN and is based on the finite-state transducer cascade, CasSys, developed by Friburger and Maurel (2004). CasEN reaches 93.2% of recall and 91.1% of weighted harmonic mean (f-score) (Friburger, 2002) on the French newspaper Le Monde corpus. CasEN obtains also high scores during the Ester2 evaluation campaign (Galliano et al., 2009). Nouvel et al. (2012) also proposed a text mining approach called mXtrK to extract informative rules from annotated texts and a stochastic model called mStrucT. This approach obtains good results, and shows that the hybrid system is very sensitive to what is provided by the knowledge module. The mXS system, initially developed for French, has been recently adapted to German (Nouvel et al., 2014).

2.3.3 Spatial Named Entity Recognition

NER methods automatically annotate different types of named entities: dates, persons, organisations, numeric values, as well as place names. There are a significant number of systems available such as OpenNLP[6] from Apache, OpenCalais[7] from Thomson Reuters, and CasEN (Friburger and Maurel, 2004).

[6] http://opennlp.apache.org/
[7] http://www.opencalais.com/

More specific methods that are solely concerned with geographical data are known as geoparsing or toponym recognition (Leidner, 2007).

The main difficulty in extracting geographical information is the ambiguity inherent in natural language. As stated in the introduction, there are actually several types of ambiguity involved in toponym resolution. In addition, a large number of spatial entity types exist: geopolitical entities (countries, administrative divisions), populated places (towns, addresses and postal codes), and natural geographical entities (parks, valleys, mountains, rivers, etc.), all of which can create ambiguities about the type of geographic objects.

Geoparsing aims at extracting keywords and keyphrases describing geographical references (place names) from unstructured text, and mapping to geographic referents according to a spatial model. Toponym resolution (Leidner, 2007) involves associating a non-ambiguous location with a place name and involves resolving the problem of ambiguities that toponyms may contain. The use of resources like gazetteers is thus very useful.

In the last few years, we have seen a number of geographical resources emerging, such as Geonames, OpenGeoData[8], Openstreetmap and Wikimapia[9]. In an open data context, and with some benefits from participative communities, these resources are expanding and being made more widely available through Web services and linked data (see Section 2.5). Some of these web-based geographic services are free, interoperable, and standardised, but the number and diversity of platforms makes using the data a complicated process. Before being able to use this mountain of data, first the most appropriate resources must be selected according to actual needs (Florczyk et al., 2010). Each resource can have different issues, for example the choice between a resource that covers a wide area but non-exhaustively and a more exhaustive resource covering a smaller area.

However, some recent works propose gazetteer-independent toponym resolution approaches, such as the work proposed by DeLozier et al. (2015). They point out the fact that gazetteers are highly incomplete and propose a method of toponym resolution using geographic word profiles based on local spatial statistics over a set of geo-referenced language models. Their method computes the overlap of geo-profiles in text without using gazetteer and obtains good results on international news (TR-CoNLL) and historical corpora (CWar). [10]

2.3.4 Toponym Disambiguation

Toponym disambiguation is defined as a subtask of toponym resolution and is complementary to the subtask of toponym recognition (Leidner, 2007). Ambiguity of toponyms is closely related to the use people make of them. According to Buscaldi and Magnini (2010), frequent toponyms are usually less ambiguous than rare toponyms. Furthermore, Smith and Mann (2003) studied the ambiguity in the Getty Thesaurus of Geographic Names (TGN) and show that almost 60% of toponyms used in North and Central America exist more than once. Obviously this percentage depends on the gazetteer used for the analysis. For instance, also in North and Central America and using Geonames the percentage of ambiguous toponyms is almost 10%. Concerning Europe almost 17% of toponyms are ambiguous on TGN against 13% on Geonames. These percentages represent only the ambiguity of names that refer to several locations.

According to Leidner (2007) in a geospatial context, there are at least three types of ambiguity: *discord*, *non-specificity* and *linguistic ambiguity*. Leidner defined the *discord* ambiguity as referring to territorial dispute between nations or divergent definitions of geographic terms by national geographic agencies. The *non-specificity* ambiguity refers to the problem of vagueness and lack of precision, and the *linguistic ambiguity* consists of three types of ambiguity: *morpho-syntactic ambiguity*, *feature type ambiguity* and *referential ambiguity*. Smith and Mann (2003) have also defined three main types of ambiguity but with different names: *referent class ambiguity*, *referent ambiguity* and *reference ambiguity*.

[8] http://www.opengeodata.fr/

[9] http://wikimapia.org/

[10] The TR-CoNLL corpus, provided by Leidner (2007) consist of 6,000 toponyms identified through 1,000 Reuter's international news articles. The CWar annotated corpus contains 341 books (the Perseus Civil War and 19th Century American Collection) written primarily about and during the American Civil War (Speriosu and Baldridge, 2013).

The *morpho-syntactic ambiguity*, defined by Leidner (2007), refers to place names that may not be names but belong to another word-class, such as common nouns. The second linguistic ambiguity, *feature type ambiguity* or *referent class ambiguity*, refers to place names that may be used in a non-geographical context (i.e., organisations or persons). Then, the *referential ambiguity* or *referent ambiguity* refers to place names that do not uniquely refer to one location, e.g., *Paris* (capital of France) and *Paris* (city of Texas, USA). Finally, Smith and Mann (2003) have proposed another type of ambiguity called *reference ambiguity* that refers to places having several names. Amitay et al. (2004) distinguishes only two types of ambiguity: *geo/non-geo ambiguity* and *geo/geo ambiguity*. The *geo/non-geo ambiguity* defined by Amitay et al. (2004) is equivalent to the *morpho-syntactic ambiguity* and the *feature type ambiguity* defined by Leidner (2007) and the *referent class ambiguity* defined by Smith and Mann (2003). The *geo/geo ambiguity* refers to *referential ambiguity* defined by Leidner (2007) and the *referent ambiguity* and *reference ambiguity* defined by Smith and Mann (2003).

Furthermore, according to Wacholder et al. (1997) there is another type of ambiguity called *structural ambiguity*. It arises when the structure of the words constituting the toponym are ambiguous. For instance, is the word 'River' part of the toponym 'River Thames' or not ? We consider this type of ambiguity as a subset of the *reference ambiguity*. However, this kind of words may be used to disambiguate toponyms using the semantic of geographical features in order to improve the context knowledge (Rauch et al., 2003; Nguyen et al., 2013).

In the remainder of the thesis we are focusing on *referent ambiguity* and *reference ambiguity*. Widely studied in recent years, the admittedly difficult task of toponym disambiguation remains a scientific problem today, which has been tackled with very different approaches (Garbin and Mani, 2005; Buscaldi and Magnini, 2010; Roberts et al., 2010; Speriosu and Baldridge, 2013). Buscaldi (2011) provides an overview about different ways of disambiguating toponyms. According to this work, the approaches can be classified in three categories: map-based, knowledge-based, and supervised or data-driven approaches.

Map-based approaches

Map-based approaches use as the context for disambiguation other unambiguous and georeferenced toponyms found on the same document. These approaches provide a score to any of the possible locations according to the distance to the unambiguous toponym, and usually do not need any information other than the coordinates of the places appearing in the context.

As stated by Smith and Crane (2001), a place is more likely to be located near other places mentioned around it. This idea is directly related to the first law of geography introduced by Tobler: "Everything is related to everything else, but near things are more related than distant things."[11] For instance, Smith and Crane give the following example: 'If *Philadelphia* and *Harrisburg* occur in the same paragraph, a reference to *Lancaster* is more likely to be to the town in Pennsylvania than to the one in England or Arizona.' Smith and Crane (2001) describe a basic implementation of a map-based method that analyses the distance from the centroid of the locations of unambiguous or already disambiguated toponyms cited in the text to a candidate location in order to choose the most appropriate disambiguated toponym. The centroid refers to document context or what they call the 'region of interest' and is calculated using the referenced areas expressed in the document. They define the local context of a toponym location as a moving window of the four toponyms mentioned before it and the four after it. With the same approach, Buscaldi and Rosso (2008b) propose a map-based disambiguation method where all possible referents (even ambiguous toponyms) are used to compute the centroid and where the context size depends on the number of toponyms considered. The proposed algorithm consists in finding the coordinates of each toponym to calculate the centroid. Then the locations that are located more than twice the standard deviation away from the centroid are removed and the centroid is recalculated over the new set of points. Finally, they calculate the distance from the centroid to each location and select the one having the minimum distance. However, this approach may be refined with multiple improvements related to the definition of the disambiguation context, the computation of distance to the candidate location or the consideration of additional physical properties.

With respect to the definition of a more refined disambiguation context, Zhao et al. (2014) propose, for instance, a GeoRank algorithm inspired in Google Page Rank algorithm for the disambiguation of

[11] https://en.wikipedia.org/wiki/first_law_of_geography

toponyms in Web resources. The idea is that other toponyms identified on the same resource vote to the alternative locations of the ambiguous toponym according to their distance in the text and the geographical distance to the tentative location.

Habib and Van Keulen (2012) show that effectiveness of toponym disambiguation is affected by the effectiveness of toponym extraction. Concerning the diversity of techniques for computing closeness to a candidate location, Habib and Van Keulen (2012) propose the use of clustering techniques to disambiguate ambiguous toponyms in holiday descriptions. The clustering approach is an unsupervised disambiguation approach based on the assumption that toponyms appearing in same document are likely to refer to locations close to each other distance-wise. Another alternative to take into account the distance closeness is the one proposed by Zhang et al. (2012). They use an Exact-All-Hop Shortest Path (EAHSP) algorithm for disambiguating road names in text route descriptions. Road name disambiguation belongs to the scope of toponym disambiguation. Their proposed EAHSP algorithm aims at finding a path that maximizes the number of crossed roads in a proximity area. Besides, it must be also acknowledged that this work faces additional difficulties since databases for road names are not so popular, and their associated geometry is not point-based.

Finally, related to the consideration of additional physical properties, Derungs and Purves (2014) propose a map-based disambiguation algorithm for toponyms found in landscapes descriptions as part of a mechanism to assign a general geospatial footprint to documents. Apart from using the Euclidean distance from ambiguous toponyms to already identified unambiguous toponyms, the proposed disambiguation algorithm exploits the topographic similarity between toponyms. The performance of their disambiguation algorithm is closely related to the availability of gazetteers with topographic information, which may be missing for some regions. Additionally, this work tries to identify the terms used to describe natural features in a region and compare the different vocabularies used for natural features in different regions.

Knowledge-based approaches

Knowledge-based approaches make profit of knowledge sources (gazetteers, ontologies, etc.) to determine if other related toponyms in this knowledge source are also referred in the document (Qin et al., 2010; Buscaldi and Rosso, 2008b; Lieberman and Samet, 2012), or additional information from the toponyms, such as importance, size, population counts (Overell and Rüger, 2008), or document creators (e.g. documents of social media (Ireson and Ciravegna, 2010)) can be exploited. Knowledge-based method aims at considering semantic relations between named entities, concepts or key terms such as social, associative or lexical relatedness and not only co-occurrence statistics of terms. Knowledge sources such as Wikipedia and WordNet (see Section 2.5), are the most widely used. For instance, Overell and Rüger (2008) used Wikipedia[12] to generate a tagged training corpus and a co-occurrence model applied to the disambiguation of toponyms in unstructured text. This method aims at using article categories and links to other articles in Wikipedia to improve the knowledge context and disambiguate toponyms.

Amitay et al. (2004) proposed a toponym disambiguation approach for their geotagging system (Web-a-Where) based on population statistics and arborescent relationships. They described a knowledge-based method using population heuristics, hierarchical relations and geographical coordinates. The main idea is that each toponym existing in a taxonomy node such as *Paris/France* or *Paris/Texas/United States* obtains a score to define the importance of the place. Depending on the number of occurrence on the document and on the level of hierarchy. Finally, they sum the score obtained for each toponym to propose a geographic focus of a webpage.

Buscaldi and Rosso (2008a) proposed a conceptual density-based approach for toponym disambiguation. With the same idea, Bensalem (2010) proposed a context-based heuristic using toponyms hierarchical path obtained from WordNet. They both refer to arborescent proximity as the distance proximity between the toponyms referents.

Batista et al. (2012) also address the *referent ambiguity* (i.e., when the same toponym refers to more than one place) and propose two mapping techniques based on semantic similarity measures. These methods do not take into account geographic features usually associated to a toponym (e.g., city, street, lake, etc.), they only use semantic similarity between concepts of the ontology. They evaluated their

[12] http://www.wikipedia.org

approach using a geospatial ontology of the Portuguese territory and a collection of geographic annotated Portuguese news articles.

Figure 2.7 shows a part of a world hierarchical tree representing the arborescent relationships between places. For example, if the toponym *Dallas* is expressed in the context of the ambiguous toponym *Paris*, knowledge-based methods disambiguate *Paris* to *Paris/Texas* instead of *Paris/France*, because *Paris* and *Dallas* have the common root *Texas*.

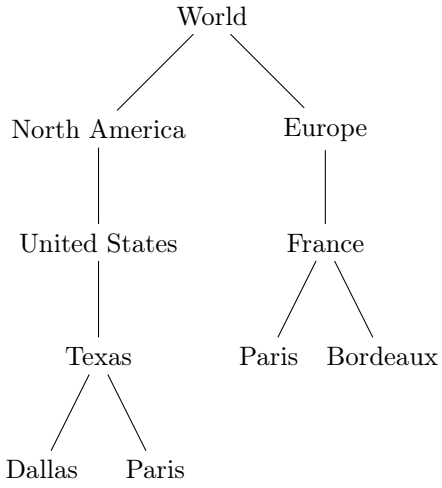


Figure 2.7: Hierarchical tree

However, this kind of information is not the most suitable for a discriminating task in the case of specific documents describing, for example hiking trails. As explained by Derungs and Purves (2014), since these documents contain usually fine-grain toponyms or natural features such as mountains, lakes, hamlets and refuges, the coverage provided by knowledge resources are very limited.

Data-driven approaches

Finally, data-driven approaches (Smith and Mann, 2003; Agrawal and Shanahan, 2010) are based on machine learning algorithms. According to Buscaldi (2011), although data-driven methods are widely used in WSD, they are not commonly used in toponym disambiguation. However, they may obtain better results than knowledge-based methods. These methods exploit non-geographical content and events to build probabilistic models, using spatial relationships between entities (i.e., persons, organisation) and places. Smith and Mann (2003) trained a Naïve Bayes classifier using toponyms already disambiguated in the training set using cues such as "Nashville, Tenn." or "Springfield, MA". Their goal was to find the corresponding state of American cities and country for foreign cities. They made experiments on different corpora (American news articles, American Memory and Civil War documents) and obtained a precision between 22% and 87%, depending on the corpus. Garbin and Mani (2005) point out that 67,82% of ambiguous toponyms found in a news article corpus (New York Times) were not associated with *local discriminator* (i.e., a feature characterising the toponym within a ±5-word window). For instance, 'TX' (abbreviation of Texas) is considered as a local discriminator for the toponym 'Dallas' in 'Dallas, TX'. They proposed a rule-based classifier for toponym disambiguation obtaining precisions between 65% and 88% also depending on the corpus. Li et al. (2006) proposed a probabilistic toponym disambiguation system using local and global context. Local context (also called *local discriminator* by Garbin and Mani (2005)) may refers to geographical terms co-occurring in close proximity, association of geographical terms such as lake or city and also population statistics. The global context refers to the occurrence of referent such as state or country that are ancestors in the hierarchy defined in geographical resources. Martins et al. (2010) used HMM to annotate place references and then a SVM to rank the possible disambiguation. They evaluated the method through gold-standard document collections in three

different languages (English, Spanish and Portuguese) annotated by humans. They obtained precisions between 25% (on the Portuguese dataset) and 68% (on the English dataset).

This section has described several approaches for NER and toponym disambiguation. For both problematic, approaches are classified in two main opposite categories, data-driven approaches versus knowledge-based approaches. One of the main drawbacks of data-driven approaches is the lack of classified collections and the need of large corpora of annotated ground truth, whereas knowledge-based approaches require only a small set of training data. With respect to our concern, in the next chapters, we describe how we use knowledge-based approaches for the automatic annotation of spatial information, and we also show how we combine map-based and knowledge-based approaches to solve the problem of toponyms ambiguity. Moreover, the following section describes markup languages for encoding and sharing spatial information.

2.4 Markup languages for encoding spatial information

2.4.1 Overview

A markup language defines a set of tags and/or a set of rules for creating tags that can be embedded in digital text to provide additional information about the text. Markup languages are fundamental to structure and display information in web browsers, they are used to transform an unstructured text into an exchangeable and interoperable format. Hypertext Markup Language (HTML) is the most familiar markup language. HTML is a descendant of SGML, used to create web pages. SGML developed in 1986 is an international ISO standard (ISO 8879:1986) for defining generalized markup languages for documents. SGML was developed to facilitate the sharing of machine-readable documents.

Extensible Markup Language (XML) was defined by the World Wide Web Consortium (W3C)[13] as a subset of SGML. Thus, XML is more restrictive than SGML and has been designed to make parsing much easier. XML is nowadays commonly used for describing and exchanging data, it provides a mechanism to impose constraints on the storage layout and logical structure. The XML specifications define an XML document as a well-formed text, i.e. the XML document satisfies rules defined in the specification such as "an XML document must contain a single *root* element that contains all the other elements". An XML document consists of *tags* (i.e., markups) and *content*. These two elements may be distinguished by a parser with simple syntactic rules. A tag is defined by a pair consisting of a *start-tag* (e.g., `<body>`) and an *end-tag* (e.g., `</body>`), except for those that do not enclose any content which are defined by a unique tag ending by a forward slash after all other characters within the brackets (e.g., `<img />`). In addition to being well-formed, an XML document may be valid according to a formal specification document (e.g., DTD, XML Schema, Relax NG).

With regard to our concern of automatic annotation of geospatial information in texts, markup languages may be categorized into three classes: spatial, spatio-temporal and generic markups. Furthermore, we can distinguish two categories of markup languages, those focused on the encoding of information and those focused on annotation of texts. The first category refers to exchange formats of data whereas the second one refers to the annotation of information in textual documents written in natural languages.

2.4.2 Spatial Markup Languages

We will now describe four different spatial markup languages, Geography Markup Language (GML), Keyhole Markup Language (KML), SpatialML and Toponym Resolution Markup Language (TRML). GML and KML are two exchange formats used in GIS for encoding spatial features. In contrast, SpatialML and TRML are annotation schemes for annotating geographic entities in text.

GML (Geography Markup Language)
GML is an XML grammar defined by Open Geospatial Consortium (OGC) (OpenGIS® Geography Markup Language Encoding Standard[14]) for describing geographic data (Burggraf, 2006). GML is also

[13] http://www.w3.org/
[14] http://www.opengeospatial.org/standards/gml/

an international ISO standard (ISO 19136:2007) which provides a set of core schema components such as feature objects (e.g., geometry, topology, temporal) with a simple semantic model between objects and properties. Features can be concrete (e.g., roads and buildings) or abstract and conceptual (e.g., political boundaries). GML describes generic geographic data sets that contain points, lines and polygons (`<Point>`, `<LineString>`, `<Polygon>`) associated with real-world coordinates. Figure 2.8 shows the GML annotation for encoding the geometric representation corresponding to the spatial expression (18).

(18) au sud de Pau
south of Pau

```
<Polygon>
   <outerBoundaryIs>
      <LinearRing>
         <coordinates>
            43.2345070972552, -0.389339593262433
            43.3061317796098, -0.392743259810513
            43.3085863498989, -0.294231809315081
            43.2369584558224, -0.290950779544868
            43.2345070972552, -0.389339593262433
         </coordinates>
      </LinearRing>
   </outerBoundaryIs>
</Polygon>
```

Figure 2.8: Example of GML markup

XML provides the extensibility to create various different languages and offers the possibility to mix different XML encodings within the same document. For example, GML can be used for integrating geospatial data within another XML grammar (see Figure 2.9).

```
<root xmlns:gml="http://www.opengis.net/gml">
   <entity>
      <name>Pau</name>
      <position>
         <gml:Point srsDimension="2"
             srsName="http://www.opengis.net/def/crs/EPSG/0/4326">
            <gml:pos>43.301667 -0.368611</gml:pos>
         </gml:Point>
      </position>
   </entity>
</root>
```

Figure 2.9: Example of an XML markup integrating GML annotations

KML (Keyhole Markup Language)
KML is another XML standard for expressing geographic annotation. KML was originally developed for use with Google Earth and became an international standard of the OGC (the OpenGIS® KML Encoding Standard[15]) in 2008. KML is a language for describing visualization of geographic information. It includes the presentation of graphical data on the globe and the control of the user's navigation in the sense of where to go and where to look. Figure 2.10 shows the KML annotation of the geometric representation of the spatial expression (18).

The main difference between KML and GML is that KML is a language for describing visualization of geographic data. It defines styles for drawing lines and shapes by setting line widths, color and fill colors. The coordinate system must be specified explicitly with a CRS (coordinate reference system) in

[15] http://www.opengeospatial.org/standards/kml/

```
<kml>
   <Document>
      <Style id="redLine">
         <LineStyle><color>ff0000ff</color><width>4</width></LineStyle>
      </Style>
      <Placemark>
         <styleUrl>#redLine</styleUrl>
         <Polygon>
            <outerBoundaryIs>
                <LinearRing>
                   <coordinates>
                      -0.389339593262433, 43.2345070972552
                      -0.392743259810513, 43.3061317796098
                      -0.294231809315081, 43.3085863498989
                      -0.290950779544868, 43.2369584558224
                      -0.389339593262433, 43.2345070972552
                   </coordinates>
                </LinearRing>
            </outerBoundaryIs>
         </Polygon>
      </Placemark>
   </Document>
</kml>
```

Figure 2.10: Example of KML markup

GML[16]. Furthermore, there are no feature types in KML, whereas it is possible to differentiate different types of roads with GML. From this perspective, KML is complementary to the GML standard.

SpatialML

Mani et al. (2008) proposed an XML annotation scheme called SpatialML for geolocating geographic entities in text. It proposes numeric representation of places such as the GML or KML standards. SpatialML has been mapped to the Generalized Upper Model (GUM) ontology (Bateman et al., 2010). It also characterizes relationships among places (including both relative and absolute locations), providing some support for qualitative reasoning about topological and orientational relations based on RCC8 relations. Topological, orientation and distance relations are expressed by qualitative spatial links (`<QSLINK>`) and relative links (`<RSLINK>`). These two non-consuming tags refer to the expression of spatial relations that are annotated by the `<SIGNAL>` tag. Locations (both nominal and named entities) mentioned in the text are marked with `<PLACE>` tags and mapped to data from geographical resources. The *latlong* attribute content of `<PLACE>` elements have no default format, it can include strings with or without decimals. This annotation can be parsed into GML or KML coordinates with appropriate coordinate systems, allowing a SpatialML markup to be transformed automatically to its visualization in Google Earth. Figure 2.11 shows the SpatialML annotation of the expression (18). In this example, the element `<PLACE id="3">` refers to the association of the place (id=2) and the spatial relation (id=1), which is defined by the `target` attribute of the `<RLINK>` element. The main lack of this annotation scheme is that it doesn't provide any support to identify spatio-temporal information such as motion, moving objects, or paths.

```
<SIGNAL id="1" type="DIRECTION">sud</SIGNAL>
<PLACE id="2" country="FR" form="NAM" latlong="43.301667N␣-0.368611W">Pau</PLACE>
<PLACE id="3" />
<RLINK id="4" direction="S" source="2" target="3" signals="1" />
```

Figure 2.11: Example of SpatialML markup – **Source:** (Palacio, 2010)

[16] Whereas GML which uses the (latitude, longitude) order, KML uses the (longitude, latitude) order. However, the order in GML can be different according to the specified CRS.

TRML (Toponym Resolution Markup Language)
Standard markup languages for encoding spatial information such as GML, KML or SpatialML only deal with geometric and thematic properties of spatial features. Other specific markup languages have been proposed, such as TRML proposed by Leidner (2006), which is particularly dedicated to the task of toponym resolution. Although TRML describes the structure of the document with different tags (e.g., `<doc>`, `<s>`,`<w>`,`<toponym>`, etc.), it is mainly focused on named entities and more specifically on toponyms. Each `<toponym>` content element contains a `<candidates>` element that contains a set of alternatives candidate referents (`<cand>`). Latitude and longitude are stored in decimal form in the *lat* and *long* attributes, respectively. Figure 2.12 shows the TRML annotation for the example (18).

```
<s id="s1">
   <w tok="au" pos="PRP" />
   <w tok="sud" pos="NN"/>
   <w tok="de" pos="PRP"/>
   <toponym term="Pau">
      <w tok="Pau" pos="NNP" ne="I-LOC"/>
      <candidates>
         <cand id="c1" src="GN" lat="43.3" long="-0.36" humanPath="Pau, France" />
         <cand id="c2" src="GN" lat="39.79" long="8.80" humanPath="Pau, Italy" />
         <cand id="c3" src="GN" lat="-15.44" long="-39.68" humanPath="Pau, Brazil" />
      </candidates>
   </toponym>
</s>
```

Figure 2.12: Example of TRML markup

2.4.3 Spatio-Temporal Markup Languages

Although the temporal dimension is an important component in the definition of spatial information, there are only few works representing both temporal and spatial dimensions. We will now describe two markup languages dealing with spatio-temporal information, Temporal Geographical Markup Language (TGML) and ISO-Space.

TGML (Temporal Geographical Markup Language)
As stated by Langran (1992) a temporal GIS should be able to supply the complete lineage of a single feature, the evolution of an area over time, and the state of a specified feature or area at a given moment. Several spatio-temporal models have been proposed such as the Object-Relationship Model proposed by Claramunt et al. (1999) and some studies used spatial markup languages for integrating temporal information of historical phenomena. For instance, Zipf and Krüger (2001) proposed an XML spatio-temporal framework called TGML. TGML uses GML in a very flexible way and provides buildings blocks for temporal structures like *intervals*, time *spans* or *instants*. It aims at modelling temporally changing spatial data by adding object and attribute time-stamping and coupling GML and temporal structures. Guerrero Nieto et al. (2010) proposed also a methodology for linking linguistic corpora and GIS. Their approach is divided into three steps: the automatic identification of temporal expression, the manual normalization of the temporal expressions with TimeML and the incorporation of TimeML into a geodatabase.

TimeML proposed by Pustejovsky et al. (2005), is a markup language for events and temporal expressions in natural language. It includes three basic tags: `TIMEX3`, `EVENT` and `SIGNAL`. `TIMEX3` is used to annotate explicit temporal expressions such as times, dates, durations, etc. The `EVENT` tag is used to annotate elements in a text that mark the semantic events. Syntactically, `EVENTs` are typically verbs such as *to begin*, *to occur*, etc. `SIGNAL` is used to annotate sections of text, typically function words, which indicate how temporal objects are related to each other (e.g., *before*, *after*, *during*). TimeML includes also other elements (e.g., `TLINK`, `ALINK`, `SLINK`) used to annotate different kind of relationships.

Furthermore, according to the problem of representing both imperfect spatial and temporal information with textual annotation, Mouna Snoussi et al. (2012) proposed to extend SpatialML and TimeML,

towards the representation of imprecision and vagueness values and degrees for spatio-temporal information. Their approach perform a semantic text analysis using a predefined list of natural language terms such as *near to*, *before*, *after*, *between*, etc, in order to identify imperfection.

ISO-Space

Pustejovsky et al. (2012) propose ISO-Space, an emerging standard for encoding spatio-temporal information in text. It provides an annotation specification for encoding spatial and spatio-temporal information as expressed in natural language texts. According to Pustejovsky et al. (2012) any specification language for encoding spatio-temporal information require the following representational mechanisms:

- the representation of locations as regions in an interpreted spatial domain;
- the representation of objects as occupying regions of space;
- the identification of appropriate topological value within the model;
- the representation of direction and orientation, both for the domain of discourse and for regions and objects;
- the representation of metric properties;
- the representation of object movement.

The ISO-Space specification incorporates annotations from SpatialML for static spatial information, and from TimeML and Spatio-Temporal Markup Language (STML) for spatio-temporal information. The STML, introduced by Pustejovsky and Moszkowicz (2008), proposes a set of motion classes based on the classifications of verbs of motion in (Muller, 1998). The ten motion classes from STML are: *move*, *move external*, *move internal*, *leave*, *reach*, *detach*, *hit*, *follow*, *deviate*, *stay*. These verb classes are mapped to RCC-8 relations and correspond to a semantic interpretation of motion concepts. Additionally, ISO-Space uses ontologies such as GUM (Bateman et al., 2010) to classify places.

The ISO-Space specification supports the identification of several kinds of spatial relations: topological, directional, orientational, metric properties, and metric value between two spatial objects. It supports also the identification and the characterization of motion of objects through time. There are eight main elements available:

- `<PLACE>` tags annotate geographic entities and regions (e.g., towns, lakes).
- `<PATH>` tags annotate locations with a focus on the potential for traversal (e.g., road, river).
- `<SPATIAL_ENTITY>` tags annotate anything that is spatially relevant (e.g., a person, a car, a building).
- `<NONMOTION_EVENT>` tags annotate TimeML events that are spatially relevant. Events can be denoted by verbs, nouns, adjectives, prepositions (e.g., in, on, at) and locative adverbs (e.g., here, there).
- `<MOTION>` tags annotates events that involve a change of location. `<MOTION>` tags are categorized as *manner* (e.g., to walk, to swim), *path* (e.g., to arrive, to leave) or *compound* and are classified into the ten classes used in STML and derived from the classifications of motion verbs (Muller, 1998): *move*, *move_ external*, *move_ internal*, *leave*, *reach*, *detach*, *hit*, *cross*, *follow* and *deviate*. Each class is defined by a *path focus* and an *event structure*.
- `<MOTION_SIGNAL>` tags specifies information about a motion-event (manner or path).
- `<SPATIAL_SIGNAL>` tags annotate the relationship (directional, topological or both) between two spatial elements.
- `<MEASURE>` tags are a special kind of spatial signal that capture distances and dimensions.

`<PLACE>`, `<PATH>` and `<SPATIAL_ENTITY>` tags can be used as non-consuming tags (i.e., null or empty string content), when a spatially relevant entity is referenced indirectly.

There are also spatial link elements used to represent the relationships between previous elements: `<QSLINK>` (qualitative spatial links), `<OLINK>` (orientation information), `<MOVELINK>` (movement link), and `<MLINK>` (defining the dimension of a location). These links give information concerning the spatial relationships between spatial elements. Furthermore, the `<METALINK>` tag is used for connecting objects in a non-spatial way, for example when an entity is referenced multiple times (coreference) or referenced

as a subset of another entity (subcoreference). Figure 2.13 shows the ISO-Space annotation for the motion expression (19).

(19) He left Biarritz for Pau.

```
<SPATIAL_ENTITY id="se1" form="NAM" countable="TRUE">He</SPATIAL_ENTITY>
<MOTION id="m1" motion_type="PATH" motion_class="LEAVE"
    motion_sense="LITERAL">left</MOTION>
<PLACE id="pl1" form="NAM">Biarritz</PLACE>
<MOTION_SIGNAL id="a1" motion_type="PATH" >for</MOTION_SIGNAL>
<PLACE id="pl2" form="NAM">Pau</PLACE>
<MOVELINK id="mvl1" trigger="m1" mover="se1" source="pl1" goal="pl2"
    goal_reached="UNCERTAIN" motion_signalID="a1"/>
```

Figure 2.13: Example of ISO-Space markup

Spatial Role Labeling (SpRL)
Kordjamshidi et al. (2011) propose an annotation scheme for tagging tokens that participate in expressing a spatial concept based on the Holistic Spatial Semantic theory (Zlatev, 2010). They defined SpRL as the task of identifying and classifying the spatial arguments of the spatial expressions mentioned in a sentence including both static and dynamic spatial semantics. They define a new framework for spatial relation extraction using machine learning approaches at the linguistic level. In contrast to ISO-Space, the annotation of spatial objects is not an isolated annotation of words, but an annotation of relations between a word and a spatial pivot. Annotation is performed at the sentence level. The semantic spatial components in Holistic Spatial Semantic (HSS) theory are: `TRAJECTOR`, `LANDMARK`, `MOTION-INDICATOR`, and `SPATIAL-INDICATOR`. `TRAJECTOR` (also called locale/figure) annotates the entity (person, object, event) whose location is described. `LANDMARK` (also called reference object) annotates the entity related with the `TRAJECTOR`, and `MOTION-INDICATOR` and `SPATIAL-INDICATOR` annotate spatio-temporal and spatial relation between `TRAJECTOR` and `LANDMARK`. `LANDMARK` tag contains a *path* attribute which labels the `LANDMARK` as the source, goal, or midpoint of a motion. This representation differs from the ISO-Space approach which considers the path as an individual annotation element. Furthermore, `<SR>` elements refer to spatial relations and describe the relation between the components constituting the considered spatial relation, trajector, landmark, spatial-indicator, motion-indicator and frame-of-reference. Figure 2.14 shows the SpRL annotation for the motion expression (19).

```
<TRAJECTOR id="1">He</TRAJECTOR>
<LANDMARK id="1" path="BEGIN">Biarritz</LANDMARK>
<LANDMARK id="2" path="END">Pau</LANDMARK>
<SPATIAL-INDICATOR id="1" general-type="Direction" specific-type="Relative"
    spatial-value="NTPP">for</SPATIAL-INDICATOR>
<MOTION-INDICATOR id="1">left</MOTION-INDICATOR>
<SR id="1" trajector="1" landmark="1" spatial-indicator="NIL" motion-indicator="1"
    frame-of-reference="ABSOLUTE" />
<SR id="2" trajector="1" landmark="1" spatial-indicator="1" motion-indicator="1"
    frame-of-reference="ABSOLUTE" />
```

Figure 2.14: Example of SpRL markup

2.4.4 Generic Markup Languages

Some others markup languages are highly modular and provide a kind of generic framework, in which each module provides a standard for specific application. We will describe in more detail the TEI proposal, which is widely used in digital humanities.

Text Encoding Initiative (TEI)

Text Encoding Initiative (TEI) aims at the creation of international standards for textual markup, it is concerned with the markup of expressions which are, or can be, expressed as text (i.e., expressions which take a *written* form). Originally, the TEI launched in 1987 was a research project within the humanities, sponsored by the Association of Computers in the Humanities (ACH), the Association for Computational Linguistics (ACL), and the Association of Literary and Linguistic Computing (ALLC). Since 2000, TEI is organized as an international consortium: TEI Consortium[17] which aims to maintain and develop the TEI standard.

The TEI Guidelines (titled *Guidelines for Electronic Text Encoding and Interchange*) provide a guide to best practices for interchange and encoding of textual material in digital form. TEI emphasizes the interchange of textual information, but other forms of information such as images and sound are also addressed. The TEI Guidelines describe an encoding scheme, which can be expressed using a number of different formal languages. The first editions of the guidelines used the SGML. But since 2002, this has been replaced by the use of the XML. The last and current version of the guidelines is the TEI P5[18] (i.e., TEI Proposal number 5).

The TEI encoding consists of a set of modules ('tagsets') and is structured as follow:

- Core scheme fragments
  - Standard components of the TEI main scheme; these are mandatory.
- Base scheme fragments
  - Basic building blocks for specific text types.
- Additional scheme fragments
  - Extra tags useful for particular purposes

A TEI-conformant text must contain a single TEI header (marked up as a `<teiHeader>` element) followed by a single text element (marked up as a `<text>`). The TEI header gives the meta-data on the TEI document. These two elements are combined together to form a single `<TEI>` element, which must be declared within the TEI namespace[19]. Figure 2.15 shows an excerpt of TEI markup.

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
   <teiHeader>
      <!-- ... -->
   </teiHeader>
   <text>
      <body>
         <!-- ... -->
      </body>
   </text>
</TEI>
```

Figure 2.15: TEI markup

TEI has an overly broad scope and for that reason it has been designed to be highly modular and extensible. TEI encoding can be applied to many different kinds of texts such as dramatic texts, early manuscripts, transcribed speech, etc. TEI was intended to be used as a set of building blocks for creating a schema suitable for a particular project. TEI P5 uses a mechanism in the TEI customization architecture, which permits a customization to define only the TEI elements to be included in a schema. Indeed, the spirit of TEI is: standardization does not mean 'do as I do', but it means 'explain to me what you do'.

TEI provides a set of 22 modules (listed in Table 2.1), each of which declares particular XML elements and their attributes according to their purpose.

The *Core* module of TEI describes elements that may appear in any kind of text such as paragraph (`<p>`), foreign words (`<foreign>`) and quotation (`<quote>`). More related to our concern, the *Core* module also provides elements for encoding proper nouns or *referring strings*. The `<rs>` tag annotates referring strings or general purpose names. The `<name>` tag annotates only proper nouns (i.e., named

[17] TEI-C http://www.tei-c.org
[18] http://www.tei-c.org/Guidelines/P5/
[19] TEI namespace: http://www.tei-c.org/ns/1.0/

Module	Description
analysis	Simple analytic mechanisms
certainty	Certainty and uncertainty
core	**Elements common to all TEI documents**
corpus	Header extensions for corpus texts
declarefs	Feature system declarations
dictionaries	Dictionaries and other lexical resources
drama	Performance texts
figures	Tables, formulae, and figures
gaiji	Character and glyph documentation
header	**The TEI Header**
iso-fs	Feature structures
linking	Linking, segmentation and alignment
msdescription	Manuscript Description
namesdates	Names and dates
nets	Graphs, networks and trees
spoken	Transcribed Speech
tagdocs	Documentation of TEI modules
tei	**Declarations for datatypes, classes, and macros available to all TEI modules**
textcrit	Text criticism
textstructure	**Default text structure**
transcr	Transcription of primary sources
verse	Verse structures

Table 2.1: The TEI modules

entities) and may be nested within a referring string. The attribute *type* can be used to characterize the element (e.g., person, place). Figure 2.16 shows an excerpt of TEI markup using the *Core* module for the expanded named entity 'ville de Pau'.

```
<rs>
   ville de <name type="place">Pau</name>
</rs>
```

Figure 2.16: Excerpt of TEI markup for names and referring strings (*Core* module)

The *Namesdates* module (i.e., names and dates) of TEI contains the definition of elements for encoding of names and other descriptive phrases of persons, places, or organizations with more details than with the elements provided by the *Core* module. With regard to our concern, we are mainly focused on the elements available for the representation of geospatial information expressed in text. The *Namesdates* module provides more elements in order to cover component parts of a place name that provide important information about the semantic or relation between space and time. For instance, the `<geogName>` element (i.e., geographical name) identifies a place name associated with some geographical features and the `<geogFeat>` element (i.e., geographical feature name) contains a common noun identifying some geographical feature contained within a geographic name (e.g., lake, valley, etc.). There are also some elements for describing relative place names such as `<offset>` and `<measure>`. Figure 2.17 shows an excerpt of TEI markup using the *Namesdates* module for the expanded names entity 'ville de Pau'.

```
<geogName>
   <geogFeat>ville<geogFeat>
        de
   <name>Pau</name>
</geogName>
```

Figure 2.17: Excerpt of TEI markup for place names (*Namesdates* module)

Although there is no default schema, TEI P5 does provide a number of example customizations and the possibility to build your own one. The recommended way to customize the TEI is to create a formal specification expressing customizations, as an XML document using TEI One Document Does it all (ODD) markup. The four modules *Core*, *Header*, *Textstructure* and *Tei* should be always included.

The ODD format includes the schema fragments, prose documentation, and reference documentation for the TEI Guidelines in a single document. This adds a series of elements which are used to specify a new schema, and modifications to the TEI element structure. A TEI schema is defined by a `<schemaSpec>` element containing explicit declarations for objects (i.e., elements, classes, or macro specifications) and references to other TEI objects. Figure 2.18 shows the simplest customization of the TEI scheme. An ODD processor will generate an appropriate schema from this set of declarations, expressed using the XML DTD[20] language or the ISO RELAX NG[21] language. DTD and Relax NG both define rules for the structure and content of an XML document. The TEI consortium developed the web-based application Roma[22] for TEI customization. Roma provides a convenient interface for the creation of ODD files following TEI guidelines.

```
<schemaSpec ident="TEI-minimal" start="TEI">
    <moduleRef key="tei" />
    <moduleRef key="header" />
    <moduleRef key="core" />
    <moduleRef key="textstructure" />
</schemaSpec>
```

Figure 2.18: Minimal schema specification

In this section we have described several solutions for encoding, storing and sharing spatial annotated texts. Then, in the following section, we describe gazetteers (i.e., geographical resources) used to retrieve information about locations and making connexions with other datasets of the Semantic Web (Linked Data).

2.5 Gazetteers

In order to find the location of place names expressed in text, and connect text with geographic space, we need to rely one external geographical resources. This task is also known as Toponym Resolution. The main types of geographical resources include: maps, online and digital sources, travel guides and gazetteers. Maps may be print or digital and are usually provided by National Mapping and Cadastral Agency (NMCA). Online map services provide up-to-date road maps and satellite images with free access (e.g., Google Maps, Google Earth, Nokia Here, Bing Maps). Travel guides are most frequently used to answer recreation questions such as vacation planning and hotel or restaurant information (e.g., Lonely Planet, Guide du Routard, etc.). Finally, a gazetteer is a geographical dictionary containing a list of geographical names and physical features. Gazetteers provide information about the location of the feature such as latitude and longitude. Some gazetteers may also provide additional information such as

[20] DTD (Document Type Definition) defines tags and attributes used in an XML or HTML document.
[21] Relax NG (Regular Language for XML Next Generation) is a ISO schema language for XML.
[22] http://www.tei-c.org/Roma/

social statistics, physical features and feature class (e.g., country, city, river, lake, etc.). Gazetteers are one of the most used types of geographical resources, often used in conjunction with mapping systems in GIS.

In this section we describe the different types of gazetteers but also the problem of coverage and granularity. Indeed, geographical resources may be categorised according to their scope or coverage. For instance, local resources can be very detailed, containing fine-grained geographical data, and global resources may have low details and cover only the most important places. Then we also introduce some standard data models such as the ISO 19112 standard for geographic identifiers and the INfrastructure for SPatial InfoRmation in Europe (INSPIRE) data specification for geographic names.

2.5.1 Gazetteer models

A gazetteer may refer to a simple dictionary containing an alphabetical list of geographical names associated with further information (also called *short-form gazetteer*), but it may also refer to a geographical ontology where spatial concepts and entities are linked with semantic relationships.

Gazetteers are used for indirect geo-referencing through place names. Hill (2000) distinguishes three components of a gazetteer: toponym which refers to the name of a location, geographical feature type (e.g., country, city, river, etc.) and spatial footprint which refers to the representation of the location such as point (latitude and longitude) and polygon. In addition, an ontology is a formal definition of types and properties of entities and relationships between them. Ontologies describe entities as instances, types as concepts and properties as attributes. For instance, Abadie and Mustière (2010) propose a taxonomy of geographical concepts built using semi-automatic analysis of textual specifications of geographic databases. Lopez-Pellicer et al. (2012) propose an ontology schema that combines administrative structure, spatial component and temporal evolution of jurisdictional domains used to create the Spanish jurisdictional model for improving historical IR systems.

Geographical ontology is a domain ontology modelling geographical information and describing spatial concepts, spatial entities and spatial relations. Spatial relations are usually represented by containment relationship (i.e., *part of*) but ontologies may also contain information about neighbouring places, such as confluent of a river. Ontology aims to define semantics of relations between concepts and is used by search engines in IR systems to disambiguate queries and compute a relevance ranking of results. An example of such systems is the SPIRIT system proposed by Fu et al. (2005), which introduced a geographical ontology to develop spatial-aware search engine. This geo-ontology integrate multiple datasets: the Seamless Administrative Boundaries of Europe (SABE) dataset and the Getty TGN. The SABE is a digital map dataset representing the geometry of administrative boundaries for Europe which contains geometry footprint of each place stored. The Getty TGN[23] contains more than 2 millions names and information about places. In such gazetteers, toponyms are structured hierarchically and each record is identified by a unique ID.

Gazetteers can suffice for tasks involving the recognition of geographical references, but other Information Retrieval tasks require additional information. Spatial data in gazetteers is usually confined to centroids, which is too limited for considering spatial relationships (Martins et al., 2005).

One of the most important issue in a gazetteer is the model that guides its design. This model can be very different from one resource to another. It can range from a very simple thesaurus (e.g., TGN thesaurus) to a very detailed application-level ontology. In the middle, there are some specific data models created for this domain such as the models provided by the ISO 19112 standard for geographic identifiers (ISO, 2003), or the INSPIRE data specification for geographic names. Both documents are related to the definition of locations (i.e., geographical names). ISO 19112 was published in 2003 and revised in 2009 and the European INSPIRE initiative was released in 2007.

The European INSPIRE[24] Directive is a standardized spatial data infrastructure to promote sharing of geospatial data throughout all levels of government, private and non-profit sectors, and academic communities (Bartha and Kocsis, 2011).

The ISO 19100 series (Geographic information standards) was selected as international standard for the technical base of INSPIRE. Although, we are mainly interested in the ISO 19112 (Spatial referencing

[23] http://www.getty.edu/research/tools/vocabularies/tgn/
[24] http://inspire.ec.europa.eu/

by geographical identifiers), there are a lot of other standards defined in the ISO 19100 series, such as ISO 19103 (Conceptual schema language), ISO 19111 (Spatial referencing by coordinates), ISO 19136 (Geography Markup Language), etc. The scope of ISO 19112 is to establish a general model for spatial referencing using geographic identifiers. This standard defines the components of a spatial reference system and the essential components of a gazetteer. Spatial referencing by coordinates is not addressed in this document; however, a mechanism for recording complementary coordinate references is included. Essentially, it defines a gazetteer as an aggregation of location instances.

The aim of the INSPIRE Data Specification (INSPIRE, 2014) is to define a common model for the spatial data theme on geographical names as defined in Annex I of the INSPIRE Directive. This theme comprises "Names of areas, regions, localities, cities, suburbs, towns or settlements, or any geographical or topographical feature of public or historical interest". The INSPIRE Data Specification provides a more detailed model for the locations instances in a gazetteer than the ISO 19112 standard. According to the description in the introduction (INSPIRE, 2014):

> "This data specification describes concepts related with geographical names, i.e. proper nouns applied to a natural, man-made or cultural real world entity. The data specification is guided by the multi-language and multi-scriptual situation in Europe: a geographic entity can have different names in one or several languages, and each name can have different spellings, i.e. spellings in different scripts. Because of this multi-language and multi-scriptual context, this specification defines a product that is feature oriented in order to enable to express which different names are used to designate one given place. In other words, the spatial objects defined in this specification are the 'named places', and the 'geographical names' are seen as information related to a named place."

2.5.2 Access to Gazetteer services

Concerning the development of an automatic processing chain in which the goal is to connect text and geographic space, one key issue is the availability and the accessibility of geographical data. Then, we will now describe how standard gazetteers are accessible.

The OGC[25] promoted the standardisation of gazetteers during last decade, and even proposed a specific interface for gazetteer services, through the Web Feature Service (WFS) specification. OGC web service specifications include also other standards such as Web Map Service (WMS) Specification, Web Map Tile Service (WMTS) Specification and Web Coverage Service (WCS) Specification. WFS is an implementation specification which defines interfaces for describing data manipulation operations of geographic features encoded in GML. The Gazetteer Service specification has been released by the OGC as an Application Profile of the WFS. The Application Profile defines the data model to access place names data from a gazetteer service (Lehto et al., 2013).

There are also some other types of interfaces such as Linked Data and RESTful services. Linked Data refers to a set of best practices for publishing structured data on a machine readable way and accessible by semantic queries (Bizer et al., 2009).

The concepts of Linked Data and Semantic Web (which are now very similar) were both introduced by the W3C[26]. Figure 2.19 shows the linking Open Data cloud diagram[27]. We can notice that geographical sources play an important role in Linked Data (GeoNames in particular, which is one of the most important dataset, but also others sources such as OpenStreetMap, LinkedGeoData, etc.). Basic rules of Linked Data are to use the RDF data model to publish structured data on the Web and use RDF links to interlink data from different data sources (Yu, 2011) such as DBpedia and Geonames. The web of data (or Web of Linked Data) is constructed with documents on the web and URIs identifies any kind of object or concept. In a geospatial context (Egenhofer, 2002), LinkedGeoData[28] is an effort to add a spatial dimension to the Web of Data. LinkedGeoData uses the information collected by the

[25] www.opengeospatial.org/
[26] http://www.w3.org/
[27] Linking Open Data cloud diagram 2014, by Max Schmachtenberg, Christian Bizer, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/
[28] http://linkedgeodata.org/

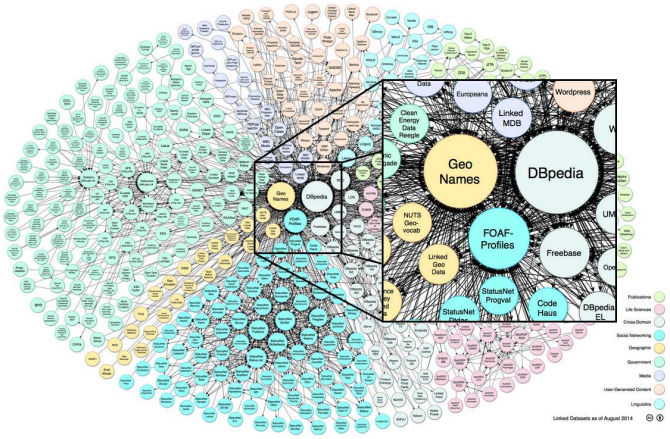


Figure 2.19: The Linking Open Data cloud diagram – **Source:** http://lod-cloud.net/

OpenStreeMap (OSM)[29] project and makes it available as an RDF knowledge base according to the Linked Data principles. Representational State Transfer (REST) is a software architectural style with high scalability, performance, and simplicity (Pautasso et al., 2014). Resources are identified using URIs, and RESTful services are implemented through a uniform interface (client-server) usually implemented by the HTTP protocol.

2.5.3 Coverage and Granularity

According to Leidner (2004), gazetteers vary in many dimensions: availability, scope, coverage, correctness, granularity, balance and richness of annotation. Gazetteers may be produced by private companies or by public sector which are providing paid access to the services, or free but under license with restrictions of use such as rate limit or maximum number of requests per day. Usually, public and official gazetteers provide a limited coverage where data are limited to a specific country or continent but with a high reliability. Official gazetteers are provided by national mapping agencies[30].

Official gazetteers provided by European countries such as BDNyme[31] (France), Nomenclátor Geográfico Básico de España[32] (Spain), and Toponimi d'Italia IGM[33] (Italy) are compliant with the INSPIRE data specification for geographic names. For instance, BDNyme which lists continental France's toponyms with their coordinates, has a coverage of more than 1.7 million entries.

Leidner (2007) proposes a comparison of several gazetteers based on these criteria and shows that the selection of the appropriate resource depends on the needs of the project. For example, in the case of a scientific study, the data must be freely available. The scope and coverage of the gazetteer is also an important criteria to assess the geographical resources. Usually, Gazetteers with a worldwide scope provide less coverage in term of density than national gazetteers. However, several resources may be used in conjunction in order to take advantage of each one. For instance, Florczyk et al. (2010) present an architectural approach for adaptive compound geocoding Web services using diverse geographical resources such as gazetteers and address geocoding services.

Additionally, when the existent gazetteers are not pertinent for the needs of a specific project, people need to populate their own gazetteers. There are some works in the literature that aim at the creation of new databases of geographic locations. For instance, Lieberman et al. (2010) present a method for generating a local spatial lexicon by processing a corpora from local newspapers, which is used later for the geocoding of documents. In principle, the method only aims at including in this spatial lexicon the

[29] http://www.openstreetmap.org
[30] http://unstats.un.org/unsd/geoinfo/ungegn/geonames.html
[31] http://www.geoportail.gouv.fr
[32] http://www.ign.es
[33] http://www.pcn.minambiente.it/GN/

disambiguated toponyms extracted from newspapers' articles and discovered in an existing gazetteer. The method used for disambiguating toponyms with possible interpretations is based on the definition of a convex hull covering nearby possible locations of toponyms. However, the method could be easily expanded to assign the geographic extent of this convex hull to all those names recognized as place names, but not found in the gazetteer.

Several works attempt to automatically structure large-scale geographic information using Flickr[34] geo-referenced images. For instance, Rattenbury et al. (2007) and Ahern et al. (2007) described a method for automatically building databases containing geographic names and their associated coordinates. They extract geographic information (place and event semantic) from a large set of tags introduced by Flickr users. They analyse the tags associated with the geo-referenced Flickr images to generate aggregate knowledge. Furthermore, Rattenbury and Naaman (2009) present different methods based on burst-analysis techniques for extracting place semantics from Flickr Tags. Training these methods with Flickr data containing high-resolution location metadata (i.e., longitude and latitude), it is possible to derive automatic associations between place name tags and explicit georeferences. As the authors claim, these methods could automate the creation of place gazetteer data. In the same line, Serdyukov et al. (2009) propose the development of a statistical language model to predict the most likely geocoding corresponding to a set of location tags associated to a Flickr photo. Additionally, there are some works applied in similar contexts to the ones proposed for the context of our experiments, i.e., narrative descriptions of places in small areas, whose objective is also to infer the spatial location of complex place names, not directly found in gazetteers. In this area Scheider and Purves (2013) have proposed the use of semantic technologies to process narrative descriptions of mountain itineraries or historic places to infer the location of complex names in terms of spatial relations with respect to well recognized landmarks.

Finally, there are some works aiming at the enrichment of existing geographic databases. Popescu et al. (2008) present a new automated technique, based on a linguistic analysis of Web documents, for creating and enriching a geographical gazetteer, called Gazetiki. Their proposal merges disparate information from Wikipedia, Panoramio[35], and web search engines in order to identify and categorize geographical names and find their geographical coordinates. Another example in this area is the work of Smart et al. (2010). They present a mediation framework to access and integrate distributed gazetteer resources to build a meta-gazetteer that generates augmented versions of place name information. In this mediation framework they include geofeature augmentation module. During the merging process of features from different sources, they detect matching of these features and create more complete and consistent information, e.g., adding hierarchical information about administrative units. Another work about the enrichment of existing toponyms could be the work proposed by Hao et al. (2010) for the enrichment of destinations in travelogues with knowledge (e.g., local topics such as beach, mountain or other features related to the toponym) mined from a large corpus using probabilistic models.

2.5.4 Description of some well-known gazetteers

Additionally, there are also some gazetteers with a worldwide coverage including names from every part of the world such as EuroGeoNames, GeoNames[36] and OSM.

EuroGeoNames

EuroGeoNames (EGN)[37] combines geographic names from the NMCAs across Europe to create a unique service and data set. ISO 19112 schema was used as the main data input and output schema and now EGN contents are compliant with the INSPIRE specifications. A central data collection of exonyms and other variant names (e.g., historical names) covering 25 national languages and regional languages has been created to support multilingualism. EuroGeoNames is still under development by EuroGeographics[38] which represents the European National Mapping, Cadastral and Land Registry Authorities.

[34] Flickr is a popular photo-sharing website that supports user-contributed tags and geo-referenced images: www.flickr.com

[35] Panoramio is platform dedicated to the sharing of geo-referenced images: www.panoramio.com

[36] http://www.geonames.org

[37] http://www.eurogeographics.org/eurogeonames

[38] http://www.eurogeographics.org/

UN/LOCODE

UN/LOCODE[39], the official gazetteer proposed by the United Nation contains over 100,000 locations in 249 countries. The United Nations Code for Trade and Transport Locations is a geographic coding scheme developed and maintained by United Nations Economic Commission for Europe (UNECE). LOCODE have five characters, the first two characters are letters and code the country and the three remaining characters code a location within that country. For instance 'FR PAR' refers to Paris in France and 'US NYC' refers to 'New York City' in the United States. The table's structure contains also other information such as name, sub-division, function, coordinates, etc. UN/LOCODE is freely available, it can be consulted online and downloaded (MS Access database, text, CSV and/or HTML files).

GeoNames

GeoNames is a project for the creation of a world geographic database which is provided mostly by official public sources such as the National Geospatial Intelligence Agency, the United States Geological Service and the U.S. Board of Geographic Names. It contains more than 8 million toponyms categorised into nine feature classes, administrative boundary, hydrographic, area, populated place, road or railroad, spot, hypsographic, undersea, and vegetation. GeoNames uses the WGS94 standard which is the reference coordinate system used by the Global Positioning System (GPS). The database can be downloaded or queried online. The first five results of a query for the toponym 'Paris' are listed in Figure 2.20. Furthermore, Figure 2.21 shows the same results on a map.

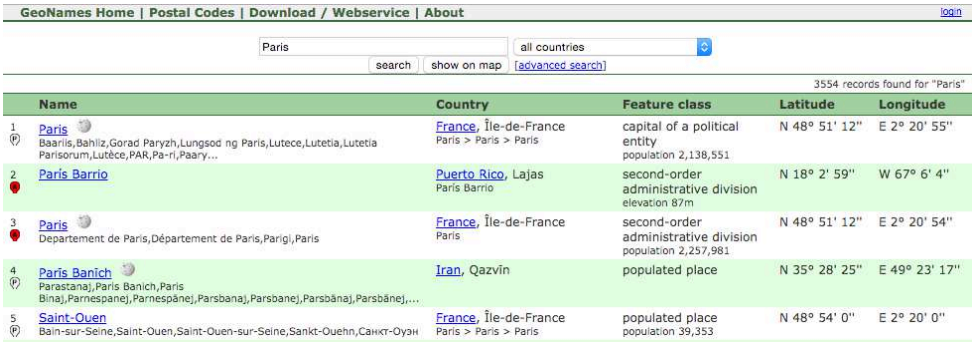


Figure 2.20: First five records for the toponym "Paris" in GeoNames

Openstreetmap

The OSM project aims to create a free digital map and to provide data concerning toponyms and spatial features. OSM is a participative project taking benefits of crowdsourcing and involving a large-scale community over the Web. The notion of Volunteer Geographic Information introduced by Goodchild (2007), raises the question of information quality. For instance Haklay (2010) noticed the lack of coverage in rural or poor areas. The results of a query for the toponym 'Paris' using the Nominatim Search API[40] are shown in Figure 2.22. Nominatim is a tool to browse OSM data by name and address. In this example results are listed in XML format but Nominatim may also provide JSON output.

WordNet[41] is a lexical database for English containing more than 100,000 concepts and expressing relations such as hyponyms, holonyms and synonyms (Fellbaum, 2012). Words (nouns, verbs, adjectives and adverbs) are grouped into sets of cognitive synonyms called synsets, each expressing distinct concepts. Relations such as *is-a*, p*art-of* or *instance-of* are very useful to define relations between geographical entities. For instance, in WordNet the toponym 'France' is an instance of the concept 'country' and 'Paris' is an instance of the concept 'national capital' and is also part of the entity 'France' (Fig. 2.23).

Synsets of WordNet are not representing alternate names like in typical gazetteers but synonym, for instance 'French capital' is considered as a synonym of the name 'Paris'. These synsets representation

[39] http://www.unece.org/cefact/locode/welcome.html
[40] http://nominatim.openstreetmap.org
[41] http://wordnet.princeton.edu/

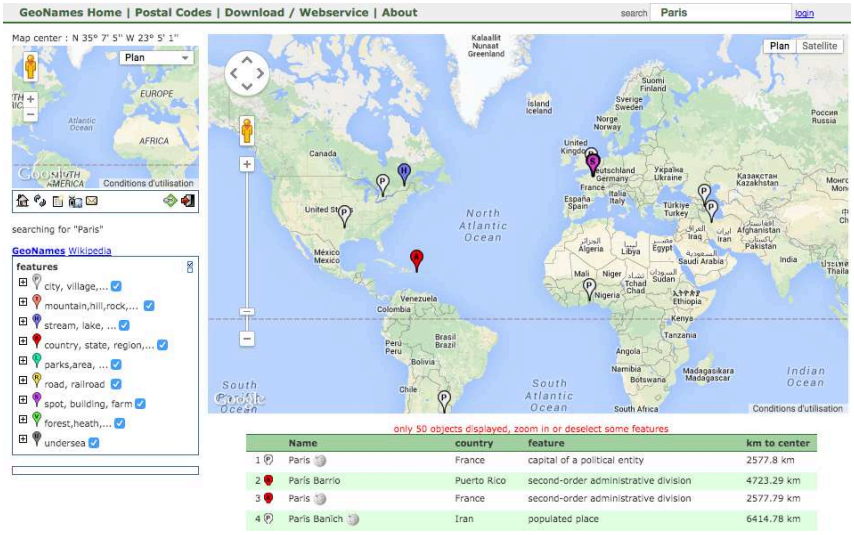


Figure 2.21: Map-based results for the toponym "Paris" in GeoNames


```
▼<searchresults timestamp="Thu, 21 May 15 13:08:41 +0000" attribution="Data © OpenStreetMap contributors, ODbL 1.0. language=fr
  <place place_id="127732055" osm_type="relation" osm_id="7444" place_rank="16"
  boundingbox="48.8155755,48.902156,2.224122,2.4697602" lat="48.8565056" lon="2.3521334" display_name="Paris, Île-de-France,
  France métropolitaine, France" class="place" type="city" importance="0.97893459932191"
  icon="http://nominatim.openstreetmap.org/images/mapicons/poi_place_city.p.20.png"/>
  <place place_id="127330920" osm_type="relation" osm_id="71525" place_rank="12"
  boundingbox="48.8155755,48.902156,2.224122,2.4697602" lat="48.85881005" lon="2.32003101155031" display_name="Paris, Île-de-
  France, France métropolitaine, France" class="boundary" type="administrative" importance="0.97893459932191"
  icon="http://nominatim.openstreetmap.org/images/mapicons/poi_boundary_administrative.p.20.png"/>
  <place place_id="63292496" osm_type="way" osm_id="33063046" place_rank="16"
  boundingbox="35.26725,35.3065029,-93.7618069,-93.6750829" lat="35.2920325" lon="-93.7299173" display_name="Paris, Logan Count
  Arkansas, États-Unis d'Amérique" class="place" type="city" importance="0.68385884862688"
  icon="http://nominatim.openstreetmap.org/images/mapicons/poi_place_city.p.20.png"/>
  <place place_id="63211936" osm_type="way" osm_id="33299478" place_rank="16"
  boundingbox="33.611853,33.738378,-95.6279279,-95.4354549" lat="33.6617962" lon="-95.555513" display_name="Paris, Lamar County
  Texas, États-Unis d'Amérique" class="place" type="city" importance="0.55374443751163"
  icon="http://nominatim.openstreetmap.org/images/mapicons/poi_place_city.p.20.png"/>
```


Figure 2.22: First four records for the toponym "Paris" in OSM

Figure 2.23: Results of WordNet for the query "Paris"

aims at building a network of word and relations. For that reason, WordNet is widely used in NLP tasks and more specifically in word sense disambiguation.

This section has described gazetteers that are used to store and query information about locations. We have seen that it exists a multitude of datasets of geographical data which can have different coverage

and granularity. These datasets can also provide access to other datasets of the Semantic Web (Linked Data) and are accessible through gazetteer services.

2.6 Summary

In this chapter, we have described the concepts of itinerary and route instruction. We have seen how space and motion are expressed and characterized in language (Talmy, 1985; Vandeloise, 1986; Aurnague and Vieu, 2015). Furthermore, as we have seen, there are works dealing with route descriptions in urban areas (Ioannis et al., 2014; Götze and Boye, 2015) and other on natural environments (Brosset et al., 2008; Sarjakoski et al., 2011). Despite their differences, all these works use expressions of landmarks and actions (Denis, 1997) which are expressed in language mainly by place names, nouns and verbs. According to these works and with respect to our concern, we describe in Chapter 3 how spatial cognition and spatial analysis are used for the reconstruction of itineraries from a descriptive texts.

According to Jackendoff, a satisfactory computational theory is an essential factor in developing suitable general algorithmic and implementation theories. Unfortunately the mainstream models of language (such as the proposals arising from Chomsky's approach (Chomsky, 1965)) do not lend themselves to being easily adapted to textual information processing. There are at least three reasons for this. One is that these models have no theory of meaning to which the language faculty is connected. A second reason is that most models have been developed in view of the generation and not for the analysis of the meaning. A large part of these models, focusing almost exclusively on the computational theory, find themselves largely out of touch with other domains. A third problem is that this frameworks do not really distinguish what has to be considered as a lexicon and could be therefore stored (in memory) from what is designed in runtime (when creating a sentence). Several frameworks such as the one proposed by Jackendoff characterized linguistic structures in terms of constraints rather than in terms of algorithms that generate sentences. The view of the lexicon shared in many of these constraint-based approaches is that there is no principled formal distinction between words, rules, and other larger assemblages. This chapter has described several NLP approaches for NER and toponym disambiguation. As we have seen these approaches are classified in two main categories, data-driven approaches and knowledge-based approaches. Each one having advantages and drawbacks. The main drawback of data-driven approaches is the lack of classified collections and the need of large corpora of annotated ground truth. Although knowledge-based methods could be time-consuming to develop, they require only a small amount of training data. Furthermore, knowledge-based methods are more adapted to approaches based on domain-specific analysis of corpus and rules are described in a readable way and are easy to modify and maintain.

Many of the toponym disambiguation methods described in Section 2.3.4 use toponyms that are geographically the closest to disambiguate the candidate toponyms. This can lead to poor results when important information is not included in the context, when the candidate toponym is not geographically close to non-ambiguous toponyms, or if it is not linked to a geopolitical entity (Speriosu and Baldridge, 2013). Some studies use the notion of event to disambiguate toponyms. For instance, Roberts et al. (2010) consider that there are three types of entities that participate in an event: people, organisations, and geopolitical entities. They use an ontology constructed from Geonames, and associate geopolitical entities with people and organisations using links from Wikipedia, but no other information or clues is used from the context. These various methods are often applied to corpora of news articles in which toponyms are associated with events, well-known figures or geopolitical entities and not with spatial relations (Garbin and Mani, 2005; Buscaldi and Magnini, 2010; Speriosu and Baldridge, 2013). In this type of discourse, toponyms are not necessarily related to each other, and are not for example linked by motion events. Moreover, Speriosu and Baldridge (2013) show that toponym disambiguation methods that are based on the text (context extraction and interpretation of spatial relations) are more effective than knowledge-based methods (using metadata about locations found in gazetteers) or heuristics that use distance calculations. Moreover, map-based methods need more context than knowledge-based methods. With respect to our concern, we describe in Chapter 4 our proposal for a toponym disambiguation method. We show how we combine map-based and knowledge-based methods to solve toponyms ambiguities in the specific context of itinerary descriptions.

Markup languages described in Section 2.4 may be categorized into three classes: spatial, spatio-

temporal and generic markups. For the specific task of spatial information representation, several XML-based languages have been proposed as syntactic approaches for encoding geospatial information. We distinguish two categories of markup languages, those focused on the encoding of information and those focused on annotation of texts. The first category refers to exchange formats of data whereas the second one refers to the annotation of information in textual documents written in natural languages. We show that both types of markup languages are complementary and that encoding markup languages can be extended or integrated within other markup languages for text annotation purposes. Furthermore, with regard to our concern of extracting spatial information in texts, we show in Chapter 5 how it is possible to provide an adaptable generic markup language (based on TEI), which can be easily used in a fully automatic process.

Chapter 3

Reconstruction of Itineraries from Text

There are paths without travelers. But there are even more travelers who don't have their own paths.

— Gustave Flaubert, *Letters to Madame Louise Colet*

Contents

3.1 Introduction

This chapter proposes an approach for the automatic geocoding of itinerary described in natural language. The contribution described in this chapter (Figure 3.1) addresses the problem of identifying waypoints and finding their sequence in order to build a geocoded representation of the route of the itinerary. We propose a generic method for the automatic reconstruction of itineraries from texts, combining information extracted from texts and information found in geographical databases.

Modelling and analysing itineraries lies in the general framework of Time-Geography and received much attention in the literature (Winter and Raubal, 2006). In particular, Spaccapietra et al. (2008) proposes a pattern for conceptually modelling itineraries and its implementation to store and query this model in a Database Management Systems (DBMS). Some others focus on analysing the trajectory to discover knowledge about movements (Laube et al., 2005). In our model, in addition to the classical key elements of waypoints and paths, a key proposal is to enrich the description with the notion of related points used to describe the itinerary (e.g., feature seen or mentioned as landmarks), leading to consider the described itinerary as a directed acyclic graph with branches linking each landmark to the place from where it is seen or mentioned.

The remainder of this chapter is structured as follows. Section 3.2 proposes a definition of the notion of itinerary to identify the key concepts involved in the description of itineraries in texts. Section 3.3

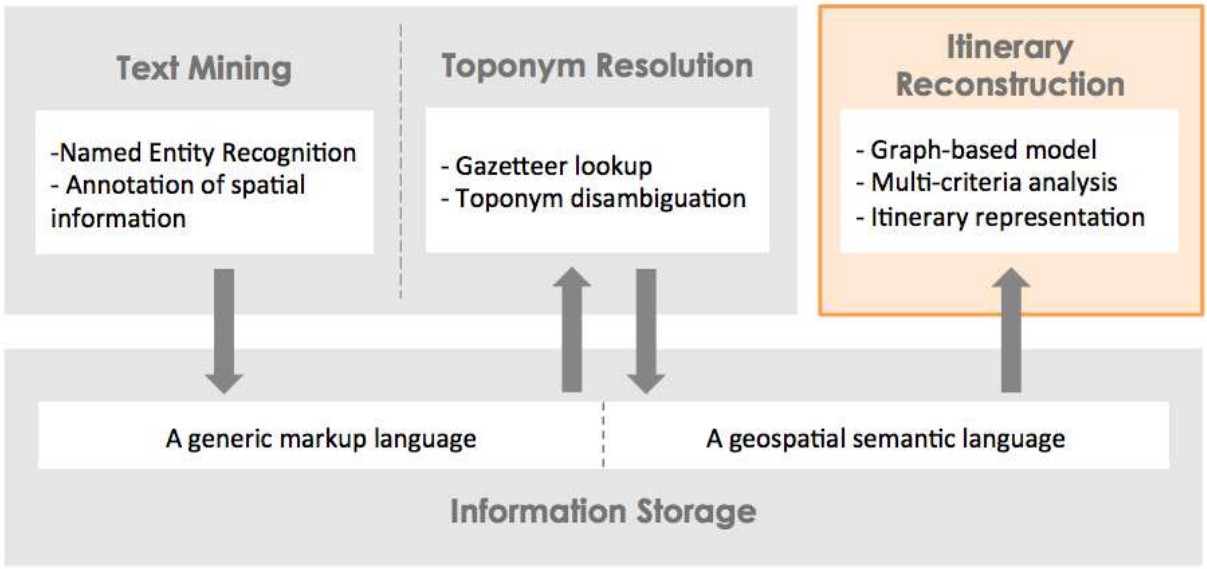


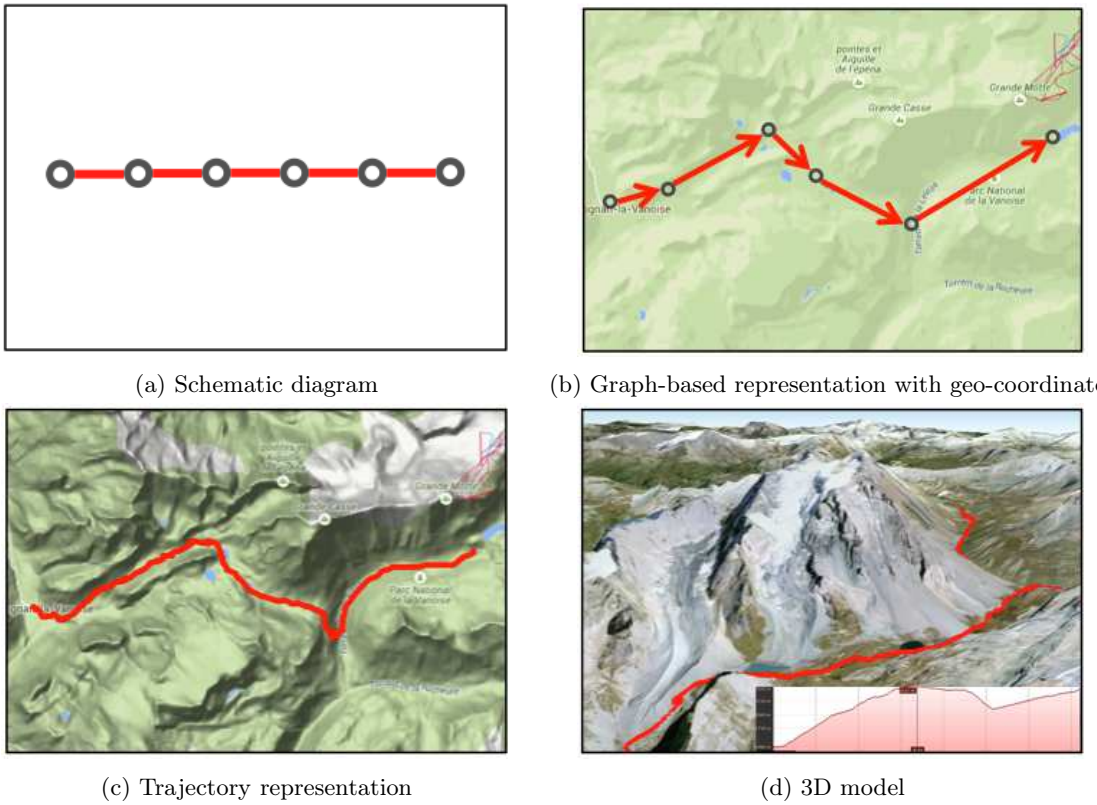
Figure 3.1: Contribution of this chapter (highlighted in orange)

describes a generic model for the representation of itineraries and an automatic process for the itinerary reconstruction. Finally, Section 3.4 describes a method for the automatic approximation of the spatial footprint of the displacement and Section 3.5 summarises and concludes this chapter.

3.2 Description and Concepts of Itinerary

Considerable amounts of geographical data are still collected not in the form of GIS data but in natural language texts form. For example, in the GIS area there are several types of narrative structures describing itineraries or displacements, each one in a different way. One type of these descriptions is narrative descriptions of real journeys (i.e., travelogues) or travel novels (e.g., Gulliver's Travels). In this kind of texts the description of itineraries is just a piece of information in a story with lots of descriptions involving persons, events and places not always related to the itinerary. The description of a displacement is usually scattered in the text along many other things. Another kind of descriptions of displacements are the ones provided in emergency calls, recorded by emergency services. These calls are made by citizens who require in-situ assistance for any kind of incident and try to describe their location using place names and motion expressions (from where they come or where they go) and using landscapes features and perception expressions (what they are able to see around them). Hiking guides belong to another category of narrative text describing itineraries. They can be considered to be a variant of route instructions, which are a composite of procedural and descriptive discourse. In this kind of descriptions all the information is related to the itinerary. Although itineraries are described using the same elements as for the other narrative categories, they are structured as instructions. The reader must follow these instructions in order to take the same directions as expressed in the described path.

From a geographical point of view, an itinerary can be represented in different ways. From the simplest to the complex one. Figure 3.2 shows different examples of representation. From the schematic diagram to the 3D model. Figure 3.2a shows the representation of an itinerary with a schematic diagram. This simple type of representation is sometimes called *topological map*, it does not show the geographic locations but rather the relative positions which are not drawn to scale. Thus, in this simplified representation waypoints are equidistant. This basic design is widely adopted for transport network maps such as the Tube Map of the London Underground. Figure 3.2b shows a graph-based representation in which vertices are associated with real-world locations. Vertices refer to waypoints and are associated to geographical information such as geographical coordinates. Edges of the graph refer to the Euclidean distance between



(a) Schematic diagram

(b) Graph-based representation with geo-coordinates

(c) Trajectory representation

(d) 3D model

Figure 3.2: Examples of itinerary representation: from topological map to a 3D model

two waypoints and may represent a first approximation of the route (using straight lines). Figure 3.2c shows a trajectory representation, which describes with high relevance the real route of the itinerary. This accurate representation uses information of terrain profile and takes into account the geometry of spatial objects. For instance, the trajectory may follow the geometry of rivers, roads (lines) and cities (polygons). Finally, one last representation may be the creation of a 3D model describing the route of the itinerary (Figure 3.2d).

In this work, we propose to build a graph-based representation (Fig. 3.2b) using geo-coordinates as vertex coordinates. We also suggest further improvements in order to build a trajectory representation (Fig. 3.2c).

3.2.1 Characterisation of Itinerary in Text

As we have seen in Section 2.2, itineraries and displacements are described in natural language using spatial named entities (i.e., toponyms), spatial relations (Bloom, 1994; O'Keefe, 1996), perception expressions with description of landmarks, motion expressions and trajectories (Talmy, 1985, 2000). An itinerary can be defined as a sequence of displacements between places called waypoints. Waypoints and routes are the two main elements involved in the description of an itinerary.

In this dissertation, we distinguish the terms *routes* and *paths*, and according to the definition given by Montello (2005) a path refers to the physical feature (pathway) upon which travel occurs (e.g., streets, trails) and a route refers to a displacement occurring on a path or across areas that contains no paths.

The following sentences are extracted from a French hiking description[42] and has been translated into English for the sake of clarifying the context of this dissertation:

[42]This hiking description corresponds to the representations shown in Figure 3.2.

(20) This hike goes from **Pralognan** to the **refuge of Leisse** passing by the impressive **Grande Casse**, all in a wild and dotted with lakes.

(21) In **Pralognan**, follow the road between **Hotel de la Vanoise** and **Hotel du Petit Mont Blanc** and go straight.

(22) Further pass on **Chanton bridge** and cross the forest. Soon after, you will reach **lake Des Vaches** [...]

(23) At a small crossroads, you can glimpse **Pointe du Creux Noir** then branch off south in the direction of **lake Long** which you will bypass from the right. [...]

(24) You can see all the way down **Croe-Vie bridge**.

(25) To reach it go all the way down, then cross it. [...]

(26) At the crossroads, do not take south towards the **refuge of Entre-Deux-Eaux**, but go north and walk one hour.

(27) Then follow **Leisse torrent** to achieve the day's stage.

(28) This last part is done in a wild and beautiful steep-sided valley. [...]

This typical example of route instructions illustrates the information used in itinerary descriptions to describe displacement between places. Firstly, it shows that place names (in bold) can have two roles: waypoint or visual cue. The place names 'Pralognan', 'refuge of lake Long' and 'lake Des Vaches' refer to waypoints. On the opposite, the place name 'Pointe du Creux Noir' is not considered as a waypoint because it is associated with the verb of perception 'glimpse', which means that this location is not reached during the displacement, but nevertheless useful because it acts as a visual landmark. Expressions of perception are an important component of the description of itineraries in texts. Indeed, the purpose of descriptions of itineraries is not only to describe displacements but also the context in which occurs the displacement. As we may notice with sentences (23) and (24) some places or features are used to describe the landscape in order to give more details about the environment or about the actual location of the narrator or a reader following instructions in the specific case of hiking descriptions. For instance, hiking trails usually offer nice panorama over mountains or lakes during a trip. Therefore the textual description must describe both the displacement with the different waypoints and locations or features describing landscape.

The place 'refuge of Entre-Deux-Eaux' is not considered as a waypoint either because it is associated with a negation expression. Intuitively we might imagine that with these kinds of association we can directly give the role of visual cue to the involved place name. But if we observe the two sentences (24) and (25) we can deduce that the place name 'Croe-Vie' has a double role. It has a visual cue role when mentioned in the sentence (24) in association with the verb 'to see'. But the sentence (25) changes his role into waypoint. Considering an automatic NLP process, the anaphorical form of its second evocation causes a problem not solved yet. Therefore the waypoint role is given by default and visual cue retained as a possibility.

The example also shows that the order of mention in the text does not always correspond to the order of achievement of the hike as in sentence (20), or in the case of perception or negation such as in sentences (23) and (26). Furthermore, motion or paths connecting waypoints are also used to describe itineraries. For instance, in sentence (23) 'branch off south in the direction of lake Long' means that arriving at a crossroads you have to take the junction going south.

Furthermore, we can notice that toponyms may be associated with one or more concepts related to the expression of location or motion in the language. These expressions may refer to spatial relations and are classified into three categories: topological, projective and metric relations. These spatial relations can express both the location of a toponym or spatial constraints between two spatial objects.

(29) passer par le nord du hameau de Friburge
get around from the north of hamlet Friburge

(30) marcher vers le sud jusqu'au lac de Grattaleu
walk south to the refuge of Lake Grattaleu

For instance, in phrase (29), the spatial relation 'the north of' is associated with the toponym 'hamlet Friburge'. In this case, the location of the waypoint is not the location of the toponym 'hamlet Friburge' but the location of the spatial object described by the whole expression 'the north of hamlet Friburge'. However, in phrase (30) the spatial relation 'south to' indicates the direction of the displacement to reach the location of the refuge of Lake Grattaleu.

Regarding motion expressions, as we have seen in Section 2.2.4, syntactic parts of speech, in particular motion verbs, characterise a motion event. We consider motion relations expressed in text as a specific type of spatial relations. Motion verbs involve a change of location and sometimes also a change of elevation (e.g., climbing). Moreover, we categorised these verbs according to their aspectual polarity. The three polarities are: initial (i.e., to leave), median (i.e., to cross) and final (i.e., to arrive). Prepositions are also playing an important role to describe motion events. More specifically, the association of a motion verb with a preposition of place and direction (e.g., *from*, *in*, *at*, *to*, *by*, etc.) can change the focus of the displacement to take on the polarity of the preposition instead of the verb.

Finally, in texts describing travel stories as well as those describing hikes, starting and ending points are almost always given. But here again considering an automatic NLP process, different problems may occur. For instance, in the previous example the ending point 'refuge of Leisse' is given in the sentence (20) and not at the end of the description. How to get this information? Either make a complete semantic analysis of sentence (20) to identify that the three places mentioned represent the beginning of the hike, an intermediate waypoint and the end, or solve the anaphora: 'day's stage' in sentence (27) which refers to the ending point 'refuge of Leisse' mentioned in sentence (20). Thus, to process and identify this information automatically, it is necessary to use discourse analysis techniques, which still nowadays are complex and unreliable.

3.2.2 Components of an Itinerary

The main focus of an itinerary is to describe displacements over space and time. We propose to identify the minimum number of information needed to represent an itinerary from its verbal description. Furthermore, for the definition of itinerary described in natural language, we follow the definition given by Loustau et al. (2008). Itinerary is a geographical concept defined by spatial, temporal and semantic components.

Moreover, as we have seen in Section 3.2.1, an itinerary is characterised in texts by spatial named entities, terms having a geographical sense, expressions of spatial relations, expressions of motion, spatial prepositions and expressions of perception. An itinerary may be considered as a set of displacements, i.e. a sequence of displacement between places, where each step is represented by waypoints and routes. Motion is represented by the combination of the spatial, temporal and semantic components. This section highlights the following properties for representing itineraries:

- spatial properties: places associated to their real locations
  - starting and ending point
  - intermediate waypoints
  - landmarks not considered as waypoints
- spatio-temporal properties: relations between points
  - motion relations
  - non-motion relations
  - path connecting each waypoint
  - length and duration of the itinerary
- semantic properties
  - general purpose of the trip
  - role of each relation: motion, perception, etc

These properties refer to the minimal number of elements needed for building a geographical representation of an itinerary from its verbal description. However, in this thesis we are focusing on spatial and spatio-temporal properties of the itinerary. The spatial component is the core component of the definition of an itinerary, it describes the relevant spatial areas in relation with the itinerary. The overall spatial

area of an itinerary consists of starting and ending points, intermediate waypoints and routes connecting all these different locations. A route is connected to two waypoints and an itinerary contains at least one route and two waypoints. Furthermore, other spatial entities are involved in itinerary description, such as landmarks used to describe the spatial context of the itinerary. These landmarks are not considered as waypoints and give details about the spatial location of waypoints and routes composing the itinerary. All waypoints (i.e., starting, ending and intermediate points) are expressed in the textual description using spatial entities. From a geographical point of view, these spatial entities may be represented by different shapes or geometries (e.g., point, line, surface). The geometry of these spatial entities are defined by their inherent spatial representation (e.g., city, country, river) and may also depend on the geographical scale.

The temporal component is also a key component of an itinerary, it is defined by dates, durations and temporal relations between each step of displacement. The temporal relations define the temporal properties of an itinerary such as date of arrival at a waypoint, date of departure from a waypoint and duration of the displacement. Each step of displacement is represented by waypoints and temporal relations. The dates of arrival and departure from the same intermediate waypoint can be different if an activity occurs in this location (e.g., eating, sleeping, taking pictures, etc). Thus the arrival date at a waypoint is less or equal to the departure date from this waypoint.

In our work we also consider some semantic aspects such as types of relations (e.g., motion, perception, etc.) and purpose of relations which may be defined by the general purpose of the trip itself (e.g., hiking, vacation trip, historical travel story, gastronomic itinerary, sporting path). These semantic aspects can give important information concerning the geographical context of the displacement. Indeed, geographical features of waypoints and landmarks may be related to the semantics of the itinerary (e.g., mountains, lakes, churches, vineyards, etc). However, in itinerary descriptions the semantic component may also describe the role of each relations between places. For instance, most relations are describing displacements and motions (see examples 31 and 32), but some are expressing perception (see examples 33 and 34).

(31) puis nous arrivons à Courmayeur
then we arrive at Courmayeur

(32) traverser le pont de la Glière
continue over the Gliere bridge

(33) ne pas prendre à droite le chemin qui mène au Moriond
do not take to the right the path that leads to Moriond

(34) une vue magnifique sur le Mont Blanc
a beautiful view of the Mont Blanc

Expressions of perception are important in some contexts of evocation, especially when the narrator wants to report some specific situations or feelings. It is frequently used in itinerary description to describe landscapes and give more details about the geographic context of the itinerary.

In this thesis, we consider relevant entities as being spatial entities and relations as displacement or perception between these spatial entities. In the remainder of this chapter, we will propose a model which considers these properties in order to build a first approximation of the route of the itinerary. Furthermore, Chapter 4 describes an automatic NLP method for identifying and annotating automatically these elements in text.

As we have seen, itinerary is a geographical concept defined by spatial, temporal and semantic component. In this thesis, we are focusing on the spatial and spatio-temporal components of the concept of itinerary. Spatio-temporal relations define a relative space, whose properties depend on the configuration of the expressed relations. Furthermore, we consider temporal relations as a specific type of spatial relations. Indeed, duration of displacement between two places may be interpreted as a specific type of distance. For instance, the following phrases 'walk for 20 minutes' and 'walk for two kilometers' express both duration and geographical distance (Kemmerer, 2005).

In order to establish the steps of an emerging route, an itinerary is defined as being a special type of spatio-temporal relation. It is a spatio-temporal sequence of steps moving between different places. The route of an itinerary could thus be considered as a succession of spatial relations.

3.3 Automatic Itinerary Reconstruction

As mentioned in the introduction of the dissertation, we divided the problem of the automatic reconstruction of itineraries from texts into three sub-problems. The first one is to find the locations of all spatial named entities expressed in the text. This problem involves the annotation, the resolution and the disambiguation of toponyms. The second problem is to find the sequence of waypoints that depict the order in which the waypoints are crossed during the displacement and build a first approximation of the representation of the itinerary. Then, the third problem is to propose a better approximation of the representation (not just straight lines between waypoints) taking into account the availability of route networks in urban areas and geographical obstacles (e.g., rivers, mountain peaks) in rural areas.

In this section, we address the second and the third problems: finding the sequence of waypoints in order to build a geocoded representation of the itinerary and propose an approximation of the real path. We propose a generic method for the automatic reconstruction of itineraries from texts, combining information extracted from texts and information obtained from geographical resources.

3.3.1 A Graph-Based Model of Itineraries

As we have seen in Section 3.2 with the definition of the concept of itinerary, itineraries consist of waypoints, routes connecting waypoints, visual cues and places not reached. Waypoints represent places reached during the displacement, and routes represent motion. We classify waypoints into three categories: starting point, ending point and intermediate points. In the case of loops, starting and ending points are the same. In addition to this, other spatial information is used to define an itinerary in a text, such as places not reached during the displacement. These places are not considered as waypoints because they are not directly involved in the route. They are used to describe landscape and can contribute to infer locations of unnamed waypoints located on the route between two other waypoints.

We propose a graph-based model for the representation of itineraries. We define an itinerary as a Directed Acyclic Graph (DAG), $G = (V, E)$ comprising a set V of vertices and a set E of edges. The edges of the graph represent route segments and the vertices represent locations (Fig. 3.3). Each vertex v of G is associated with its real-world location and each two consecutive vertices are connected by an edge. The leaves[43] of G represent the starting point and ending point and also the points that are not considered as waypoints. This graph contains a main edge representing the displacement (solid lines in Fig. 3.3) and secondary edges representing the relations between waypoints and places not reached during the displacement (dashed lines in Fig. 3.3), such as places seen or described by the narrator. In a more formal way a displacement can be represented as a sequence of waypoints (locations). Each sequence has the form $(w_1, \cdots, w_n)$ where for each $i < j$, the w_i waypoint is reached before w_j. When a location is involved several times in the itinerary, for instance in the case of loops, we consider several vertices representing the same location in order to avoid cycles.

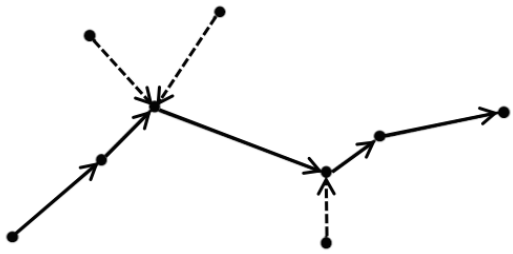


Figure 3.3: Example of graph-based representation of an itinerary

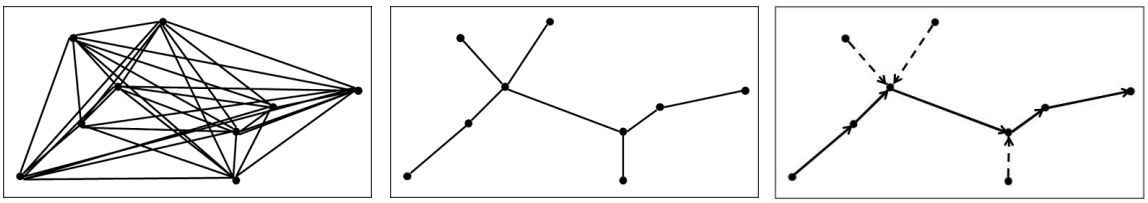
We propose to consider a set of information needed for the automatic construction of a DAG that represents the described itinerary. Some of this information can be extracted from the textual description

[43] Leaves are defined as vertices having only one incident edge (terminal vertices).

of the displacement: sequence of place names in the text, temporal relations (e.g., 'after', '2 hours later'), spatial relations (e.g., 'south of', '2 km', 'in the direction of'), polarity of the displacement (e.g., 'to leave', 'to arrive'), and the use of a place name with a perception or negation expression (e.g., 'to see', 'don't go to'). Other information can be obtained from external geographical resources: geographical distance or terrain profile between two places.

The purpose of the reconstruction of the itinerary is to interpret and link spatial information in order to reconstruct the route which refers to the described displacement. Our proposal is to combine the use of all this information, when available, as criteria in order to find the most likely route linking each step of the displacement. Since the target is to provide a generic method that can deal with all types of narrative structure describing itineraries, we make the assumption that we do not know the starting and ending points of the itinerary and the sequence of waypoints either. Therefore, the challenge is to find the itinerary that is closer to the real route intended by the authors who wrote the text.

We start by building a complete graph $K_n = (V, E)$ where all vertices v are connected, then we propose to use a multi-criteria analysis approach to compute and assign a weight to each edge of this graph (Fig. 3.4a). The weights represent the probability for an edge to be in the final route. Once we have a complete weighted graph, we compute a minimum spanning tree in order to get an undirected acyclic graph connecting all vertices (Fig. 3.4b). Then we transform this tree into a partially directed acyclic graph in order to identify the sequence of waypoints and build the DAG representing the itinerary (Fig. 3.4c).



(a) Complete graph (b) Minimum Spanning Tree (c) DAG representing the itinerary

Figure 3.4: Example illustrated of the process: from a complete graph to a DAG

3.3.2 Multi-Criteria Analysis Approach

The first step of our approach is to build a weighted complete graph, where each vertex represents a location extracted from the description of the itinerary. Then, we propose to use a multi-criteria analysis approach to assign a weight to each edge of the complete graph. Our approach combines local information extracted from the text with physical features extracted from external sources such as gazetteers or datasets providing digital elevation models. This combined spatial and textual analysis aims at resolving some ambiguities and reconstructing the geocoded representation of the route of the itinerary. The aim is to identify waypoints and find the most probable itinerary linking them with a minimal *length*. The term length is not referring only to geographical distance, but to an aggregated value that takes into account different criteria whose weight is going to be minimized. This length is a combination of contextual information extracted from the description and geometric information like terrain profile or geographical coordinates. Finding this optimal itinerary should help to remove ambiguities or places appearing in the text but not actually crossed. This naturally leads to the notion of minimal *weighted spanning tree*. The minimum weighted spanning tree of a set of vertices is the tree connecting all the vertices together with the minimum weight, this weight being the sum of the weights of the edges linking vertices. As we are looking for the minimal spanning tree, all the criteria have to be minimized, that is to say, the lower the values are, the better it is. The criteria used in the proposed approach are described in this section and the approach to combine these criteria is explained in Section 3.3.3.

Sequence of the displacement (C_1)
The first information that can be easily extracted from the text is the sequence in which the places

appear. However, the sequence of places in the text is not the same as the sequence of the itinerary. Indeed, ordering places as they occur in the text is not effective most of the time. In many cases, the discourse is not linear and the sequence of place names in the text can be totally different from the real sequence of displacements. In such cases, the order of place names should not be taken into account in our decision process or with a lower weight in comparison with the other criteria. Anyway, this can be an important information to help taking decision among several alternatives. For example, in the specific case of hiking descriptions, sequence of place names in the text is often close to the real one. In this case, the order of place names in text is an imprecise but useful information for the decision process. We use this information to define a criterion as the distance between two place names in the textual description, in other words it represents the number of place names appearing between those two place names. Each place name is associated with a number s_i equals to its order of apparition in the text, with $i \in [1, n]$ where n is equal to the total number of place names in the text. The value of the *text distance criterion* for an edge (i, j) is the distance between two place names in the text: $C_1 = |s_i - s_j|$.

Geographical distance (C_2)
We compute the orthodromic distance (or great circle distance), which is the shortest distance between two points on a sphere. We use the haversine formula (Sinnott, 1984) to calculate this distance shown in equation (3.1): d is the distance between the two points A and B; r is the radius of the sphere[44]; lat_A, lng_A and lat_B, lng_B are the latitude and longitude of points A and B respectively. Although the earth is not a perfect sphere, this formula gives a good approximation of the distance between two places (using straight lines) for small distances.

$$d_{AB} = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{lat_B - lat_A}{2}\right) + \cos(lat_A)\cos(lat_B)\sin^2\left(\frac{lng_B - lng_A}{2}\right)}\right) \tag{3.1}$$

This quantitative criterion based on geographical data and spatial analysis is important to fix errors introduced by the other criteria, and gives information even when other information are not available in the textual description such as spatial relations or expression of motion and perception.

Effort (C_3)
In addition to considering the geographical distance, we propose to consider the effort needed to go from one location to another one taking the slopes of the route into account. To obtain an approximation of the effort of the displacement made by a pedestrian during hikes and treks that are often occurring in mountains, we propose to take into account the elevation profile of the route. Indeed, hikes and treks occur most of the time in mountain areas where paths have ups-and-downs. Furthermore, the elevation gain is commonly used to describe the difficulty and estimate the duration of treks (Naismith's rule). This information is used to determine the steepness of a trail. It is an important factor to assign a difficulty rate. We compute the cumulative elevation gain and the cumulative elevation loss between two locations. For that purpose we compute the sum of elevation gain (pE) and the sum of elevation loss (nE) according to the terrain elevation profile. To determine the value of the effort criterion (ef), we compute the equation (3.2) widely recognized by experienced ramblers and hikers (equation (3.2)).[45]

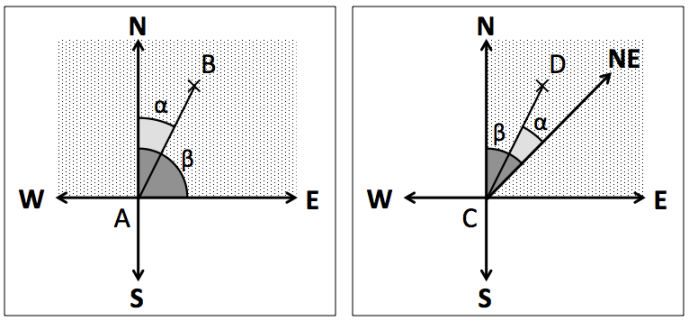
$$ef = (0.01 * pE) + (0.003 * nE) \tag{3.2}$$

Orientation (C_4)
Projective relations attempt to formalize relations expressed in natural language by orientation and cardinal relations (Clementini, 2009) such as: north of, in the direction of, etc. For example, if it is written in the text that after one place we are going north, then we compare this information with geographical coordinates and assign a lower important weight to edges connected to the places that are north than places that are south. We also take into consideration binary relations expressing motion in the direction of a place. In this work we focus on directional and cardinal relations between two place names.

[44] We consider the radius of the earth equal to 6 378 kilometers.
[45] http://www.adorr.fr/francais/elements_techniques.htm

We introduce the criterion (C_4) called *orientation criterion* and used to compare projective relations (north, south, in the direction of) extracted from the text and associated to a place with the locations of the other places. We use a projection-based calculus of directions known as *projection-based method* Frank and Mark (1991) or *cardinal algebra* Ligozat (1998) and defining nine basic cardinal relations (n, ne, e, se, s, sw, w, nw, eq). We calculate the angle α between the alternate locations and the azimuth representing the orientation relation (*north* = 0°, *east* = 90°, etc). Then, to normalize this angle we divided α by β, where β is equal to 90° when the orientation is expressed by a cardinal direction (north, east, south, west, etc.) or 45° when it is expressed by an ordinal direction (northeast, southwest, etc), or a relative direction (in direction of a specific place). Indeed, we assume that the use of cardinal directions in natural language is fuzzier than the use of ordinal directions or azimuths.

For example, Figure 3.5a can be the representation of the phrase 'Leave A and go to the north ...' and Figure 3.5b the representation of the phrase 'From C walk north-east to ...'. In these examples, B and D are two alternatives that can be reached from A and C, respectively. The questions are: 'how much' B is north of A ? And 'how much' D is north-east of C ? The value of the orientation criterion for the edge (A, B) (Fig. 3.5a) is $C_4 = \alpha/90$ and $C_4 = \alpha/45$ for the the edge (C, D) (Fig. 3.5b). C_4 is normalized between 0 and 1, and the smaller the value is, the more consistent the alternate location is with the orientation relation expressed in the text.



(a) Cardinal relation (b) Ordinal relation

Figure 3.5: Illustration of the calculation of the orientation criterion

Elevation (C_5)

We are also dealing with another kind of spatial relation called *elevation relation* (C_5), which can be denoted in the text by verbs (to climb, to come down). The *elevation relation* criterion is used to assign a specific weight to the edges connecting places associated with verbs that convey the sense of change of elevation. We use a trilean value for this criterion. If there are no such verbs expressed in the text, the value is equal to 0.5. When elevation relations are expressed in the text, we compare this information with the elevations of all the other places. If the elevation between two places is consistent according to the elevation relation expressed in the text the value is equal to 0 and 1 otherwise.

Temporality (C_6)

We define a criterion called *temporality criterion* (C_6) based on temporal relations automatically extracted from the text (Muller and Tannier, 2004) such as temporal prepositions (e.g., before, after, then). Elements used to express motion in language are very important for the analysis and the reconstruction of an itinerary. Motion can be denoted by verbs (to go, to leave) and prepositions (from, to). If a temporal relation is expressed between two places, this helps us to determine that these two places are likely to be consecutive. We use this information to set a boolean value: 0 if two places are linked in the text by a temporal relation, and 1 otherwise.

Perception or negation (C_7 and C_8)
We propose to use the information that a place name is associated in the text with a perception or negation expression. The use of perception or negation expression with a place name implies that this place name is not reached during the displacement: it is only seen or used as a landmark to go somewhere else. Perception verbs are frequently used in itinerary descriptions to describe landscapes that we can see far away, such as mountains or lakes. This can be interpreted as a special kind of spatial relations. Information of perception can help to infer locations using the information that during the displacement between two places we are able to see a specific lake or mountain peak. However, in this current work we are not using perception information to infer new locations, but to decide whether a place name is not directly involved in the trajectory because it is not reached during the displacement. The value of the perception (C_7) and negation (C_8) criteria between two places is equal to 1 if at least one of the two places is associated with a perception or negation expression, and 0 otherwise.

3.3.3 Building an Edge-Weighted Complete Graph

We have defined the different criteria that characterize an itinerary. Table 3.1 shows the list of these criteria and their range of values. All these criteria are defined using information extracted from the textual description of the itinerary or they can be computed using geographical data. We use these criteria to decide over a number of alternatives for the successive displacements in order to reconstruct the route. Some criteria are quantitative, such as text distance, geographical distance and effort, and the other are qualitative. Qualitative criteria refer to different types of spatial relations expressed in the language. Following the proposal of Aurnague and Vieu (2015), we consider these criteria as having both functional properties (Landau and Jackendoff, 1993; Talmy, 2000) and geometrical constraints (Vandeloise, 1986).

Criteria	Description	Range of values
C_1	Text distance	$\mathbb{N}_{\geq 0}$
C_2	Geographical distance	$\mathbb{R}_{\geq 0}$
C_3	Effort	$\mathbb{R}_{\geq 0}$
C_4	Orientation	between 0 and 1
C_5	Elevation	0, 0.5 or 1
C_6	Temporality	0 or 1
C_7	Perception	0 or 1
C_8	Negation	0 or 1

Table 3.1: Criteria and their range of values

We propose to use the Weighted Sum Model (WSM), which is a well-known method for multi-criteria decision in decision theory (Triantaphyllou, 2000). It combines criteria into a single criterion by multiplying each criterion with a weight and summing up the weighted criteria. Weighted summation methods are designed to process only quantitative data. However, with the use of a method of standardisation the information become comparable. The WSM method prioritises criteria by assigning weights and reduces the amount of information by summing the weighted standardized criteria.

The data input are a set of criteria $C = \{C_0 \cdots C_n\}$, a list of alternatives $A = \{A_0 \cdots A_m\}$, and a set of weights $W = \{w_0 \cdots w_n\}$, where n represents the number of criteria and m the number of alternatives. In our case, the alternatives represent the location of place names, and a_i represents the cost to go from one place to another (A_i) and a_{ij} the cost of traversing edge a_i according to criterion j. We use the sum of the weighted criteria (equation 3.3) to assign a weight a_i to each edge of the complete graph

representing all the possible connections between places.

$$a_i = \sum_{j=1}^{n} w_j a_{ij} \qquad \forall i \in [1, m] \tag{3.3}$$

The weights of criteria have been assigned according to the Analytic Hierarchy Process (AHP) indirect method (Saaty, 1999) for deriving priorities of criteria. This method is based on the pairwise comparisons of criteria. Firstly, the criteria are compared, two by two, with respect to their importance to reaching the goal of establishing the right sequence of waypoints. This importance is assigned using a fundamental AHP scale (Table 3.2) that ranges from 1 (both criteria have equal importance) to 9 (favoring one criterion over another is of the highest possible order of affirmation).

Intensity of importance[a]	Definition
1	Equal importance
3	Moderately favor one criterion over another
5	Strongly favor one criterion over another
7	Very strongly favor one criterion over another
9	Favoring one criterion over another is of the highest possible order of affirmation

[a] Intermediate values can be used to express intermediate intensities

Table 3.2: The fundamental values can be used to express intermediate intensities

Secondly, the results of these comparisons are entered into a matrix, whose principal right eigenvector will be used to derive the relative strengths of criteria. Taking into account the context of hiking descriptions, Table 3.3 shows the matrix with the pairwise comparisons of criteria that were used to derive the priorities.

Criteria	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	**AHP Priority**
C_1	1	3	4	4	4	2	1/3	1/3	**0.14**
C_2	1/3	1	2	2	2	1/2	1/5	1/5	**0.06**
C_3	1/4	1/2	1	1	1	1/3	1/6	1/6	**0.04**
C_4	1/4	1/2	1	1	1	1/3	1/6	1/6	**0.04**
C_5	1/4	1/2	1	1	1	1/3	1/6	1/6	**0.04**
C_6	1/2	2	3	3	3	1	1/4	1/4	**0.10**
C_7	3	5	6	6	6	4	1	1	**0.29**
C_8	3	5	6	6	6	4	1	1	**0.29**
							Sum of priorities		1
							Inconsistency		0.022

Table 3.3: AHP priorities of criteria, and range of values for measuring each criterion

Table 3.4 shows the pairwise comparisons of criteria and the followed principles (together with a short summary of the motivation the intensity of importance). Anyway, as discussed later in section 3.5 the pairwise comparisons should be adjusted to the different types of texts to be processed, or alternative methods for deriving priorities could be considered.

	1st criterion		2nd criterion		Motivation
	Id - Definition	**Imp**	**Id**	**Imp**	
1	C_1 – **text distance** (important in the specific context of hiking descriptions and route directions)	3.00	C_2	1.00	C_1 is more important than C2 in hiking descriptions.
2		4.00	C_3	1.00	The effort (C_3) is difficult to estimate because several parameters such as distance, elevation terrain profile, way of travel, nature of the ground, etc can influence.
3		4.00	C_4	1.00	Although C_4 and C_5 are important for the reconstruction of itineraries, in some situations, due to the ambiguity of the language, C_4 and C_5 can introduce errors.
4		.00	C_5	1.00	
5		2.00	C_6	1.00	C_6 defines clearly the sequence of waypoints, but is not always available and is hard to extract and interpret from natural language.
6		1.00	C_7	1.00	C_7 and C_8 are strongly more important than the other criteria, but some errors of interpretation may arise. For instance, in the case of the use of anaphora it can happen that a location is first expressed in association with a perception or negation expression, and later it is refered as a waypoint with an anaphora. All the other criteria can contribute to solve an error of interpration.
7		1.00	C_8	1.00	
8	C_2 – **geographical distance** (a key criterion for the reconstruction of itineraries)	2.00	C_3	1.00	See comment on line 2.
9		2.00	C_4	1.00	See comment on line 3-4.
10		2.00	C_5	1.00	
11		1.00	C_6	2.00	When temporal expressions are available, C_6 becomes a more important criterion.
12		1.00	C_7	5.00	See comment on line 6-7
13		1.00	C_8	5.00	

Table 3.4: Comparison of criteria with respect to the goal using the AHP fundamental scale (Id: identifier of criterion; Imp: intensity of importance)

	1st criterion		2nd criterion		Motivation
	Id - Definition	**Imp**	**Id**	**Imp**	
14	C_3 – **effort** (difficulty of the trail, contributes to evaluate the duration of the trip))	1.00	C_4	1.00	The estimation of effort (C_3) is considered in the same level of difficulty as the detection of expressions of orientation and elevation.
15		1.00	C_5	1.00	
16		1.00	C_6	3.00	See comment on line 11.
17		1.00	C_7	6.00	See comment on line 6-7
18		1.00	C_8	6.00	
19	C_4 – **orientation** (orientation expressions are important in the description of displacements, but not always available and difficult to interpret due to the ambiguity of the language)	1.00	C_5	1.00	See comment on line 14-15.
20		1.00	C_6	3.00	See comment on line 11.
21		1.00	C_7	6.00	See comment on line 6-7.
22		1.00	C_8	6.00	
23	C_5 – **elevation** (expressions of change of elevation between two locations such as "climb" are important clues, but sometimes human perception may differ from reality)	1.00	C_6	3.00	See comment on line 11.
24		1.00	C_7	6.00	See comment on line 6-7.
25		1.00	C_8	6.00	
26	C_6 – **temporality** (an important criterion when available as it informs about the sequence of waypoints reached, but hard to extract from natural language and interpret)	1.00	C_6	6.00	See comment on line 6-7.
27		1.00	C_7	6.00	
28	C_7 – **perception** (perception expressions and negation expressions (C_8) associated with place names are very important clues to determine that these locations are not referring to waypoints)	1.00	C_8	1.00	The detection of expressions of perception and the detection of expressions of negation have an equivalent level of difficulty.

Table 3.4: Comparison of criteria with respect to the goal using the AHP fundamental scale (continued)

Additionally, we normalize the criteria C_1 to C_4, whose values are beyond the range $[0-1]$ in order to make the criteria comparable with each other, using the formula in equation (3.4) with $k \in \{1-4\}$ (also known as 'fraction of the sum' normalization (Barba-Romero, 2001)).

$$a_{ik} = \frac{a_{ik}}{\max(a_{ik})}, \qquad \forall i \in [1, m] \tag{3.4}$$

And finally, we sum up the weighted criteria with equation 3.3 and assign the values to each edge of the complete graph.

3.3.4 Minimum Spanning Tree (MST)

We are working with a connected, weighted, complete undirected graph, where all the weights are positive numbers. We use Prim's algorithm (Prim, 1957) to find a minimum spanning tree for a connected, weighted, undirected graph (see Algorithm 1). This algorithm build the tree (T) one vertex at a time. It starts by adding randomly a vertex (v_0) to the set of nodes (Q) and removes it from the input weighted graph (G). Then, at each step it adds the edge with the minimum weight connecting a vertex v already inserted in Q with a new vertex w, not included in Q yet. To solve the problem of duplicate nodes, introduced to avoid cycles when the same location appear several times, we assign an infinite weight to the edges connecting two duplicate nodes. The advantage of this approach is that we do not need a directed graph and we do not need to know which are the starting and ending points. Algorithm 1 shows a simple version of Prim's algorithm to facilitate the understanding of its applicability in this context. However, if efficient structures such as heaps are used for the storage of weighted edges, the overall time complexity of this greedy algorithm is linearithmic $O(m \log n)$, where n is the number of vertices and m the number of edges.

Algorithm 1: Minimum Spanning Tree

Input: undirected connected weighted graph $G = (V, E)$
being V a list of vertices V = $\{v_0 \cdots v_n\}$
Output: tree T representing the set of edges composing an MST of G

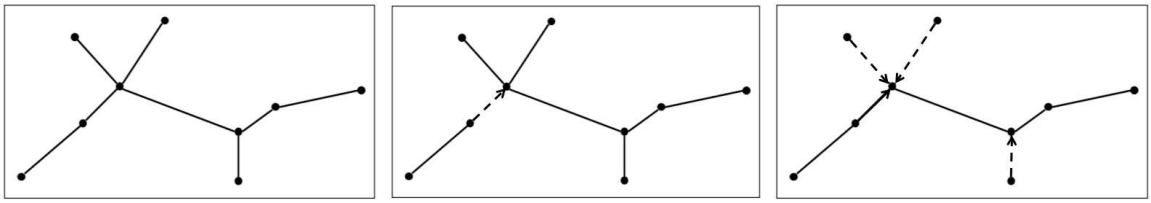
```
1  Q ← empty list;
2  Insert(Q,v0) ;                                   // with v0 chosen randomly
3  Remove(V,v0);
4  while V != empty do
5      minWeight ← ∞;
6      foreach vertex v ∈ Q do
7          foreach vertex w ∈ V do
8              if weight(e_{v,w}) < minWeight then
9                  bestEdge ← e_{v,w} ;             // weight according to equation 3
10                 w' ← w;
11             end if
12         end foreach
13     end foreach
14     Insert(T,bestEdge);
15     Insert(Q,w');
16     Remove(V,w');
17 end while
18 return(T);
```

The result is a connected acyclic undirected graph, also called 'path graph', which means that this tree has no root. Furthermore, any two vertices in this tree are connected by a unique simple path. This tree represents the described itinerary, with vertices representing waypoints of the displacement and also vertices representing locations involved in the description of the itinerary but not reached during the displacement, such as visual cues (e.g., mountain peaks, lakes, etc.).

3.3.5 Building a DAG from the minimum spanning tree

The last step of the process is to build a DAG representation of the described itinerary. We propose to find the longest path on the minimum spanning tree (i.e., maximum number of vertices between two leaves) in order to identify which leaves are the starting and ending points, and remove vertices that are not part of the displacement. This problem is the equivalent of finding the largest sub-graph having a Hamiltonian path or finding the topological order of a DAG.

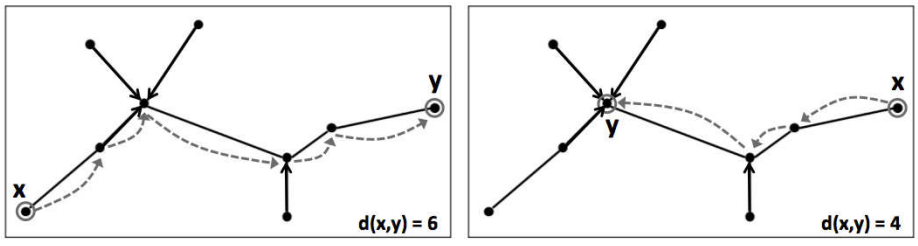
We transform the tree into a Partially Directed Acyclic Graph (PDAG) also called Chain Graph. The class of chain graphs was introduced by Lauritzen and Wermuth (1989) as a generalization of both undirected graphs and acyclic directed graphs and admit both directed and undirected edges. Let $G = (V, E)$ be a chain graph with a finite vertex set V and an edge set $E \subseteq V \times V$. An edge $(v, w) \in E$ is directed if $(w, v) \notin E$ and undirected if $(w, v) \in E$. We denote a directed edge (v, w) by $v \rightarrow w$ and an undirected edge (v, w) by $v - w$. If $(v, w) \in E$, then v and w are adjacent.



(a) (b) (c)

Figure 3.6: From an undirected acyclic graph to a partially directed acyclic graph

To transform undirected edges into directed edges we use spatio-temporal relations that are expressing motion. As we have seen in Section 3.3.2, motion can be denoted by verbs and prepositions and play an important role in the description of itineraries. It gives information concerning the polarity of the displacement (Slobin, 1996; Aurnague, 2011). When such relations are available, we use it to transform undirected edges into directed edges (Fig. 3.6b). We also transform undirected edges connecting locations and visual cues into directed edges (Fig. 3.6c). Considering a location represented by a vertex v and a visual cue represented by a vertex w, we transform the undirected edge $v - w$ into a directed edge $w \rightarrow v$. Then to find the longest path in the chain graph and assign a direction to all edges, we use a Depth First Search (DFS) algorithm (Tarjan, 1972; Karger et al., 1997). We apply a DFS starting from every leaf (vertices having only one connected edge) except those that represent locations associated with perception or negation expressions. We compare the resulted distance of each DFS to identify the longest one.



(a) (b)

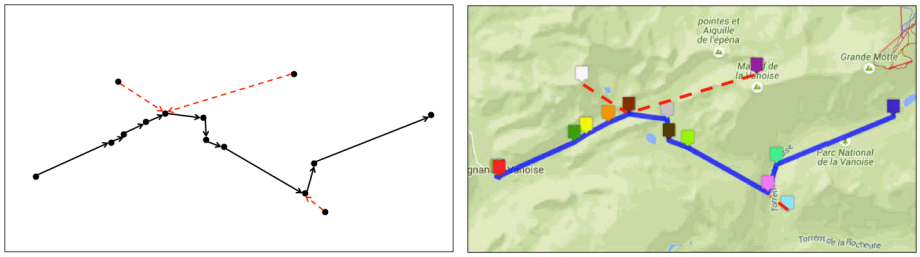
Figure 3.7: Illustration of the method to find the longest path

For example, Figure 3.7 shows the two possible paths, and in this example Figure 3.7a shows the longest path of the chain graph. The leaves x and y (Figure 3.7a) are considered as being the starting and ending points respectively. All other leaves are not considered as waypoints in the DAG representation

of the displacement. When there are no spatial relations available to transform undirected edges into directed edges, we are still able to find the longest path but we cannot distinguish starting and ending points.

3.4 Approximation of the Spatial Footprint of an Itinerary

Previous section (Section 3.3) has proposed a method to order the sequence of waypoints and build a first approximation of the representation of the itinerary using a multi-criteria approach and a graph-based model representation. Vertices of the graph represent waypoints and places involved in the description of the itinerary. Each vertex is associated to the real-world location of the waypoint and each edge between two vertices represents one route between two waypoints. Thus, the first approximation of the representation of the itinerary, uses straight lines to represent routes between waypoints (Fig. 3.8). In Figure 3.8 the DAG representing the itinerary is directly mapped using a geographical coordinate system to get a map-based representation.



(a) DAG representation of the itinerary (b) Spatial representation of the itinerary

Figure 3.8: Approximation of the spatial footprint of the itinerary

Then, a following step would be to propose a better approximation of the representation (not just straight lines) taking into account the availability of route networks in urban area and geographical obstacles (e.g., rivers, mountain peaks, etc.) in rural areas. Now that the sequence of waypoints is known we propose to apply new criteria between each pair of waypoints to give a better representation of the spatial footprint of the real route between these two waypoints. There are two cases where we propose to improve the precision of the representation of the itinerary, the representation of waypoints and the representation of routes.

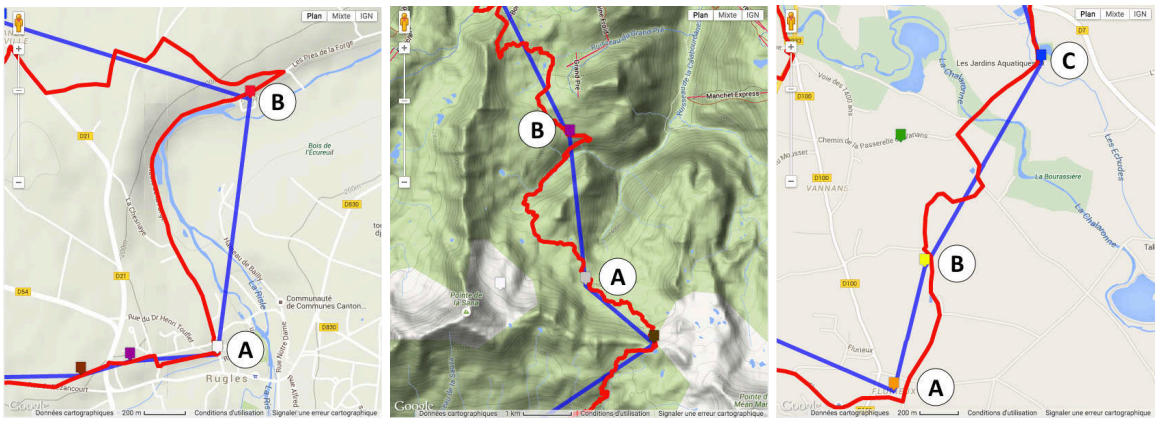
We saw in Section 3.2.2, that for the first case the problem is not how to represent waypoints, but to find their good location. Indeed, a waypoint refers to the actual location where the route is going through during the displacement. For instance, if a waypoint is represented in the text with the name of a lake, obviously the actual location of the waypoint is not the point representing the center of the lake but a point somewhere next to the lake. Some methods have been proposed by Loustau (2008) to infer locations of spatial entities using spatial relations. His solution takes into account the inherent geometry type or shape of the spatial object that refers to the spatial entity, and infers the location of the waypoint using spatial relations (south of, right of, near, at 200 meters, etc.) expressed in the textual description of the displacement.

The second case refers to the representation of the routes and assumes that it is better to circumvent an obstacle rather than to pass over. We propose to add several "intermediate virtual waypoints" to obtain a spatial representation with an expected higher precision. In order to add these "intermediate virtual waypoints", we propose to consider natural obstacles such as river, chasm, mountainous topography, and route networks when available.

Our work could thus be extended by taking into account more information extracted from the text, and more information coming from geographic databases describing feature shapes. For example, if the itinerary description mentions "cross the forest", "follow the river" or "walk one hour", that information

could be used to define some new criteria, if they are crossed with databases describing forests and rivers or digital elevation models, and with spatial analysis defining the notion of "cross", "follow" or approximate distance from walking durations.

Figure 3.9 shows three examples of cases where the precision of the automatically built approximated route are not very accurate and could be improved. Whereas the first (Fig. 3.9a) and second (Fig. 3.9b) examples refer to natural obstacles, the third example (Fig. 3.9c) refers to a route describing a displacement in urban area. According to these three examples, it may be noticed that even if the waypoints are well located, the approximation of the route may be far away from the real route representing the itinerary. Figure 3.9a shows an example of route between two waypoints (A and B) that follows a riverbank. The red line represents the real route described in the text and the blue line represents the automatic reconstruction build with the method proposed in Section 3.3. We can notice that the real route is following the river. However, the automatically generated route is going straight without paying attention to the geographical context of the area. Figure 3.9b shows a route between two waypoints (A and B) in a mountainous terrain. In this case, whereas the real route (red line) bypasses the mountain, the automatically generated route (blue line) passes through. Figure 3.9c illustrates a route between three waypoints (A, B and C) that occurs in an urban area. The real route of the displacement (red line) is following roads and streets. In this case, road networks have to be taken into account to improve the representation of the spatial footprint of the itinerary. In the specific case of urban area, we propose to use route planning methods such as those proposed for urban transportation (Delling et al., 2009) and dealing with the Travelling salesman problem. Experiments (see Chapter 7) show that descriptions having a lot of waypoints obtain a higher accuracy. The approximation made by the automatically generated route is better when there are a lot of waypoints and the locations of these waypoints are relatively near from each other. For that reason, our proposal is to add *intermediate virtual waypoints* along the route to create short steps of displacement and obtain approximations with higher accuracy.



(a) (b) (c)

Figure 3.9: Examples of proposed improvements

In order to consider natural obstacles, we propose to add two criteria. The first criterion aims at minimising the effort needed to reach the next waypoint. So when the effort needed to cross an obstacle is too high, we try to find the shortest way to avoid the obstacle and reach the destination point. We could use the terrain profile model available between the two waypoints to establish the new trajectory to reach the destination. The second criterion is based on the Visibility Calculus introduced by Tarquini et al. (2007) and De Felice et al. (2011). They define relations such as visible, occluded and partially visible, describing if an object is visible from an observer point taking into account an obstacle. Regarding the visibility problem, Fogliaroni and Clementini (2015) propose a qualitative 3D frame of reference used to model human visual perception. The proposed frame of reference aims to model projective and visibility information as ternary relations. This approach can be applied to a landmark–based route instruction system. The criterion based on visibility relations is used to infer new intermediate waypoints in order

to obtain a more accurate trajectory of the route and to avoid too large distances between waypoints.

3.5 Summary

In this chapter we proposed a formal model for representing an itinerary as described in a text. A DAG is used foreseen, where the vertices of the graph represent locations and the edges represent segment between two locations. These elements are the core of most models representing routes or itineraries with a graph approach (Werner et al., 2000; Spaccapietra et al., 2008; Vasardani et al., 2013). The model is original in that in addition to taking into account the classic elements (routes and waypoints), it emphasizes other elements describing an itinerary: features seen or mentioned as landmarks. To go further, this model could be enriched with other elements for a more precise description. In particular, texts may describe the itinerary at several levels of granularity (some times a global description of the whole itinerary, and some times a precise description of particular pieces of the itinerary) and modelling multiple granularities would then be useful, as proposed by Hornsby and Egenhofer (2002). Other key elements could be to rely on linear referencing principles to model events appearing within a particular route (Güting et al., 2006), or allowing the modelling of fuzzy information like in sentences (35) and (36). Additional notions like the one of "entry, course and exit" emphasized by the Route Graph model would also be useful to represent orientation elements on how the itinerary enters or exits waypoints, and how it is related to other elements along routes (Werner et al., 2000; Krieg-Brückner and Shi, 2006).

(35) Further pass on Chanton bridge and cross the forest. Soon after, you will reach lake Des Vaches.

(36) Then follow Leisse torrent to achieve the day's stage.

(37) In Pralognan, follow the road between Hotel de la Vanoise and Hotel du Petit Mont Blanc and go straight.

(38) At the crossroads, do not take south towards the refuge of Entre-Deux-Eaux, but go north and walk one hour.

Moreover, a comprehensive approach is proposed for automatically identifying the sequence of waypoints from a geoparsed text and building an approximation of a plausible sequence of the described itinerary. The feasibility of this approach has been tested for the automatic approximate geocoding of itineraries described in a corpus of hiking descriptions (see Chapter 7). This feasibility study also allowed us to illustrate that combining quantitative and qualitative criteria, based on knowledge extracted from the text and knowledge extracted from geographic databases, improves the approximation of a described itinerary. However, there are still some limits and some possible improvements would be to improve the geoparsing by adding a deeper linguistic processing and a deeper spatial analysis to take into consideration new categories of spatial relations and to annotate unnamed locations. Moreover, it would be also possible to consider a method to integrate more information coming from geographic databases describing feature shapes. For example, if the itinerary description mentions "follow the road/river" (sentences (37) and (36)), "cross the forest" (sentence (35)), or "walk one hour" (sentence (38)), that information could be used to define some new criteria, if they are crossed with databases describing forests and rivers or digital elevation models, and if some spatial analysis tools are defined to approximate the notion of "cross", "follow" or to approximate distance from walking durations. Some other information in the text describe relations between parts of the itinerary, like "go straight" (sentences (37)). In order to handle that, we cannot directly define new criteria to weight each edge of the graph, but we should extend the notion of minimal spanning tree and constrain the search so that those relations between parts are fulfilled.

Another limit of our approach concerns the setting of the multi-criteria approach. One key issue in our multi-criteria approach is how to define weights and the combination strategy. For such a problem of setting the weights of a multi-criteria combination, or for setting the suited model of combinations, machine learning is a widely used approach that we could follow. In particular, machine learning is used in the natural language processing domain for approaches to entity tagging that are based on probabilistic models such HMM (Rabiner, 1989) or, approaches to extracting semantic relationships between entities (Béchet et al., 2014). However, whatever machine learning technique is used, a key issue is to get a sufficient number of examples and to precisely define the learning task (Mitchell, 1997). Those examples

cannot be direct examples for our task, as we have seen that the text alone is never sufficient to reconstruct the actual precise itinerary. However, we may expect that a huge number of examples could overcome some of those difficulties, if one tries to learn weights that minimise the proposed evaluation distance.

Our approach aims to reconstruct the sequence of displacement taking account the geographical area of achievement. Another possible improvement is to approximate the actual footprint of the displacement. To do that, we may extrapolate from a very small amount of information present in the text. Some external knowledge has to then be introduced, like displacement habits: for example, hikers do not cross rivers and may minimise they effort. Other information could come from other itinerary descriptions, in any format (text or geolocalised paths). This would require to introduce some mechanisms for merging itinerary descriptions, as proposed by (Belouaer et al., 2013).

The next chapter describes an approach for the automatic annotation of information in texts that can be used as input of the current proposal for the automatic itinerary reconstruction.

Chapter 4

Text Mining and Toponym Resolution

Words are the source of misunderstandings
— Antoine de Saint-Exupéry, *Le Petit Prince*

Contents

4.1 Introduction

If the global understanding of a text is still considered an unattainable task to the current capabilities of computer systems, partial understanding with a predefined view has recently become a feasible task. Generally this task is called text mining (Kao and Poteet, 2006). The objective is not to do extensive analysis of the textual contents of documents but tracking through indices for certain informational patterns. The interpretive process is not only led by the text, but it is also guided by a priori knowledge of the sought information. Present states of the art shows that linguistics analysis can be very accurate but still remains local. This allows both to master the complexity, and facilitates the portability of these systems on different natural languages. A first level of analysis, always triggered by the presence of specific keywords, builds a first interpretative structure. The first structure is integrated in a process to build on it more complex and richer structures, but the principle remains local. This focus and specialization allow to build high-performance extraction systems.

In this chapter we address the problem of automatically annotating passages in the text that describe the various trips making up the itinerary. This problem involves the annotation, the resolution and the disambiguation of toponyms but also the annotation of spatial or motion semantic relatedness between phrases and named entities.

The objective is to propose a method for establishing a processing chain to support the geoparsing and geocoding of text documents describing events strongly linked with space and with a frequent use

of fine-grain toponyms. The geoparsing part of the workflow is a Natural Language Processing (NLP) approach which combines the use of a part of speech tagger and a cascade of finite-state transducers.

The main contribution of this chapter is the geoparsing and geocoding part of the method. The geoparsing process relies on the automatic annotation of spatial information based on the definition of expanded named entities and spatial role labeling. The geocoding algorithm is an unsupervised algorithm that takes profit of clustering techniques to provide a solution for disambiguating those toponyms found in external geographical resources, and at the same time estimating the location of those other fine-grain toponyms not found.

The contribution described in this chapter (Figure 4.1) provides the input of the method of reconstruction of itinerary described in Chapter 3.

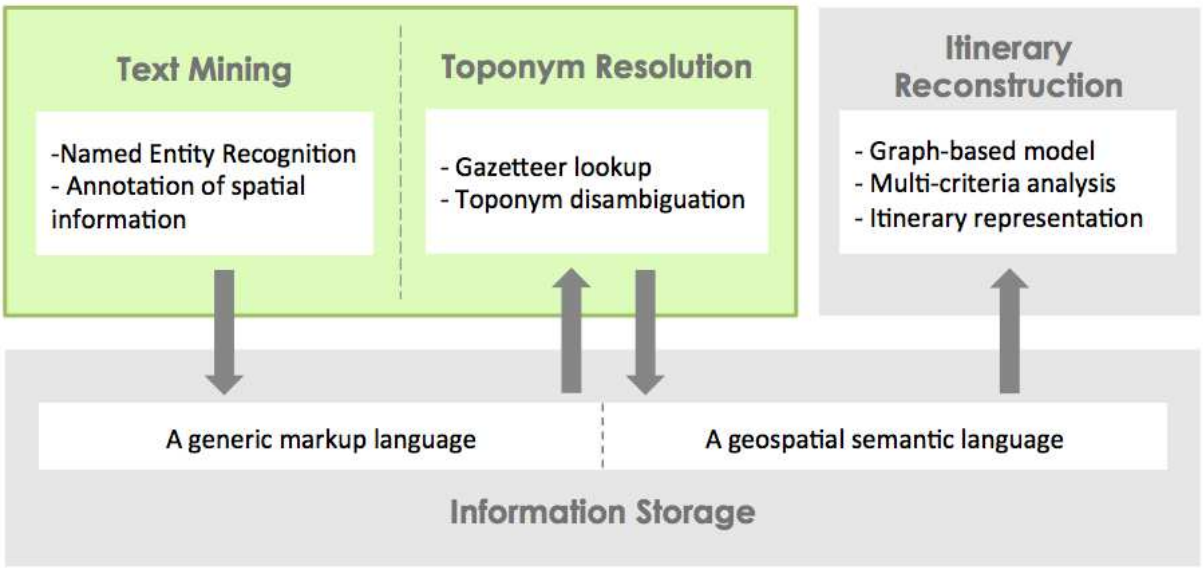


Figure 4.1: Contribution of this chapter (highlighted in green)

The remainder of this chapter is structured as follows. Section 4.2 describes our proposed method for annotating expanded named entities and spatial information such as motion expressions and spatial relations. Section 4.3 presents a method for the classification of named entities based on toponym resolution and describes our proposal for the problem of toponym disambiguation. Finally, Section 4.4 summarises and concludes this chapter.

4.2 Named Entity Recognition and Spatial Role Labeling

4.2.1 Overview

In our work, we are not only dealing with spatial named entities and toponym recognition but we are also trying to identify and classify spatial expressions mentioned in texts, as in Spatial Role Labeling (SpRL). Kordjamshidi et al. (2012) define the SpRL as the task of identifying and classifying the main components of the spatial semantics from natural language such as trajectories, landmarks and spatial indicators. Our aim is to identify geographical information in a text, as well as any textual clues that allow us to link some spatial named entities and exclude other information that should not be taken into consideration in the itinerary description.

As mentioned in the literature review (Section 2.2), spatial relations are an important factor in language to express space and motion. They are also important for the disambiguation of toponyms. Particularly, they are relevant in forms of discourse that can be found in a corpus of description of itineraries such as hiking or travelogues, and those of same categories where spatial relations exist on

several levels of granularity and can be applied to the discourse at different scales. In this chapter, we will examine two levels of granularity. The first one involves local spatial relations that are part of a spatial named entity. And the second level involves spatial relations that make connection between different spatial named entities. To illustrate this idea, let us take example (39). In the spatial named entity 'north of hamlet Friburge' the spatial relationship contained ('north of') needs to be interpreted in order to find the actual location of the spatial entity considered. In this case, the referent point that we are trying to localise is not 'hamlet Friburge' but 'north of hamlet Friburge', so we need to interpret this spatial relation in order to find the spatial footprint corresponding to the full expression.

(39) Traverser Champagny-le-Haut et contourner par le nord du hameau de Friburge.
Cross Champagny-le-Haut and get around through the north of hamlet Friburge.

(40) Vous apercevrez le Lac de la Plagne, puis marcher vers le sud en direction du refuge du lac Grattaleu.
You will see Lac de la Plagne, then walk south to the refuge of Lake Grattaleu.

In the case of hiking trails, the object under consideration is the participant in the motion event. In sentence (39) the spatial relations are 'the north' and 'around', but also the motion verbs 'to cross' and 'to get'. This is a description of a motion event relative to spatial named entities. Then in sentence (40), the spatial relation 'south to' is associated with the motion verb 'walk' and refers to the direction of the displacement whereas the spatial relation 'the north' of sentence (39) is associated with as spatial entity and refers to the location of the waypoint. The toponym 'Lac de la Plagne' is associated with a perception verb 'to see', which means that the toponym is not really part of the itinerary taken but a visual landmark. Moreover, the term 'lake' serves to precisely identify the toponym, removing all ambiguity from the geographic object being referred to, i.e. it is most definitely a lake, and not a town for example.

The first step in our approach is a system whereby spatial information described in textual documents is automatically annotated. The method combines the notions of marking and extracting of named entities and spatial information, through the use of local grammars (Gross, 1997) and external resources (lexicons, gazetteers, etc.). These grammars are lexicalised graphs that make use of dictionaries and have the advantage of being able to be applied directly to texts (Constant, 2003) for a syntagmatic analysis.

The core of our method of annotation upgrades on the previous work done by Loustau (2008) and Nguyen (2012). Loustau (2008) proposed a task driven conceptual model for the conception of pedagogical activities involving interpretation of itineraries and Nguyen (2012) proposed a method for geographical ontology enrichment through the automatic extraction of topographical terms. The experimental approach proposed by Loustau (2008) introduced the prototype πR for the interpretation of itineraries in travelogues. Nguyen (2012) proposed a formalisation of relationships between verbs, terms and spatial entities called *VT structures*. They introduced the key concepts of a formal model and linguistic rules for the annotation of spatial information in text.

We propose to enrich and adapt this approach of annotation for the extraction of spatial information in order to reconstruct itineraries from verbal descriptions. Our proposed model includes a better interpretation of motion verbs (semantic and polarity), spatio-temporal relations and negation expressions. Furthermore, in addition to taking into account the classic elements (path and waypoints) our model allows to represent the other elements describing an itinerary such as feature seen or mentioned as landmarks. Finally, according to the annotation part of our method, we build our model and our system to be compatible with Romance languages and not only with French.

4.2.2 Finite-State Transducers Cascade

Our approach is based on a domain-specific corpus analysis. In order to solve the problem of itinerary reconstruction from text, we first defined and analysed all the components of an itinerary (see Section 3.2). Thanks to the analysis of the representation of these components in the language, we obtain a manageable number of rules for our annotation model. Moreover, according to the works done by Loustau (2008) and Nguyen (2012), we propose a finite-state transducer cascade that annotates spatial information.

For flexibility and portability of the system in different languages, we propose to use the same patterns for different languages. For that purpose, we build a transducer cascade with reduced lexicons and manageable numbers of rules easily adaptable for foreign languages. Only, lexicons and few rules have to be translated in order to adapt our transducers cascade to each language. Experiments described in Chapter 7 using a corpus of documents in French, Spanish and Italian show that our proposed system obtains comparable results for the three languages.

Furthermore, another advantage of this knowledge-based method is that we do not need to create a huge learning corpus to train a model. As mentioned in the state of art (see Section 2.3.2), transducers take decisions locally exploiting morpho-syntactic patterns and lexicons (Abney, 1996). Finite-state transducers produce a compact and accurate description of local syntactic constraints that can be viewed as local rules or local grammars describing linguistic clues (Gross, 1997). Transducers annotate phrases when patterns are recognised in the text. The output alphabets defined the annotations made by the transducers. Finite-state transducers are the most effective and readable way to encode morpho-syntactic rules and patterns. The 'knowledge' defined by rules is described in a readable way and grammars are easy to modify and maintain. Additionally grammars and lexicons are easy to translate and adapt for foreign languages. Furthermore, different patterns can be processed at the same time using a cascade of transducers applied one after the other.

Although our approach is based on a cascade of transducers, we propose a hybrid solution. Unlike the CasEN system proposed by Maurel et al. (2011) which uses dictionaries to assign lemmas and categories of words and combines the use of dictionaries and rules for the classification of named entities, we propose an approach based on a Part-of-speech (POS) analysis and we combine the results of our cascade with external resources for the named entity classification.

The matching of words with lexical entries from dictionaries without taking accounts the context increase the number of ambiguity because the same word may have different meaning. For instance in French the word 'est' may have several lexical entries in a dictionary, it may be a verb, a noun or an adjective. Furthermore, with the POS analysis as a pre-processing we eliminate the problem of the availability of dictionaries for foreign languages. The POS tagger assigns lemma and grammatical categories to each word. This pre-processing step is language dependent and is encapsulated in order to provide the same tags (i.e., identifiers of word categories) for each language (see Section 6.2.1 describing the implementation). This allows our transducers to use a 'pivot' format to describe POS tags used by rules and patterns. Table 4.1 shows the result of the POS tagging on the French sentence (41).

(41) Contourner par le nord du hameau de Friburge.
Get around through the north of hamlet Friburge.

word	**POS**	**lemma**
Contourner	V	contourner
par	PREP	par
le	DET+ART	le
nord	N	nord
du	PREPDET	du
hameau	N	hameau
de	PREP	de
Friburge	NPr	Friburge
.	PUN	.

Table 4.1: Output of POS tagging

The tag PREP refers to prepositions, DET+ART refers to determiners (articles), N refers to common nouns, PREPDET refers to prepositions plus articles, NPr refers to proper nouns and PUN refers to punctuations. Further details about the implementation are given in Section 6.2.

The POS pre-processed text is given as input of our cascade of transducers. The use of an already tagged text allows us not to use additional resources such as dictionaries usually needed to identify the category of words. For instance, the CasEN system requires several dictionaries such as *dela-fr* to identify categories of words and *Prolex* used for the NE classification. The DELA dictionaries describe simple and compound lexical entries with their grammatical, semantic and inflectional information[46], and Prolex is a dictionary of proper names proposed by Tran and Maurel (2006). Thus, the CasEN system combines the use of rules and dictionaries (e.g., one for toponyms, another one for firstnames, etc) for NE classification.

We propose to reduce the need of integrated resources in order to propose transducers the most generic as possible, easily adaptable for different languages without having to produce and maintain huge dictionaries. Our goal is to combine the use of transducers with up-to-date external resources (such as linked data available online) for the NE classification. External resources can be chosen depending on the language and also depending on the specific needs or context of the annotation. For instance, for a French text describing a foreign country a static dictionary of French toponyms is needless whereas a resource describing toponyms corresponding to this country would be very useful.

Furthermore, our objective is to obtain as output of the cascade a generic annotation that can be used to any kind of further treatment. We apply this method in a specific context of hiking reconstruction and we mainly focus our work on spatial information. However, we annotate named entities and spatio-temporal information in a generic way. After the automatic annotation process by the cascade we make the toponyms classification and resolution. This post-processing may disambiguate or resolve errors introduced during the annotation step.

Hereafter we describe the different elements characterising space and motion in text annotated by our system.

4.2.3 Space and Motion in Text

According to the literature (see Section 2.2), a description of an itinerary provides route instructions and gives information about direction and distance of travel and about the environment and the locations of places and landmarks related to the itinerary. We rely on cognitive and linguistic works such as (Talmy, 1985), (Vandeloise, 1986), (Boons, 1987) and (Landau and Jackendoff, 1993) to identify the different concepts that are used in the language to express space and motion. We focus our work on four main concepts used to express space and motion in text: spatio-temporal relations, named entities, expressions of motion and expressions of perception.

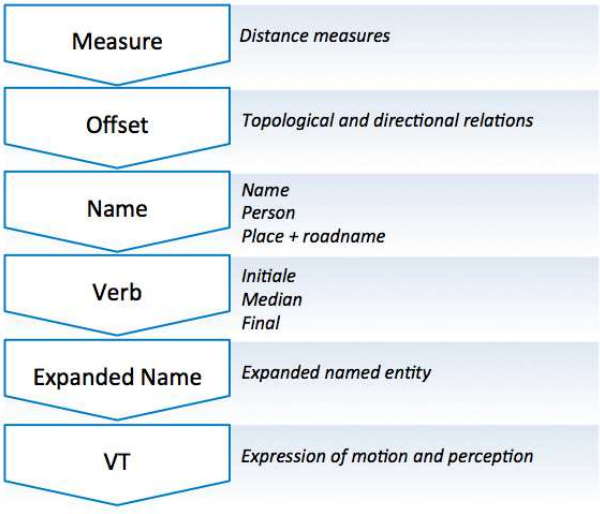


Figure 4.2: Main transducers of our cascade

[46] http://infolingu.univ-mlv.fr/

We propose a cascade of transducers composed of six main transducers (Fig. 4.2). The two first transducers deal with spatio-temporal relation expressions. The first one annotates distance measure expressions and the second one annotates temporal and spatial (topological and directional) relations. Then the other transducers annotate named entities, expanded named entities, verbs and expressions of motion and perception.

However, our transducers make only a first step of classification and do not aim to disambiguate named entities. Most of the time, there are not enough evidences to classify named entities. For instance a Named Entity (NE) like 'St Paul' (or 'Saint-Paul') may belong to different types of NE. 'St Paul' may refer to St Paul's Cathedral in London. However, St Paul is also the name of several cities or village in France (Saint-Paul), is also the name of a Canadian island and is obviously the name of the apostle. The problems of NE classification and ambiguity are addressed in the Section 4.3.

4.2.3.1 Spatial Relations

In order to establish the steps of an emerging route, itinerary is defined as being a special type of spatial relation. It is a spatio-temporal sequence of steps and movements between different places. An itinerary could thus be thought of as a succession of spatial relations. We follow the proposal of Gaio et al. (2012) for the recognition of spatial references and spatial relations in language, as well as using a hybrid approach (Gaio et al., 2008) combining the three main categories of spatial relations: topological relations (Egenhofer and Franzosa, 1991), distance relations, and directional relations (Frank, 1991).

We integrate spatial and temporal relations into a more generic concept called *offset* allowing a geographic object to be addressed indirectly. Offsets can be a part of the concrete entity, their role being to specify location, and grammatically they can belong to different word classes (prepositions, adverbs or adjectives). Parts of speech such as spatial adverbs of location (Borillo, 2004) are annotated to reveal spatio-temporal sequences in the discourse. These are prepositions of place, which occur frequently in hiking guides (e.g., 'here', 'there', 'near of', 'from the left', etc.). These prepositions structure the discourse by describing a spatial sequence (a step in a journey) and/or a temporal sequence (a succession of events). We also propose to annotate some temporal relations that are used very often such as 'after', 'then' (Muller and Tannier, 2004). More specifically, the temporal information that we propose to annotate refer to relationships between two motion events (see phrases 42 and 43).

(42) Plus loin traverser le pont de Chanton. **Peu après** vous attengnez le lac des Vaches.
Further pass on Chanton bridge. ***Soon after****, you will reach lake Des Vaches.*

(43) Dirigez-vous vers le nord **puis** suivre le torrent de la Leisse.
Go north, ***then*** *follow Leisse torrent.*

Our method annotates distance measures. This type of spatial relations is considered as named entity and belongs to the NUMEX category defined by the named entity task of MUC-6 (Grishman and Sundheim, 1996). The 'measure' transducer recognises amounts related to space with patterns such as *100 meters*, *four kilometers* etc. These patterns identify phrases containing a value (numerical value or words) and a unit (e.g., meters, miles, km, etc.).

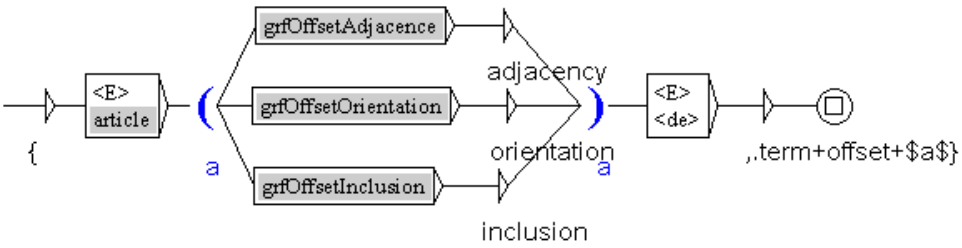
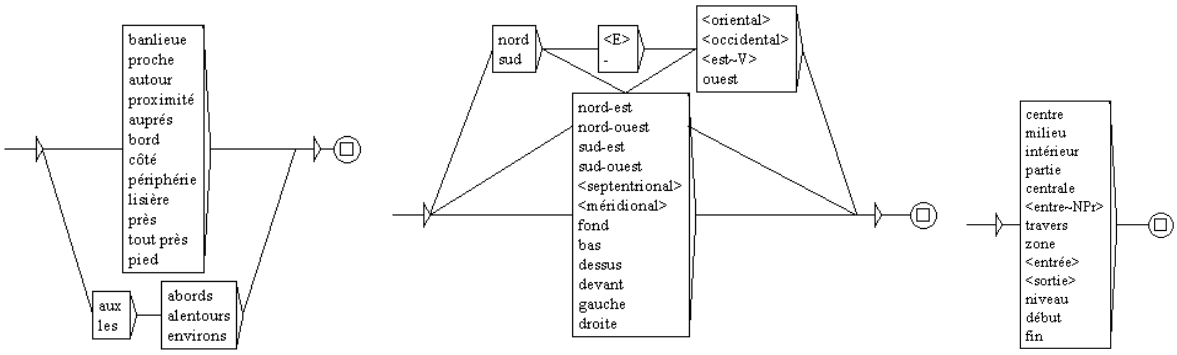


Figure 4.3: Transducer annotating French topological and directional spatial relations

The second main transducer of our cascade (shown on Figure 4.3) annotates topological and directional spatial relations. The greyed boxes refer to sub-graphs describing patterns and lexicons of topological and directional expressions. We distinguish two types of topological relations, adjacency and inclusion. Figure 4.4 shows the three sub-graphs used to annotate spatial relations. These sub-graphs consist of lexicons describing adjacency (Figure 4.4a), orientation (Figure 4.4b) and inclusion (Figure 4.4c) relations. Patterns describing expressions with words such as 'near', 'around' and 'suburb' refer to adjacency relations (sentence (44)). And patterns defined by expressions such as 'part of', 'inside' or 'into' refer to inclusion relations (sentence (45)). Finally, directional relation patterns are described using cardinal or local relations with words such as 'north', 'south-east' and 'left' in association with prepositions like 'of' (sentence (46)).



(a) Adjacency (b) Orientation (c) Inclusion

Figure 4.4: Sub-graphs identifying French spatial relations

(44) Passer **à proximité de** la chapelle Saint-Aubin.
*Passing **near** the chapel Saint-Aubin.*

(45) Se rendre **dans** la commune d'Attignat Oncin.
*Enter **into** the city of Attignat Oncin.*

(46) On prend sur **la gauche** en direction du **Sud**.
*We take on **the left** towards **the South**.*

According to the output alphabet, the transducer produces the following annotation on sentence (44):

Passer **{**à proximité de**,.term+offset+adjacence}** de la chapelle Saint-Aubin.

Spatial relations expressed by prepositions like 'from', 'to', 'towards' and expressions like 'in direction of' are not considered as offset. They are used in the last main transducer of our cascade to link expressions of motion with named entities and are very useful to determine the polarity of motion relations.

4.2.3.2 Named Entities

Although we are mainly interested in spatial NEs, we must identify other types of entities in order to discard them. For that purpose, transducers of our cascade are implementing a first step of classification using internal and external evidences (McDonald, 1996) to identify, annotate and classify NEs. Evidences may refer to definitive or heuristic criteria such as known terms, abbreviations (e.g., *Mr.* for *Mister*) or known names found in lexicons. For instance, the NE 'Mr. Martin' is classified as a person, whereas 'Martin St.' is classified as a place with a sub-category 'road name'. Figure 4.5 shows the transducer to annotate named entities.

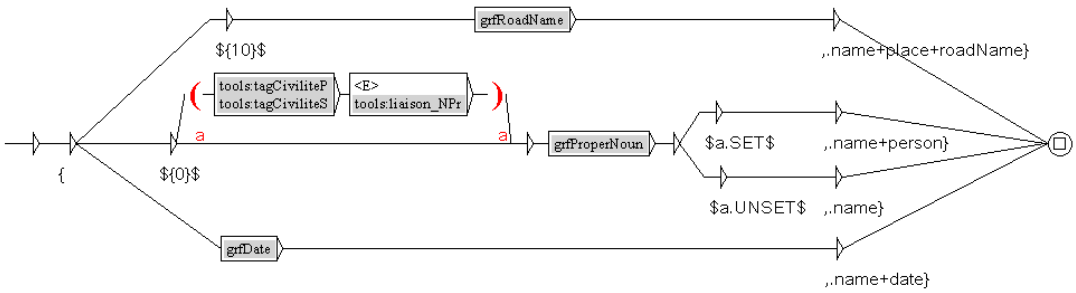


Figure 4.5: Transducer annotating French named entities

Road Names

Road or street names, also called odonyms, obey to a numbering scheme to classify and identify roads. Two main categories of road may be distinguished, motorways and non-motorways. Usually road names start with a letter or an abbreviation which represent the road category followed by a number. For instance, in England road names start with a single letter (A, B, C, etc.) followed by a number of one to four digits (e.g., A505, B125). However, in France, road names begin with an acronym of one or two letters followed by a numerical value separated with a space (e.g., RN 12, A 64, D 117). Italy road names consist of an acronym followed by a number (e.g., SS 5, SR 439, SP217) and may end with one or more letters (e.g., SP5ter, SP23a). In Spain, road names start with one or two letters followed by a number separated with a hyphen (e.g., N-152, A-1). Spanish highways are classified into several categories: 'A' (Autovía), 'AP' (Autopista) and 'R' (Autopista radial).

A lot of different types of road exist, for instance in France, 'A' is the abbreviation for 'autoroute' which is the equivalent of freeways or motorways such as toll roads. 'N' or 'RN' is the acronym for 'Route Nationale' (i.e. national road) which is the equivalent of the British primary routes. 'D' or 'RD' stands for 'Route Départementale' which refer to local roads or B-road in the United Kingdom. Concerning non-motorways, in France 'GR' is the acronym for 'Grande Randonnée'. GR hiking paths refer to long-distance footpaths and hiking trails.

Furthermore, there is also a numbering system developed in Europe called the *international E-road network* referring to roads crossing national borders (e.g., E 80).

The transducer recognising French road names implements the following regular expression:

```
^(GR|RD|D|N|A|E|RN)([0-9])+
```

However, according to the ambiguous nature of language the transducer is more complex. For instance, we must anticipate whether people use a different separator than the official one such as a space or a hyphen.

Another transducer identifies address patterns (sentence (47)) containing words such as 'road', 'street' and 'lane', associated with a named entity. In this case, we notice the importance of the order of the sequence of transducers in the cascade. Indeed, several road names are built with the name of a famous person such as 'avenue du Général de Gaulle' in France or 'avenida de Juan Carlos I' in Spain. Person named entities have to be recognised before road names.

(47) Rejoindre le centre de Malaucène par la **rue Guimety**.
Reach Malaucène's city center by **Guimety Street**.

(48) Quitter Sourdon par la **D 14**.
*Leave Sourdon by **D 14**.*

According to the output alphabet, the transducer recognising road names produces the following annotation on sentence (48):

Quitter Sourdon par **{**D 14**,.name+place+roadName}**.

Person
According to Friburger and Maurel (2004) a lot of person names (about 45%) are preceded by a context containing a title or an occupation name such as Mr., Dr., or President. And moreover, more than 90% of person names are associated to a left context in journalistic texts. Person names may be composed of a first name followed by a patronymic name or patronymic name alone.

As we are focusing on spatial named entities, we are only doing a first step of classification using simple rules (easy to adapt for foreign languages) taking account of some internal or external evidences, when the decision can be taken without ambiguity. Otherwise we annotate named entities without classification. And we propose to make the main step of classification with the toponym resolution after the execution of our cascade of transducers using up-to-date external resources.

Date
As mentioned in previous chapters, time is a component involved in the description of an itinerary. Although in this thesis we are focusing on the spatial component, we propose to annotate simple time information such as date or duration. These elements belong to the TIMEX category defined by the named entity task of MUC-6. Our transducers are built with lexicons of days and months and regular expressions describing dates or durations. The annotation of temporal information plays an important role in the reconstruction of the itinerary and more specifically to find the good sequence of waypoints. As mentioned in Chapter 3, taking into consideration temporal relationships between steps of the displacement is very important in order to automatically reconstruct the route.

4.2.3.3 Expanded Named Entities

We define an Expanded Named Entity (ENE) as an entity built from a proper name and Expanded Spatial Named Entity (ESNE) as specific Expanded Named Entity (ENE) with spatial denotation. We also define the notion of relative Expanded Spatial Named Entity (ESNE) when the ESNE is associated with one or more concepts relating to the expression of location in the language (spatial relations). We will give further details about the definition and the representation of the concept of ENE in Section 5.2.3. This definition is based on the work done by Loustau (2008), who defined the concepts of relative and absolute spatial named entities, and Nguyen (2012), who proposed the topographic subtyping of place named entities, i.e., concrete entities may be denoted by topographical terms and spatial relations. For example (49) and (50) are two relative ESNEs built from proper names (*Friburge* and *Grattaleu*), associated with concepts having a geographical sense (*hamlet*, *refuge* and *lake*), and spatial relations (*the north of* and *south of*).

(49) le nord du hameau de Friburge
the north of hamlet Friburge

(50) le refuge au sud du Lac Grattaleu
the refuge south of Lake Grattaleu

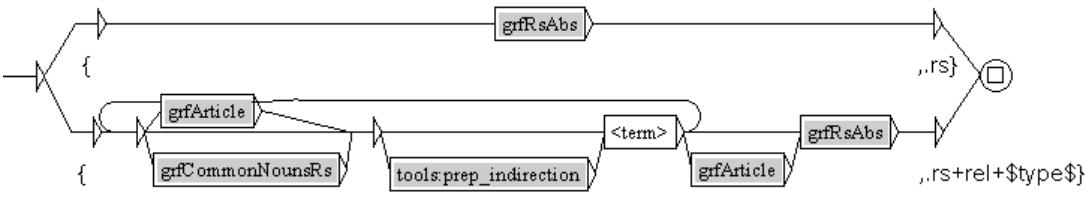


Figure 4.6: Transducer annotating French absolute and relative ENEs

Figure 4.6 shows the transducer that annotates both absolute and relative ENEs and Figure 4.7 shows the transducers to annotate absolute ENEs. The transducer that annotates absolute ENEs (grfRsAbs) is used as a sub-graph in both transducers. The element `<term>` shown in Figure 4.6 refers to the annotations

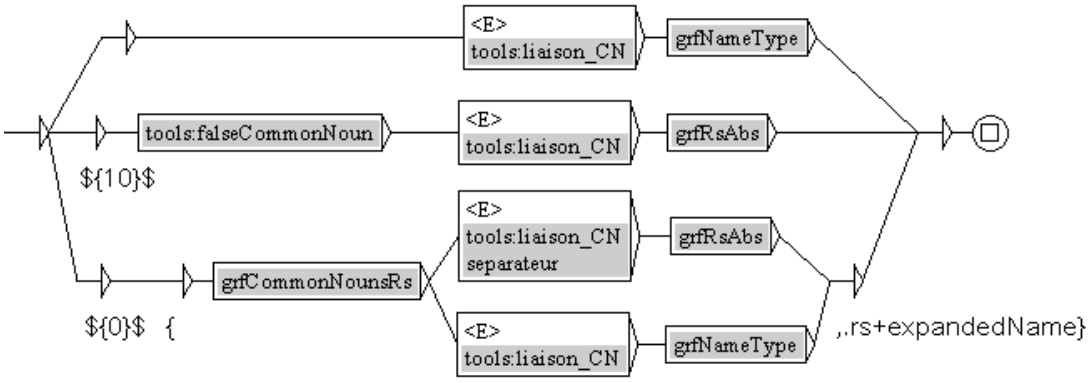


Figure 4.7: Transducer annotating French absolute ENEs (grfRsAbs)

of spatial relations (see Section 4.2.3.1). The sub-graph *grfCommonNounsRs* annotates common nouns (i.e., concepts) contained within the ENE, with the output alphabet `term+N`.

Concepts contained within ENEs play an important role to classify and disambiguate named entities. Common nouns refer to the type of the named entity. In the case of relative spatial named entities, type or spatial relations give important information concerning the location of the object according to the point of reference. Indeed, spatial relations associated with the place name give information about the actual location described. For instance, the ESNE (49) refers only to one part (the north) of the spatial object 'hamlet Friburge' and not to the whole object. Thus, according to the location of the spatial object described by the spatial named entity 'hamlet Friburge' we can infer and found the location of the whole spatial object described. Furthermore, according to the output alphabet, these transducers produce for the ENE (49) the annotation shown in Figure 4.8.

{
  {le nord du**,.term+offset+orientation}**
  {
    {hameau**,.term+N}**
    de **{**Friburge**,.name}**
  ,.rs+expandedName}
,.rs+rel}

Figure 4.8: Result of the annotation of the ENE (49)

4.2.3.4 Expressions of Motion and Perception

Based on the polarity of prepositions and classification of verbs (Laur, 1993) (see related work on Section 2.2.4), we are able to establish simple linguistic rules in order to extract the source or target named entities in a motion event. We classified verbs into categories: motion verbs (e.g., *to go*, *to leave*, etc.), location verbs (e.g., *to locate*, *to be*, etc.) (Borillo, 1990), verbs of visual perception (e.g., *to glimpse*, *to see*, etc.), and verbs we refer to as topographic (e.g., *to converge*, *to overhang*, etc.). The classification of verbs may depend on the context and on the type of object associated with these verbs. This is particularly the case, for location verbs such as *to be* and verbs we refer as topographic such as *to converge*. According to the local context these verbs can have a different meaning (e.g., 'I am at the office'. vs 'I am late' and 'the rivers converge' vs 'our soldiers converge').

Motion verbs are also categorized into three sub-categories: initial (*to leave*), median (*to cross*) and final *to arrive*. The use of location verbs, from a syntactic point of view, is very similar to that of motion verbs as previously described. They are often associated with prepositions of location or place names in hiking guides. They can be used, for example, to describe a step or stop in a journey and may be

associated with specific action such as eating or sleeping. They also allow for better spatial representation and facilitate the location of the different events relative to each other. As mentioned in Chapter 3, verbs of visual perception are very useful in the description of itineraries. We use this information as a criteria in our method of automatic itinerary reconstruction. Spatial named entities associated with verbs of visual perception are not considered as waypoints but refer to visual cues describing landmarks or landscapes. Finally, topographic verbs are used when the narrator is describing a place using its topographical features.

In order to mark and formalise the relations between expanded named entities, topographical terms, spatial relations and verbs of motion or perception, we use and adapt the *VT structures* (Fig. 4.9), previously described by Gaio et al. (2012) and Nguyen et al. (2013).

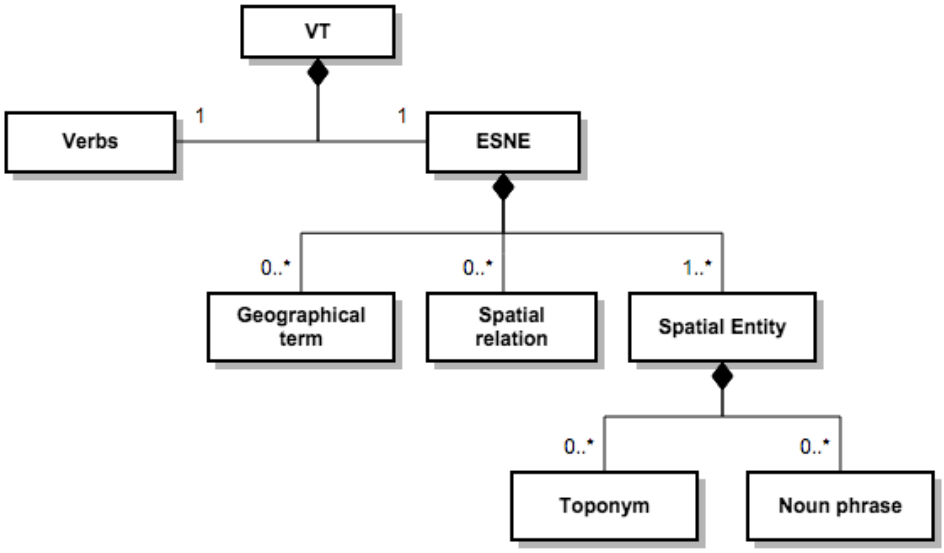


Figure 4.9: UML diagram of the *VT* structure – **Source:** (Gaio et al., 2012)

VT structures (Fig. 4.9) are formally defined as V, S, G, T: groups of classified verbs, spatial relations, geographical terms, and toponymic names, respectively. Classified verbs, also called *verbs of travelogue* by Nguyen (2012), are composed of motion verbs and verbs of perception. *VT* is defined as a pair (v, t) where v is an instance of V and the t set is defined as $t = (g, s, to|t)$, in such a way that g is an instance of G, s is an instance of S and to is an instance of T. The symbol $to|t$ indicates that the third t group can be made up of either t (recursion) or to.

Here are some examples of *VT structures*:

(51) Partir de Malaucène au nord-ouest vers le col de la Chaîne
Leaving from Malaucène to Col de la Chaîne northwest

(52) Suivre la route entre l'hotel de la Vanoise et l'hotel du Petit Mont Blanc
Follow the road between Hotel de la Vanoise and Hotel du Petit Mont Blanc

(53) Passer près de la vielle chapelle Saint Jean
Passing near the old chapel Saint Jean

(54) Suivre le torrent de la Leisse
Follow Leisse torrent

We improve the *VT structures* by taking into account other parts of speech revealing spatio-temporal sequences in the discourse, such as spatial adverbs of locations (Borillo, 2004) and temporal expressions (Schilder and Habel, 2001). These are prepositions of place or time, which occur frequently in the description of itineraries (e.g., *here*, *there*, *near of*, *then*, *after*, *before*, etc.). These prepositions structure the discourse by describing a spatial sequence (a step in a journey) and/or a temporal sequence (a succession of events). We also propose more complex *VT structures* taking into consideration ternary

relationships between a verb of motion and two spatial entities. For instance, example (51) shows that the verb 'to leave' is associated with two named entities 'Malaucène' and 'Col de la Chaîne'. This example also shows the important role played by prepositions. Indeed, the prepositions 'from' and 'to' give information concerning the polarity of the displacement associated to each place names. Our transducer is able to identify that Malaucène is the departure point and that 'Col de la Chaîne' is the arrival point of the displacement described by this *VT structure*. Figure 4.10 and 4.11 show the transducers to annotate classified verbs and *VT structures*.

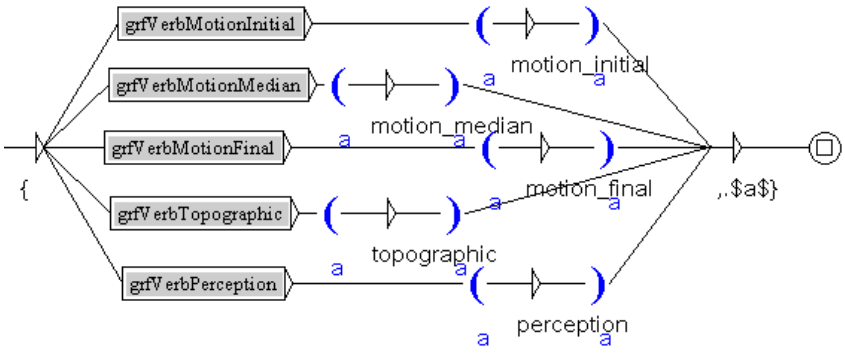


Figure 4.10: Transducer annotating French classified verbs

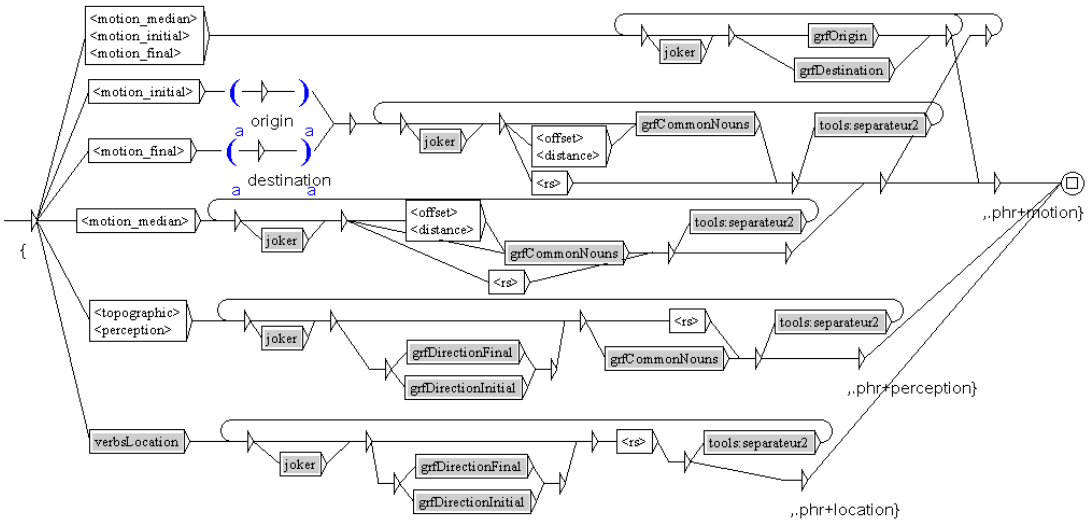


Figure 4.11: Transducer annotating French *VT structures*

According to their output alphabet, these transducers produce the annotation for the ENE (54) shown in Figure 4.12.

4.3 Recognition and Resolution of Spatial Named Entities

4.3.1 Overview

Our goal is to extract and interpret spatial information in order to automatically reconstruct a described itinerary. For that purpose, we propose to focus the classification task on spatial named entities. This task is also known as the *toponym recognition*.

{
{suivre**,.motion_median}**
{
le **{**torrent**,.term+N}**
de la **{**Leisse**,.name}**
,.rs+expandedName}
,.phr+motion}

Figure 4.12: Result of the annotation of the ENE (54)

Our cascade of transducers produces a generic annotation of NEs, i.e., NE boundaries are identified but not classified except those associated with internal or external evidences such as persons or road names.

Our approach of classification is based on the classic *gazetteer lookup* method. However, many of the toponyms contained in itinerary descriptions are fine-grain and are not stored in gazetteers. Due to the problem of incompleteness of geographical resources, we query several gazetteers in conjunction. These gazetteers have a different scope and coverage. Furthermore, the annotation of ENEs helps us to classify named entities as spatial named entities even if they are not found in gazetteers.

We combine the classic *gazetteer lookup* method with the sub-typing of place names. Our approach is to analyse information contained within the ENEs. We lookup into geographical ontologies or lexicons to determine if the sub-type contains within the ENE matches a geographical concept. This method is based on the approach described in Nguyen et al. (2013). It relies on ENEs recognition expressed in terms of semantic features and combines the use of specific intra-sentential lexico-syntactic relations and external resources like gazetteers, thesauri, or ontologies.

As mentioned before, we propose a hybrid solution combining subtyping of place names and gazetteer lookup. We query geographical resources to classify NEs as spatial named entities and also to find their geo-coded representation. Indeed, our goal is not only to know that the name refers to a place but to be able to locate this place in order to reconstruct the plausible footprint of the itinerary.

We proposed to query several geographical resources in order to obtain a better coverage and increase the number of toponyms found. However, querying several geographical resources increases also the problem of toponym ambiguity. Indeed, some toponyms may be stored in several gazetteers and sometimes their coordinates are not exactly the same. To solve this problem we apply a radius (buffer) in order to remove near duplicate toponyms. We apply this method on toponyms having the same name and coming from different geographical resources.

Furthermore, toponym resolution Leidner (2007) refers to the association of a non-ambiguous location with a place name. Thus this involves the problem of toponym disambiguation and, as we have seen in Section 2.3.4 of the related work, during this process a lot of ambiguities may arise.

Once spatial named entities are extracted and spatial named entities boundaries are identified, the main issue to solve for the resolution is the ambiguity contained in place names. The problem of solving toponyms ambiguities is known as toponym disambiguation (Buscaldi and Rosso, 2008b). As mentioned in the state of art (Section 2.3.4), we consider two main types of toponym ambiguities: *reference ambiguity* and *referent ambiguity*.

Reference ambiguity defined by Smith and Mann (2003) refers to places having several names. For example, this happens when the name has changed over the time, or when the name commonly used by people is different from the official name. Apart from these clear cases of reference ambiguity, our method is focused on the problem of the inclusion or not inclusion of subtypes within the official name of a toponym (*structural ambiguity*). Section 4.3.2 describes our proposal concerning the disambiguation of structural ambiguity.

Referent ambiguity defined by Smith and Mann (2003) refers to place names that represent several geographical places. *Referent ambiguity* is also known as *referential ambiguity*, which Leidner (2007) considers as a subset of the *linguistic ambiguity*. Some well-known examples of toponyms are usually used to illustrate this class of ambiguity. For instance, the toponym 'Paris' refers to hundreds different

geographic places around the world such as the capital of France and cities in different countries like in the United States (Texas), Canada, Togo, Pananama, etc. Section 4.3.3 describes our proposal to solve this kind of toponym ambiguity.

Furthermore, with respect to the problem of toponym resolution another problem arises: the incompleteness of geographic resources. Indeed, frequent occurrences of toponyms (especially fine-grain toponyms) are not stored in geographic resources. As far as we know, the incompleteness of geographic resources has not been considered in the literature as a type of toponym ambiguity. In this thesis, we propose to consider the incompleteness of geographical resources as a type of toponym ambiguity called the *unreferenced toponyms ambiguity*. Section 4.3.4 describes the proposed method for the geocoding of unreferenced toponyms in order to partially solve the problem of incompleteness.

4.3.2 Subtyping of Place Named Entities

The use of contextual elements such as words that have a geographical denotation (e.g., downtown, valley, ridge) is very important in toponym disambiguation (Hollenstein and Purves, 2010) and allows the ambiguity to be removed from the type of the spatial entity in consideration. In the remainder of this section we distinguish two concepts: *type* and *subtype*. The type refers to the geographical nature of the spatial object in consideration, whereas the subtype refers to the expression of the type within the textual description if it does exist.

A large number of spatial entity types exist: geopolitical entities (e.g., countries, administrative divisions), populated places (e.g., towns, addresses and postal codes), and natural geographical entities (e.g., parks, valleys, mountains, rivers), all of which can create ambiguities about the type of geographic object in question. As described by Rauch et al. (2003), we propose to use the local linguistic context, when available, to identify subtypes associated with toponyms (e.g., city, lake, river) and then filtered out irrelevant references. Then, we propose the use of the annotation of ESNE to extract the local context associated with toponyms (subtype). Thanks to the concept of ENE and to our automatic annotation system we are able to distinguish the proper name and the subtype that are part of the toponym. Figure 4.13 shows the annotation of the ESNE obtained with our processing chain following the TEI guidelines (see Chapter 5 for further details). The `<geogName>` element refers to the ESNE, the `<geogFeat>` element refers to the subtype of the toponym and the `<name>` element annotates the proper name.

(55) hamlet of Fontanettes

```
<geogName>
   <geogFeat>hamlet</geogFeat>
   of
   <name>Fontanettes</name>
</geogName>
```

Figure 4.13: Annotation of the ESNE 'hamlet of Fontanettes'

First, our gazetteer lookup method query geographical resources with the full name of the toponym (including subtype and name) and if there is no record for the full name we query a second time only with the name. Then we compare the subtype extracted from the text with the metadata associated with each record to match corresponding references and filter out irrelevant ones. One problem is that the local context is not always available in the textual description.

Due to the ambiguous nature of natural language and more particularly to the phenomenon of underspecification (which holds that values are predictable), a spatial named entity may be expressed in texts without any subtype, in that case, the subtype is implied and refers to the intrinsic or default type of the spatial entity. For instance, 'France' is a country, and 'Pau' is a city. Furthermore, sometimes toponyms are stored in gazetteers with the so-called *full name*, which means that the name consists of a subtype and a proper name. For instance, Figure 4.14 shows the result for the query 'Lac de la Rocheure' using

the Nominatim API of Openstreetmap. We can notice that this ESNE built with the geographical term 'Lac' (i.e., *lake* in english) and the proper name 'Rocheure' is found with its full name. Although there is no ambiguity with the name of the toponym, we propose to verify that the result is consistent by comparing the subtype extracted from the text with the metadata available on the resource thank to a matching system based on an ontology. For instance, this system is able to determine that the subtype 'lac' (i.e., lake) matches the type 'water' defined by the resource (Figure 4.14). However, it may happen that the reference found in resources does not correspond to the one we are looking for, which is frequent for buildings like hotels or restaurants having the name of closed locations.

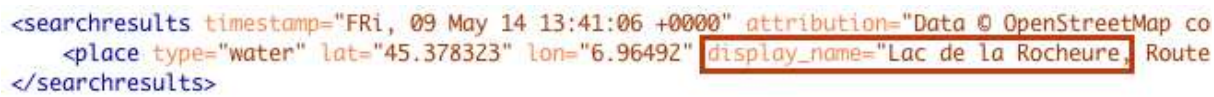


Figure 4.14: Example of results for the query 'Lac de la Rocheure' using the Nominatim API of Openstreetmap

However, even if the toponym is stored in the resource with its full name, it may have several records that involve ambiguities. We can notice on Figure 4.15 that there are several types of ambiguity: *structural* and *referent*. Figure 4.15 shows the results for the query 'Mont Blanc' which is the highest mountain peak in Europe and is located on the French-Italian border. This unusual location shared by two countries is shown on the two first records of Figure 4.15. The two first references refer to the same location (according to their geographical coordinates), the first one in Italy and the second one in France. Then, we notice that all the records do not have the same type. Thus, according to the problem of *structural ambiguity*, we propose to use the matching of subtype to remove irrelevant references. But as we can see on this example, there are several records with the same value for the type attribute ('peak'). This means that at the end of the first step of disambiguation dealing with *structural ambiguity*, it may still remain some ambiguities. This problem of referent ambiguity will be addressed in Section 4.3.3.

Figure 4.15: Example of results for the query 'Mont Blanc' using the Nominatim API of Openstreetmap

Most of the time, although subtypes of toponyms are part of the name expressed in the textual description, in many cases toponyms are just stored in gazetteers using the proper name. However, they are usually associated with metadata describing the type of feature (hamlet, city, stream, etc.). For example, the ESNE 'hamlet of Fontanettes', which consists of the subtype 'hamlet' and the proper name 'Fontanettes', is not found in geographical resources.

The ENE 'hamlet of Fontanettes' is not found in resources with its full name but only with the proper name 'Fontanettes'. Figure 4.16 shows the results for the query 'Fontanettes'. We can notice that the value for the attribute *display_name* contains only the proper name 'Fontanettes' and not the full expression of ENE 'hamlet of Fontanettes'. Moreover, we can notice in Figure 4.16 that several results exist when we query the gazetteer with the proper name 'Fontanettes'. One way to know which result matches our query is to compare the metadata of the results and more especially the attribute *type* with the subtype (*geogFeat*) of the toponym extracted from the text. In this example (Figure 4.16) only one

record matches our query. But in some cases, it may remain referent ambiguities when several results have the same type.

```
<searchresults timestamp="Fri, 17 Oct 14 08:32:29 +0000" attribution="Data © OpenStreetMap cor
    <place type="hamlet" lat="45.3805943" lon="6.7336952" display_name="Fontanettes, Albertvil
    <place type="bus_stop" lat="46.1812175" lon="5.9821148" display_name="Fontanettes, Rue des
    <place type="bus_stop" lat="46.1808313" lon="5.9820612" display_name="Fontanettes, Route d
    <place type="bus_stop" lat="45.3211483" lon="6.5367973" display_name="Fontanettes, Les For
</searchresults>
```

Figure 4.16: Example of results for the query 'Fontanettes' using the Nominatim API of Openstreetmap

To summarize, our method is to find the subtype of toponyms using contextual information described in the textual descriptions in order to compare this subtype with the value of the metadata available with each record of the toponym in the geographical resources.

Additionally, the fact of querying different gazetteers also expands the probabilities of the method to find a matching with one of the possible names of a place. The implementation of this proposal is described in Section 6.2.3. Furthermore according to results of the experiments shown in Section 7.6.1, one toponym may have both reference and referent ambiguities and in this case this approach of disambiguation is not sufficient. Another problem is that the subtype is not always given in the text and other toponyms with the same name and type may be also stored in the gazetteer.

4.3.3 Density-Based Spatial Clustering

In many cases the disambiguation approach based on subtyping of place names presented in section 4.3.2 is not enough because external geographical resources may contain several toponyms with both the same name and type. In the case of *referent ambiguity* we need a mechanism to distinguish the good group of toponyms associated with the real trajectory of the hiking description.

As stated by Smith and Crane (2001), a place is more likely to be located near other places mentioned around it. This is particularly true in the case of hiking descriptions. Indeed, each place is related to another one by a motion event or is related to the route by perception expressions in order to describe landmarks and more generally the spatial context of the hike.

The effect of referent ambiguity may be very significant. For instance, Figure 4.17 shows a map of the geographic distribution of the result of the geocoding of the toponyms expressed in phrases (56) to (61).

(56) Emprunter successivement **rue des Capucins** et **rue de Compostelle** [...]
Walk down **Capuncins Street** and **Compostelle Street** [...]

(57) Après l'entrée de **l'usine de Fontanille** prendre à gauche. [...]
After the entry of the **factory Fontanille** turn left [...]

(58) Suivre jusqu'à **Saint-Privat** [...]
Follow to **Saint-Privat** [...]

(59) Atteindre **La Roche** [...]
Reach **La Roche** [...]

(60) Traverser le **hameau Lic** [...]
Cross **hamlet Lic** [...]

(61) Suivre la route jusqu'à la **chappelle Saint-Roche** [...]
Follow the road to the **chapel Saint-Roche** [...]

Each point on the map refers to a location and each colour refers to a toponym. These phrases are extracted from a French hiking description and refer to a hiking trail occurring in the French Alps. This example illustrates the high level of ambiguity. A large number of points are located all over France and most of the toponyms of this example are fine-grained toponyms very common in France. Indeed, a lot

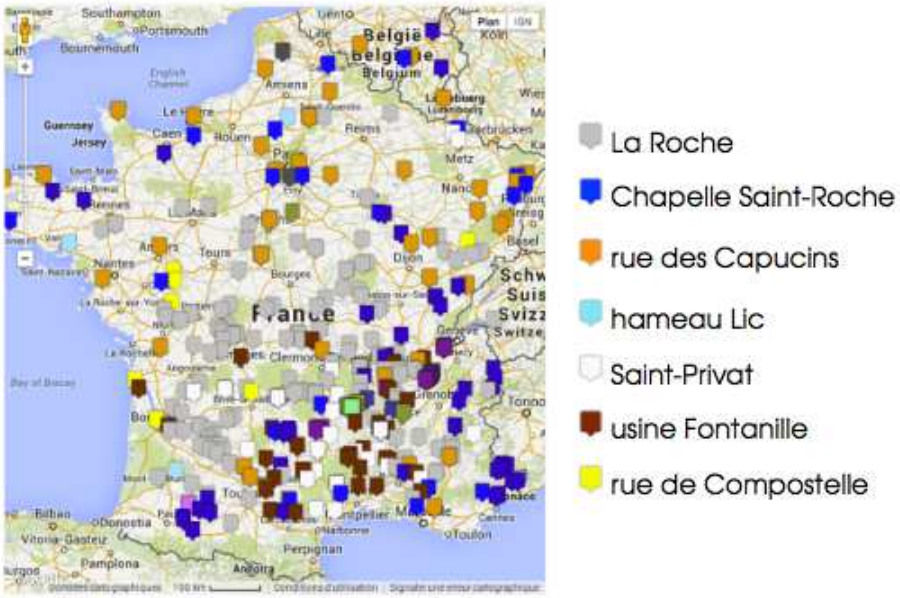


Figure 4.17: Illustration of the referent ambiguity

of small villages (settlements), churches, hotel or streets have the same name. We give further details about the amount of ambiguous toponyms with respect to the corpus of experiments in Section 7.6.

The fact that our main objective is to reconstruct itineraries helps in the disambiguation process. Indeed, the main difficulty in several works dealing with the problem of toponym disambiguation (see Section 2.3.4) is that the method needs to find some kind of relationships between toponyms in order to discard irrelevant references using other toponyms or unambiguous toponyms. For instance, relationships between toponyms may refer to geographic distance (Buscaldi and Rosso, 2008b) defining an area of interest or arborescent proximity (Bensalem, 2010) using conceptual matching.

As we are working with specific texts describing displacements, we cannot use standard methods of toponyms disambiguation that are usually applied to a corpus of news articles. Each corpus of documents have pros and cons, and disambiguation methods have to be chosen according to the specific context of evocation of toponyms. For example, in news articles the notion of event is very important and can help to disambiguate toponyms, and methods commonly use semantic relationships between different types of entities (company, person and places). However, it is different in the case of a travel description. Standard knowledge-based approaches (see Section 2.3.4) are not suitable because it is common to find toponyms referring to geographical entities of varying size, and occurrences of toponym have nothing to do with importance in terms of population heuristics. We can easily imagine a hiking trail starting from a big city, and then leaving the urban area and continue in forests or mountains. Another problem is the fact that it is common to have hiking trails or travelogue describing displacements occurring on borders and going from one country to another. For example, this is common in the Pyrenees between France and Spain, or in the Alps between France, Italy and Switzerland, etc. To illustrate these constraints we can cite the well-known *Camino de Santiago* (Way of St. James) which refers to several pilgrimage routes in Europe crossing different countries and different kinds of geographical areas (urban area, countryside). These routes include typical landmarks such as cathedral, church or chapel but also a lot of different landmarks located all along the route. Furthermore, techniques based on hierarchical relations are also difficult to apply in the specific case of a displacement described by documents containing usually fine-grain toponyms or natural feature (mountains, lakes, refuges) because the coverage provided by knowledge resources is very limited.

Similar to the works of Intagorn and Lerman (2011) or Feuerhake and Sester (2013), we propose to use clustering algorithms to find collections sharing a spatial property, and in our case, these collections enable us to find clusters of the most likely geospatial points belonging to a hiking trail. Our proposal to

disambiguate *referent ambiguities* can be classified as a map-based approach. In particular, we use the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) clustering algorithm introduced by Ester et al. (1996). It uses the concept of density to determine the neighbourhood of a point, that is, what constitutes a cluster. This map-based approach can deal with disparate toponyms in terms of importance without considering population or social statistics and can also deal with toponyms located in different countries. DBSCAN uses two parameters to define the density concept: *Eps* and *MinPts*. *Eps* (epsilon radius) determines the area of a neighborhood and *MinPts* determines the minimum number of points that have to be contained in that neighborhood to deem it a cluster. Figure 4.18 illustrates the DBSCAN algorithm, circles refer to the maximum epsilon radius between two points. Then, green and blue circles highlight points having neighbours, whereas red dots represent outliers, which are not close enough from other points.

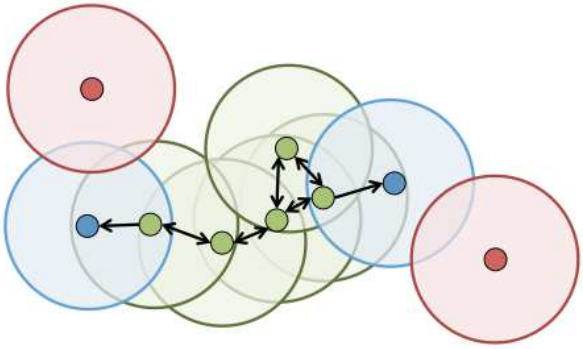


Figure 4.18: Illustration of DBSCAN (Ester et al., 1996)

In our current methodology, the values of DBSCAN parameters have been empirically adjusted according to the features of the linking dataset used in the experiments. An automatic method based on machine learning such as the one proposed by (Anders and Sester, 2000) or (Ester et al., 1998) is discussed in the conclusion of this chapter (Section 4.4).

DBSCAN can deal with the problems of data with noise, i.e. DBSCAN has the ability to detect outliers. In our context, an outlier is a point that does not belong to the hiking trail cluster. Additionally, since hiking trails may have many points describing different trajectory shapes, DBSCAN can find arbitrarily shaped clusters. The output of the DBSCAN is a set of clusters of toponyms whose footprints are close. Every cluster represents a possible set of points describing the hiking trail. Then we need a way to identify the best cluster matching with the set of points in the hiking trail. The heuristic is defined as follows: given a set of cluster C_1, C_2,... ,C_n generated by the clustering algorithm, the best cluster C_b is the one containing the largest number of distinct toponyms. In other words, the best cluster identifies the area with the largest co-occurrence of toponyms.

The implementation of this proposal is described in Section 6.2.3 and according to results of the experiments shown in Section 7.6.2 the proposed method reduces considerably the referent ambiguities. Now the remaining problem is how to deal with unreferenced toponyms ambiguity, i.e. how to find spatial locations for toponyms that are not stored in gazetteers.

4.3.4 Geocoding for Unreferenced Toponyms

Whatever the resource is selected for geocoding, the problem that usually arises is its completeness. Indeed, we noticed during the experiments (see Chapter 7) that incompleteness of geographical resources is an important factor involving toponyms ambiguity. We introduce the notion of *unreferenced toponyms ambiguity* and we propose a method to approximate the spatial footprint of those unreferenced toponyms.

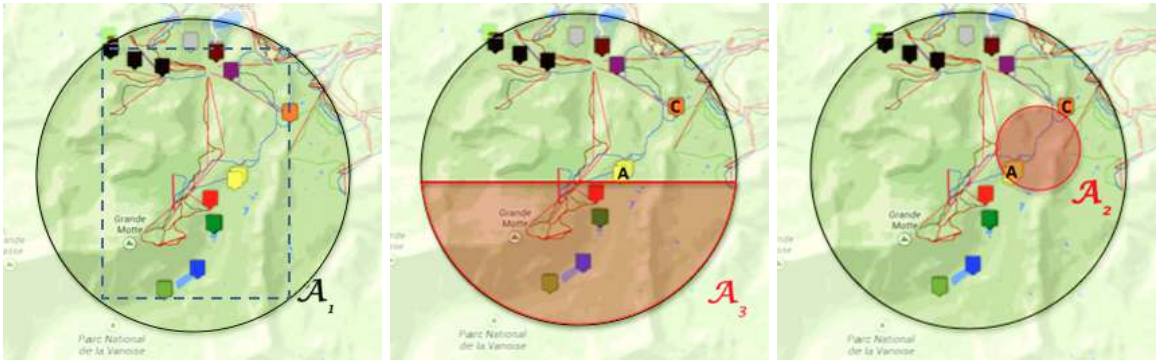
Currently, the present Web geocoding public market is dominated by geocoding services for average users, i.e. users that can accept low Quality of Service in terms of resolution (Florczyk et al., 2010). For instance, geocoding of administrative units, street names in urban areas, or names of well-known touristic sites are the main needs. But there are contexts, even for public citizens or casual users, where

the completeness of resources is crucial, in particular as regards the geocoding of fine-grain toponyms. For instance, in a corpus of narrative descriptions of places in a small area, it is common to find toponyms referring to geographical entities of varying size. Additionally, there are frequent occurrences of micro-toponyms, which are not usually found in gazetteers thought for broader audiences.

With respect to the map-based disambiguation part, our work is similar to the method proposed by Habib and Van Keulen (2012) as we also use a clustering technique. The difference is that they do not face the problem of not finding toponyms in their gazetteer. Additionally, the granularity of their spatial footprint is higher than ours: their objective is to identify to which country a set of toponyms belongs to. The disambiguation part may also have some similarities with the work of Derungs and Purves (2014) as the types of input documents to be processed are similar: descriptions of natural landscapes could be considered as a superset of hiking descriptions. However, their aim is not to try to geolocate toponyms not found in a backend gazetteer. Anyway, the vocabulary for natural features that they analyse could probably intersect with the toponyms (or part of the place names) that we try to geolocate in this work.

With respect to the creation of new toponyms, our work also has some similarities with the one proposed by Lieberman et al. (2010). Our purpose is also to create a gazetteer or spatial lexicon for an intended audience in small areas. Additionally, the proximity of locations associated with ambiguous toponyms is the main criterion to discard alternatives. Furthermore, some ideas of the works of Scheider and Purves (2013) and Hao et al. (2010) could be used to provide additional information to discovered fine-grain toponyms. The annotation obtained after the toponym extraction process could be used to inform about a more precise geolocation, or the topics associated to this toponym. This additional contribution about the estimation of the location is closely related to the step of spatial inference proposed by Leidner and Lieberman (2011) in their 'Reference model for processing textual geographic references'. Our proposal is to infer locations from the locations of previously disambiguated toponyms using spatial relation contained in the textual description such as 'south of', 'north of' and 'between'. However, these spatial inferences cannot be as precise as points with geographical coordinates (latitude/longitude). These spatial inferences are represented by a geographical area which can be refined depending on various spatial informations contained in the textual descriptions.

Our approach is a hybrid solution that combines map-based disambiguation with information about spatial relations extracted from the textual description for the assignement of georeferences for new toponyms.



(a) (b) (c)

Figure 4.19: Refining spatial inferences according to the context

For example, Figure 4.19 shows three different cases of inferences. In the first case (Fig.4.19a) there is no other explicit spatial information in the text linked with the unreferenced toponyms. In this case, when there is no information concerning the context of a toponym, we define a geographical area that contains all well-located toponyms thanks to the clustering method previously described. Indeed, in the specific context of the description of an itinerary, toponyms are related to each other and they are located in the same area. The second case is when explicit spatial relations are associated with the unreferenced toponym. For example, if we know that the unreferenced toponym is somewhere south of A (Fig.4.19b),

then we can define a new area smaller than the previous one. A third case arises when we have even more information available in the textual description. In this case we can define a much smaller area. For example, if we know that the unreferenced toponym is somewhere between two other toponyms, we can define a small area between these two toponyms (Fig.4.19c). Spatial relations are very important to determine the context of unreferenced toponyms.

The implementation of this proposal is described in Section 6.2.3 and the results of the experiments are shown in Section 7.6.3.

4.4 Summary

This chapter has proposed a processing chain for the geoparsing and geocoding of texts containing travel descriptions,

We proposed a linguistic approach based on syntactic-semantic patterns for Named Entity Recognition and Spatial Role Labeling. Our proposed approach is a hybrid solution combining a POS analysis pre-processing, a cascade of finite-state transducers and the interrogation of external gazetteers for the annotation of named entities and spatial information such as spatial relations and expression of motion and perception. Furthermore, this chapter makes a special emphasis on two main problems related to the geocoding part of the method: the existence of ambiguous toponyms, and the lack of gazetteers with enough coverage for fine-grain toponyms. The solution proposed for addressing these two problems has been based on the use of clustering techniques. On the one hand, the definition of clusters provides a map-based disambiguation approach to identify the clusters with the highest number of candidate toponyms in terms of distance. On the other hand, the bounding polygon of these clusters can be used as an estimate to define the location of those fine-grain toponyms not found in gazetteers.

The only requirement for the application of the proposed method to a set of travel descriptions in a new language is the customization of the geoparsing part and the availability of gazetteers for the geographic area covered by the texts in this new language.

However, some refinements could be included in the proposed method to increase its performance. On the one hand, additional heuristics could be taken into account to address the problem of reference ambiguity (i.e., several place names for the same place). In the current work we have considered the problem of structural ambiguity (i.e., finding or not finding sub-types within the toponym names in gazetteers). But other problems could be taken into account: variants of toponym names and types in other languages, abbreviations, etc. On the other hand, with respect to the adjustment of parameters in the application of the DBSCAN clustering method to deal with the problem of referent ambiguity, a possible refinement would be the automatic definition of parameter values by means of machine learning techniques as it is proposed in other works using clustering techniques (Anders and Sester, 2000) and (Ester et al., 1998).

Finally, it must be noted that the proposed processing chain for geoparsing and geocoding could be applied to other types of text corpora. The proposed method is general enough to be applied for any kind of narrative descriptions in a open and natural area.

Chapter 5

A Multi-Scale Markup Language:

A Case Study of Geospatial Semantic Language

The limits of my language mean the limits of my world.
— Ludwig Wittgenstein, *Tractatus Logico-Philosophicus*

Contents

5.1 Introduction

5.1.1 Overview

In this chapter we describe a model for marking semantically an unstructured text, such as in semantic role parsing (Gildea and Jurafsky, 2002; Pradhan et al., 2003) (Figure 5.1). We define a formal representation of unstructured text written in natural language that can be applied for the task of Named Entity Recognition (NER) and Spatial Role Labeling (SpRL).

The objective of this chapter is to propose a multi-scale annotation process based on a core generic layer, which can be adapted into a more specific layer depending on the need of any project. The proposed markup language is based on the TEI Guidelines[47] (TEI P5, 2014) to propose a generic and extensible markup language. This language is particularly dedicated for the text mining task and can be transformed into a more specific markup language by adding more semantic relationships between elements of the text.

[47] http://www.tei-c.org/Guidelines/P5/

Although our proposal is still at an early stage of development, the proposed markup language was applied for the problem of automatic information extraction and toponym resolution described in Chapter 4 and for the problem of automatic itinerary reconstruction described in Chapter 3. We show the feasibility of this proposal from a generic annotation of texts describing itineraries toward a geospatial semantic annotation.

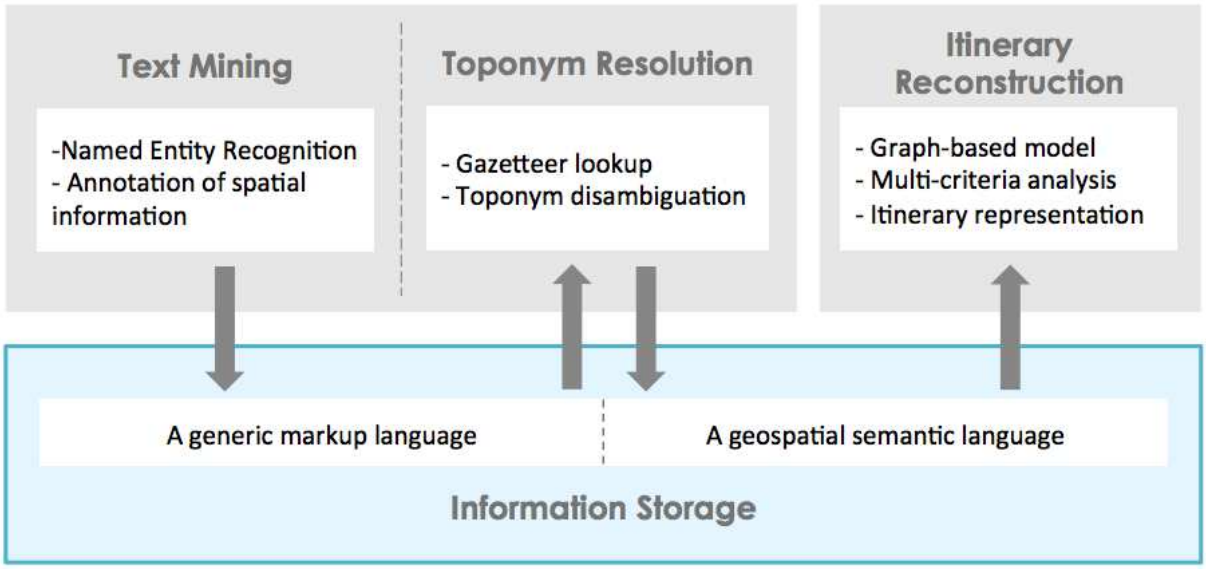


Figure 5.1: Contribution of this chapter (highlighted in blue)

5.1.2 Motivation and Background

Our proposal for the automatic reconstruction of itinerary described in Chapter 3 uses elements annotated from the textual description combined with information found in external geographical resources. With regard to our concern, we need a markup language able to annotate information in text in order to reconstruct the described itinerary.

As we have seen in the state of art (Section 2.4), we consider two categories of markup languages, those dedicated to the encoding of spatial data (e.g., GML, KML) and those dedicated to the annotation of spatial or spatio-temporal information in texts (e.g., SpatialML, ISO-Space). SpatialML (Mani et al., 2008) is more focused on the annotation of static spatial information and does not provide any support to identify spatio-temporal information such as motion. ISO-Space (Pustejovsky et al., 2012) annotates spatio-temporal information and has been also designed for capturing implicit spatial information. Although ISO-Space seems a comprehensive standard for capturing spatio-temporal information, some of its elements remain complex and are not really suited for a fully automatic process. Furthermore, these markup languages are more focused on the specification of relations (spatial or spatio-temporal) than NEs. Standard markup languages consider NEs as being only composed of a pure proper name, whereas we consider more complex expressions that we call expanded named entities (ENEs). For that reason, we define more deeply the concept of ENE (Section 5.2.3) and we propose the specification of a markup language adapted to the annotation of both ENEs and spatial/spatio-temporal relations.

Furthermore, as we have seen in the state of art (Section 2.4.4), TEI is a standard for textual markup and provides a guide to best practices for interchange and encoding textual information. TEI has been designed to be the more generic and adaptable as possible. It provides a generic framework particularly adapted for the customization. Then, we decided to propose a TEI-compliant markup language based on a core generic layer, dedicated to the annotation of NEs, which can be used to share pre-processed corpus of documents. Furthermore, our proposal of a multi-scale markup language is designed to be compatible with an automatic process of annotation.

The remainder of this chapter is structured as follows. Section 5.2 describes the global attributes and XML elements defined by the TEI Guidelines for text segmentation (sentence, word, etc.) and for the annotation of NEs (name, referring string, etc.) applied for the definition of the generic layer of our multi-scale markup language. Then, Section 5.3 describes the adaptation of the generic language for a specific semantic role. In particular, we describe the elements and their attributes, dedicated to geospatial semantic information, which we use for the specific layer of our multi-scale markup language specification. Finally, Section 5.4 summarises and concludes this chapter.

5.2 A Generic Markup Language for Expanded Named Entity Representation

This section describes the 'core generic layer' of our multi-scale annotation language. We define the specification of a generic markup language based on the TEI P5 guidelines[48]. TEI is an international standard for textual markup defined by the TEI Consortium. It provides a guide to best practices for interchange and encoding of textual material in digital form (see Section 2.4 for further details).

The main objects of our proposed generic markup specification are categorized in three groups. The first one refers to the global attributes defined by the TEI Guidelines which are available for all TEI elements, the second one refers to the standard elements describing the text segmentation and the third one refers to the elements contained within expanded named entities.

5.2.1 Global Attributes

TEI Guidelines provide attributes common to all elements in the TEI encoding scheme. They are described in the *tei* module (the TEI Infrastructure). Table 5.1 lists all the global attributes optionally available for any TEI element.

Attribute name	Description
`xml:id`	unique identifier
`n`	number or label
`xml:lang`	language
`rend`	indicates how the element was rendered or presented in the source text
`style`	style definition which defines the rendering or presentation used in the source text
`rendition`	points to a description of the rendering or presentation used in the source text
`xml:base`	base URI reference to resolve relative URI references into absolute URI references
`xml:space`	intention about how white-space should be managed by applications

Table 5.1: Global attributes for every TEI element

The `xml:id` attribute provides a unique identifier for the element and the format of the value of this attribute is defined in the XML recommendations provided by the W3C. The `xml:id` attribute has to be unique within a single document, and `xml:id` values shall begin with a letter. The `n` attribute also provides an identifier but without any restriction in the format of the value. The `xml:lang` attribute indicates the natural language and writing system of the element. The `rend`, `rendition` and `style` attributes are used to give information about the physical presentation of the text. For instance, they provide information about font-style (e.g., italic, bold, etc.). The `xml:base` attribute belongs to XML namespace like the `xml:id` and `xml:lang` attributes, and it is used to set a context for all relative URLs. Finally, the `xml:space` attribute provides a mechanism for indicating how white-spaces should be managed. This last attribute is used in very specific cases.

However, in the current version of our proposal, we are using only two attributes: `xml:id` and `xml:lang`.

[48] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html

5.2.2 Text segmentation

This section describes the elements that refer to the text segmentation. Structuring of textual information operates on various levels of discourse. They describe the segmentation of a text into traditional linguistic categories such as sentences, words and punctuation marks. These elements are defined by the TEI Guidelines and belong to the *Analysis* module (Simple analytic mechanisms) of TEI.

(62) On parvient ensuite au refuge du Col de la Vanoise.
Then we reach the refuge of Col de la Vanoise.

Sentence: <s> (s-unit)
The <s> element contains a sentence-like division of a text. It may be used to annotate sentences, or any other non-overlapping segments such as complete and non-nesting segmentation of a text. Figure 5.2 shows the result of the annotation of sentence (62).

```
<s>On parvient ensuite au refuge du Col de la Vanoise.</s>
```

Figure 5.2: Example of annotation of sentence.

Word: <w>
The <w> element represents a grammatical word. The word segmentation depends on which characters are defined as words dividers. Whereas for Indo-European languages the word separator is typically a blank space, the East Asian languages such as Chinese and Japanese differ, i.e., words are not explicitly delimited. However, in our work we are mainly focused on Indo-European languages and more particularly on French, Spanish and Italian.

Attribute name	Description
lemma	Canonical form of the word
type	Part-of-speech
subType	Semantic sub-categorization

Table 5.2: Attributes for <w> tag

Table 5.2 shows the current set of <w> attributes defined in our specification and Figure 5.3 shows the result of the annotation of sentence (62).

```
<s>
  <w lemma="on" type="PRO">On</w>
  <w lemma="parvenir" type="V" subtype="motion_final">parvient</w>
  <w lemma="ensuite" type="ADV">ensuite</w>
  <w lemma="au" type="PREPDET">au</w>
  <w lemma="refuge" type="N">refuge</w>
  <w lemma="du" type="PREPDET">du</w>
  <w lemma="col" type="N">Col</w>
  <w lemma="de" type="PREP">de</w>
  <w lemma="le" type="DET">la</w>
  <w lemma="Vanoise" type="NPr">Vanoise</w>.
</s>
```

Figure 5.3: Example of annotation of words.

The lemma attribute for <w> is optional and may contain the canonical form of the word. The type attribute for <w> is mandatory and contains the lexical categories of word (part-of speech). Although there are significant variations depending on the language, common parts of speech are: noun, verb, adjective,

adverb, pronoun, preposition, conjunction, interjection, numeral, article or determiner. The label for each category is usually defined by abbreviations such as N, V, ADJ, ADV, etc. Different POS tagsets have been defined and used in well-known projects such as the Brown Corpus[49], the Penn Treebank[50] (Santorini, 1990) and the French Tree Bank[51] (Abeillé et al., 2003). Further information concerning the tagset used in our work will be given in Section 6.2. The `subType` attribute for `<w>` is optional and may contain semantic information. For instance, in the current version of the language, `subType` is used to classify verbs. The possible values are: *motion_initial*, *motion_median*, *motion_final*, *perception* and *topographic*.

Grammatical phrase: `<phr>`
The `<phr>` element defined by the TEI Guidelines in the *analysis* module represents a grammatical phrase. The `type` attribute may be used to indicate the type of phrase such as noun phrases, prepositional phrases, etc. Figure 5.4 shows the result of the annotation of sentence (62).

```
<w lemma="on" type="PRO">On</w>
<phr type="verb">
  <w lemma="parvenir" type="V" subtype="motion_final">parvient</w>
  <w lemma="ensuite" type="ADV">ensuite</w>
  <w lemma="au" type="PREPDET">au</w>
  <w lemma="refuge" type="N">refuge</w>
  <w lemma="du" type="PREPDET">du</w>
  <w lemma="col" type="N">Col</w>
  <w lemma="de" type="PREP">de</w>
  <w lemma="le" type="DET">la</w>
  <w lemma="Vanoise" type="NPr">Vanoise</w>.
</phr>
```

Figure 5.4: Example of annotation of a `<phr>` element

Punctuation: `<pc>`
The `<pc>` element contains a character or string of characters regarded as constituting a single punctuation mark. Table 5.3 shows the current set of `<pc>` attributes.

Attribute name	Value
`force`	*strong*, *weak*, *inter*
`type`	*declarative*, *imperative*, *interrogative*, *exclamatory*

Table 5.3: Attributes for `<pc>` tag

All the attributes of the `<pc>` element are optional. The `force` attribute for `<pc>` indicates if the considered punctuation mark is a separator for words or phrases. The `type` attribute for `<pc>` indicates the kind of punctuation. Figure 5.5 illustrates the annotation of punctuation using the TEI `<pc>` element.

```
<pc>,</pc>
<pc force="strong" type="interrogative">?</pc>
<pc force="strong">.</pc>
```

Figure 5.5: Example of annotation of punctuation characters.

[49] http://www.comp.leeds.ac.uk/ccalas/tagsets/brown.html
[50] http://www.cis.upenn.edu/~treebank/
[51] http://www.llf.cnrs.fr/Gens/Abeille/French-Treebank-fr.php

5.2.3 Expanded Named Entity Representation

5.2.3.1 Principles

In this section, we will describe in more details the notion of ENE and then we will describe the TEI elements and their attributes used to represent ENEs.

As we have described in Section 4.2 and according to Jonasson (1994), we have defined an ENE as an entity built from a proper name and that can be composed of one or more concepts. According to Jonasson (1994) there are two categories of proper names: pure and descriptive. Pure proper names can be simple (i.e., composed of a single lexeme) or complex (i.e., composed of several lexemes) and are only composed of proper names. Descriptive proper names refer to a composition of proper names and common names (i.e., expansion). In other words, descriptive proper names overlap pure proper names. Descriptive proper names refer to a NE built with a pure proper name and a descriptive expansion. This expansion can change the type (e.g., location, person, etc) of the initial pure proper name.

In our work, we consider both categories of proper names (i.e., pure and descriptive), whereas most of works of NER are usually only considering pure proper names. We define several levels of overlapping, (0, 1, 2, etc.) for the representation of ENEs. Each level is encapsulated in the previous level.

Level 0
ENEs of Level 0 refer to pure proper names. It can be seen as the core component of an ENE. Thus, we consider NE as a special kind of ENE. The following examples (63 - 65) show some examples of entity of level 0:

(63) Sète → one entity (location)

(64) Balaruc-le-Vieux → one entity (location)

(65) Charles de Gaulle → one entity (person)

Level 1
ENEs of Level 1 refer to descriptive proper names composed of a pure proper name (i.e., an entity of level 0) and a common noun (i.e., expansion). The following examples (66 - 68) show the representation of ENEs. We can notice that in these cases, descriptive expansions do not change the intrinsic nature of the object described by the proper name.

(66) commune de Balaruc-le-Vieu

(67) région Aquitaine

(68) général Charles de Gaulle

However, when the associated term is not equal to the intrinsic or default type of the pure proper name, it defines a new entity that overlaps the pure proper name. The following examples (69 - 72) illustrate that an entity may contain the name of another entity, and that the new entity may have a different type. For instance, a sentence containing the expanded named entity “maire de Sète” (maire = mayor) is referring to the person and not necessarily to the city.

(69) port de Sète → two entities, **Sète** (location) and **port de Sète** (location)

(70) étang de Thau → two entities, **Thau** (location) and **étang de Thau** (location)

(71) château de [Versailles] → two entities, **Versailles** (location) and **château de Versailles** (location)

(72) maire de [Sète] → two entities, **Sète** (location) and **maire de Sète** (person)

Level 2
ENEs of Level 2 refer to a descriptive proper name composed of another descriptive proper name. ENEs of Level 2 are built with ENEs of level 1 and with a descriptive expansion.

The following examples (73 - 75) show some ENEs of Level 2. The behaviour is the same as for the previous level, the expansion can change the nature of the object described by the ENE of level 1. For instance, in example (75) the person ENE 'général Charles de Gaulle' becomes a reference to a location with the expansion 'avenue' having a geographical sense.

(73) hinterland du [port de [Sète]]

(74) maire de [la ville de [Sète]]

(75) avenue du [général [Charles de Gaulle]]

Level 3
ENEs of Level 3 are built with ENEs of level 2 plus a descriptive expansion. Actually, there is not really a limit to the overlapping. However, it is really rare to find an ENE of level 3 or more.

The following examples (76 - 77) show some ENEs of Level 3.

(76) les bâtiments de [l'exploitation agricole du [domaine de [la Brunelie]]]

(77) la valle glaciale [del lago di [monte [Acuto]]]

The proposed hierarchy of overlapping of ENEs introduces more detailed entities and allows the annotation of more fine-grain NEs and less errors of classification.

(78) le propriétaire du restaurant du lac de Neuvic.
the owner of the restaurant of Neuvic Lake.

To illustrate the advantage of using the introduced concept of ENE, we have tested four well-known English NER tools (Stanford, Open Calais, Illinois and FreeLing) with example (78). The results are shown below:

- Stanford Named Entity Recognizer[52] annotates one entity:
  - 'Neuvic Lake' as **location**
- Open Calais[53] annotates two NEs:

[52] http://nlp.stanford.edu/software/CRF-NER.shtml
[53] http://new.opencalais.com/

  - 'Neuvic Lake' as **natural feature**
  - 'restaurant of Neuvic Lake' as **facility**
- NER tool of the Cognitive Computation Group[54] (Ratinov and Roth, 2009) of the University of Illinois annotates only one NE:
  - 'Neuvic Lake' as **organization**
- FreeLing[55] annotates one NE:
  - 'Neuvic Lake' as **geographical location**

Our proposal considers ENE (78) as a whole entity whereas standard NER tools consider only the entity 'Neuvic Lake', except Open Calais that also annotates 'restaurant of Neuvic Lake' as a NE (facility). However in this case, example (78) refers to a person (the owner of the restaurant) and not to a spatial entity (the lake). Furthermore, the problem of wrong classification due to bad boundaries detection occurs also for smaller named entities such as for example (79). Among the four NER tools previously tested (i.e., Stanford, Open Calais, Illinois and FreeLing), only one (Open Calais) succeeds to detect the ENE as a person. All the other, consider only the named entity 'Paris' as a location.

(79) la maire de Paris.
the mayor of Paris.

Our approach annotates each level of the ENE. For instance, considering example (78), our NER method described in Chapter 4 produces the following results[56]:

- 'Neuvic' as **proper name** (ENE level 0)
- 'Neuvic Lake' as **geographical name** (ENE level 1)
- 'restaurant of Neuvic Lake' as **place name** (ENE level 2)
- 'owner of the restaurant of Neuvic Lake' as **non-spatial** (ENE level 3)

5.2.3.2 TEI Based Annotation

According to our proposal of a TEI compliant XML markup language, we will now describe the TEI elements and their attributes used to represent ENEs.

Name: `<name>` (proper name)
The `<name>` element provided by TEI annotates proper names or noun phrases, which are equivalent to ENEs of level 0. Table 5.4 shows the current set of `<name>` attributes defined in our customized specification.

Attribute name	Description
`type`	category of NE (location, person, organization, etc.)
`subType`	semantic sub-categorization

Table 5.4: Attributes for `<name>` tag

In our specification, all the attributes of the `<name>` element are optional and the `type` attribute may refer to the category of the NEs (location, person, organisation, etc.) such as the ENAMEX types proposed in the MUC-6 typology or such as those defined by Sekine et al. (2002); Tran and Maurel (2006) and Ehrmann (2008). The `subType` attribute for `<name>` indicates a second level in the classification such as *roadName* if the `type` attribute value is equal to 'location'. Figure 5.6 illustrates the annotation of pure proper names (i.e., ENEs of level 0) using the TEI `<name>` element.

In our annotation scheme, we also take advantage of tags provided by TEI for dates and time. These tags are described in the *Core* module of the TEI Guidelines and are available for all TEI documents. They refer to the expressions of date, time or duration in texts. Figure 5.7 shows some examples of annotation of date and time entities.

[54] http://cogcomp.cs.illinois.edu/page/demo_view/NER
[55] http://nlp.lsi.upc.edu/freeling/
[56] the example has been translated into English for the need of the comparison with the other NER tools.

```
<name xml:id="n1">
  <w type="NPr">Vanoise</w>
</name>
---
<name xml:id="n2" type="location" subtype="roadName">
  <w type="NPr" >GR55</w>
</name>
```

Figure 5.6: Example of annotation of NEs

```
<date when="2015-07">July 2015</date>
<time when="2015-07-29T20:42:00-05:00">Jul 29 2015 at 8 pm</time>
<time dur="PT20M">twenty minutes</time>
```

Figure 5.7: Example of annotation of date and time

Term: `<term>`
The `<term>` element, defined by the TEI Guidelines, contains a single-word, multi-word, or symbolic designation, which is regarded as a technical term. In the current generic layer of the language, we use the `<term>` element for several purposes such as annotating common nouns, which refer to the descriptive expansion part of ENEs (i.e., common nouns associated with a proper name). Table 5.5 shows the current set of `<term>` attributes defined in our customized specification.

Attribute name	Description/value
type	*N*, *offset*, *measure*
subType	semantic sub-categorization

Table 5.5: Attributes for `<term>` tag

Table 5.5 shows the current set of `<term>` attributes defined in our customized specification and Figure 5.8 shows some examples of annotation of `<term>` elements.

```
<term xml:id="t1" type="N">
  <w lemma="refuge" type="N">refuge</w>
</term>
---
<term xml:id="t2" type="offset" subtype="orientation">
  <w lemma="au" type="PREPDET">au</w>
  <w lemma="nord" type="ADJ">nord</w>
  <w lemma="de" type="PREP">de</w>
</term>
---
<term xml:id="t3" type="offset" subtype="direction_final">
  <w lemma="jusque" type="PREP">jusqu</w>
  <w lemma="au" type="PREPDET">au</w>
</term>
---
<term xml:id="t4" xml:lang="en" type="measure">
  <w type="NUM">200</w>
  <w lemma="meter" type="N">meters</w>
</term>
```

Figure 5.8: Example of annotation of `<term>` elements

In our proposed customization of TEI, the `type` attribute for the `<term>` element is mandatory and its value must be: *N*, *offset* or *measure*. The *N* value means that the `<term>` element contains a common noun and may refer to the descriptive expansion part of an ENE. Furthermore, in the current version of

the language, the *offset* type refers to the expression of spatial or temporal relations. For instance, in the case of an *offset* type, the value of the `subtype` attribute may be: *orientation*, *adjacency*, *inclusion*, *direction_initial* or *direction_final*, depending on the nature of the spatial relations. Finally, the *measure* type annotates distance measures.

Referencing string: `<rs>`
As we have seen in the definition of the concept of ENE, there are several levels of expansion. Each level can encapsulate the ENE of lower level. According to the TEI Guidelines, the `<rs>` element defined in the *Core* module, contains a general purpose name or referring string. In the generic layer specification of our multi-scale markup language, we use the `<rs>` element for annotating ENEs. Furthermore, in our customized specification, a `<rs>` element is either composed of a `<term>` element and another `<rs>` element; or it consists of a `<term>` element and a `<name>` element. `<rs>` elements interpret ENE in a broad manner and can encapsulate all types of `<term>` elements.

Attribute name	Value
`type`	*expandedName*, *relative*, *sequence*

Table 5.6: Attributes for `<rs>` tag

We specify the `<rs>` element as having only one attribute: `type` (Table 5.6). We define the attribute `type` as optional and its value must be equal to: *expandedName*, *relative*, or *sequence*. `<rs>` elements having an *expandedName* type refer to the expression of ENEs (see `<rs xml:id="rs1">` in Figure 5.9). They must be composed of a `<name>` element or another `<rs>` element and may contain a `<term>` element with a type value equal to *N*. Furthermore, we use the global attribute `n` to specify the level of encapsulation (e.g., 0, 1, 2, etc.).

```
<rs xml:id="rs2" type="relative">
  <term type="offset" subtype="adjacency">
    <w lemma="pres" type="ADV">pres</w>
    <w lemma="du" type="PREPDET">des</w>
  </term>
  <rs xml:id="rs1" n="1" type="expandedName">
    <term type="N">
      <w lemma="chalet" type="N">chalets</w>
    </term>
    <w lemma="de" type="PREP">de</w>
    <w lemma="le" type="DET">la</w>
    <name xml:id="n1">
      <w type="NPr">Gliere</w>
    </name>
  </rs>
</rs>
```

Figure 5.9: Example of annotation of `<rs>` element.

The *relative* type refers to the expression of a proper name (i.e., `<name>`) or of an ENE of level > 0 (i.e., `<rs type="expandedName">`), associated with a modifier, i.e. a `<term>` element having an *offset* value for the `type` attribute (see `<rs xml:id="rs2">` in Figure 5.9). Finally, the *sequence* value for the `type` attribute refers to a sequence of several `<rs>` elements (Fig. 5.10). A `<term>` element contained by a `<rs type="sequence">` is applied to all the `<rs>` elements. For instance, the term 'les bourgs' (i.e., small villages) is associated with both *Barioz* and *Bieux* entities (which are the names of small villages).

All these elements used for the text segmentation and the representation of ENE define the generic layer of the multi-scale markup language. They have been applied for the text mining and NER tasks defined in Chapter 4. Now we will describe the second layer of the multi-scale markup language, which is dedicated to the annotation of geospatial semantic textual information.

```
<rs type="sequence">
  <term xml:id="t1" type="N">
    <w lemma="le" type="DET">les</w>
    <w lemma="bourg" type="N">bourgs</w>
  </term>
  <w lemma="de" type="PREP">de</w>
  <rs xml:id="rs1" n="0" type="expandedName">
    <name>
      <w type="NPr">Barioz</w>
    </name>
  </rs>
  <w lemma="et" type="CONJC">et</w>
  <w lemma="de" type="PREP">de</w>
  <rs xml:id="rs2" n="0" type="expandedName">
    <name>
      <w type="NPr">Bieux</w>
    </name>
  </rs>
</rs>
```

Figure 5.10: Example of annotation of <rs> element.

5.3 Towards a Geospatial Semantic Markup Language

5.3.1 Overview

In this section, we describe the adaptation to transform the generic core layer towards a geospatial semantic markup language. We propose some guidelines for a TEI compliant markup language for encoding spatial information. Some elements belonging to the generic layer of our multi-scale markup language (<term> and <rs>) are turned into more specific elements embedding geospatial semantics and provided by the TEI Guidelines in the *Namesdates* module.

Geographical feature name: <geogFeat>
The content of <geogFeat> elements is defined by the TEI Guidelines as a common noun identifying some geographical feature (e.g., valley, mount, etc.) contained within a spatial NE. This is the equivalent of the definition of a descriptive expansion part of an ENE (i.e., <term type="N"> element) having a geographical denotation and associated with a spatial NE. Thus, in our customized specification, the <term type="N"> elements having a geographical sense (i.e., city, lake, river, etc.) which are used in conjunction with <name> elements are turned into <geogFeat> elements. Figure 5.11 shows two examples of <geogFeat> elements.

```
<geogFeat>
  <w lemma="lac" type="N">lac</w>
</geogFeat>
<name>
  <w type="NPr">Grattaleu</w>
</name>
---
<geogFeat>
  <w lemma="torrent" type="N">torrent</w>
</geogFeat>
<w lemma="de" type="PREP">de</w>
<w lemma="la" type="DET">la</w>
<name>
  <w type="NPr">Leisse</w>
</name>
```

Figure 5.11: Example of annotation of geographical feature names

Geographical name: `<geogName>`
According to the *Namesdates* module of the TEI Guidelines, the `<geogName>` element identifies a name associated with some geographical feature such as 'River Thames' or 'col de la Vanoise'. Thus, `<rs type="expandedName">` elements which refer to geographical names are turned into `<geogName>` elements. Figure 5.12 shows an example of annotation of a `<geogName>` element.

```
<geogName type="T" subtype="PASS">
  <geogFeat>
    <w lemma="col" type="N">col</w>
  </geogFeat>
  <w lemma="de" type="PREP">de</w>
  <w lemma="le" type="DET">la</w>
  <name>
    <w lemma="Vanoise" type="NPr">Vanoise</w>
  </name>
</geogName>
```

Figure 5.12: Example of annotation of geographical names

As we have seen in the definition of the concept of ENE and in the description of the `<rs>` element, we defined several level of encapsulation. The global `n` attribute may be also used to indicate the level of encapsulation of `<geogName>` elements (in the same way that it was done for the `<rs>` element). Figure 5.13 shows an example of encapsulation of two `<geogName>` elements.

```
<geogName type="S" subtype="RHSE" n="2">
  <geogFeat>
    <w lemma="refuge" type="N">refuge</w>
  </geogFeat>
  <w lemma="du" type="PREPDET">du</w>
  <geogName type="T" subtype="PASS" n="1">
    <geogFeat>
      <w lemma="col" type="N">Col</w>
    </geogFeat>
    <w lemma="de" type="PREP">de</w>
    <w lemma="le" type="DET">la</w>
    <name>
      <w lemma="Vanoise" type="NPr">Vanoise</w>
    </name>
  </geogName>
</geogName>
```

Figure 5.13: Example of annotation of encapsulation of `<geogName>` elements

According to our customized specification, the attributes `type` and `subtype` are optional and their values refer to the nature of the geographical feature such as lake, mountain, valley, city, etc. In the current version of the language, we propose to follow the classification introduced for the GeoNames Ontology[57]. The nine categories of feature classes of GeoNames are shown in Table 5.7. The *code* column lists the nine possible values for the `type` attribute and the *feature code* column shows some examples of feature classes, which are used for the `subtype` attribute. And according to our customized specification, these two attributes can be also used for the `<name>` element when it is not included in a `<geogName>` element.

[57] http://www.geonames.org/ontology/

Name	Code	Feature code
Administrative boundaries	A	first-order administrative division (ADM1), ...
Area	L	locality (LCTY), park (PRK), ...
Hydrographic	H	stream (STM), lake (LK), canal (CNL), ...
Hypsographic	T	mountain (MT), valley (VAL), pass (PASS), ...
Populated place	P	populated place (PPL), farm village (PPLF), ...
Road / Railroad	R	trail (TRL), street (ST), road (RD), ...
Spot	S	school (SCM), church (CH), resthouse (RHSE), ...
Undersea	U	canyon (CNYU), bank (BNKU), reef (RFU), ...
Vegetation	V	forest (FRST), cultivated area (CULT), ...

Table 5.7: GeoNames feature classes

Offset: `<offset>`
According to the TEI Guidelines, the `<offset>` element marks that part of a relative temporal or spatial expression which indicates the direction of the offset. Thus, the generic `<term type="offset">` elements referring to spatial relations are turned into `<offset>` elements in the geospatial semantic layer of our proposal.

According to our specification of the geospatial semantic layer, the `<offset>` element annotates spatial relations expressed in texts. As we have seen in the previous chapters of this dissertation, we distinguish three categories of spatial relations: topological relations (Egenhofer and Franzosa, 1991), directional relations (Frank, 1991) and distances. In this first version of the language we consider two main types of topological relations: adjacency and inclusion. Furthermore, distance relations are described with the `<measure>` element in the paragraph below.

Attribute name	**Description/value**
`type`	*orientation*, *adjacency*, *inclusion*, *direction_initial*, *direction_final*
`subtype`	sub-categorization

Table 5.8: Attributes for `<offset>` tag

Table 5.8 shows the current set of `<offset>` attributes defined by our customized specification and Figure 5.14 shows three examples of annotation of `<offset>` element.

```
<offset type="adjacency" subtype="near">
  <w lemma="pres" type="ADV">pres</w>
  <w lemma="du" type="PREPDET">des</w>
</offset>
---
<offset type="orientation" subtype="north">
  <w lemma="au" type="PREPDET">au</w>
  <w lemma="nord" type="ADJ">nord</w>
  <w lemma="de" type="PREP">de</w>
</offset>
---
<offset type="direction_final">
  <w lemma="jusque" type="PREP">jusqu</w>
  <w lemma="au" type="PREPDET">au</w>
</offset>
```

Figure 5.14: Example of annotation of `<offset>` elements

The `type` attribute is mandatory and its value shall be: *orientation*, *adjacency*, *inclusion*, *direction_initial* or *direction_final*. The value of the optional `subtype` attribute depends on the value of the `type` attribute. For the *orientation* `type`, the value of the `subtype` attribute shall be: *south*, *east*,

northwest, *above*, *behind*, etc. For the *adjacency* `type`, the value of the `subtype` attribute shall be: *next*, *near*, etc. And for the *inclusion* `type`, the value of the `subtype` attribute shall be: *in*, *inside*, etc.

Measure: `<measure>`
According to the *Core* module of the TEI Guidelines, the `<measure>` element contains a word or phrase referring to some quantity, usually comprising a number, a unit, and a commodity name. Thus, according to our specification of the geospatial semantic layer, the `<term type="measure">` elements referring to distance relations are turned into `<measure>` elements in the geospatial semantic layer of the language.

Attribute name	Description/value
`type`	*distance*
`unit`	unit identifier
`quantity`	numeric value

Table 5.9: Attributes for `<measure>` tag

Table 5.9 shows the current set of `<measure>` attributes defined in our customized specification. All attributes are optional. The `unit` attribute indicates the units used for the measurement. The value shall be expressed in the International System Units (SI)[58]. The value of the `quantity` attribute shall be a numeric value. Figure 5.15 shows an example of annotation of `<measure>` element.

```
<measure xml:lang="en" type="distance" unit="m" quantity="200">
  <w type="NUM">two</w>
  <w type="NUM">hundred</w>
  <w lemma="meter" type="N">meters</w>
</measure>
```

Figure 5.15: Example of annotation of `<measure>` element

Place name: `<placeName>`
The `<placeName>` element is defined by the TEI Guidelines in the *Namesdates* module as containing an absolute or relative place name. Thus, according to our specification of the geospatial semantic layer, the `<rs>` elements of the generic layer of our proposal referring to geographical places (i.e., containing a `<geogName>` or `<name type="place">`) are turned into `<placeName>` elements. Furthermore, we specify that `<geogName>` elements must be included into `<placeName>` elements.

Attribute name	Value
`type`	*absolute* or *relative*

Table 5.10: Attributes for `<placeName>` tag

We distinguish between two types of `<placeName>`: absolute and relative and we define the `type` attribute as optional (Table 5.10). Absolute `<placeName>` elements refer to standard spatial ENEs and relative `<placeName>` elements refer to spatial ENEs associated with spatial relations (i.e., `<offset>` and `<measure>` elements). In other words, the `<rs type="expandedName">` elements defined in the generic layer are turned into `<placeName type="absolute">` and `<rs type="relative">` elements are turned into `<placeName type="relative">`. Figure 5.16 and Figure 5.17 show an example of annotation of an absolute and a relative `<placeName>` element respectively.

[58] http://www.bipm.org/en/publications/si-brochure/

```
<placeName type="absolute">
  <geogName type="S" subtype="RHSE" n="2">
    <geogFeat>
      <w lemma="refuge" type="N">refuge</w>
    </geogFeat>
    <w lemma="du" type="PREPDET">du</w>
    <geogName type="T" subtype="PASS" n="1">
      <geogFeat>
        <w lemma="col" type="N">Col</w>
      </geogFeat>
      <w lemma="de" type="PREP">de</w>
      <w lemma="le" type="DET">la</w>
      <name>
        <w lemma="Vanoise" type="NPr">Vanoise</w>
      </name>
    </geogName>
  </geogName>
</placeName>
```

Figure 5.16: Example of annotation of an absolute <placeName>

```
<placeName type="relative">
  <measure type="distance" unit="m" quantity="200">
    <w type="NUM">200</w>
    <w lemma="mètre" type="N">mètres</w>
  </measure>
  <offset type="orientation" subtype="north">
    <w lemma="au" type="PREPDET">au</w>
    <w lemma="nord" type="ADJ">nord</w>
    <w lemma="de" type="PREP">de</w>
  </offset>
  <name type="place">
    <w type="NPr">Pau</w>
  </name>
</placeName>
```

Figure 5.17: Example of annotation of a relative <placeName>

Place: <place>
The <place> element is defined by the TEI Guidelines as a generic element containing data about a geographic location. In the current version of our specification of the geospatial semantic layer, we use this element to replace the generic <rs type="sequence"> element. Thus, we consider that the <place> element refers to the definition of a spatial area from the association of several locations (Figure 5.18). According to our specification, <place> elements can contain the <type> and <subtype> attributes described in the <geogName> element and referring to the feature class of the spatial object.

Grammatical phrase: <phr>
We customize the <phr> element described in the generic layer of our multi-scale markup language for annotating motion events and perception expressions.

Attribute name	Description
type	type of the phrase
subtype	semantic sub-categorization (e.g., *motion* or *perception*)
function	a second level of semantic sub-categorization

Table 5.11: Attributes for <phr> tag

According to our specification of the geospatial semantic layer, the *subtype* attribute (optional) in-

```
<place type="P" subtype="PPLS">
  <geogFeat>
    <w lemma="le" type="DET">les</w>
    <w lemma="bourg" type="N">bourgs</w>
  </geogFeat>
  <w lemma="de" type="PREP">de</w>
  <placeName xml:id="pn2">
    <name>
      <w type="NPr">Barioz</w>
    </name>
  </placeName>
  <w lemma="et" type="CONJC">et</w>
  <w lemma="de" type="PREP">de</w>
  <placeName xml:id="pn1">
    <name>
      <w type="NPr">Bieux</w>
    </name>
  </placeName>
</place>
```

Figure 5.18: Example of annotation of a <place> element

dicates the semantic of the phrase, its value shall be: *motion* or *perception*. The *function* attribute (optional) indicates the motion class. Furthermore, in the current version of our proposal we consider a set of six motion classes based on the classifications of verbs of motion proposed by Muller (1998): *leave*, *hit*, *reach*, *external*, *internal*, and *cross*.

Figure 5.19 shows the result of the annotation of sentence (62) using all the elements available in the geospatial semantic layer of the proposed multi-scale language.

```
<s>
  <w lemma="on" type="PRO">On</w>
  <phr type="verb" subtype="motion" function="reach">
    <w lemma="parvenir" type="V" subtype="motion_final">parvient</w>
    <w lemma="ensuite" type="ADV">ensuite</w>
    <w lemma="au" type="PREPDET">au</w>
    <placeName type="absolute">
    <geogName type="S" subtype="RHSE" n="2">
      <geogFeat>
        <w lemma="refuge" type="N">refuge</w>
      </geogFeat>
      <w lemma="du" type="PREPDET">du</w>
      <geogName type="T" subtype="PASS" n="1">
        <geogFeat>
          <w lemma="col" type="N">Col</w>
        </geogFeat>
        <w lemma="de" type="PREP">de</w>
        <w lemma="le" type="DET">la</w>
        <name>
          <w lemma="Vanoise" type="NPr">Vanoise</w>
        </name>
      </geogName>
    </geogName>
  </placeName>
  </phr>
  <pc force="strong">.</pc>
</s>
```

Figure 5.19: Example of annotation of a <phr> element

5.3.2 Encoding Geometric Properties of Spatial Features

We will now describe the elements and their attributes used for encoding geometric properties of spatial features

Location: `<location>` and `<geo>`
The TEI Guidelines describe also some elements for encoding geometric properties of spatial features. According to the *Namesdates* module, the `<location>` element defines the location of a place as a set of geographical coordinates and the `<geo>` element contains any expression of a set of geographic coordinates. However, geographic coordinates such as latitude and longitude values are not often available directly in the textual description and must be retrieved from external geographic resources.

```
<placeName type="absolute">
  <name>
    <w type="NPr">Pau</w>
    <location>
      <country key="FR" />
      <bloc type="continent" key="EU" />
      <geo>43.301667 -0.368611</geo>
    </location>
  </name>
</placeName>
```

Figure 5.20: Example of annotation of the `<location>` element

Figure 5.20 shows an example of annotation using the `<location>` and `<geo>` elements. The `<bloc>` and `<country>` elements (optional) indicate the continent and the country respectively to which the location belongs.

As we have defined the concept of ENE as an encapsulation of several levels of expansion (see Section 5.2.3), according to our specification the `<location>` and `<geo>` elements can be nested in various elements depending on the ENE to which it refers. For instance, in Figure 5.20 the `<location>` element is nested in the `<name>` element and refers to the location of the spatial NE 'Pau', whereas in Figure 5.21 the `<location>` element is nested in `<placeName>` element which refers to the ENE 'sud de Pau'.

Furthermore, a location may be specified by using a non-TEI XML vocabulary such as GML and KML. Then, we also propose a mapping via gazetteer unique identifiers, i.e. the use of RDF identifiers to interlink with resources of the Web of Linked Data such as Geonames or DBpedia which define RDF data models. Figure 5.21 shows an example of annotation using the GML standard for encoding the spatial boundaries of the ENE 'sud de Pau'.

5.3.3 Indication of Uncertainty

Certainty: `<certainty>`
The `<certainty>` element indicates the degree of certainty associated with some aspects of the text markup. This element is described in the *Certainty* module of the TEI Guidelines.

Attribute name	Description/value
`target`	URI data pointer
`locus`	*name*, *start*, *end*, *location*, *value*
`assertedValue`	alternative value
`degree`	degree of confidence

Table 5.12: Attributes for `<certainty>` tag

Table 5.12 shows the current set of `<certainty>` attributes defined in our specification. The `target` attribute indicates the element to which the certainty is applied using the URI syntax. The `target`

```
<placeName xml:id="pn1" type="relative">
  <offset type="orientation" subtype="south">
    <w lemma="sud" type="ADJ">sud</w>
    <w lemma="de" type="PREP">de</w>
  </offset>
  <name type="place">
    <w type="NPr">Pau</w>
  </name>
  <location>
    <geo>
      <gml:Polygon>
        <gml:outerBoundaryIs>
          <gml:LinearRing>
            <gml:coordinates>
              -0.389339593262433, 43.2345070972552
              -0.392743259810513, 43.3061317796098
              -0.294231809315081, 43.3085863498989
              -0.290950779544868, 43.2369584558224
              -0.389339593262433, 43.2345070972552
            </gml:coordinates>
          </gml:LinearRing>
        </gml:outerBoundaryIs>
      </gml:Polygon>
    </geo>
  </location>
</placeName>
```

Figure 5.21: Example of annotation of <location> with GML

attribute is optional and if it is not expressed, the certainty relies to its parent element. The locus attribute indicates the aspect concerning which certainty is being expressed. The locus attribute is mandatory and its value shall be one of the following: *name*, *start*, *location*, *value*. The *name* value indicates that the uncertainty relies on the name of the element to which the <certainty> element refers. The *start*, *end* and *location* values indicate respectively whether the start, the end or both the start and the end of the element are correctly identified. The *value* value indicates that the uncertainty concerns the content of the element. The assertedValue attribute provides an alternate value for the aspect of the considered markup. For instance, when the value of the locus attribute is equal to *name*, the value of the <assertedValue> refers to the alternate value of the name of the element in question. And finally, the degree attribute indicates the degree of confidence expressed by the <certainty> element.

With respect to the problem of NE classification, the <certainty> element can be used to indicate the degree of certainty of the type assigned to a NE. Figure 5.22 shows an example of a <certainty> element applied to a <placeName> element. The *name* value of the locus attribute and the *rs* value of the assertedValue attribute indicate that the <placeName> element may be a <rs> element.

```
<placeName xml:id="pl1">
  <certainty target="#pl1" locus="name" assertedValue="rs" degree="0.6" />
  <name>
    <w type="NPr">Paris</w>
  </name>
</placeName>
```

Figure 5.22: Example of annotation using the <certainty> element

Furthermore, the <certainty> element can be also used to indicate the certainty degree of the toponym disambiguation task. In this case, the <certainty> element must be associated with the <geo> element. Figure 5.23 shows an example in which the uncertainty relies on the geographical location of the spatial entity.

```
<placeName xml:id="pn1" type="absolute">
  <name>
    <w type="NPr">Pau</w>
    <location>
      <geo xml:id="geo1">43.301667  -0.368611</geo>
    </location>
  </name>
</placeName>
<certainty target="#geo1" locus="value" degree="0.8" />
```

Figure 5.23: Example of annotation using the <certainty> element

5.4 Summary

This chapter has proposed a multi-scale markup language for encoding textual information, particularly adapted for the annotation of NEs. The main idea is to provide a general framework for people to create their own specific markup language based on a core generic layer for the NER task. Our proposal relies on the TEI standard, which is widely used in digital humanities and linguistics for the encoding and interchange of textual documents in digital form. Thus, the objective is to define several specific languages, each one adapted to a specific need and all based on the same generic core layer. The proposed generic core layer may be used to create and share pre-processed corpus.

Table 5.13 shows a summary of the different elements defined by the generic and the geospatial layers of our multi-scale markup language.

Textual elements	**Tagset of the Generic layer**	**Tagset of the Geospatial layer**
word	<w>	
sentence	<s>	
punctuation	<pc>	
spatial and temporal relations measure expressions expansion of ENEs	<term>	<offset> <measure> <geogFeat>
NE *(level 0)*	<name>	
ENE *(level > 0)*	<rs>	<geogName> <placeName> <place>
VT structure	<phr>	

Table 5.13: Summary of the tagset defined for the Generic and the Geospatial layer of our multi-scale markup language

Unlike a lot of works dealing with NER that are usually considering only pure proper names or very few entities that we defined as ENEs of level 1 such as Eiffel Tower and River Thames, we consider in our proposal both categories of proper names (i.e., pure and descriptive).

Figure 5.24 shows the result of the annotation of sentence (62) using all the elements and attributes defined in our customized specification of a geospatial semantic language.

We provide specification files for the proposed PERDIDO customization of TEI. This specification defines the constraints of the geospatial layer of our multi-scale markup language. As we have seen in the state of art (Section 2.4.4), the TEI consortium has developed the Roma[59] web-interface for the creation of TEI customizations. The Roma interface helps in the creation of a formal specification ODD (One Document Does it all) file which can be used to generate appropriate schemas (DTD, RelaxNG, XML Schema). The ODD specification file adds a series of elements, which are used to specify a new schema, and modifications to the TEI element structure. For instance, we can customize the list of elements, their

[59] http://www.tei-c.org/Roma/

```
<s>
  <w lemma="on" type="PRO">On</w>
  <phr xml:id="phr1" type="verb" subtype="motion" function="reach">
    <w lemma="parvenir" type="V" subtype="motion_final">parvient</w>
    <w lemma="ensuite" type="ADV">ensuite</w>
    <w lemma="au" type="PREPDET">au</w>
    <placeName xml:id="pn1" type="absolute">
      <certainty target="#pn1" locus="name" assertedValue="rs" degree="1.0"/>
      <location>
        <geo xml:id="geo1">51.969604 -2.893146</geo>
        <country key="FR" />
        <bloc type="continent" key="EU" />
        <certainty target="#geo1" locus="value" degree="1.0"/>
      </location>
      <geogName type="S" subtype="RHSE" n="2">
        <geogFeat>
          <w lemma="refuge" type="N">refuge</w>
        </geogFeat>
        <w lemma="du" type="PREPDET">du</w>
        <geogName type="T" subtype="PASS" n="1">
          <geogFeat>
            <w lemma="col" type="N">Col</w>
          </geogFeat>
          <w lemma="de" type="PREP">de</w>
          <w lemma="le" type="DET">la</w>
          <name>
            <w lemma="Vanoise" type="NPr">Vanoise</w>
          </name>
        </geogName>
      </geogName>
    </placeName>
  </phr>
  <pc force="strong">.</pc>
</s>
```

Figure 5.24: Example of annotation

name, the list of their attributes and their value. The ODD file[60], the XML Schema[61] and the DTD[62] of the PERDIDO customization are available online.

Furthermore, these specification files help for the creation of annotated corpus. They can be used to validate that annotations are correct according to the PERDIDO specifications.

We applied our proposal of a multi-scale markup language for the annotation of geospatial information from textual descriptions of itineraries (see Chapter 4). Then, we use this encoding of geosptial information in order to automatically reconstruct the described itinerary. The automatic reconstruction of itineraries described in Chapter 3 is based on a calculation (multi-criteria approach) using all the elements annotated from the textual description. Furthermore, at this stage of the development, the information concerning the reconstructed itinerary is not introduced in the annotation. However, the objective of the multi-scale markup language is also to allow users to defined new layers according to their own needs. For instance, users can define a third layer derived from the second layer of the multi-scale markup language using non-TEI elements and also non-consuming tags which may interlink already tagged elements such as NEs or ENEs with verbs or any other kind of element (e.g., spatial or temporal relations, etc.). For example, we are planning to specify a third layer of the multi-scale markup language, introducing more semantic and dedicated to the representation of itineraries or motion using some ISO-Space tags such as the spatial link elements (<QSLINK>,<OLINK>,<MOVELINK>,<MLINK>).

[60] ODD: http://erig.univ-pau.fr/PERDIDO/ns/Perdido_odd.xml
[61] XML Schema (XSD): http://erig.univ-pau.fr/PERDIDO/ns/Perdido.xsd
[62] DTD: http://erig.univ-pau.fr/PERDIDO/ns/Perdido.dtd

Chapter 6

Integration of the Processing Chain on a Web-Based Architecture

Most good programmers do programming not because they expect to get paid or get adulation by the public, but because it is fun to program.

— Linus Torvalds

Contents

6.1 Introduction

This chapter describes the design and implementation of our proposed processing chain for the automatic reconstruction of itineraries from descriptive texts. We propose a web-based application highly modular, in which each module is independent and can be replace by another one. The processing chain described in this chapter, implements the three main contributions of this dissertation, i.e. the automatic annotation of geospatial information, the toponym disambiguation and the automatic itinerary reconstruction.

The remainder of this chapter is structured as follows. Section 6.2 describes the workflow of our processing chain and the main modules referring to the three main contributions described in the previous chapters (i.e., the automatic annotation of geospatial information in texts, the toponym disambiguation and the itinerary reconstruction). Section 6.3 describes the design of web services for the automatic annotation of texts according to the methods described in Chapter 4 and based on the markup language introduced in Chapter 5. Then, Section 6.4 describes the features of a demonstration tool available online which was developed for the annotation of geospatial information in text and the reconstruction of itineraries with a map-based representation. Finally, Section 6.5 summarises and concludes this chapter.

6.2 Processing chain

A processing chain is defined as a sequence of processes which are all connected by input and output data. We developed our processing chain in a highly modular way in order to adapt more easily the chain according to new constraints. For instance, the chain is easily adaptable for the treatment of a new corpus written in a new language. Indeed, we started with a linguistic process designed for French and then we have duplicated and adapted the components for the Spanish and Italian languages. Another advantage is that we can experiment different solutions for the same tasks in order to choose the more accurate.[63] Furthermore, depending on the needs of the user, we can stop the process after the automatic annotation, or on the contrary we can start the process of the automatic itineraries reconstruction with an already annotated corpus.

The proposed components implement the methods proposed in previous chapters. In this dissertation, we addressed the problem of automatically annotating information in texts that describe displacements making up an itinerary. With respect to the problem of NE (or ENE) recognition, the first components (block (a) on Figure 6.1) implement a linguistic sub-processing chain for the automatic annotation of geospatial information. This first main block of the workflow contains a pre-processing step described in Section 6.2.1 and the core of the proposed method for the automatic annotation is based on a cascade of transducers (Section 6.2.2). Then, the second main block of the chain (block (b)) implements the proposed method for the toponym resolution task described in Chapter 4, i.e. the classification of annotated ENE (into spatial or non-spatial entities) combined with a method for toponym disambiguation. This component is described in more detail in Section 6.2.3. Finally, the last main component of our chain is described in Section 6.2.4 and addresses the problem of the automatic itinerary reconstruction described in Chapter 3.

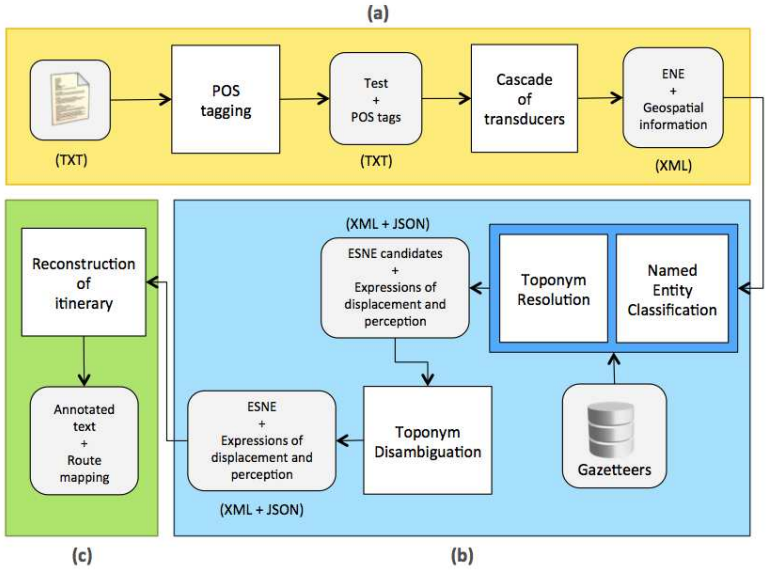


Figure 6.1: Block diagram of our processing chain

To implement the methods and algorithms proposed in the previous chapters, we develop a software toolkit called PERDIDO. The core of the PERDIDO system is designed as a component based architecture highly modular for more flexibility, robustness and scalability. The PERDIDO system was implemented in Java and is designed to handle the workflow and make connections between all the components shown in Figure 6.1. The system architecture of PERDIDO is shown in Figure 6.2. This layered architecture can be divided into three layers. The *Data Layer* provides access to external data such as raw or annotated texts and gazetteers. The *Business Logic Layer* contains the integration framework and the components. The

[63] see Section 7.5.1 describing the evaluation and the comparison of different POS analysers

integration framework provides an abstraction level of the representation of data and linguistic markup. It makes connections between the components mixing information contained in annotated XML files with information found in gazetteers. The components implement the core functionalities of the system from the pre-processing of text documents to the itinerary reconstruction. Finally, the *Application Layer* provides web services and tools for performance evaluation.

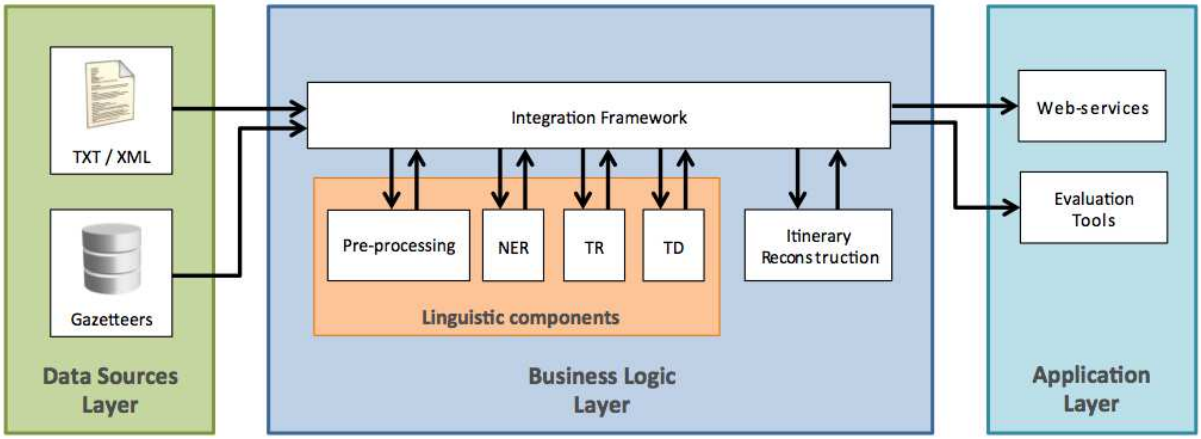


Figure 6.2: Layered architecture of the PERDIDO system

6.2.1 Pre-processing

The pre-processing component of the PERDIDO system transforms and pre-annotates raw texts with different process: sentence splitting, tokenisation, lemmatisation and POS tagging. These shallow linguistic tasks are usually done by standards POS taggers. Furthermore, this pre-processing step is language dependent. Thus, the integration framework (see Figure 6.2) is designed to handle the output provided by different POS taggers. However, each POS tagger uses different tagsets to assign grammatical categories of words (see Appendix B). Then, the integration framework implements a generic transformation to standardize tagsets in order to be use by the next components of the processing chain.

For the need of our experiments (see Chapter 7) we integrate three different POS taggers in the pre-processing component of our chain. We selected TreeTagger and FreeLing which are well-known POS analysers available for French, Spanish and Italian and Talismane which is by default only available for French.

Treetagger

TreeTagger[64] is a probabilistic POS tagger using decision trees (Schmid, 1994). It provides POS and lemma information and is available for several languages (German, English, French, Spanish, Italian, etc) and adaptable to others. Figure 6.3 shows an excerpt of the POS output.

```
Poursuivre      VER:infi        poursuivre
par             PRP             par
le              DET:ART         le
pont            NOM             pont
de              PRP             de
la              DET:ART         le
Glière          NAM             <unknown>
.               SENT            .
```

Figure 6.3: Excerpt of TreeTagger POS output

[64] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

FreeLing
FreeLing[65] is an open source language analysis tool suite (Padró and Stanilovsky, 2012), it provides a POS tagger and lemmatizer for several languages including French, Spanish and Italian. Figure 6.4 shows an excerpt of the POS output.

```
Comenzamos comenzar VMIP1P0 0.641985
la el DA0FS0 0.972269
ruta ruta NCFS000 1
en en SPS00 1
la el DA0FS0 0.972269
base base NCCS000 0.981707
de de SPS00 1
el el DA0MS0 1
puente puente NCMS000 1
de de SPS00 1
el el DA0MS0 1
Tercer_Milenio tercer_milenio NP00000 1
. . Fp 1
```

Figure 6.4: Excerpt of FreeLing POS output

Talismane
Talismane[66] (Tool for the Analysis of Language Inferring Statistical Models from the Annotation of Numerous Examples) is a syntax analyser (Urieli, 2013) which consists of four main modules which transform a raw text into a series of syntax dependency trees: phrase boundary detection, tokenising, pos-tagging and parsing (generation and labeling of syntax dependencies between words). The task of each module is resolved statistically by training a probabilistic model on annotated corpus. By default, Talismane is only available for French and was trained on the French Treebank (Abeillé et al., 2003). Figure 6.5 shows an excerpt of the POS output.

```
0       Poursuivre      poursuivre      VINF    v       _
1       par     par     P       P       _
2       le      le      DET     DET     g=m|n=s
3       pont    pont    NC      nc      g=m|n=s
4       de              de              P       P       _
5       la      la      DET     DET     g=f|n=s
6       Glière  _       NPP     _       _
7       .       .       PONCT   PONCT   _
```

Figure 6.5: Excerpt of Talismane POS output

According to figures 6.3, 6.4 and 6.5, we can notice that each POS tagger uses different tagsets to assign grammatical categories of words. POS tags differ from one POS tagger to another but also from one language to another. The different POS tags used by TreeTagger (for French, Spannish and Italian), FreeLing and Talismane are shown in Appendix B. Then, the integration framework implements a generic transformation to standardize tagsets in order to be use by the next components of the processing chain. A parameter file specifies which POS tagger is assign for the analysis of documents depending on the language. The integration framework calls the corresponding POS tagger with as input the raw text provided by the *Data Sources Layer*. Then, the POS output is mapped with the PERDIDO POS tagset shown in Table 6.1. The PERDIDO tagset is a simplified version of the tagsets used for the different POS taggers shown in Appendix B.

[65] http://nlp.lsi.upc.edu/freeling/
[66] http://redac.univ-tlse2.fr/applications/talismane/talismane_en.html

Tag	Description	Tag	Description
A	adjective	PREP	preposition
ABR	abbreviation	PREPDET	preposition + determinant
ADV	adverb	PUN	punctuation
CONJC	conjunction	PRO	pronoun
DET	determinant	PRO+POS	possessive pronoun
N	noun	PRO+REL	relative pronoun
NPr	proper name	SYM	symbol
NUM	numerical	V	verb

Table 6.1: POS tags used by PERDIDO

6.2.2 Automatic Annotation of Named Entities and Geospatial Information

The second main component of our processing chain, called NER in Figure 6.2, deals with the automatic annotation of NEs and geospatial information. This component implements the solution described in Chapter 4, which addresses the problem of automatically annotating passages in the text, that describe the various trips making up the itinerary.

We have proposed a hybrid solution based on a cascade of finite-state transducers combined with external resources for the named entity classification. The proposed cascade of finite-state transducers which annotates spatial information and ENEs was developed using the CasSys program (Friburger and Maurel, 2004) available in the Unitex platform[67] (Paumier, 2003) (Figure 6.6).

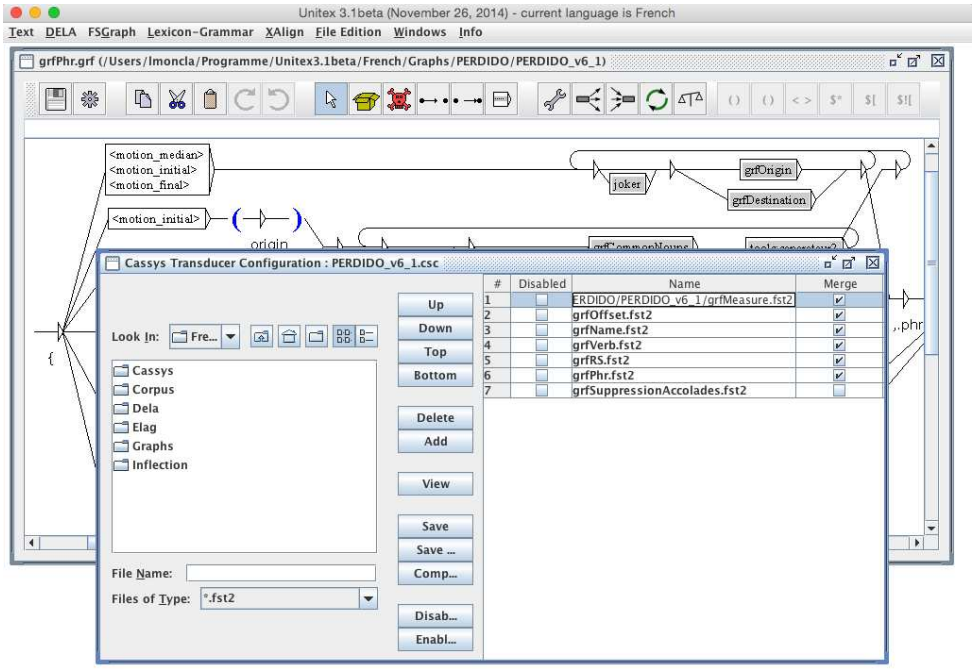


Figure 6.6: CasSys program in the Unitex platform

The cascade uses as input the results provided by the pre-processing component. However, the output format of the POS taggers is not compatible with Unitex and according to figures 6.3, 6.4 and 6.5, we can notice that the output format of the three POS taggers differs. Then, the integration framework implements a transformation process in order to provide a standardized output format for POS processed

[67] http://www-igm.univ-mlv.fr/~unitex/

texts. Then, the output of each POS tagger is turned into the same format compatible with Unitex. Furthermore, Unitex accepts two types of input: raw texts or already tagged texts. To process raw texts, Unitex applies pre-processing graphs and dictionaries. However as we have seen in Chapter 4 our method relies only on POS processed texts and does not use additional dictionaries. A tagged text compatible with Unitex is a text containing words with lexical tags enclosed in braces, defined as follows:

{word,lemma**.**POS**}**

Figure 6.7 shows the format of POS processed input of our cascade using sentence (80).

(80) Marcher 10 km jusqu'au refuge des Barmettes.
Walk for 10 km until the refuge of Barmettes.

```
{Marcher,marcher.V} {10,.NUM} {km,kilomètre.ABR} {jusqu',jusque.PREP} {au,au.PREPDET}
    {refuge,refuge.N} {des,du.PREPDET} {Barmettes,.NPr}
```

Figure 6.7: Excerpt of PERDIDO POS processed output

For the development of the PERDIDO system, we have followed the principles introduced for the development of the CasEN system (Maurel et al., 2011), which implements a combination of two cascades of transducers. The first one called *analysis cascade* is the core of the annotation process, it executes a sequence of transducers which annotate elements in a specific order. Elements annotated by previous transducers of the cascade may be involved in the annotation of bigger elements by the following transducers. The second cascade called *synthesis cascade* transforms the output of the first cascade (XML-CasSys) into the desired format. In our case the analysis cascade executes the main transducers described in Section 4.2.3 and provides a specific XML output defined by the CasSys program (XML-CasSys). This first cascade is language dependent and is adapted to each language covered by the PERDIDO system (French, Spanish and Italian). However, the differences between the three versions of the analysis cascade are limited to the translation of lexicons (contained in sub-graphs) and rules and patterns described in the main transducers remain the same. Then, the synthesis cascade transforms the XML-CasSys output into the TEI-compliant XML markup language described in Chapter 5. In contrast to the analysis cascade, the synthesis cascade is language independent and is used for the three languages covered by the PERDIDO system.

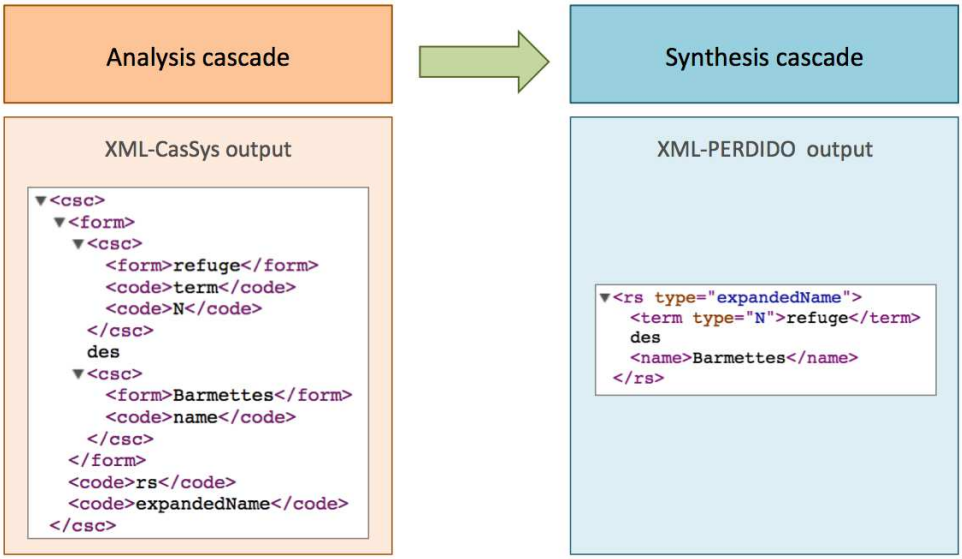


Figure 6.8: Illustration of the output of the cascades of analysis and synthesis

Figure 6.8 illustrates the two cascades with their corresponding output XML. The XML-CasSys output format consists of four elements: `<csc>`, `<form>`, `<lem>` and `<code>`. The `<csc>` element refers to an annotation introduced by the cascade. The `<form>` element refers to the content value concerned by the annotation. The `<lem>` element specifies the canonical form of words and the `<code>` elements indicate information added by the transducers.

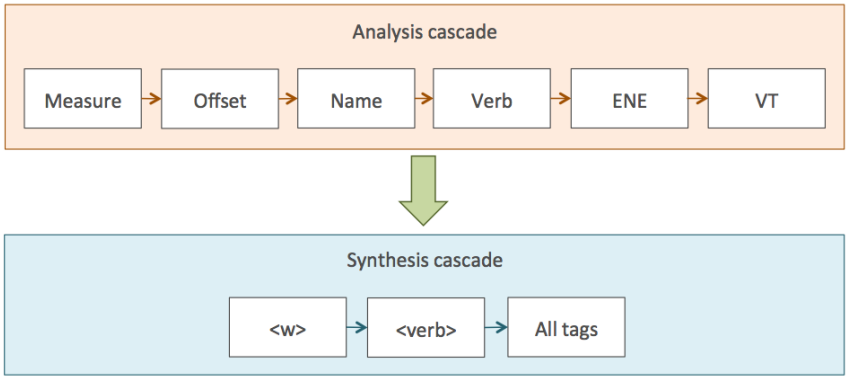


Figure 6.9: Main transducers of the two cascades

Figure 6.9 shows the main transducers of the analysis and synthesis cascades. Furthermore, we distinguish two categories of transducers: the main transducers (shown on Figure 6.9), which annotate elements by adding information (tags) to the textual content; and *tool graphs* (or sub-graphs), which are used by the main transducers and refer to lexicons, lists, and patterns (e.g., regular expressions). The implementation of the six main transducers of the analysis cascade shown in Figure 6.9 is described in Section 4.2.3. The first one annotates distance measure expressions and the second one annotates spatial relations. Then, the other transducers annotate NE, ENE, verbs and expressions of motion and perception. The output alphabets of the transducers are based on the description of the proposed markup language described in Chapter 5. For instance, Figure 6.10 shows a visual representation of the elements annotated by the PERDIDO cascade using sentence (80).

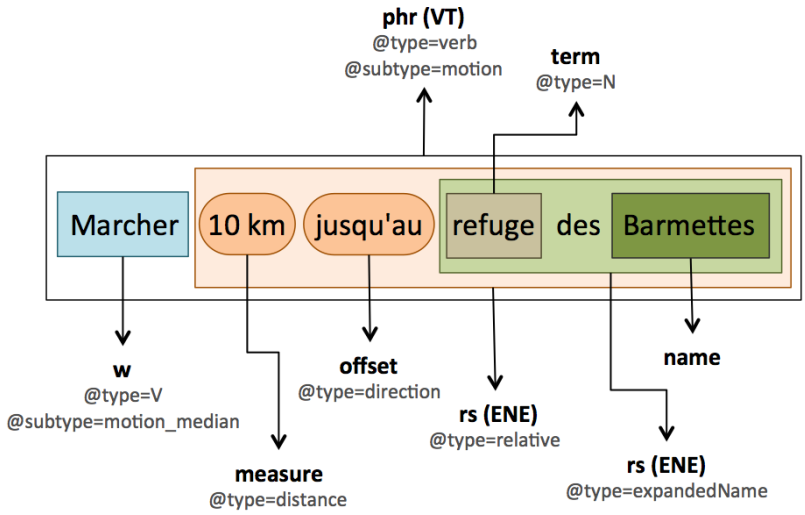


Figure 6.10: Illustration of the elements annotated by the cascade with phrase (80)

Then, Figure 6.11 shows a simplified version (i.e., without word elements) of the XML output generated by the analysis cascade (Figure 6.11a) and by the synthesis cascade (Figure 6.11a). The synthesis cascade is composed of three main transducers (Figure 6.9). Their role is to transform the XML output

```
<csc>
 <form>
  <csc>
   <form>Marcher</form>
   <lem>marcher</lem>
   <code>V</code>
   <code>motion_median</code>
  </csc>
  <csc>
   <form>
    <csc>
     <form>10 km</form>
     <code>measure</code>
     <code>distance</code>
    </csc>
    <csc>
     <form>jusqu'au</form>
     <code>offset</code>
     <code>direction</code>
    </csc>
    <csc>
     <form>
      <csc>
       <form>refuge</form>
       <code>term</code>
       <code>N</code>
      </csc>
      des
      <csc>
       <form>Barmettes</form>
       <code>name</code>
      </csc>
     </form>
     <code>rs</code>
     <code>expandedName</code>
    </csc>
   </form>
   <code>rs</code>
   <code>relative</code>
  </csc>
 </form>
 <code>phr</code>
 <code>verb</code>
 <code>motion</code>
</csc>
```

(a) Analysis cascade (CasSys-XML)

```
<phr type="verb" subtype="motion">
 <w type="V" subtype="motion_median">Marcher</w>
  <rs type="relative">
  <measure type="distance">10 km</measure>
  <offset type="direction">jusqu'au</offset>
  <rs type="expandedName">
   <term type="N">refuge</term>
   des
   <name>Barmettes</name>
  </rs>
 </rs>
</phr>
```

(b) Synthesis cascade (PERDIDO-XML)

Figure 6.11: XML outputs of the cascades of transducers

of the first cascade into the PERDIDO markup language following the guidelines presented in Chapter 5. The first transducer transforms annotation concerning word elements. For instance, Figure 6.12 shows the annotation done by the analysis cascade (Figure 6.12a) and the result provided by the synthesis cascade (Figure 6.12b).

```
<csc>
  <form>refuge</form>
  <lem>refuge</lem>
  <code>N</code>
</csc>
```

(a) Analysis cascade (CasSys-XML)

```
<w lemma="refuge" type="N">refuge</w>
```

(b) Synthesis cascade (PERDIDO-XML)

Figure 6.12: Transformation of word elements

The second transducer transforms annotation concerning classified verbs, adding a `subtype` attribute to `<w>` elements to specify the semantic role of the verb. (Figure 6.13b)

```
<csc>
  <form>Marcher</form>
  <lem>marcher</lem>
  <code>V</code>
  <code>motion_median</code>
</csc>
```

(a) Analysis cascade (CasSys-XML)

```
<w lemma="marcher" type="V"
    subtype="motion_median">Marcher</w>
```

(b) Synthesis cascade (PERDIDO-XML)

Figure 6.13: Transformation of verb elements

Finally, the third transducer of the synthesis cascade transforms all the other types of elements, such as names, ENEs, spatial relations and VT structures.

```
<csc>
  <csc>
    <form>
      <w lemma="refuge"
          type="N">refuge</w>
    </form>
    <code>term</code>
    <code>N</code>
  </csc>
  <w lemma="du"
     type="PREPDET">des</w>
  <csc>
    <form>
      <w lemma="Barmettes"
          type="NPr">Barmettes</w>
    </form>
    <code>name</code>
  </csc>
  </form>
  <code>rs</code>
  <code>expandedName</code>
</csc>
```

(a) Analysis cascade (CasSys-XML)

```
<rs type="expandedName">
   <term type="N">
     <w lemma="refuge"
         type="N">refuge</w>
   </term>
   <w lemma="du" type="PREPDET">des</w>
   <name>
     <w lemma="Barmettes"
         type="NPr">Barmettes</w>
   </name>
  </rs>
```

(b) Synthesis cascade (PERDIDO-XML)

Figure 6.14: Transformation of the other elements

6.2.3 Toponym Resolution and Disambiguation

As we described in Chapter 4, we propose to do the classification of named entities outside the cascade as a post-processing using up-to-date external resources and linked data. Then, the problems of NE classification, Toponym Resolution and Toponym Disambiguation are addressed in the TR and TD components shown in Figure 6.2. With respect to our concern, the objective of the NE classification is to determine if NEs refer to spatial entities or non-spatial entities. For that purpose, our proposal implements a gazetteer lookup method. However, as we have seen with the definition of the different types of toponym ambiguities (Section 2.3.4), the fact that a NE is found with the gazetteer lookup method is not an absolute answer to the question: "Is this entity a spatial entity?". Then, we have also proposed some toponym disambiguation methods in order to reduce the number of ambiguities. As we have seen in the state of art of this dissertation, we consider two main types of toponym ambiguities defined by Smith and Mann (2003): *structural ambiguity* and *referent ambiguity*. The structural ambiguity which is part of the reference ambiguity is the problem of inclusion or not inclusion of subtypes within the name of entities. The referent ambiguity refers to place names that represent several geographical places.

The gazetteer lookup method aims at classifying spatial and non-spatial ENE and associating geographical coordinate to spatial ones. Due to the problem of incompleteness of geographical resources, we query several gazetteers in conjunction. We have selected gazetteers provided by national mapping agencies BDNyme (France), Nomenclátor Geográfico Básico de España (Spain) and Toponimi d'Italia IGM (Italy) and we also query well-known gazetteers having a world-wide coverage: GeoNames and OpenStreetMap. Our gazetteer lookup method looks for the toponym name in all available fields of the backend database of gazetteers containing official or alternative names. The Spanish[68] and Italian[69] official national gazetteers provide access to discrete geographic features through Web Feature Services (WFS). Furthermore, the GetFeature operation returns a selection of features including geometry and attribute values. The French national gazetteer[70] is not directly accessible online and have to be downloaded in a shapefile format. We have imported this shapefile in a local PostGIS database[71] in order to query this database through the toponym resolution component of our processing chain. Then we also query the REST web services provided by GeoNames[72] and the nominatim API[73] to search GeoNames and OSM data by name, respectively.

Figure 6.15 shows the activity diagram of the gazetteer lookup and subtyping method. Once we have queried the gazetteers we apply a "proximity filter" in order to remove duplicate points introduced by the different gazetteers. Then, if there is no result and that the ENE belongs to the level 0 then the process ends and the considered ENE is classified as non-spatial, but if the ENE belongs to a higher level (i.e., level >0) then we query again with the ENE of lower level. However, if there is only one result after the "proximity filter" and if the ENE belongs to the level 0, then the process ends and the ENE is classified as spatial, otherwise if the ENE belongs to a higher level we apply our subtyping matching method (described in Section 4.3.2) if the subtype of the ENE matches a geographical concept or the type of spatial object expressed in the metadata provided by the gazetteer then the process ends and the ENE is classified as spatial. Finally, if there are more than one result after the "proximity filter", the subtyping matching is applied for ENE of level > 0.

At the end of the process each ESNE is associated with a unique ID, and each result provided by the gazetteer lookup is stored in a JSON file with metadata describing the ESNE: ID, name, country, continent, type (from the gazetteer), subtype (from the text), latitude, longitude, elevation and the source (i.e., name of the gazetteer in which the result was found).

Furthermore, after the process described in Figure 6.15, we also apply the subtyping method to the ENEs of level > 0 that have not been found in gazetteers. This is part of the classification process. For instance, the ENEs (81) and (82) are not found in gazetteers, neither with their ENEs of level 0 ('Giménez Abad' and 'Ruchère'). However, the terms *puente* (bridge) and *col* (mountain pass) are part of the multilingual ontology of geographical concepts developed for the subtyping module. Thus these

[68] http://www.ign.es/wfs-inspire/ngbe?service=WFS&request=GetCapabilities
[69] http://wms.pcn.minambiente.it/ogc?map=/ms_ogc/wfs/Toponimi_2011.map&service=wfs&request=GetCapabilities
[70] http://professionnels.ign.fr/bdnyme
[71] http://postgis.net/
[72] http://www.geonames.org/export/web-services.html
[73] http://nominatim.openstreetmap.org/search

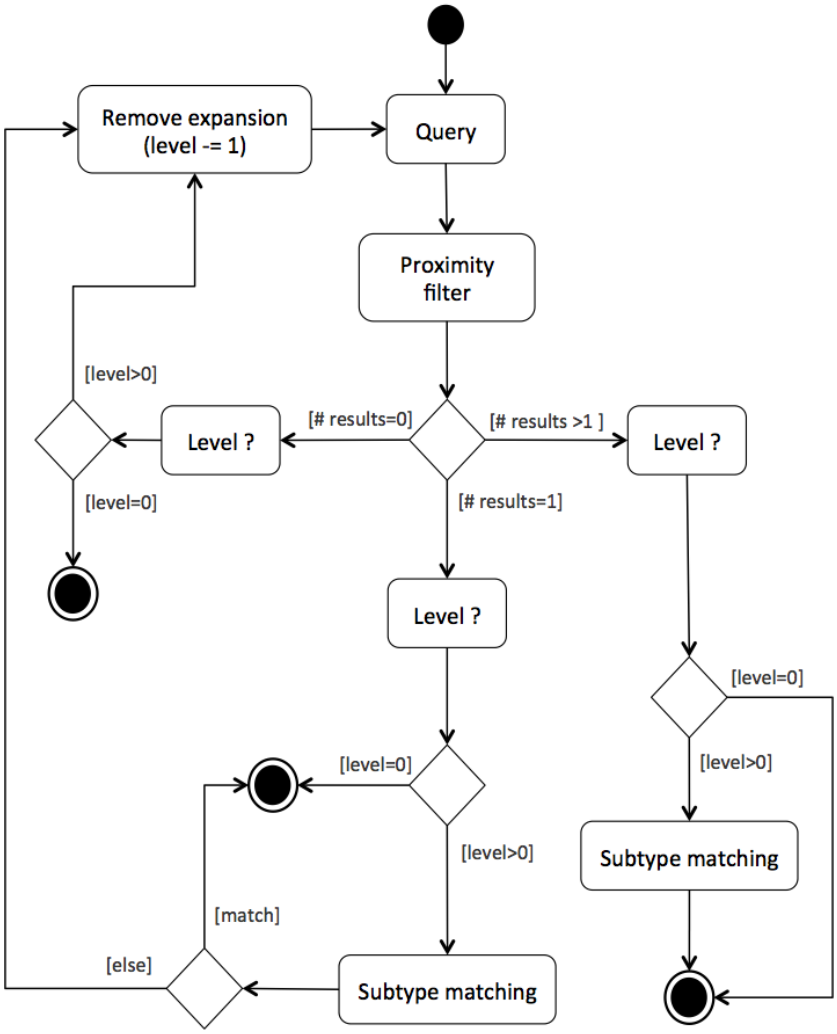


Figure 6.15: Activity diagram of the first steps of the toponym resolution

two ENE are recognized as spatial entities (i.e., ESNE) and are considered to be unreferenced toponyms.

(81) el puente Giménez Abad
Giménez Abad Bridge

(82) le col de la Ruchère
the Ruchère Pass

This means that we are able to determine that these ENEs refer to spatial entities, even if we do not have their corresponding geographical coordinates. Evaluation of experiments of the classification process is described in Section 7.5.2.

After the gazetteer lookup and the subtyping disambiguation process, we apply a density-based spatial clustering as described in Section 4.3.3. Indeed, in many cases after the first steps of the toponym resolution approach, shown in the activity diagram (Figure 6.15), it still remains several ESNEs with both the same name and type. Thus, we have proposed to use the DBSCAN clustering algorithm to determine good group of ESNEs. Evaluation results of the experiments are shown in Section 7.6.2. Furthermore, we have also proposed a method based on spatial inferences for disambiguating unreferenced toponyms (i.e., actual toponyms that are not found in gazetteers). Results of experiments are described in Section 4.3.4.

6.2.4 Itinerary Reconstruction

With respect to the problem of itinerary reconstruction, we use the results of the automatic annotation of geospatial information in texts as the input of our proposal of an automatic method for the reconstruction of itinerary described in Chapter 3. This section describes the customization of the method proposed in Chapter 3 for the adequate running of the experiments and taking into account the corpus of itineraries that has been selected. Firstly, it must be noted that the text mining method described in Chapter 4 for the generation of the geographically annotated corpus extracts different kinds of information such as motion expressions, spatial relations, perception expressions and not only spatial named entities. As mentioned in Chapter 3, we use this information as criteria to weight the edges of the graph connecting each two places in order to take decisions and find the most likely route between places. Additionally, apart from the criteria extracted from texts, there are also some criteria that are described using information coming from digital elevation datasets. The proposed method of automatic reconstruction of itineraries implemented in the *itinerary reconstruction* component (Figure 6.2) is connected to the *integration framework* and relies on a combination of geospatial information encoded in XML files (following the guidelines described in Chapter 5) and on information found in external sources such as digital elevation datasets.

We propose to detail hereafter a typical case of hiking description extracted from our French corpus of experiment that shows the strength of the proposed approach. The following phrases[74] summarize the textual description of the hike:

(83) From ***Malaucène*** to ***Col de la Chaîne*** northwest [...]

(84) where you can admire a beautiful view of the ***Dentelles de Montmirail*** [...]

(85) go south in the direction of ***Sainte-Madeleine Abbey*** [...]

(86) we head straight to the ***castle of Barroux*** [...]

(87) the view extends on the ***Dentelles de Montmirail*** [...]

(88) passing near the old ***chapel Saint Jean*** and the ***Abbey of N.-D. de l'Annonciation*** [...]

(89) return to ***Malaucène***.

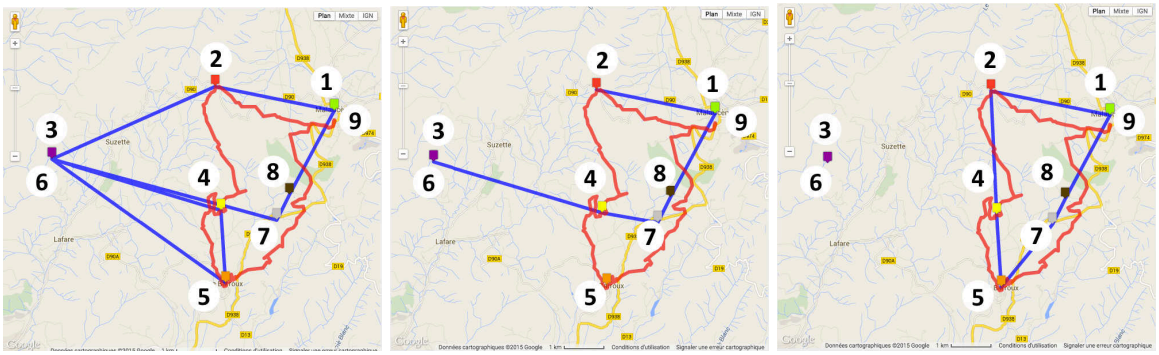
	Place name	Latitude	Longitude	Elevation
1	Malaucène	44.1741	5.1322	331
2	col de la Chaîne	44.1793	5.0966	466
3	les Dentelles de Montmirail	44.1638	5.0478	347
4	Abbaye Sainte-Madeleine	44.1529	5.0983	364
5	le château du Barroux	44.1373	5.0996	300
6	les Dentelles de Montmirail	44.1638	5.0478	347
7	chapelle Saint-Jean	44.1508	5.1150	334
8	Abbaye N.-D. de l'Annonciation	44.1562	5.1187	364
9	Malaucène	44.1741	5.1322	331

Table 6.2: Geographical coordinates and elevation of place names

This hiking trail is a loop, where the place name *Malaucène* is both the starting and the ending point. Table 6.2 shows the list of place names extracted from this hiking description. Place names are ordered as they appear in the text. They are associated with geographical coordinates (latitude, longitude) and elevation. The place names *Malaucène* and *Dentelles de Montmirail* appear twice in the description, and the place *Dentelles de Montmirail* is associated with expressions of perception (phrases (84) and (87))

Figure 6.16 shows the result of the itinerary reconstruction of this trail, using the *text distance* and the *geographical distance* criteria independently and using the proposed multi-criteria approach (Fig. 6.16c).

[74] translated into English for the sake of clarity



(a) Text distance (C_1) (b) Geographical distance (C_2) (c) Multi-criteria

Figure 6.16: Results of automatic itinerary reconstruction using different criteria

The red line represents the real GPS trajectory of the displacement, and the blue line represents the approximation of the route computed automatically. In this example, we can notice that none of these two criteria taken independently can solve the problem of itinerary reconstruction, neither *text distance* (Fig. 6.16a) nor the geographical distance between places (Fig. 6.16b). The multi-criteria approach (Fig. 6.16c) taking into account all the criteria give better results than criteria taken independently.

Edge	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	Multi-criteria
2-1	**0.14**	**0.42**	**0.37**	**0.81**	**0.5**	**0**	**0**	**0**	**0.11**
2-3	0.14	0.62	0.48	0.80	0.5	1	1	0	0.52
2-4	**0.29**	**0.43**	**0.21**	**0.04**	**0.5**	**1**	**0**	**0**	**0.19**
2-5	0.43	0.68	0.24	0.05	0.5	1	0	0	0.22
2-6	0.57	0.62	0.48	0.80	0.5	1	1	0	0.58
2-7	0.71	0.51	0.26	0.36	0.5	1	0	0	0.27
2-8	0.86	0.45	0.17	0.48	0.5	1	0	0	0.29
2-9	1.00	0.42	0.22	0.91	0.5	1	0	0	0.33

Table 6.3: Weight values for all edges connected to the vertice 2

To illustrate the multi-criteria approach used to weight the complete graph, Table 6.3 shows the value of the weights for each criterion and for all edges connected to the vertex 2 (*col de la Chaîne*). The multi-criteria column shows the weights assigned to each edge using the formula described in equation (3.3) of section 3.3.3. According to these weights the vertice 2 should be connected to vertices 1 and 4, as we may notice in Figure 6.16. Further experiments and results are described in Section 7.4.

6.3 Web Services

We have designed web service for each component of the *linguistic components* (PERDIDO API) shown in Figure 6.2. These web services accept both POST and GET requests. The proposed PERDIDO API is accessible after a registration process and give access to the following list of web services:

- POS processing: return POS processed text in the Unitex-compliant PERDIDO format
- NER: apply the PERDIDO NER process
- NERC: apply the PERDIDO geospatial annotation process
- GetToponyms: return the list of toponyms with their geo-location

6.3.1 POS Processing

The PERDIDO POS web service[75] returns the results of the POS processing described in Section 6.2.1. Table 6.4 describes the parameters and their value, and Figure 6.17 shows an example of the results provided by the PERDIDO POS web service. The three parameters (content, lang and api_key) are required. The *content* parameter refers to the input of the POS process web service and the *lang* parameter specify the language of the content in order to apply the corresponding POS analyser. Then, the *api_key* refers to the key automatically generated and attributed during the registration process.

Parameter	Description
content	textual content
lang	may be either of the following values: French, Spanish, Italian
api_key	your API key

Table 6.4: Required parameters of the PERDIDO POS web service

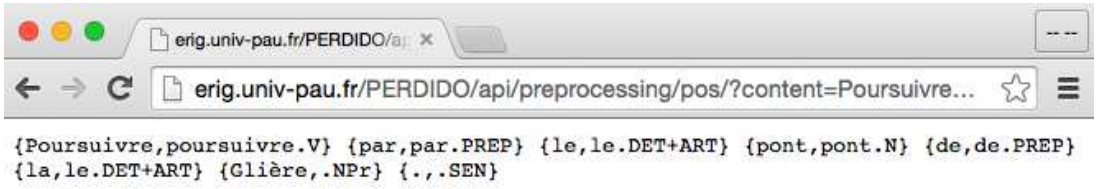


Figure 6.17: Example of result PERDIDO POS web service

6.3.2 Named Entities Recognition

The PERDIDO NER web service[76] returns the TEI-compliant XML results of the generic annotation of ENEs following the guidelines described in Chapter 5. This annotation relies on the core generic layer of our proposed multi-scale markup language for the detection of ENEs and annotates also geospatial information such as spatial relations, expressions of motion and expressions of perception. This kind of annotated text can be shared in order to be used in any further treatments. Table 6.5 describes the required parameters and their value, and Figure 6.18 shows an example of the results provided by the PERDIDO POS web service.

Parameter	Description
content	textual content
lang	may be either of the following values: French, Spanish, Italian
api_key	your API key

Table 6.5: Required parameters of the PERDIDO NER web service

[75] http://erig.univ-pau.fr/PERDIDO/api/preprocessing/pos/
[76] http://erig.univ-pau.fr/PERDIDO/api/parsingTEI/generic/

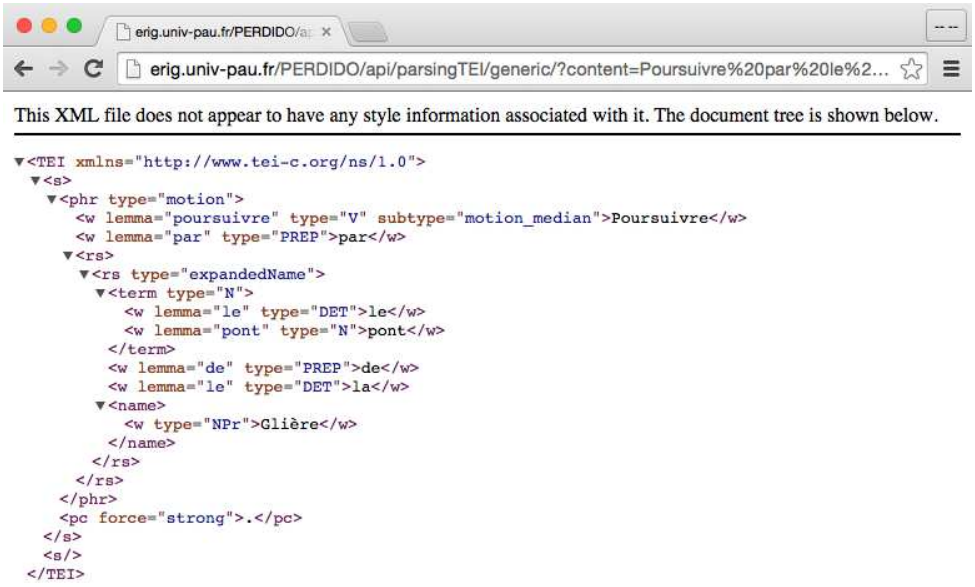


Figure 6.18: Example of result PERDIDO POS web service

6.3.3 Named Entities Classification and Toponym Resolution

The PERDIDO NERC web service[77] requires the same parameters as the previous ones (i.e., *content*, *lang* and *api_key*). It returns the XML annotations of ESNEs and spatial information and relies on the geospatial layer of the proposed multi-scale markup language described in Chapter 5. Figure 6.19 shows an example of result provided by the PERDIDO NERC web service.

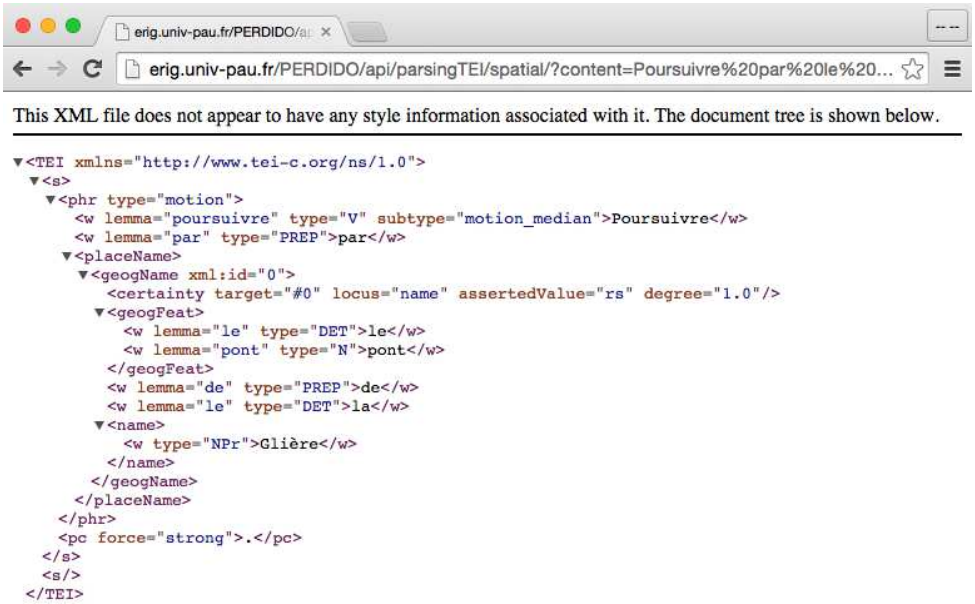


Figure 6.19: Example of result PERDIDO NERC web service

[77] http://erig.univ-pau.fr/PERDIDO/api/parsingTEI/spatial/

6.3.4 Get Toponyms

The PERDIDO GetToponyms web service provides the list of toponyms detected in the text input associated with metadata such as name, geographical coordinates, feature type, and source. This last web service also requires the same parameters as the previous ones (i.e., *content*, *lang* and *api_key*). Additionally, it supports two possible types of outputs, XML or JSON, which are specified on the URL of the service according to the following pattern:

http://erig.univ-pau.fr/PERDIDO/api/toponyms/[ouput]/

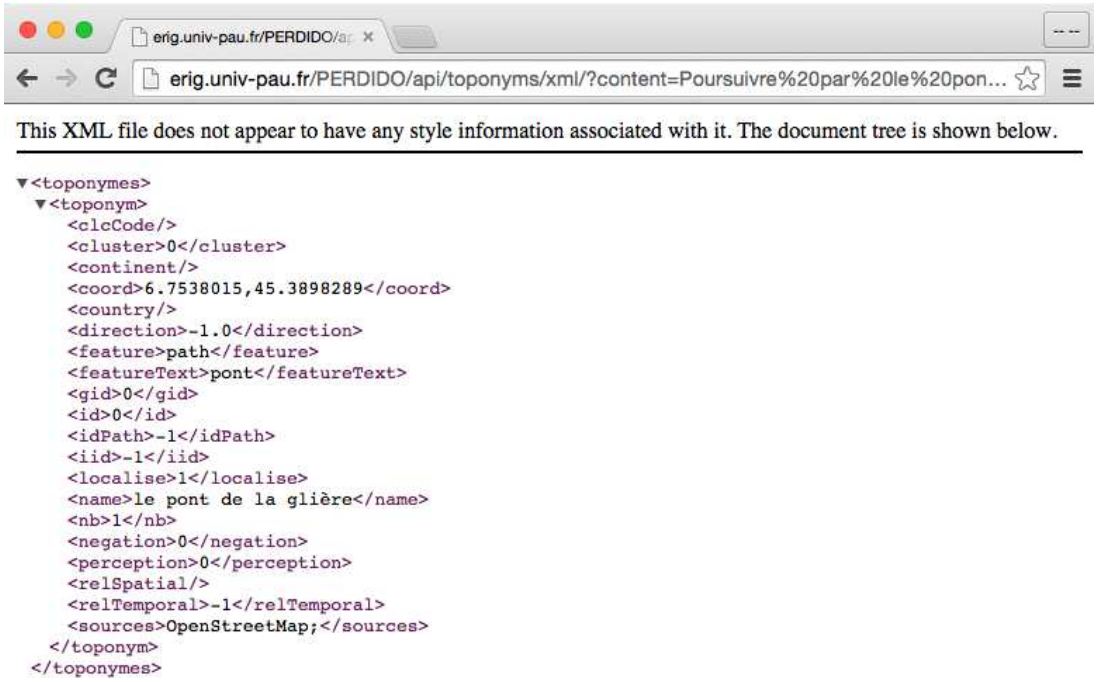


Figure 6.20: Example of result PERDIDO GetToponyms web service

6.4 Online Demonstration Tool

In addition to the web services, we have also developed an online demonstration tool[78], developed with Java Servlets technology, for the automatic reconstruction of itineraries extracted from narrative texts. Some examples are accessible to illustrate and show the result of the automatic annotation of text combined with the automatic reconstruction of itinerary described.

Figure 6.21 shows the homepage of the web-based application and Figure 6.22 shows the typical page that contains the result of the automatic process of itinerary reconstruction. The left panel contains the result of the automatic annotation of geospatial information (described in Chapter 4) with highlighted captions and the right panel shows a map containing the result of the geocoding of waypoints and the result of the automatic method of itinerary reconstruction described in Chapter 3.

This web-based application offers several functionalities. The user can directly type a text with an input form, but he can also upload a text file containing the textual content to be annotated. Furthermore, the user can also upload a GPX file corresponding to the real trajectory of the trip in order to compare it with the automatic reconstruction built with our proposed method.

[78]PERDIDO demonstration tool: http://erig.univ-pau.fr/PERDIDO/
demo video: http://erig.univ-pau.fr/PERDIDO/demo/Perdido.mp4

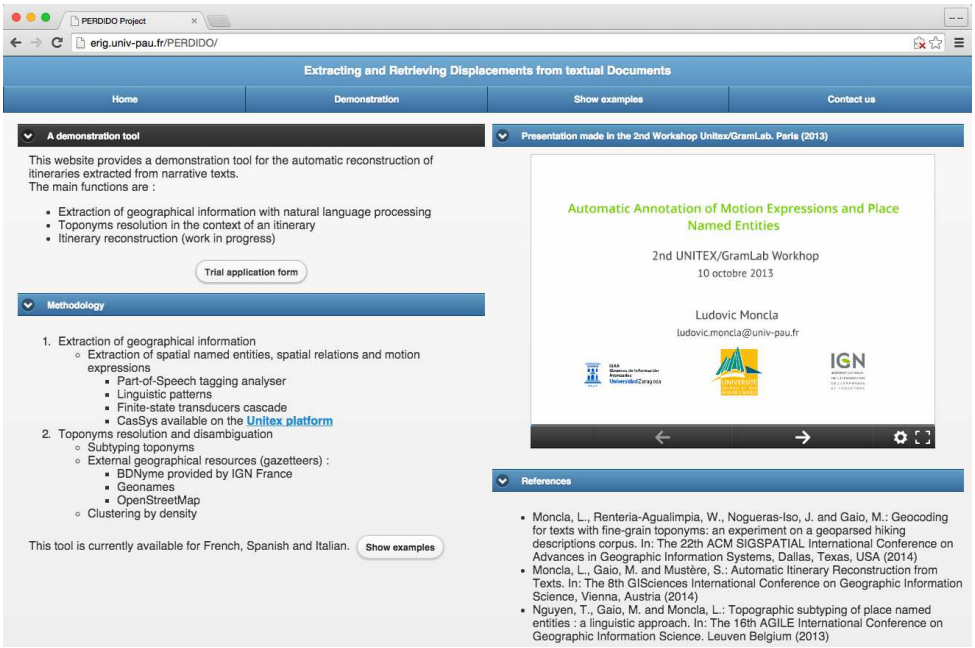


Figure 6.21: Homepage of the online demonstration tool: http://erig.univ-pau.fr/PERDIDO/

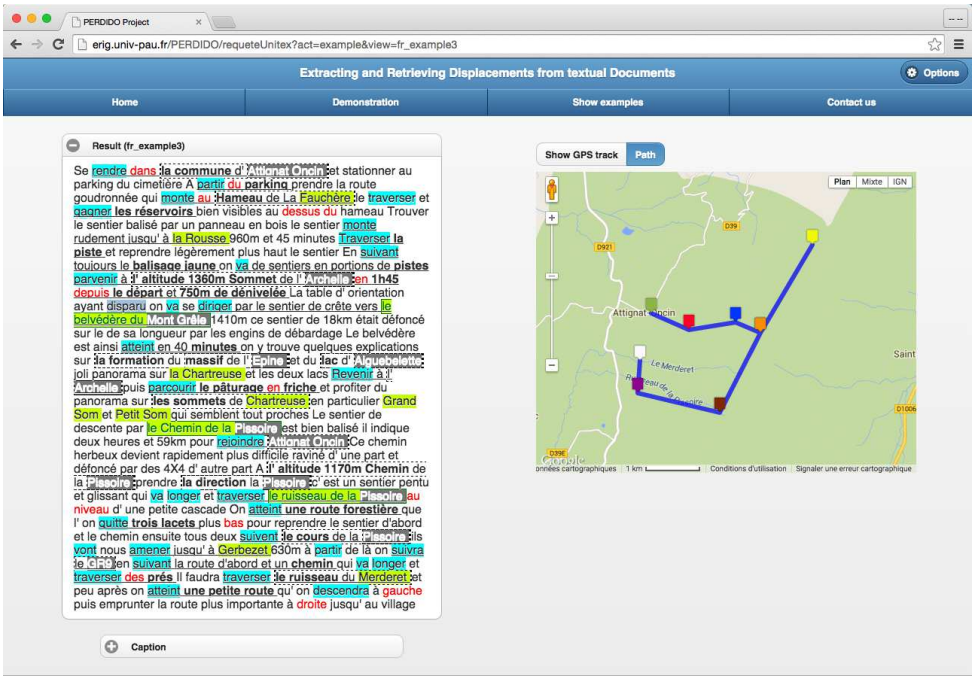


Figure 6.22: Visualization of annotations and itinerary reconstruction

We have used the Google Maps API[79] for the development of the map functionality, and we have integrated the official French maps provided by the IGN Geoportail[80] through Web Map Tile Service (WMTS). Furthermore, the user have the possibility to show or hide the representation of the reconstruction and have also the possibility to show or hide the visualization of the GPS track (if available).

[79] https://developers.google.com/maps/
[80] http://api.ign.fr/tech-docs-js/fr/developpeur/wmts.html

6.5 Summary

In this chapter we have described the design and implementation of the proposed solution for automatic reconstruction of itineraries from texts, and for the automatic annotation of geospatial information in texts. We have proposed the design of a highly modular processing chain. The first main part of our processing chain deals with linguistic process designed for French, Spanish and Italian languages and implements components described in Chapter 4. Then, the second main part of the processing chain integrates the components dedicated to the automatic reconstruction of itineraries as described in Chapter 3.

We have also described the different web services and the demonstration tool available in order to show the feasibility of our proposals and the potential of our contributions. Furthermore, we have also developed some evaluation tools, which are used to evaluate the experiments over a corpus of hiking descriptions. The results of these experiments over real data are described in the next chapter.

Chapter 7

Evaluation

No amount of experimentation can ever prove me right; a single experiment can prove me wrong.

— Albert Einstein

Contents

7.1 Introduction

This chapter describes the evaluation of the three main proposals of this dissertation. In Chapter 3 we proposed an approach for the automatic geocoding of itinerary described in natural language. This method relies on an annotated text input provided by the automatic annotation process described in Chapter 4. The described automatic annotation process addresses the problem of the automatic annotation, the resolution and the disambiguation of toponyms but also the annotation of spatial and motion relations between phrases and named entities. In this chapter, we propose the evaluation of each part of our contribution, i.e. the automatic reconstruction of itinerary, the automatic annotation and the disambiguation of toponyms.

The remainder of this chapter is structured as follows. Section 7.2 describes the dataset used for the experiments. Section 7.3 introduces the methods for evaluating the result of the experiments. Section 7.4 describes the evaluation of our proposed method of automatic geocoding of itinerary. Section 7.5 and Section 7.6 describe the experiments and the results of the Text Mining and Toponym Disambiguation tasks, respectively. Finally, Section 7.7 summarises and concludes this chapter.

7.2 Dataset

7.2.1 Overview

There are several textual corpora available for English language and oriented to the annotation of spatial information such as TR-CoNLL, SpatialML, LGL and MotionBank. TR-CoNLL (Leidner, 2007) contains nearly 1,000 news articles (REUTERS) from the named entity recognition shared task of CoNLL 2003 (Conference on Natural Language Learning). This corpus was particularly developed for the task of Toponym Resolution. About 7,000 toponym instances are annotated with latitude/longitude coordinates of nearly 14,000 distinct unique candidate referents. The SpatialML project proposed by Mani et al. (2008) provides a corpus of annotated Automatic Content Extraction (ACE) documents released by the Linguistic Data Consortium[81]. SpatialML annotates named and nominal spatial entities and their mapping to geo-coordinates. SpatialML also annotates orientation and topological relations between places. The LGL (Local-Global-Lexicon) dataset introduced by Lieberman et al. (2010) consists of articles from local news sources. These articles are intended for more geographically localized audiences, and concern local events that mention small places. MotionBank (Pustejovsky and Yocum, 2013) is a subcorpus of ISO-SpaceBank consisting of 50 entries of travel blog, which describe displacements across the Americas.

Some other corpora exist for French language such as the French newspaper *L'Est Républicain* provided by the Ortolang project[82] (Outils et Ressources pour un Traitement Optimisé de la LANGue) or the ESTER[83] (Campagne d'évaluation des Systèmes de Transcription Enrichie d'Émissions Radiophoniques) evaluation campaign providing a corpus of about 100 hours of manually transcribed French radio broadcast news and articles from the French newspaper *Le Monde*.

Furthermore, a recent dataset more closely related to our work is provided by the SpaceEval[84] task of the International Workshop on Semantic Evaluation (SemEval-2015). The SpaceEval dataset adopts the annotation specification from ISO-Space: natural languages are filled with particular constructions for talking about spatial information, including toponyms, spatial nominals, locations that are described in relation to other locations, and movements along a path.

7.2.2 A Gold-Standard Corpus of Hiking Descriptions

However, none of these datasets are both dedicated to the description of displacements and available for French, Spanish and Italian. Thus, we decided to build our own corpus in order to experiment and evaluate our processing chain. We chose hiking descriptions for building a multilingual corpus (French, Spanish and Italian) of narrative descriptions of places in an open and natural area. Hiking descriptions are a specific type of documents describing displacements using geographical information, such as toponyms, spatial and motion relations, and natural features or landscapes. Furthermore, the main advantage of this kind of text is that there are specialized websites providing textual descriptions associated with the real GPS trajectory of each trail.

We developed a web-crawler specially configured for some websites hosting hiking descriptions. Thousands of hiking descriptions were automatically extracted from specialized websites in French[85], Spanish[86],

[81] http://www.ldc.upenn.edu/
[82] http://www.ortolang.fr/
[83] http://www.afcp-parole.org/ester/
[84] http://alt.qcri.org/semeval2015/task8/
[85] http://www.visorando.com – http://www.pyrandonnees.fr/ (fr)
[86] http://senderos.turismodearagon.com (es)

and Italian[87]. Appendix A shows some examples of hiking descriptions taken from these websites. Each document of the corpus describes one trail and is associated with the real trajectory (GPS) of the route. Each trail is only described by one document. In our work real GPS trajectories are only used for the evaluation of the results of the automatic process of itinerary reconstruction described in Chapter 3.

To evaluate our processing chain we built a gold-standard corpus called 'PERDIDO corpus', consisting of 90 hiking descriptions manually annotated following the TEI compliant Guidelines described in Chapter 5. As the manual annotation task of documents is very time consuming, we developed a controlled manual tagging tool to help and control the manual annotation.

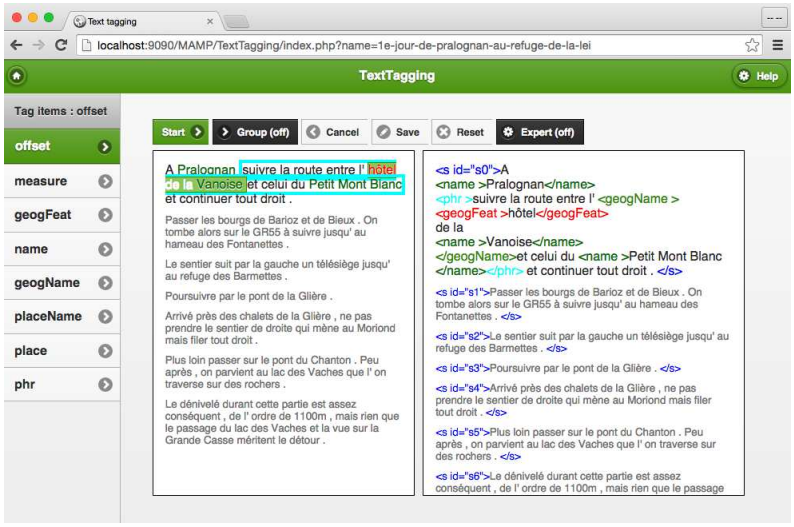


Figure 7.1: Interface of the controlled manual tagging tool

Figure 7.1 shows the interface of the current version of our controlled manual tagging tool[88]. This web-based application (still under development) automatically extracts rules and constraints from an XML Schema and allows the user to validate the XML annotation. We designed this application to be the most simple, intuitive and generic as possible. It can works with any valid XML Schema. The user provides the XML Schema file, then it just have to select one tag on the list of the available tags (left panel) and apply it to the text by clicking on words. The application shows the XML result in the right panel in real-time. Then the user can fill the values of attributes. Once the annotation is made (sentence by sentence), the user can ask for validation. According to the XML Schema the application returns the list of errors to correct or create the XML file.

	French	Spanish	Italian
# of documents	30	30	30
# of words	11297	5549	15724
# of ESNE	638	416	475
Avg. # of ESNE	21	13.9	15.8

Table 7.1: Document sets

Among all the documents available, we selected 30 hiking descriptions for each language, which are representative of the whole corpus in terms of distribution of words and NEs. In the remainder of this chapter, we consider these three selected sets as three bodies of reference (i.e., one per language) in order to evaluate differences in the results of the experiments. Indeed, the precision to identify expressions in natural languages may vary from one language to another.

[87] http://www.parks.it/parco.alpi.marittime/ (it)
[88] Demo video: http://erig.univ-pau.fr/PERDIDO/demo/TextTagging.mp4

Table 7.1 shows some features of each body of reference of the PERDIDO corpus. We can notice that those corpora have different characteristics, such as the average number of words or the average number of ESNEs per document.

An evaluation of the content of each body of reference is detailed hereafter, regarding in particular the proportion of motion and perception expressions associated with ESNEs.

French

- 97% of NEs refer to ESNE (638 occurrences).
- 12% of ESNE refer to roadNames such as 'GR55' or 'D24'.
- 12.7% refer to the name of a linear spatial object (e.g., rivers, streets, etc.).
- 7 occurrences of sequences of ESNE such as in sentences (90) and (91).
- 38% of the occurrences of ESNE are associated with a verb of motion.
- 6% of the ESNE are associated with an expression of perception (see example (91)).
- 1% of the occurrences of ESNE are associated with negation expressions.
- 54% of the occurrences of ESNE are associated with feature types (see the most frequent terms in Table 7.3).

(90) Traverser successivement les hameaux de **Tallode**, **Liac** et **Lic**.
Cross successively the hamlets of ***Tallode***, ***Liac*** *and* ***Lic****.*

(91) Profitez d'une magnifique vue sur les glaciers d'**Estelette**, de la **Lée Blanche**, du **Miage**, du **Brouillard**, de **Frêney** et du **Brenva**.
Enjoy a magnificent view over the glaciers ***Estelette***, ***Lée Blanche***, ***Miage***, ***Brouillard***, ***Frêney*** *and* ***Brenva****.*

Spanish

- 99% of NEs refer to spatial NEs (416 occurrences).
- 10% of ESNEs refer to roadNames such as 'A-1213' or 'TE-13'.
- 8% refer to the name of a linear spatial object (e.g., rivers, streets, etc.).
- 3 occurrences of sequences of ESNE such as in sentences (92).
- 52% of the occurrences of spatial NEs are associated with a verb of motion.
- 3% of the spatial NEs are associated with an expression of perception.
- None of the occurrences of ESNE are associated with negation expressions.
- 44% of the occurrences of ESNE are associated with feature types (see the most frequent terms in Table 7.3).

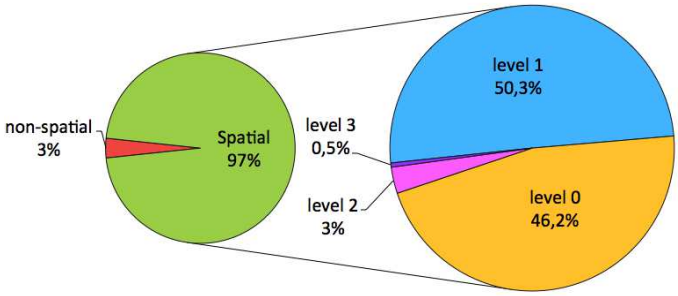
(92) Pasaremos por las estaciones ferroviarias de **Ainzón**, **Albeta**, **Bureta**, **Alberite de San Juan**.
Pass by the railway stations of **Ainzón**, **Albeta**, **Bureta**, **Alberite de San Juan**.

Italian

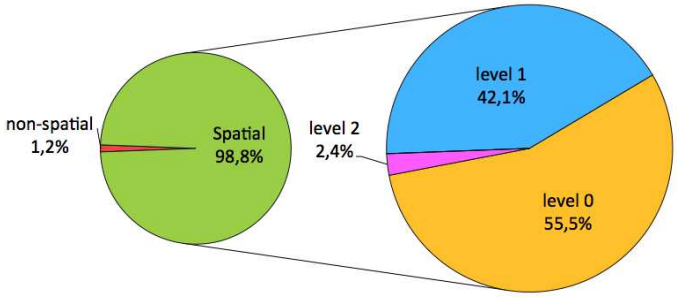
- 99% of NEs refer to spatial NEs (471 occurrences).
- 14% of spatial NEs refer to roadNames such as 'sentiero n 11'.
- 1.5% refer to the name of a linear spatial object (e.g., rivers, streets, etc.).
- 12 occurrences of sequences of ESNE such as in sentences (93).
- 50% of the occurrences of spatial NE are associated with a verb of motion.
- 11% of the spatial NE are associated with an expression of perception.
- 0.2% of the occurrences of ESNE are associated with negation expressions.
- 44% of the occurrences of ESNE are associated with feature types (see the most frequent terms in Table 7.3).

(93) Si costeggiano i masi **Cialaruns**, **Andrac**, **Alfarëi**, **Ruac** e **Fussè**.
Go along the farms of **Cialaruns**, **Andrac**, **Alfarëi**, **Ruac** and **Fussè**..

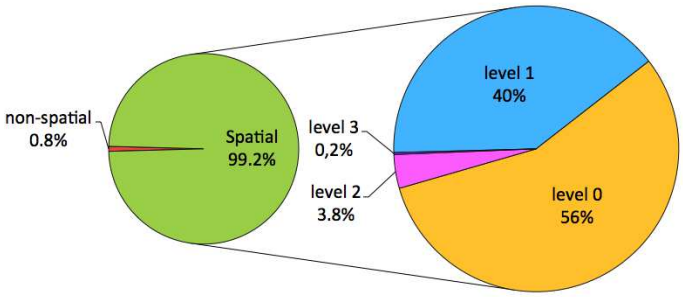
Furthermore, Table 7.2 and Figure 7.2 show the distribution of ENEs for our three corpus of experiments.



(a) French corpus



(b) Spanish corpus



(c) Italian corpus

Figure 7.2: Distribution of ENEs

	French		Spanish		Italian	
Total # of ENEs	660		421		475	
# of ESNEs	638	(97%)	416	(99%)	471	(99%)
- level 0 (% of ESNEs)	295	(46.2%)	231	(56%)	264	(56%)
- level 1 (% of ESNEs)	321	(50.3%)	175	(42%)	188	(40%)
- level 2 (% of ESNEs)	19	(3%)	10	(2%)	18	(4%)
- level 3 (% of ESNEs)	3	(0.5%)	-	-	1	(0%)

Table 7.2: Distribution of ENEs

We can notice that only few NEs (about 2%) of our gold-standard corpus are not referring to spatial entities. Furthermore, we can notice that almost 50% of spatial NEs are built with pure proper names

(*level 0*) and that the other 50% of spatial NEs are actually ESNEs built with descriptive proper names (*level >0*). More, a very few number of ESNEs (3%) are built with more than one expansion (*level >1*). See Section 5.2.3 for further details about the definition and the representation of ENEs.

Table 7.3 shows the list of the ten most frequent terms contained by ESNEs with their number of occurrence. These ten most frequent terms represent about 48% of the ESNE occurrences (*level >0*).

French		Spanish		Italian	
col	20	puente	17	rifugio	20
village	20	rio	17	monte	19
hameau	20	pueblo	12	villagio	17
route	17	iglesia	10	masi	15
sentier	15	camino	9	castello	9
chalet	13	barranco	8	lago	8
refuge	11	parque	5	passo	7
pont	11	castillo	3	foce	7
lac	8	barrio	3	chiesa	6
chapelle	8	casa	2	via	5

Table 7.3: The ten most frequent terms associated with ESNEs

	French		Spanish		Italian	
Total # of verbs	1694		867		805	
# of motion verbs (% of total # of verbs)	1101	(65%)	456	(53%)	428	(53%)
- initial (% of motion verbs)	66	(6%)	74	(16%)	34	(8%)
- median(% of motion verbs)	710	(64%)	216	(47%)	255	(60%)
- final(% of motion verbs)	325	(30%)	166	(37%)	139	(32%)
# of perception verbs (% of total # of verbs)	41	(2%)	36	(4%)	36	(4%)
# of topographic verbs (% of total # of verbs)	51	(3%)	54	(6%)	49	(6%)

Table 7.4: Distribution of Verbs

Table 7.4 shows the distribution of verbs for each corpus of reference. About 57% of verbs refer to motion verbs. And we can notice that median and final motion verbs are the most frequent ones. Furthermore, only about 3% of verbs refer to verbs of perception. However, many expressions of perception are not built with verbs but with nouns such as in examples (94) to (96).

(94) une **vue magnifique** sur le glacier de Bionassay
*a **beautiful view** of the glacier of Bionassay*

(95) vous avez de là un **panorama** sur les montagnes du Vercors
*from there you have a **panorama** of the Vercors mountains*

(96) unas **bellísimas vistas** del valle de la Huecha
*some **beautiful views** of the Huecha Valley*

Furthermore, Table 7.5 shows the list of the ten most frequent motion verbs for each corpus of reference. These most frequent motion verbs represent about 64% of the motion verbs occurrences.

French		Spanish		Italian	
prendre[a]	188	llegar	64	proseguire	44
suivre	100	recorrer	34	seguire	41
traverser	78	seguir	34	raggiungere	29
arriver	71	pasar	31	arrivare	24
continuer[a]	64	tomar	28	attraversare	22
descendre	61	continuar[a]	27	salire	22
passer	60	visitar	21	continuare[a]	20
monter	51	salir	20	scendere	18
rejoindre	44	dirigir	19	portare[a]	18
partir	35	ir	17	percorrere	15

[a] verbs expressing motion when associated with geographical feature such as 'prendre le chemin'

Table 7.5: The ten most frequent motion verbs

7.3 Evaluation Methodology

This section describes the metrics used for evaluating the results of the experiments and the types of errors introduced by the different steps of our processing chain.

7.3.1 Evaluation Metrics

The metrics *Precision* (P), *Recall* (R) and *F1-Measure* (F_1) are widely used in IR and NLP to evaluate the quality of methods for automatic retrieval and annotation. We use these metrics to assess the results of different steps of our processing chain, i.e. POS tagging, NER, NE Classification and Toponym Disambiguation.

Precision
The precision P is the ratio of the number of correct positive results divided by the number of all positive results. For instance, in the context of NER, the precision P is the ratio of the number of relevant NEs annotated (NE_C) to the total number of NEs annotated (both true positive and false positive).

$$P = \frac{NE_C}{NE_C + NE_{FP}} \tag{7.1}$$

Recall
The recall R is the ratio of the number of correct positive results divided by the number of positive results that should have been returned. For instance, in the context of NER, the recall R is the ratio of the number of relevant NEs annotated (NE_C) to the number of relevant NE instances existing (both true positive and false negative).

$$R = \frac{NE_C}{NE_C + NE_{FN}} \tag{7.2}$$

F1-Measure
Precision and recall are complementary metrics and Jardine and van Rijsbergen (1971) proposed a combined measure for measuring accuracy using precision and recall. F1-Measure F_1, which is also known as F1-Score, is the harmonic mean of precision and recall (Jardine and van Rijsbergen, 1971) and is defined

as follows:

$$F_1 = \frac{2PR}{P+R} \tag{7.3}$$

Slot Error Rate

The Slot Error Rate (SER) (Makhoul et al., 1999) is the ratio of the total number of slot errors, i.e. insertions, deletions and substitutions, divided by the total number of relevant results in the reference (N). In the context of NER, metrics such as precision and recall are not enough to distinguish all types of errors such as errors of classification and errors of boundary detection. SER computes the following types of errors:

- insertions (I): refer to annotated entities that are not actually NEs (false positive);
- deletions (D): refer to NEs that have not been detected (false negative);
- substitutions: refer to different types of errors, errors of classification (C), errors of position of the entity boundaries (B), and both (CB).

Insertions and deletions errors are weighted 1 whereas classification and boundary errors, 0.5. Then, the SER is defined as follows:

$$SER = \frac{I + D + 0,5C + 0,5B + CB}{N} \tag{7.4}$$

7.3.2 Error Propagation

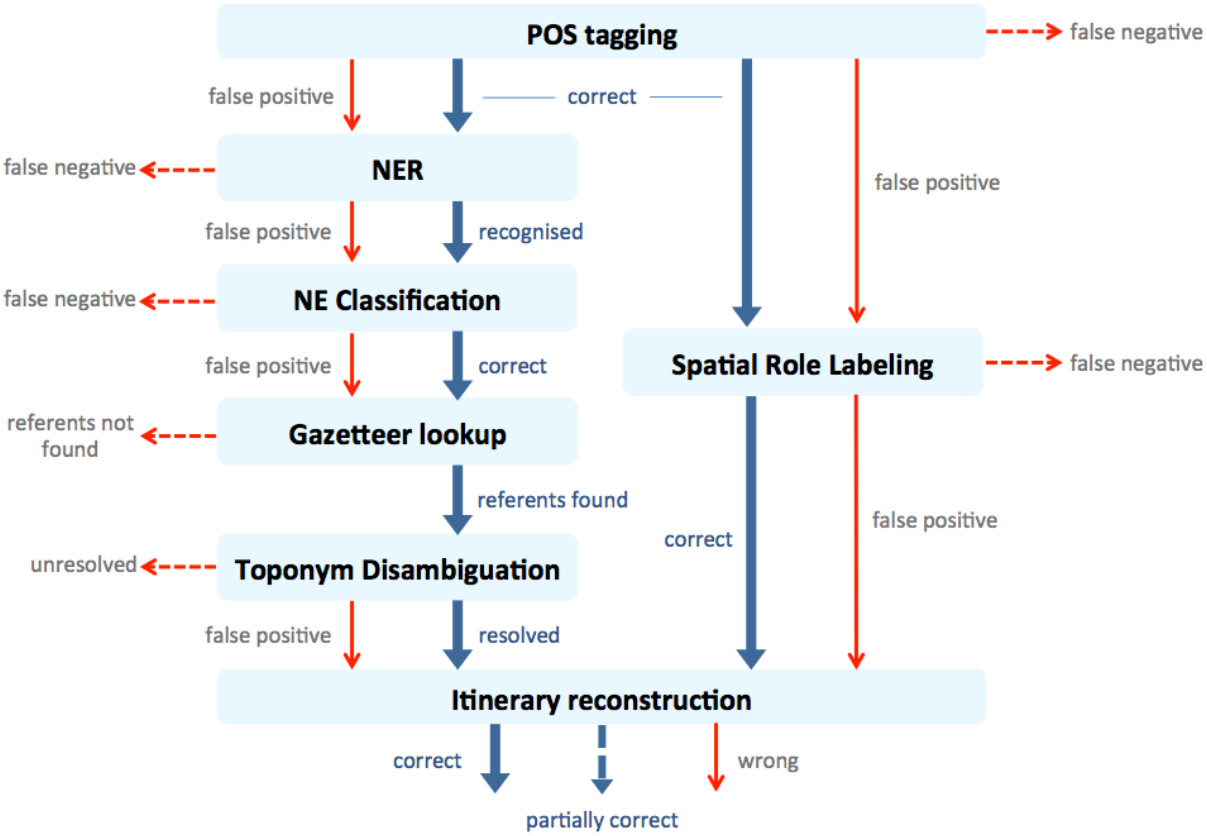


Figure 7.3: Errors propagation

Our proposed approach is designed in a modular way. Each sub-process included in our processing chain can be applied independently.[89] Errors can be introduced by each module of our automatic processing chain. Then, errors which are introduced as output of one module will be propagated by all the following modules. Figure 7.3 shows the error propagation in our proposed workflow.

Firstly, the POS tagger analysis can introduce errors concerning the grammatical category of words. In this case there are two types of errors: false positive and false negative. For instance, proper names are very useful for the step of NER. A false positive error means that the POS analysis assigns the proper name tag to a non-proper name word. This error introduced during the POS tagging will involve the wrong detection of a NE during the NER task. On the contrary, a false negative error may involve the non-detection of the NE built with the unrecognised proper name. Obviously, the problem occurs also for other grammatical categories of words such as verbs (useful for SpRL), common nouns (useful for ENE detection), etc. Errors may be also introduced during the NER task of the processing chain. Some of these errors may be involved by false positive errors previously introduced during the POS analysis. Then, false positive errors introduced by the NER module refer to recognised NEs that are actually not NEs. The NE Classification task may also introduce errors in the workflow process. In this thesis we are focusing on a binary classification of NEs: *spatial* versus *non-spatial*. Thus, false positive errors refer to non-spatial NEs which are classified as spatial and false negative errors refer to spatial NEs which are classified as non-spatial. At this point, non-proper names (e.g., common noun, preposition, etc.) and non-spatial NEs could be mis-recognised as spatial NEs, which may lead to the retrieval of several referents for non-spatial NEs by the gazetteer lookup module. Indeed, spatial and non-spatial NEs may have both the same name and introduce a different type of error called toponym ambiguity. During the gazetteer lookup task, a list of referents is found. Some referents may not exist or may not be retrieved due to incompleteness and inconsistency of gazetteers. Furthermore, some of the retrieved candidate referents may not refer to the actual toponym we are looking for. For that reason, the next module of the processing chain deals with toponyms disambiguation.

The *Toponym Disambiguation* module is designed to resolve ambiguities introduced by the previous modules of the processing chain. If the gazetteer does not retrieve any candidate referent for a given toponym, then the toponym disambiguation task tries to infer an approximate location.[90] Even if there is only one retrieved candidate referent the toponym disambigatuon module tries to resolve ambiguities anyway. For instance, the candidate referent may not refer to the spatial entity we are looking for (i.e., *referent ambiguity*) or the detected toponym is a false positive error introduced by the NE classification. Otherwise, if there are several retrieved candidate referents the toponym disambiguation module tries to determine if one of the referents refers to the good location of the spatial entity in question.[91] This module may also introduce errors, unresolved toponyms and false positive errors. False positive errors refer to resolved toponyms that are actually wrong.

Finally, the itinerary reconstruction task is different from the previous modules of the processing chain. We consider that the result of the proposed reconstruction may be correct, partially correct or totally wrong. Although this module does not deal with text mining and standard information retrieval, it may also introduce false negative and false positive errors. Indeed, some spatial NEs may be identified as waypoints whereas they were not and vice versa. Instead of using classic metrics such as precision and recall, we can compare the geometric representation of the reconstructed routes with the real ones provided by GPS tracks.

It is important to consider this problem of error propagation in the evaluation process described hereafter. For that purpose we propose two configurations for the evaluation of our processing chain. The first configuration executes the chain in a fully automatic way to obtain the results in real conditions, and in the second configuration we manually corrected the input of each module in order to evaluate their performances independently.

[89] For further information about the implementation of our processing chain see Chapter 6.
[90] see Section 4.3.4 about the geocoding for unreferenced toponyms
[91] see Section 4.3 about the toponyms resolution

7.4 Reconstruction of Itineraries

In this thesis we address the problem of the automatic reconstruction of itineraries from texts. For the evaluation of the proposed multi-criteria approach described in Chapter 3 we use the implemented components described in Chapter 6. As described in Chapter 3 we only consider the spatial component of an itinerary and we consider displacements as geometric lines. Furthermore, in order to evaluate the proposed method of itinerary reconstruction without introducing errors as input, we use our manually annotated gold standard corpus described in Section 7.2.2. Thus we assume that inputs of the process of reconstruction of itinerary are 100% correct.

We propose to use two different methods to evaluate the proposed approach. The first one (e_1) makes the comparison of the edges of the DAG obtained automatically with the edges manually built (Section 7.4.1). The second approach (e_2) makes the comparison of the real trajectory (GPS), associated with each description of the corpus with the DAG built automatically (Section 7.4.2).

7.4.1 Comparison with Manually Produced Trees (e_1)

We propose to use the precision and recall metrics to evaluate our approach and compare it to gold standard itineraries generated manually. Edges were manually built according to the textual descriptions and taking into account the comparison with the real trajectory of the route from GPS data. The precision represents the length of the relevant edges obtained automatically over the length of all the edges automatically built. And the recall represents the length of the relevant edges obtained automatically over the length of the edges manually built. Furthermore, with respect to our proposal of a multi-criteria analysis approach we evaluate the reconstruction by taking into account different combination of criteria. Table 7.6 shows the list of the criteria introduced in Chapter 3 of this dissertation for the definition of our proposal of automatic reconstruction of itineraries based on a multi-criteria approach. See Section 3.3.2 for further details about each criterion. Criteria are used to build an edge-weighted complete graph which is used to take decision over a number of alternatives for the successive displacements in order to reconstruct the most likely route. In other words, from a given location this method aims to answer the question: "Which location is the next waypoint ?". Furthermore, our proposed approach is able to make the distinction between waypoints and other locations which are not part of the route.

Criteria	Description
C_1	Text distance
C_2	Euclidean distance
C_3	Effort
C_4	Orientation
C_5	Elevation
C_6	Temporality
C_7	Perception
C_8	Negation

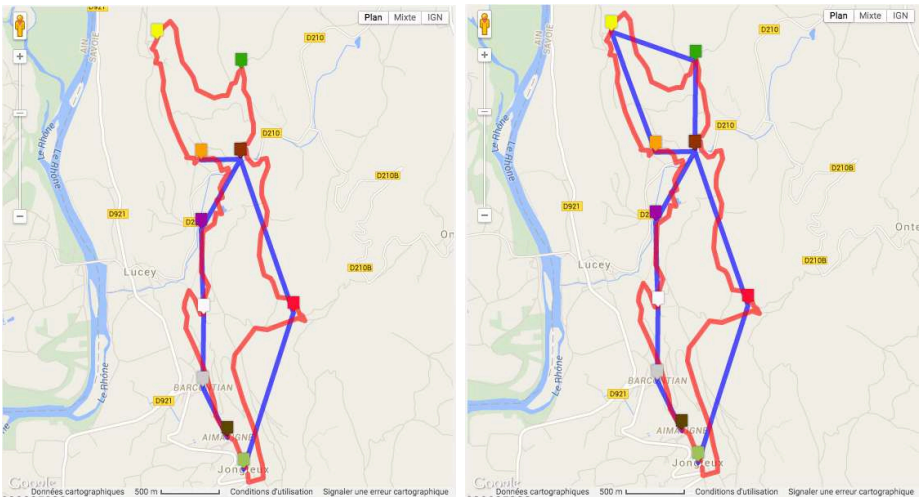
Table 7.6: Criteria

Table 7.7 shows the global precision and recall for the corpus of experiments depending on the combination of the criteria defined in the proposed multi-criteria analysis approach. Table 7.7 highlights the fact that each new criterion improves the accuracy of the automatic reconstruction. We can notice that, even if the qualitative information is not always expressed in the text, it improves significantly the accuracy. Indeed, the overall accuracy of the method (line 7) is equal to 96.1% against 84.2% for the combination of the two quantitative criteria *text distance* and *geographical distance* (line 3). Line 4 shows that taking into account the effort (C_3) improve the accuracy of the automatic reconstruction. Line 5 shows the contribution of the spatial and elevation criteria. Although the score 86,3% is not significantly

	Criteria	Precision	Recall	F1-Measure
(1)	C_1	89.5%	73.8%	80.9%
(2)	C_2	71.2%	51.2%	59.6%
(3)	$C_1 + C_2$	88.3%	80.5%	84.2%
(4)	$C_1 + C_2 + C_3$	89.1%	82.9%	85.9%
(5)	$C_1 + C_2 + C_3 + C_4 + C_5$	90.2%	82.8%	86.3%
(6)	$C_1 + C_2 + C_3 + C_4 + C_5 + C_6$	93.0%	86.0%	89.4%
(7)	$C_1 + C_2 + C_3 + C_4 + C_5 + C_6 + C_7 + C_8$	96.3%	95.8%	96.1%

Table 7.7: Evaluation of the precision and recall of edges obtained of the corpus of experiment

higher with respect to previous 85.9%, comparing the information expressed in the text such as spatial relations (north of, in the direction of, etc) or expressions referring to a change of elevation (to climb, to go down) with the geographical information found in gazetteers, improves the accuracy of the automatic reconstruction. Figure 7.4 illustrates the need of such criteria. Indeed, this spatial or elevation information are not always expressed in textual descriptions, but if it is expressed, some problems of wrong itinerary reconstruction can be solved by several criteria. In the hiking description illustrated by Figure 7.4 it is written that, from the yellow point (on the top left of the figure) the path is going south. We can notice that the use of this information improves the automatic reconstruction of the itinerary (Fig. 7.4b). Line 7 of Table 7.7 shows that the perception and negation expressions ($C_7 + C_8$) are very useful to identify which places are not waypoints.



(a) without C_4 criterion (b) with C_4 criterion

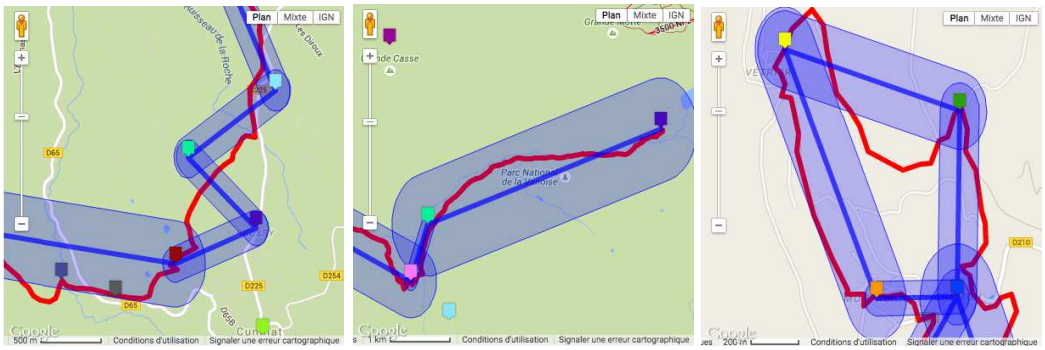
Figure 7.4: Illustration of the contribution of the spatial relation criterion (C_4)

7.4.2 Comparison with Real GPS Trajectories (e_2)

The evaluation corpus provides a ground truth of real trajectories (GPS) associated with each hiking description. To evaluate the proposed method of automatic reconstruction of itinerary, we propose to compare the automatically computed route with the real route (GPS) available with each document of our gold-standard corpus.

The implementation of the proposed method for the automatic reconstruction of itinerary builds an approximation of the route using straight lines and without taking into account road networks or geographical obstacles (rivers, mountains,...). The shape of the resulting route is obviously different from the real one. Anyway, we propose to make a comparison in order to evaluate the overall adequacy of the proposed reconstruction.

We propose to use a method taking into account an error margin to compare the similarity between the two lines. We create a buffer around the proposed path and we calculate the ratio of the real path that is included in this buffer.



(a) (b) (c)

Figure 7.5: Illustration of the buffer method

The radius of the buffer (Figure 7.5) is proportional to the length of each segment of the proposed route. Experimentally, we set the value of the radius buffer to 15% of the distance between two waypoints. The nearer two places are, the thinner the buffer is; and the farther two places are, the larger the buffer is. We use this buffer to calculate the ratio of the length of the real route that is included in the buffer of the proposed route. The average ratio of real routes included in buffers for all the documents of the corpus is equal to 72%.

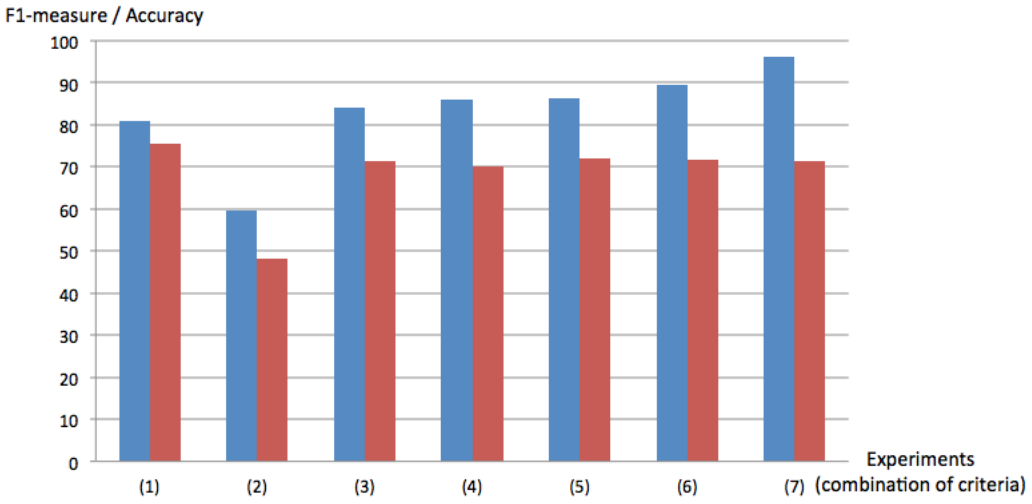
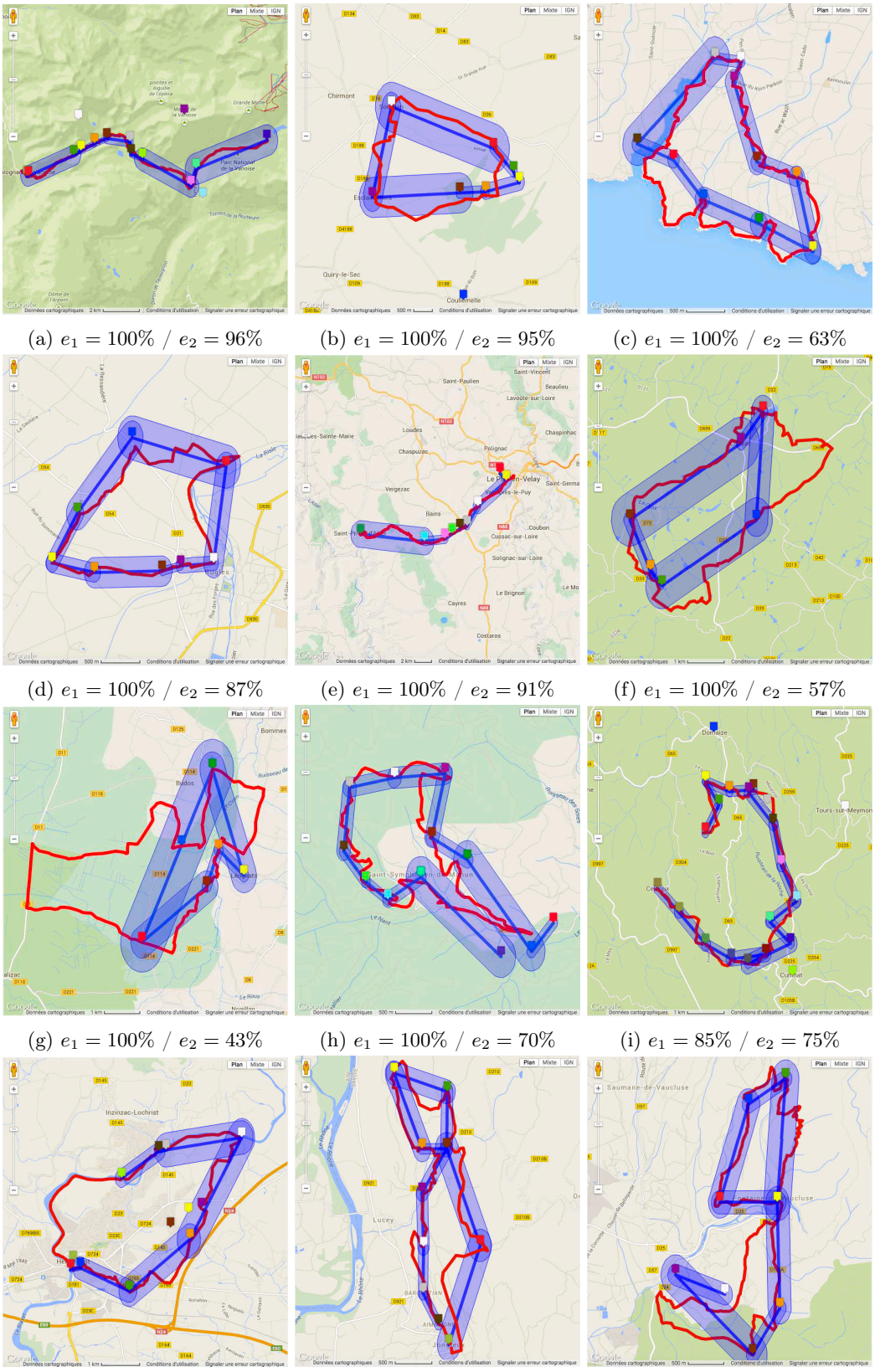


Figure 7.6: Comparison of measures obtained by the two evaluation methods on the seven experiments described in Table 7.7: blue bars show the F1-measure (e_1); red bars show accuracy (e_2)

Figure 7.6 shows the results of the two methods of evaluation (e_1 and $e2$), and Figure 7.7 shows some visual examples of results and make the comparison between the automatically reconstructed path



(a) $e_1 = 100\%$ / $e_2 = 96\%$ (b) $e_1 = 100\%$ / $e_2 = 95\%$ (c) $e_1 = 100\%$ / $e_2 = 63\%$

(d) $e_1 = 100\%$ / $e_2 = 87\%$ (e) $e_1 = 100\%$ / $e_2 = 91\%$ (f) $e_1 = 100\%$ / $e_2 = 57\%$

(g) $e_1 = 100\%$ / $e_2 = 43\%$ (h) $e_1 = 100\%$ / $e_2 = 70\%$ (i) $e_1 = 85\%$ / $e_2 = 75\%$

(j) $e_1 = 100\%$ / $e_2 = 71\%$ (k) $e_1 = 100\%$ / $e_2 = 75\%$ (l) $e_1 = 100\%$ / $e_2 = 68\%$

Figure 7.7: Comparison of the automatic reconstruction (blue) with the real trajectory (red)

and the real GPS trajectory. Each sub-figure of Figure 7.7 is associated with the score of the two methods of evaluation. We can notice that the overall results (Fig. 7.6) are decreasing using the second evaluation approach (95.9% to 72%). Indeed, as we are proposing an approximation of the path, even if the reconstruction is correct in comparison with the manually ground truth (sequence of waypoints), it may be not correct in comparison with the real path of the displacement. For instance, the reconstructions of itineraries shown in figures 7.7f and 7.7g, are correct with the evaluation e_1 (100%) but have bad scores with the evaluation e_2 (57% and 43%). These scores are explained by the fact that some locations are not named in the descriptions and that other information is expressed in terms of road names or relative directions (e.g. "turn left on road RN 12"). Another typical case of issue of reconstruction is shown in figures 7.7j and 7.7l, the problem in these descriptions of itineraries, is that at the end of the descriptions the names of the ending locations are not mentioned, and are supposed to be the same as the starting points.

7.5 Text Mining

In this dissertation, we have also addressed the problem of automatically annotating passages in the text that describe displacements making up an itinerary. Then, the first main tasks of our processing chain deal with text mining, in order to retrieve relevant information needed to build a representation of the described itinerary. As we described in Chapter 4, our module of NER and SpRL is based on a cascade of finite-state transducers and needs a POS processed text input.

7.5.1 Part-Of-Speech Tagging (Preprocessing)

The POS analysis assigns lemma and grammatical categories to each word. This pre-processing step is language dependent. Several POS taggers are freely available (at least for academic use) and we decided to compare three of them in order to choose the more efficient for each language. We selected TreeTagger and Freeling which are available for French, Spanish and Italian and Talismane which is by default only available for French. See Section 6.2.1 for more information about those three POS analysers.

Table 7.8 shows the precision, recall and F1-measure of TreeTagger, FreeLing and Talismane for the POS tagging of proper names (Table 7.8a) and verbs (Table 7.8b) for our corpus in French language.

	TreeTagger	FreeLing	Talismane
Precision	94.3%	90.5%	94.3%
Recall	87.3 %	98.1%	96.0%
F1-Measure	90.6%	94.2%	95.1%

(a) Proper names

	TreeTagger	FreeLing	Talismane
Precision	97.4%	96%	99.5%
Recall	99.8%	99.8%	99.5%
F1-Measure	98.6%	97.8%	99.5%

(b) Verbs

Table 7.8: Comparison of the French POS taggers

Then, Tables 7.9 and 7.10 show the precision, recall and F1-measure of TreeTagger and FreeLing for the POS tagging of proper names and verbs for our corpus in Spanish and Italian languages respectively.

	TreeTagger	FreeLing
Precision	99.5%	96.3%
Recall	51.2%	99.7%
F1-Measure	67.6%	98%

(a) Proper names

	TreeTagger	FreeLing
Precision	93.7%	99.2%
Recall	98.8%	99.5%
F1-Measure	96.2%	99.4%

(b) Verbs

Table 7.9: Comparison of the Spanish POS taggers

	TreeTagger	FreeLing
Precision	78.6%	91.3%
Recall	23.4%	99.3%
F1-Measure	36.1%	95.2%

(a) Proper names

	TreeTagger	FreeLing
Precision	95.9%	97.1%
Recall	98.8%	99.2%
F1-Measure	97.4%	98.1%

(b) Verbs

Table 7.10: Comparison of the Italian POS taggers

We can notice that all three POS taggers (TreeTagger, FreeLing and Talismane) obtain better results for detecting verbs than for detecting proper names. However, this comparison shows that for French (Tables 7.8), Talismane obtain better results than the two other POS taggers. And that for Spanish and Italian, FreeLing obtains better results than TreeTagger, either for detecting proper names or verbs. That comparison allows us to choose the POS analyser the more efficient for each language concerning the annotation of proper names and verbs in order to be integrated in our automatic processing chain. See Section 6.2.1 for further details about the integration of the POS pre-processing step in our automatic processing chain.

7.5.2 Named Entity Recognition and Classification

The NERC task detects and classifies NEs from texts. However, in this thesis, we consider only two types of NEs: spatial and non-spatial. Furthermore, we consider ENEs and ESNEs as described in Chapter 5.

For the evaluation of the NERC task, we evaluate the results of our proposed method (see Chapter 4) either using manual POS processed texts (POS 100% corrected) or using a fully automatic process (automatic POS processed texts). According to the results of the POS analysis (Section 7.5.1), we use Talismane for the POS tagging of the French documents and FreeLing for the Spanish and Italian documents. We call hereafter Perdido I the configuration for experiments done with manually corrected POS and Perdido II the configuration for experiments done with POS automatically processed. That way, we can show the percentage of errors introduced during the pre-processing step of our method (see Figure 7.3). We will now describe the results of the NERC task for the French, Spanish and Italian corpus thanks to the comparison with the gold-standard corpus.

French

To evaluate our proposed method for French, we decided to compare our results with those obtained with the CasEN system (Friburger and Maurel, 2004). The Quaero version of the CasEN system[92] uses the following tags to annotate NEs: *pers* (person), *loc* (location), *org* (organization), *prod* (production), *func* (function), *event*, *amount*, *date* and *time*. These tags are associated with sub-categories (Ehrmann, 2008). Among these tags we consider *loc*, and some *org* as referring to spatial NEs and the others to non-spatial NEs. We processed the 30 documents of our French gold-standard corpus of hiking descriptions with the cascade of transducers of the Quaero version of CasEN. As we have seen in the state of art (Section 2.3), the CasEN system performs well and obtains very good results on a corpus of French newspapers.

	N	CasEN		Perdido I		Perdido II	
level 0	304	149	49%	235	77%	244	80%
level 1	332	146	44%	302	91%	280	84%
level 2	20	2	0.1%	16	80%	17	85%
level 3	4	0	0%	0	0%	1	25%
total	660	297	45%	553	84%	542	82%

Table 7.11: Number of well detected ENEs with CasEN, Perdido I and Perdido II (French)

[92] http://tln.li.univ-tours.fr/Tln_CasEN.html

	N	Insertion (I)			Deletion (D)			Classification (C)			Boundaries (B)			CB		
		(a)	(b)	(c)	(a)	(b)	(c)	(a)	(b)	(c)	(a)	(b)	(c)	(a)	(b)	(c)
level 0	304	54	2	30	105	8	6	13	50	41	18	9	11	4	2	2
level 1	332	4	1	4	90	2	21	7	17	16	61	9	10	16	2	5
level 2	20	0	0	0	5	0	0	0	3	2	13	1	1	0	0	0
level 3	4	0	0	0	2	0	0	0	4	2	1	0	1	1	0	0
total	660	58	3	34	202	10	27	20	74	61	93	19	23	21	4	7

Table 7.12: Number of errors with (a) CasEN, (b) Perdido I and (c) Perdido II (French)

Table 7.11 shows the number of ENEs of the French body of reference well detected by CasEN, Perdido I and Perdido II without any error (deletion, classification, boundaries or classification and boundaries). The column 'N' shows the number of reference of ENEs in our French gold-standard corpus. Then, Table 7.12 shows the number of slot errors (insertions, deletions, classifications, boundaries and both classifications and boundaries) introduced by CasEN (a), Perdido I (b) and Perdido II (c).

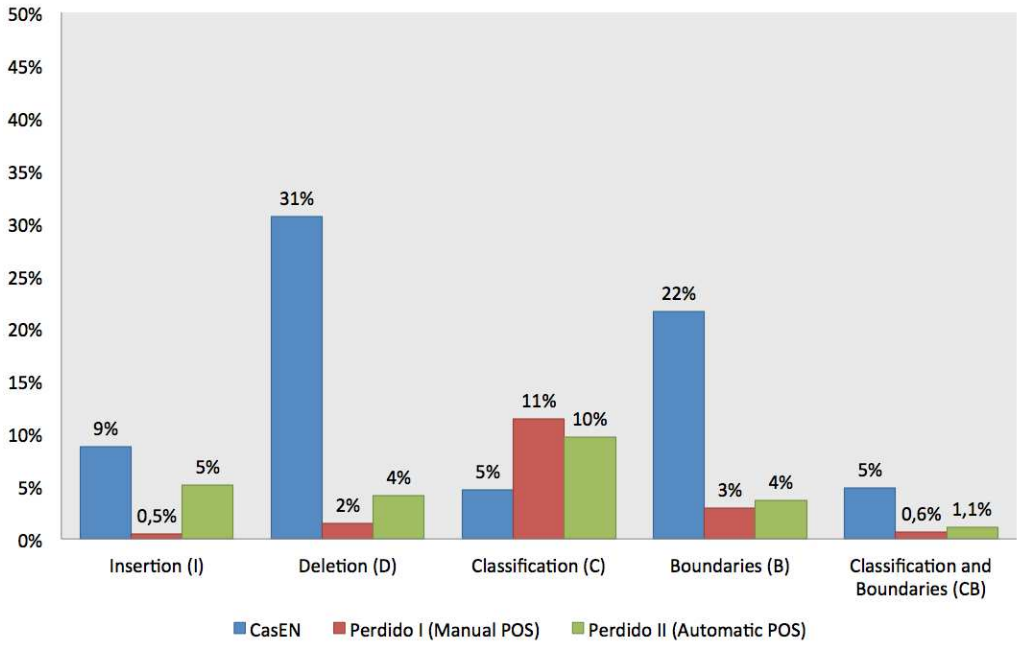


Figure 7.8: Comparison of the percentage of slot errors of CasEN, Perdido I and Perdido II (French)

Figure 7.8 shows the comparison of the percentage of slot errors of the CasEN and Perdido NER tools. Each bar of this chart refers to the percentage of errors, thus, the lower the percentages are, better are the results. Concerning errors of insertion (i.e., false positive), we can notice that both CasEN and Perdido II make more errors than Perdido I. This can be explained by the fact that Perdido I is based on a manually corrected POS pre-processing, thus there is no ambiguity or mistake concerning which words are proper names or not. Then, we can notice that CasEN makes much more errors of deletions than our method (either with a manual or automatic POS analysis). CasEN makes 202 deletion errors whereas Perdido I makes only 10 and Perdido II only 27. Our French gold-standard corpus contains a total of 660 ENEs, which means that CasEN does not detect 31% of the ENEs whereas our method does not detect only 4% of ENEs. We can also deduce from Table 7.12 that CasEN does not detect 34% of ENEs of level 0. This can be explained by the fact that CasEN uses static dictionaries of proper names for the NER task and that most of the ENEs contained in our corpus are fine-grain and do not exist in these

dictionaries. Concerning the classification errors (i.e., spatial or non-spatial), we can notice that CasEN obtains better performances than Perdido. However, the percentage of classification errors refers to the number of errors over the number of detected entities (i.e., deletion errors are not taken into account in the calculation). In addition, we have seen with the number of deletion errors that CasEN detects a fewer number of ENEs than Perdido. Figure 7.9 shows some examples of errors done by the CasEN system. Examples (1) and (2) refer to classification errors. Whereas CasEN considers these two entities as organizations, we consider these entities as spatial. Examples (3) and (4) refer both to classification and boundaries errors.

```
(1) - <event><kind>Sommet</kind> de l'<name>Archelle</name></event>
(2) - panorama sur la <func.ind><kind>Chartreuse</kind></func>
(3) - <loc.adm.town>Villers</loc> <pers.ind><name.first>St-Martin</name></pers>
(4) - Ruisseau de <pers.ind><name.first>Pierre</name> <name.last>Brune</name></pers>
```

Figure 7.9: Examples of errors of classification done by CasEN

We can notice according to Figure 7.8 and Table7.12 that CasEN makes a lot of errors of boundaries detection and more particularly for ENEs of level > 0. Furthermore, Table 7.11 shows that CasEN detects 44% of ENEs of level 1 without any errors and only 0.1% of ENEs of level > 1, mainly due to boundaries detection errors.

	level 0	level 1	level 2	level 3	**total**
SER	58.7%	43.4%	57.5%	87.5%	**51.1%**
Recall	60.5%	69.3%	75%	50%	**65.3%**
Precision	77.3%	98.3%	100%	100%	**88.1%**
Precision classification	70.2%	88.5%	100%	50%	**79.7%**
Precision boundaries	68.1%	65.4%	13.3%	0.0%	**64.8%**

(a) CasEN

	level 0	level 1	level 2	level 3	**total**
SER	13.6%	5.4%	10%	50%	**9.6%**
Recall	97.4%	99.4%	100%	100%	**98.5%**
Precision	99.3%	99.7%	100%	100%	**99.5%**
Precision classification	81.9%	94%	85%	0.0%	**87.6%**
Precision boundaries	95.6%	96.4%	95%	100%	**96%**

(b) Perdido I

	level 0	level 1	level 2	level 3	**total**
SER	31.1%	13%	7.5%	37.5%	**16.7%**
Recall	98%	93.7%	100%	100%	**95.9%**
Precision	90.9%	98.7%	100%	100%	**94.9%**
Precision classification	77.7%	92.1%	90%	50%	**84.7%**
Precision boundaries	86.9%	94%	95%	75%	**90.4%**

(c) Perdido II

Table 7.13: Evaluation of the NERC task (French)

Table 7.13 shows the overall results of the evaluation of the NERC task with CasEN (Table 7.13a), Perdido I (Table 7.13a) and Perdido II (Table 7.13c). We use the SER metric (see Section 7.3.1) proposed by Makhoul et al. (1999) which represents the total number of slot error rate taking into account the

different types of errors seen before (i.e., insertion, deletion, classification, boundaries detection and both classification and boundaries detection) and shown independently in previous tables.

Spanish
Table 7.14 shows the number and the percentage of ENEs of the Spanish body of reference well detected by Perdido I and Perdido II without any error (deletion, classification, boundaries or classification and boundaries).

	N	Perdido I		Perdido II	
level 0	231	212	92%	208	90%
level 1	174	157	90%	129	69%
level 2	12	9	75%	6	50%
level 3	-	-	-	-	-
total	421	378	90%	343	81%

Table 7.14: Number of well detected ENEs with Perdido I and Perdido II (Spanish)

	N	Insertion (I)		Deletion (D)		Classification (C)		Boundaries (B)		CB	
		(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)
level 0	231	0	10	2	2	11	16	0	1	6	4
level 1	174	0	2	0	0	12	13	1	10	4	22
level 2	12	0	0	0	0	2	2	0	2	1	2
total	421	0	12	2	2	25	31	1	13	11	28

Table 7.15: Number of errors with (a) Perdido I and (b) Perdido II (Spanish)

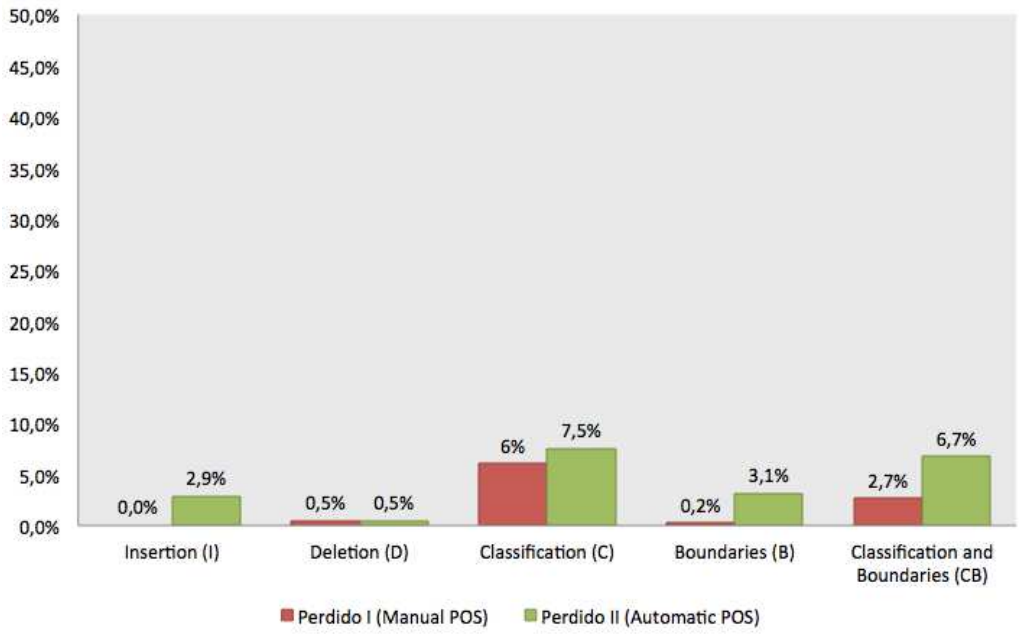


Figure 7.10: Comparison of the percentage of slot errors of Perdido I and Perdido II (Spanish)

Table 7.15 shows the number of slot errors and Figure 7.10 shows the comparison of the percentage of

slot errors introduced by Perdido I and Perdido II. We can notice that there are very few insertions, deletions and boundaries detection errors. Furthermore, as for French the main errors are due to classification (spatial / non-spatial). Not surprising, Perdido I obtains better results than Perdido II.

	level 0	level 1	level 2	**total**
SER	5.8%	6%	16.7%	**6.2%**
Recall	99.1%	100%	100%	**98.6%**
Precision	100%	100%	100%	**100%**
Precision classification	92.6%	90.8%	75%	**91.3%**
Precision boundaries	97.4%	97.1%	91.7%	**97.1%**

(a) Perdido I

	level 0	level 1	level 2	**total**
SER	10.6%	20.4%	33.3%	**15.2%**
Recall	99.1%	100%	100%	**98.6%**
Precision	95.8%	98.9%	100%	**97,2%**
Precision classification	87.5%	79%	66.7%	**83.4%**
Precision boundaries	93.2%	80.7%	66.7%	**87.6%**

(b) Perdido II

Table 7.16: Evaluation of the NERC task (Spanish)

Table 7.16 shows the overall results of the evaluation of the NERC task with Perdido I (Table 7.16a) and Perdido II (Table 7.16b)

Italian

As for French and Spanish, Table 7.17 shows the number and the percentage of ENEs of the Italian body of reference well detected by Perdido I and Perdido II without any error.

	N	Perdido I		Perdido II	
level 0	265	214	81%	199	75%
level 1	191	148	77%	133	70%
level 2	18	9	50%	7	39%
level 3	1	1	100%	1	100%
total	475	372	78%	346	71%

Table 7.17: Number of well detected ENEs with Perdido I and Perdido II (Italian)

	N	Insertion (I)		Deletion (D)		Classification (C)		Boundaries (B)		CB	
		(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)
level 0	265	1	73	2	4	47	58	1	1	1	3
level 1	191	0	2	2	4	34	38	2	9	5	7
level 2	18	0	0	1	1	4	5	3	3	1	2
level 3	1	0	0	0	0	0	0	0	0	0	0
total	475	1	75	5	9	85	101	6	13	7	12

Table 7.18: Number of errors with (a) Perdido I and (b) Perdido II (Italian)

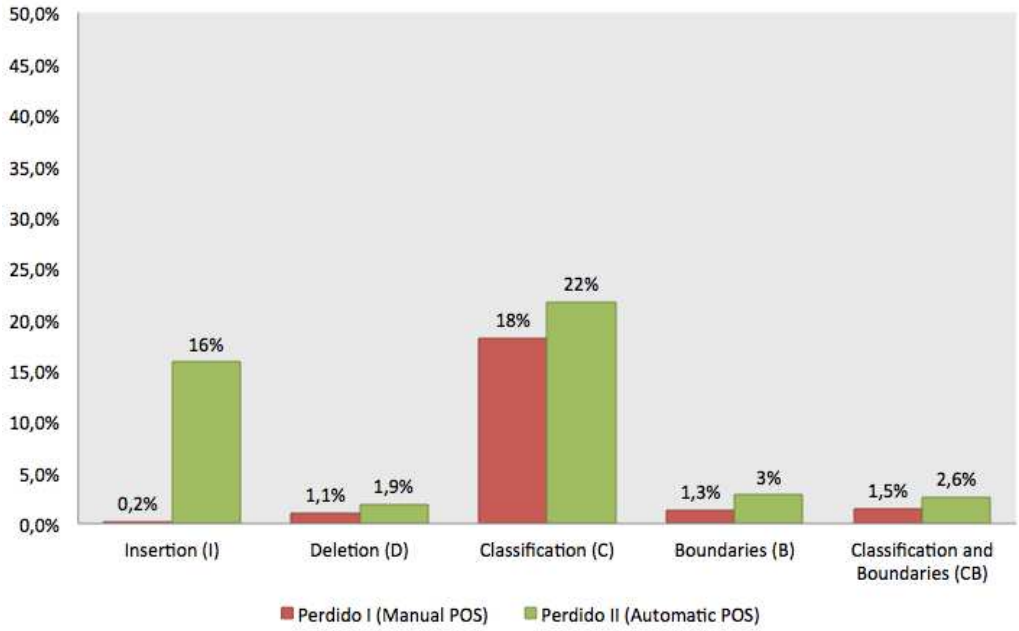


Figure 7.11: Comparison of the percentage of slot errors of Perdido I and Perdido II (Italian)

	level 0	level 1	level 2	level 3	**total**
SER	10.6%	13.1%	30.6%	0%	**12.3%**
Recall	99.3%	99%	94.4%	100%	**99%**
Precision	99.6%	100%	100%	100%	**99.8%**
Precision classification	81.4%	79.4%	70.6%	100%	**80.3%**
Precision boundaries	98.9%	96.3%	76.5%	100%	**97%**

(a) Perdido I

	level 0	level 1	level 2	level 3	**total**
SER	41.3%	19.1%	38.9%	0%	**32.2%**
Recall	98.5%	97.9%	94.4%	100%	**98.1%**
Precision	78.1%	98.9%	100%	100%	**86.1%**
Precision classification	59.9%	75.1%	58.8%	100%	**65.3%**
Precision boundaries	77%	90.5%	70.6%	100%	**81.5%**

(b) Perdido II

Table 7.19: Evaluation of the NERC task (Italian)

Figure 7.11 shows the comparison of the percentage of slot errors introduced by Perdido I and Perdido II. Then, Table 7.18 shows the number of slot errors introduced by Perdido I (a) and Perdido II (b). Finally, Table 7.19 shows the overall results of the evaluation of the NERC task with Perdido I (Table 7.19a) and Perdido II (Table 7.19b).

7.5.3 Summary

To summarize, in this section we have seen that the POS pre-processing step introduced errors as input of the NERC task. Although the results for the three languages are better with a manual POS pre-

processing (Perdido I) than with an automatic one (Perdido II), results stay comparable. According to these experiments we can notice that POS errors have a direct influence over insertion and deletion errors. For instance, for French we have 94.3% of precision for the annotation of proper names with Talismane for 5% of insertion errors, and 96% of recall for 4% of deletion errors. Furthermore, thanks to the evaluation of our NERC method, we also showed that our proposal obtains good results with few insertion, deletion or boundaries detection errors, and that most of the errors are due to the classification and particularly on ENE of level 0. This can be explain by the problem of incompleteness of gazetteers, especially for fine-grain toponyms. Furthermore, according to the comparison with the CasEN system, we can deduce that our system obtains high performances thanks to the domain-specific corpus dealing with geospatial entities and spatial relations and to the formalization and detection of ENE. Indeed, expansion parts of ENE of level > 0 are useful for the classification process.

In addition, this evaluation shows that each language have their own particularity, for instance Italian corpus obtains an SER of 32.2% (Perdido II) whereas French and Spanish obtain 16.7% and 15.2%, respectively.

7.6 Toponym Disambiguation

In this section we describe the evaluations related to the problem of toponym disambiguation. The proposal of a toponym disambiguation method is one of the main contributions of this dissertation. We have proposed an hybrid approach[93] based on a gazetteer lookup method and on the subtyping of toponyms combined with an unsupervised algorithm that takes profit of clustering techniques.

7.6.1 Subtyping of Place Named Entities

For the geocoding experiments we are using gazetteers provided by national mapping agencies: BDNyme[94] (France), Nomenclátor Geográfico Básico de España[95] (Spain), and Toponimi d'Italia IGM[96] (Italy). Additionally, we also use two well-known gazetteers: Geonames and OpenStreetMap. For further details about geographical resources see Section 2.5.

As we have seen in Section 4.3.2, the use of contextual elements such as words that have a geographical denotation (e.g., downtown, valley, ridge) is very important in toponym disambiguation. We proposed to use the local context of toponyms, when available, to solve structural and referent ambiguities. We use a subtyping method as described in (Nguyen et al., 2013) that uses the hierarchical annotations of ESNEs to identify toponyms (equivalent of ESNE of level 0) and subtypes (expansions associated with proper names or ESNE of lower level). Thus, according to the hierarchical classification of ENEs, all the ESNEs of level > 0 have a geographical subtype available.

In the remainder of this section we are considering *ESNE candidates*. As we developed an automatic processing chain, the errors introduced during the NER step are not corrected and are given as input of the next module of the chain. Indeed, errors introduced at each level of the workflow are propagated along the process (see Figure 7.3). Thus, the term *ESNE candidates* means that some of the ESNEs might not refer to spatial entities.

Then, as we have seen in Section 6.2.3, for each ESNE candidate extracted from the text our gazetteer lookup method query geographical resources with the so called 'full name'. If there is no result, then the method makes a new query with the ESNE of lower level, etc. For instance, the ESNE (97) is not found in resources with its full name but only with the proper name 'Fontanettes' (ESNE of level 0).

(97) hameau de Fontanettes
hamlet of Fontanettes

Table 7.20 shows the amount of ESNEs candidates (level 0 or level > 0) found in gazetteers with the Perdido I (Table 7.20a) and the Perdido II (Table 7.20b) configurations.

[93] This proposal is described in Chapter 4 of this dissertation.
[94] http://professionnels.ign.fr/bdnyme
[95] http://www.ign.es
[96] http://www.pcn.minambiente.it/GN/

		French			Spanish			Italian		
		(x)	(y)	(z)	(x)	(y)	(z)	(x)	(y)	(z)
Level 0	Full name query	218	160	73%	189	172	91%	223	141	63%
	National Gazetteer		121	55%		158	84%		68	30%
	Geonames		94	43%		149	79%		87	39%
	OpenStreetMap		128	59%		164	87%		118	52%
Level > 0	Full name query	368	95	26%	182	47	26%	182	17	9%
	National Gazetteer		31	8%		21	12%		0	0%
	Geonames		11	3%		6	3%		7	4%
	OpenStreetMap		86	23%		37	20%		15	8%
	Sub-toponym query		179	49%		119	65%		114	63%
	National Gazetteer		135	37%		95	52%		65	36%
	Geonames		104	28%		95	52%		70	38%
	OpenStreetMap		157	43%		109	60%		92	51%
	Sub-total		274	74%		166	91%		131	72%
Total		586	434	74%	371	338	91%	405	272	67%

(a) Perdido I

		French			Spanish			Italian		
		(x)	(y)	(z)	(x)	(y)	(z)	(x)	(y)	(z)
Level 0	Full name query	326	244	75%	255	207	81%	351	227	65%
	National Gazetteer		136	42%		182	71%		98	28%
	Geonames		112	34%		164	64%		134	38%
	OpenStreetMap		211	65%		194	76%		196	56%
Level > 0	Full name query	358	77	22%	152	23	15%	173	10	6%
	National Gazetteer		25	7%		3	2%		0	0%
	Geonames		11	3%		1	0%		2	1%
	OpenStreetMap		68	19%		21	14%		10	6%
	Sub-toponym query		197	55%		109	72%		115	66%
	National Gazetteer		135	38%		89	59%		47	27%
	Geonames		112	31%		82	54%		73	42%
	OpenStreetMap		172	48%		100	66%		98	57%
	Sub-total		274	77%		132	87%		125	72%
Total		684	518	76%	407	339	83%	524	352	67%

(b) Perdido II

Table 7.20: Number of ESNE candidates found in gazetteers

For this experiment, ESNEs candidates categorized as 'roadnames' during the NER process are not taken into account. Column (x) shows the number of ESNE candidates annotated from the text documents for each language. Column (y) shows the number of ESNE candidates having one or more results in the gazetteers. Then, column (z) shows the percentage of ESNE candidates having one or more results according to column (x). The first line (level 0) refers to the queries for ESNE candidates of level 0 (pure proper names), and the second line (Level >0) distinguishes the number of ESNE candidates having results with full name queries and those with sub-toponym queries. We can notice that about 25% of French and Spanish ESNE candidates of level > 0 are found in gazetteers with their full names (9%

for Italian). Although this number seems low, it shows the importance of considering the hierarchical classification of ENEs.

This table shows that each corpus obtain different results. Spanish ESNE obtains better results (83%) than French (76%) and Italian (67%) ones. This problem is related to the errors introduced by the classification process and can be explain by the incompleteness of gazetteers (especially the Italian's ones) and by more fine-grain ESNE expressed in French and Italian documents. Furthermore, we can notice that the gazetteers complement each other. Indeed, for instance in the case of French texts with Perdido I configuration, whereas there are 73% of ESNE candidates of level 0 found with the three gazetteers, there are only 55% found with BDNyme, only 43% found with Geonames and only 59% found with OSM.

7.6.2 Density-Based Spatial Clustering

As we have seen before, even if referent toponyms are found in gazetteers, there are still some remaining ambiguities. For instance, in the case of French hiking descriptions and depending on the used gazetteer, between 45 to 70 % of found toponyms are ambiguous. This means that for these toponyms the gazetteers provide more than one result. Figure 7.12 shows the distribution of the percentage of referent toponyms found in gazetteers for our corpus of experiments (three languages combined). We can notice that many toponyms (between 30% and 40%) have between 1 and 20 results found in gazetteers.

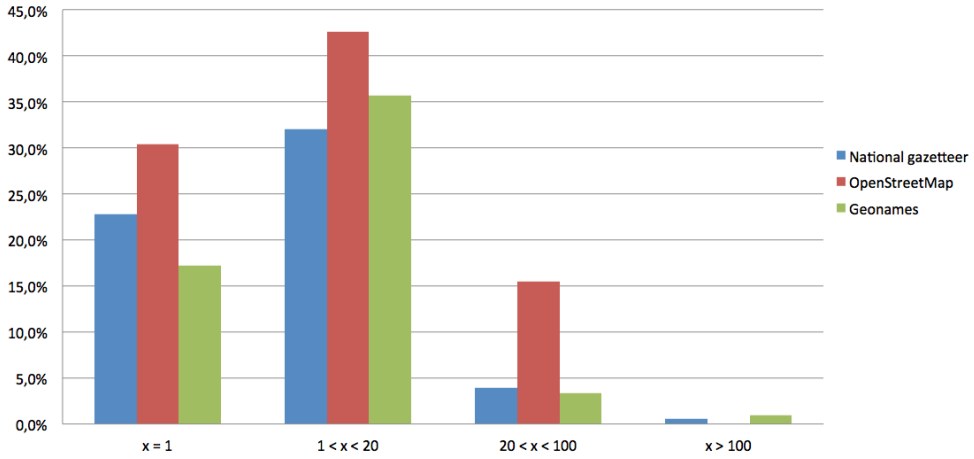


Figure 7.12: Distribution of the percentage of referents found in gazetteers

For technical reasons and in order to compare the results coming from the three gazetteers, we set a maximum limit of results per query. The limit is set to 100 results which concern less than 2% of toponyms. Table 7.21 shows that gazetteers give an average of about 10 referent locations for each toponym depending on the language.

	French	Spanish	Italian
Total # of toponyms	595	376	409
# Retrieved toponyms	518	339	352
# Results	3167	4488	5402
Avg. # Results	6.1	13.2	15.3

Table 7.21: Number of toponyms (and results) found in gazetteers

In order to avoid this ambiguity we have proposed to use the DBSCAN clustering algorithm, which uses the concept of density to determine the neighbourhood of a point. At the end of the process, every cluster represents a possible set of points describing the hiking trail. Once the clustering is done, our method chooses the 'best cluster' based on the heuristic that the best cluster is the one containing

the largest number of distinct toponyms. Our proposal, based on a density-based spatial clustering, is described in more detail in Section 4.3.3.

Figure 7.13: Illustration of the referent ambiguity

Figure 7.13 shows an example of the result of the toponym disambiguation on a French hiking description. The map on the left shows all the referent locations found in gazetteers (each color refers to an ESNE), and then the map on the middle shows the result after disambiguation (black dots refer to referent ambiguities). Finally, the focus on the right shows that the remaining color dots match the itinerary (the black line represent the real GPS trajectory of the route). Furthermore, the validity of the best cluster for every document proposed by our algorithm was evaluated by comparing the similarity between the point set of each generated cluster and the original point set of the trajectory described in a GPS Exchange Format (GPX) file. To measure the similarity we computed the convex polygon of the original point set of the trajectory and every cluster with the ST_ConvexHull PostGIS function, and then, we calculated the distance between these point sets using the ST_Distance PostGIS function (Eldawy et al., 2013; Aji et al., 2013).

In 29 out of the 30 French cases, 30 out of the 30 Spanish cases and 29 out of the 30 Italian cases the best cluster suggested by our method is the correct one, that is, the cluster with the best matching with respect to the real points in the trajectory. Table 7.22 shows the number of results before and after the toponym disambiguation. We can notice that numbers after disambiguation are less than the number of retrieved ESNEs, which means that some of the retrieved ESNEs are not located inside the best clusters.

	French	Spanish	Italian
# retrieved ESNEs	518	339	352
Before disambiguation	4816	4488	5402
After disambiguation	303	230	186

Table 7.22: Number of results before and after the toponym disambiguation

However, thanks to the comparison with respect to the real trajectory, our experiment has shown that ESNE that are not included in the best clusters (missing points) were actually not retrieved from gazetteers. Indeed, some of these missing points may have one or more referents found in gazetteers but not the actual referent we are looking for. This means that only the points included in the best clusters are well located and that all the points located outside the best clusters refer to referent ambiguities. Table 7.23 shows the comparison between the number of toponyms retrieved from gazetteers and the number of well-located toponyms.

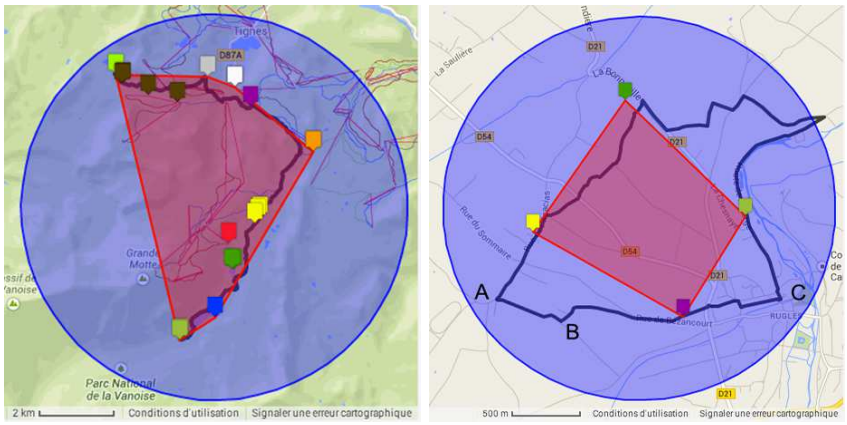
This points out the problems derived from the lack of coverage in gazetteers and the need to assign a geographic reference to those toponyms not found. For each analysed case, the missing points were associated with fine-grain toponyms.

	French	Spanish	Italian
Well-located BC	29 of 30	30 of 30	29 of 30
Retrieved ESNEs	518 (87%)	339 (90%)	352 (86%)
Well-located ESNEs	303 (58%)	230 (68%)	186 (53%)

Table 7.23: Number of best clusters (BC) and ESNEs well located

7.6.3 Geocoding for Unreferenced Toponyms

As described in section 4.3.4, in addition to the problem of automatic reconstruction of itineraries we proposed an approach to infer locations for the unreferenced toponyms. The proposed solution combines map-based disambiguation with information about spatial relations extracted from the textual description of the itinerary. This solution provides an approximate and fuzzy spatial footprint for unreferenced toponyms which are not used for the automatic reconstruction of itinerary but can be used to create or improve gazetteers. We have implemented two approaches to define a geographic area where the unreferenced toponyms are supposed to be. The first approach takes into account the geometric outline of the displacement by implementing the convex hull (in red in figure 7.14) computed with all the toponyms included in the best cluster. The second approach implements the circumscribed circle (in blue in figure 7.14) around the rectangle of the bounding box and does not take into account the geometric outline of the displacement.



(a) (b)

Figure 7.14: Refining spatial inferences according to the context

We manually reviewed all the unreferenced toponyms of each document in the corpus. We sought into different resources like web pages or detailed geographical maps to find the real locations of the unreferenced toponyms. When the toponyms were impossible to find, we also used the GPS track available with each document of our gold-standard corpus. Furthermore, we have removed from the total number of automatically annotated toponyms the ones, which were not toponyms (referent class ambiguity) and also the ones which were associated with an expression of perception. These toponyms can be far from the real trajectory described and our proposed method is not adapted to locate them.

After running the experiments, we identified different cases of spatial inference. The first one was the perfect case, that is, when all toponyms cited in the textual description were found and were all well located (Fig.7.14a). The second case was when there were some unreferenced toponyms, and their real locations were located inside the convex hull. Then the third case happened when there were several

unreferenced toponyms and when the real locations of these toponyms were not included in the convex hull but were included in the circumscribed circle (see points A,B, and C in Fig.7.14b). Last there were also some cases in which the real locations of unreferenced toponyms were located neither in the convex hull nor in the circumscribed circle.

Table 7.24 shows the number of unreferenced toponyms that were actually located inside the area defined by the convex hull or the circumscribed circle. We can notice that using the circumscribed circle, results are better (in term of number of toponyms found) than with convex hull. Although the number of new toponyms found increases with the use of the circumscribed circle, the precision of the approximated locations are better with the convex hull (because the spatial area covered is less important).

	French	Spanish	Italian
unreferenced toponyms	200	123	162
Convex hull	140 (70%)	75 (53.6%)	52 (32.1%)
Circumscribed circle	180 (90%)	102 (77.2%)	120 (74.1%)

Table 7.24: Numbers of unreferenced toponyms found in the convex hull or in the circumscribed circle

To summarize the experiments of disambiguation of the full processing chain, Table 7.25 shows some global results: the initial numbers of toponyms manually and automatically annotated (excluding toponyms associated with expressions of perception or errors); the number of toponyms located by gazetteers after the cluster-based disambiguation; the number of toponyms located by our spatial inference method; and the number of toponyms still unlocated at the end of the process. Additionally, the table shows the percentages of located (unlocated) toponyms for each body of reference. The experiment shows that the description of itinerary is very important and helps finding an approximate location for missing toponyms. Furthermore, we also noticed from our experiments that we need at least 4 or 5 well located toponyms in order to find the best cluster and to propose a good geographic area for unreferenced toponyms.

Toponyms	French	Spanish	Italian
manually annotated	542	362	350
automatically annotated	540	360	349
located by gazetteers	303 (56%)	230 (64%)	186 (53%)
located by inferences	180 (33.33%)	102 (28.33%)	120 (34.38%)
unlocated	57 (10%)	28 (8%)	43 (12%)

Table 7.25: Global results of our processing chain

7.7 Summary

In this chapter, we have presented the results of the three main contributions of this thesis: the automatic reconstruction of itineraries, the ENE recognition and classification and the ESNE disambiguation. Additionally, we have implemented the proposed solutions and integrated them into an automatic processing chain (see Chapter 6). For the evaluation we use some well-known evaluation metrics such as precision and recall, F1-measure and SER.

With respect to the problem of itinerary reconstruction, in this thesis we have proposed a method for automatically identifying the sequence of waypoints from geoparsed text and building an approximation of a plausible sequence of the described itinerary. We have proposed two evaluation methods. The first method (e_1), comparing edges manually built with those automatically generated, shows that our multi-criteria approach obtains 96% of F1-measure and also shows that each criterion improves the results of the reconstruction. Then, the second method (e_2), comparing the similarity between the automatically generated routes and the real GPS trajectories, highlights the fact that the proposed reconstruction is an approximation (using straight lines) and provides an overall accuracy of 72%.

Then with respect to the problem of automatic annotation of geospatial information, we have developed a multilingual cascade of transducers. In this chapter, we have given evaluation results for the NER task in two configurations, Perdido I which is based on a POS processed input manually corrected (100% corrected) and Perdido II which is based on a fully automatic POS processed text input. Furthermore, we also made some experiments and evaluations to choose the more efficient POS tagger for each language (French, Spanish and Italian). Additionally, the feasibility of the proposed method has been tested for the geoparsing and geocoding of ESNEs in a corpus of hiking descriptions obtaining good results in terms of accuracy, precision and recall. Moreover, it must be noted that the method performs well in a multilingual environment. The results obtained for the three different tested languages (French, Spanish and Italian) are comparable. For instance, the French corpus obtains 16.7% of SER (which takes account of insertions, deletions, boundaries detection and classification errors) and 15.2% for the Spanish and 32.2% for the Italian. It must be noted that the Italian process introduces two times more classification errors.

Finally, with respect to the problem of disambiguation, we have described the evaluation of the proposed methods described in Chapter 4. Our referent disambiguation approach failed only in 2 out of 90 cases. In these cases the problem was caused by the too few number of results retrieved. For example, in one of these cases only two ESNEs were retrieved from the gazetteer and the rest were not found (they were missing points). Moreover, the retrieved ESNEs were referencing places not related to the location of the hiking descriptions (referent ambiguity) such as bigger locations in other countries or regions. Furthermore, the evaluation has shown that our proposal of the use of a density-based spatial clustering performs well in the specific context of itinerary descriptions. It must be noted that all ESNEs that do not have any referent after the disambiguation process is because there are not stored in gazetteers. Thus, we have also described some experiments of our proposal of geocoding for unreferenced toponyms. This method provides an approximation of the location of about 80% of the unreferenced toponyms found in the documents of our gold-standard corpus.

Chapter 8

Conclusions and Future Work

I haven't been everywhere, but it's on my list.

— Susan Sontag

Contents

8.1 Summary of Contributions

The main challenge of this thesis was to connect text with geographic space and to provide a map-based representation of itineraries described in textual documents. This cross-disciplinary thesis involves three main fields of research: computer science, linguistics and geography. We have proposed an approach for the automatic geocoding of itineraries described in natural language. Our proposal aims to turn textual information written in natural language into GIS data. We divided this problem into three tasks: the annotation of geospatial information in texts (geoparsing), the toponym resolution (geocoding) and the reconstruction of the itinerary itself.

8.1.1 Reconstruction of Itineraries from Text

In Chapter 3 we described the main contribution of this PhD, i.e. an approach for the automatic geocoding of itineraries. According to a corpus analysis and related works, we have analysed the concepts defining an itinerary expressed in natural language. Itineraries and displacements are described in natural language using spatial knowledge such as spatial named entities, spatial relations, perception expressions with description of landmarks, motion expressions and trajectories. Waypoints (also known as 'decision points') and routes are the two main elements involved in the description of an itinerary and we defined an itinerary as a sequence of displacements between waypoints. Moreover, we have proposed an approach for automatically distinguishing waypoints from other types of locations and identifying the sequence of waypoints from a geoparsed text. Then, using the sequence of waypoints, our method builds a first approximation of a plausible footprint of the described itinerary. For that purpose, we have introduced a

graph-based model for representing an itinerary as described in a text. Although waypoints and routes are the two main elements describing an itinerary, our model emphasizes other elements such as features seen or mentioned as landmarks. We have proposed to define an itinerary as a Directed Acyclic Graph (DAG) in which the edges represent route or perception segments and vertices represent locations (both waypoints and visual cues). A DAG has some helpful properties such as reachability and transitive reduction (i.e., the sub-graph with the fewest edges that represents the same reachability of the graph). We can also apply algorithms for topological ordering as any DAG has at least one topological ordering. Additionally, we have shown that these properties are well adapted for the representation of an itinerary. For instance, directed edges can represent the direction of the motion between two waypoints. If a waypoint is visited several times in the itinerary, then there are several vertices in the graph to represent each time this waypoint has been visited. That way, the DAG representation respects the sequence of displacements and for each waypoint we know the previous and the next waypoint. Furthermore, our proposed method for the reconstruction of the itinerary is based on a multi-criteria approach combining quantitative and qualitative criteria, based on knowledge extracted from the text and data coming from gazetteers. The combination of criteria (i.e., text distance, geographical distance, effort, orientation, elevation, temporality, perception, and negation) is used to decide over a number of alternatives for the successive displacements.

8.1.2 Geoparsing and Geocoding

Chapter 4 has described a processing chain for the automatic geoparsing and geocoding of descriptive texts. The results provided by this processing chain can be used as input for the automatic reconstruction of itineraries described in Chapter 3. With respect to the annotation of named entities and spatial information used to describe itineraries, we have proposed a hybrid solution combining POS analysis, a cascade of finite-state transducers and the interrogation of external gazetteers. The linguistic component is based on the work of (Loustau, 2008) and (Nguyen, 2012), who have proposed a linguistic processing chain for French built with Prolog rules in the Linguastream platform[97]. We have transformed and adapted these rules into a cascade of finite-state transducers. Additionally, we have adapted our transducers for Spanish and Italian languages and shown our transducers are adaptable for romane languages with very few modifications. Our cascade of transducers is composed of six main transducers annotating measures (e.g., distance), offsets (topological and directional relations), names (named entities), verbs (classified verbs: motion, perception etc.), Expanded Named Entities and VT structures (expressions of motion and perception). Transducers annotating named entities make a first step of classification using external evidences when available and unambiguous. Then, with respect to the geocoding part of our method, we propose an approach for the named entity classification task. As we have seen in this dissertation, we consider only two types of entity: spatial and non-spatial. We have proposed a gazetteer lookup method for the classification of ENEs and for the geocoding of spatial ones. Furthermore, this chapter has made a special emphasis on two main problems: the existence of ambiguous toponyms, and the lack of gazetteers with enough coverage for fine-grain toponyms. The solution proposed for addressing these two problems has been based on a hybrid method combining subtyping of toponyms and the use of clustering techniques. The subtyping toponyms approach aims at resolving structural ambiguities and takes advantage of the annotation of ENEs. Indeed, expansions contained within ENEs of level > 0 refer to the subtype of the entities. This local linguistic context, when available, is used for classification and disambiguation (Rauch et al., 2003; Hollenstein and Purves, 2010). Then with respect to the referent ambiguities, the definition of clusters provides a map-based disambiguation approach to identify the clusters with the highest number of candidate toponyms in terms of spatial density. Finally, we proposed a method of spatial inference, which combines the bounding polygon of these clusters with spatial relations expressed in text to define approximate locations of those fine-grain toponyms not found in gazetteers.

8.1.3 A Multi-Scale Markup Language

Chapter 5 has described a model for marking semantically raw texts. We have defined a formal representation of text documents written in natural language that can be applied for the task of Named Entity

[97] http://www.linguastream.org/

Recognition (NER) and Spatial Role Labeling (SpRL). Our proposal relies on a multi-scale annotation process based on a core generic layer, which can be freely adapted into more specific layers depending on the intended goal. The proposed markup language is based on the TEI Guidelines and proposes a generic and extensible markup language. This language is particularly dedicated for the text mining task and can be customized and adapted. Although there are markup languages more focused on the specification of relations, the first layer of our proposal is more focused on the specification of NEs. For that purpose, we have introduced and defined the concept of Expanded Named Entity (ENE). In contrast to NER works usually considering NEs as being composed only by pure proper names, we consider both categories of proper names, i.e. pure and descriptive (Jonasson, 1994). We have defined an ENE as an entity built from a proper name and that can be composed of one or more expansions. As we have seen in the disambiguation approach, this hierarchical overlapping of NEs is very useful for the classification process. Indeed many errors are introduced during the Named Entity Classification process because of incorrect boundaries detection. For instance, standard methods annotate only 'Paris' instead of annotating the ENE 'mayor of Paris'. The problem is whereas 'Paris' refers to the city, 'mayor of Paris' refers to the person who can be located far away from Paris depending on the description made. Furthermore, the proposed generic core layer may be used to create and share pre-annotated corpus. Although our proposal of a multi-scale markup language is still at an early stage of development, we have shown the feasibility of this proposal from a generic annotation of texts describing itineraries toward a geospatial semantic annotation. This proposal has been applied for the problem of automatic geospatial information extraction described in Chapter 4 and was used as the input format for the automatic itinerary reconstruction method described in Chapter 3.

8.1.4 Design and Implementation

Chapter 6 has described the design and implementation of a processing chain for the automatic reconstruction of itineraries from descriptive texts. We have designed and integrated each contribution proposed in this PhD in a highly modular web-based architecture. The first part of the processing chain deals with linguistic processes designed for French, Spanish and Italian languages and implements the components described in Chapter 4. Then, the second part of the chain implements the components dedicated to the automatic reconstruction of itineraries as described in Chapter 3. Furthermore, we proposed several web services providing access to different tasks. Apart from a service for POS processing, which is an encapsulation of existing POS taggers, we created services for new proposals described in this dissertation such as named entity recognition, named entity disambiguation and toponym resolution which are new proposals described in this dissertation. Additionally, we also proposed a demonstration tool[98] available online and providing access to a web-interface for testing the PERDIDO processing chain and to several computed examples of hiking descriptions.

8.1.5 Evaluation

Chapter 7 has described the evaluation with real data of the three main contributions described in this dissertation: the automatic reconstruction of itineraries, the ENE recognition and the toponym disambiguation. Additionally, this chapter has also described our gold-standard corpus of hiking descriptions called *PERDIDO corpus*, and has also briefly described the web-based application that we designed to help and control the manual annotation of textual documents. The PERDIDO gold-standard corpus consists of 90 hiking descriptions (30 for each language: French, Spanish and Italian). Chapter 7 gives some statistics about our corpus such as the distribution of ENEs and classified verbs.

Automatic Reconstruction of itineraries

Concerning our proposal of a method for the automatic reconstruction of itineraries, we have used two complementary methods for the evaluation. The first one (e_1) makes the comparison of the edges of the DAG obtained automatically with edges manually built. Additionally, we evaluated the automatic reconstruction by taking into account different combinations of criteria and we have shown that each new

[98] http://erig.univ-pau.fr/PERDIDO/

criterion involved in the reconstruction process improves the results. Whereas the reconstruction process obtains a F1-measure of 96.1% with the combination of all criteria, it obtains 80.9% considering only the *text distance*, and only 59.6% considering only *geographical distance*. Furthermore, even if qualitative information is not always expressed in texts, it improves significantly the accuracy of the reconstruction. The second evaluation method (e_2) is based on the comparison of the real GPS track (associated with each hiking description of our gold-standard corpus) with the DAG automatically built. As we have seen in Chapter 3, our approach builds an approximation of the route using straight lines and without taking into account road networks or geographical obstacles. Thus, the shape of the resulting route is obviously different from the real one. For this comparison, we have proposed a method using a buffer to compare the similarity between the two lines. Then, with the e_2 evaluation method, we obtained an accuracy of 72%. This evaluation is complementary to the results of the e_1 method and shows that even if the reconstruction is correct according to the manually built edges, there are some missing points not expressed in the textual descriptions, and that using straight lines is not sufficient to build an accurate representation of the itinerary.

Expanded Named Entity Recognition and Toponym Resolution

Concerning the ENE recognition and the toponym disambiguation evaluations, we proposed two configurations: Perdido I, which is based on a POS processed input manually corrected (100% correct); and Perdido II, based on a fully automatic POS processed text input[99]. Additionally, we have also made some experiments and evaluations to determine the most efficient POS tagger for each language (French, Spanish and Italian). For the evaluation of the NER task we have used the Slot Error Rate (SER) metric. We distinguish several types of slot errors: insertion, deletion, classification, boundaries detection, and both classification and boundaries detection. We obtained an SER of 16.7% for the French corpus, 15.2% for Spanish and 32.2% for Italian. Although corpora of evaluation are different, according to the state of art our approach of NER obtains better results. Furthermore, the evaluation has shown that most of the errors are due to the classification and particularly for ENEs of level 0. Indeed, with the Perdido I configuration 64% of errors are classification errors and 42% with Perdido II. Furthermore, according to the comparison with CasEN, we have shown that PERDIDO makes very few deletion and boundaries detection errors, especially on ENEs of level > 0. These results can be explained by the fact that most of ENEs refer to fine-grain entities not referenced by dictionaries used by CasEN and that CasEN is not fully designed to support the detection of ENE of level > 0. Then, with respect to the disambiguation task, we have shown that spatial named entities contained in the PERDIDO corpus are highly ambiguous with an average of 10 referent locations for each place name. However, the experiments have shown that the proposed method based on a spatial density clustering performs well and solves referent ambiguities. Furthermore, these experiments have highlighted the problem of incompleteness of gazetteers. Indeed, about 60% of toponyms, expressed in the PERDIDO corpus, have no relevant references in gazetteers.

8.1.6 Publications Resulting from this Thesis

- The method for the reconstruction of itineraries described in Chapter 3 and the corresponding experiments described in Chapter 7 have been accepted for publication in the paper titled 'Reconstruction of itineraries from annotated text with an informed spanning tree algorithm' in the International Journal of Geographical Information Science (Moncla et al., 2015).

- The proposed method for the automatic annotation of geospatial information describing itineraries has been presented in the 8th GIScience conference (Moncla et al., 2014a).

- The toponym disambiguation approach described in Chapter 4 and the corresponding experiments and evaluation results described in Chapter 7 have been presented in the 22nd ACM SIGSpatial conference (Moncla et al., 2014b).

[99] According to the comparison of POS tagger described in Section 7.5.1, the automatic POS tagging obtains about 4% of errors on proper names recognition and 1% of errors on verbs recognition.

8.2 Work in Progress

As we have seen in Chapter 5, we have made a first proposal of the specification of a multi-scale markup language for encoding textual information based on the TEI Guidelines. The generic layer of our proposal is focused on the annotation of ENEs, and the additional layer introduced more semantic annotations. One of our short-term objectives is to specify a third layer for encoding spatial and spatio-temporal semantic information with emphasis on the annotation of relationships between the different already detected elements. For that purpose, our proposal can be combined with other existing markup languages such as ISO-Space, which seems to be a comprehensive model for annotating spatio-temporal information.

Additionally, as we have seen in Chapter 7, we have started the design of a web-based application to help and control the manual annotation of textual documents. We have used the first version of this tool[100] to build the PERDIDO gold-standard corpus described in Chapter 7. This tool provides a WYSIWYG[101] interface (Figure 7.1) to interact with the content, in order to add elements and attributes. Furthermore, it is designed to automatically extract rules and constraints from any valid XML Schema and guide the user during the annotation process. At the end of the annotation the system control the validity of the annotation and create the XML file. A short-term objective is to add functionalities such as taking as input pre-annotated documents based on the guidelines of the multi-scale markup language proposed in Chapter 5. Indeed, we plan to integrate some NLP tasks such as POS-processing, or ENE recognition, which obtain accurate results, in order to add annotations that belong to the generic core layer. Then, users can correct the automatic pre-annotations and continue the annotation following the guidelines of semantic layers.

Finally, in addition to the availability of web services described in Chapter 6, another short-term objective is to provide access to the components of our processing chain as an open source software. Moreover, we also plan to adapt each module of our processing chain to be integrated in well-known frameworks for NLP such as Gate[102] or UIMA[103].

8.3 Future Work

In this thesis we have proposed different contributions from the automatic annotation of geospatial information towards the automatic reconstruction of itineraries. Although we have shown the feasibility of our proposal on real data, there are still some limits and some possible improvements for further works.

Big Data

The evolution of technologies and the increasing availability of large amounts of data have led to a new multidisciplinary field of research known as 'Big data', which aims to adapt standard mechanisms such as analysis and visualization for huge amount of data:

- Much of these data are spatially and temporally referenced and offer many possibilities for enhancing geographical understanding (Kitchin, 2013). However, usually these spatial/temporal references are not explicitly annotated and require a specialized processing. The proposed methods in this dissertation could help to extract and relate the spatial references managed in a Big Data environment. The contributions of this dissertation can be also seen as a proposal for the enhancement of the cultural heritage. Indeed, we have proposed a map-based representation of textual documents describing displacements.

- Furthermore, the issue of digital preservation can be combined with concerns addressed by Big Data studies. A lot of cultural heritage documents have been digitized in last years and Big Data mechanisms are needed to exploit all these digitized documents. For instance, we can cite international projects have been developed for the enhancement and accessibility of cultural heritage

[100] Demo video: http://erig.univ-pau.fr/PERDIDO/demo/TextTagging.mp4
[101] WYSIWYG: what you see is what you get
[102] http://gate.ac.uk/
[103] http://uima.apache.org/

to support knowledge economy and improve education and research such as Europeana[104] developed by the European Commission and Memory of the World[105] established by UNESCO.

Geoparsing and Expanded Named Entity Recognition and Classification

With respect to the text mining part of our proposal described in Chapter 4 several improvements are possible:

- A first work would be to improve the geoparsing by adding a deeper linguistic processing, a deeper semantic analysis and to annotate unnamed locations.
- Another short-term objective is also to adapt our processing chain for other languages such as English language in order to profit from the availability of corpora and evaluation campaigns. The objective would be to compare our method of NER and toponym disambiguation with other well-known approaches (e.g., Stanford NER, Open Calais, etc.).

Furthermore, with respect to the evaluation process, as we have seen in Chapter 7, we distinguish several types of slot errors. The evaluation of the NER task has shown that most of the errors are due to the classification. Moreover, we have shown that the hierarchical overlapping of ENEs is very helpful to detect a local context associated with NEs, which improves significantly the classification results even when proper names are not found in gazetteers. Several further alternatives could be explored:

- Concerning ENEs of level 0, the problem is linked to the incompleteness of gazetteers (spatial ENEs not found) but also to the fact that we query only geographical resources. Thus, an improvement would be to query additional resources like other gazetteers with different coverage and more adapted to the type of documents and also other resources referencing non-spatial entities, such as dictionaries of proper names (Maurel et al., 2014).
- Concerning ENEs of level > 0, we could also improve the results of classification thanks to the query of additional resources such as ontologies and linked data (e.g., dbpedia[106], wordnet[107], etc.). Indeed, it seems easier to find domain-specific ontologies than comprehensive proper names dictionaries.
- Additionally, another area of improvement could be to combine our approach with the proposal of Maurel et al. (2011), which obtains very good results for the classification of French named entities. Although their approach uses only dictionaries and is for the moment only available for French, we could imagine to use some of their transducers at the beginning of our cascade in order to detect and classify non-spatial entities.

Machine Learning and Automatic Itinerary Reconstruction

With respect to the problem of the automatic reconstruction of itineraries addressed in Chapter 3 several improvements are possible:

- For instance, one limit concerns the setting of the proposed multi-criteria approach. One key issue in our multi-criteria approach is how to define weights and the combination strategy. For such a problem of setting the weights of a multi-criteria combination, or for setting the suited model of combinations, machine learning is a widely used approach that we could follow. In particular, machine learning is used in the natural language processing domain for approaches of entity tagging that are based on probabilistic models such as HMM (Rabiner, 1989) or, approaches to extracting semantic relationships between entities (Béchet et al., 2014). However, whatever machine learning technique is used, a key issue is to get a sufficient number of examples and to precisely define the learning task (Mitchell, 1997).

[104] http://www.europeana.eu/
[105] http://www.unesco.org/new/en/communication-and-information/flagship-project-activities/memory-of-the-world/
[106] http://wiki.dbpedia.org/
[107] http://wordnet.princeton.edu/

- Another possible improvement, and an interesting challenge, is to build a more accurate approximation of the actual footprint of the described itinerary. For that purpose, it would be possible to consider a method to integrate more information coming from geographical resources describing feature shapes, land cover or digital elevation.

Appendix A

Examples of websites hosting hiking descriptions

This appendix shows some examples of websites hosting hiking descriptions for French, Spanish and Italian languages. These websites have been used to create the PERDIDO corpus of hiking descriptions described in Chapter 7. Figure A.1 and A.2 show examples of French hiking descriptions. We can notice that these websites provide the textual description of the itineraries, a map-based representation, the elevation profile and also some metadata such as duration, difficulty, etc. Figure A.3 and A.4 show examples of Spanish and Italian hiking descriptions. These two websites provide the textual description of the itineraries and an interactive map showing the route and pictures.

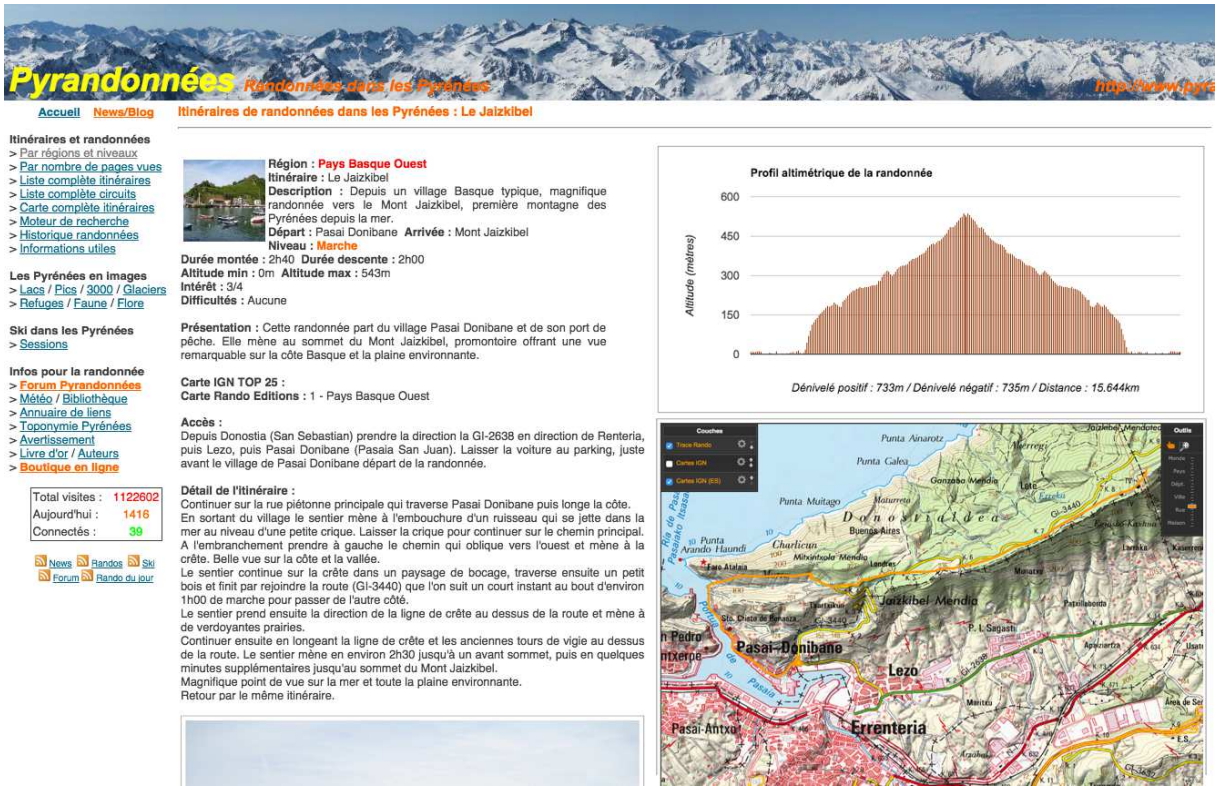


Figure A.1: Le Jaizkibel – **Source:** http://www.pyrandonnees.fr/

Figure A.2: De Pralognan au refuge de la Leisse – **Source:** http://www.visorando.com/

Figure A.3: GR 11 Sallent de Gállego - Balneario de Panticosa – **Source:** http://senderos.turismodearagon.com/

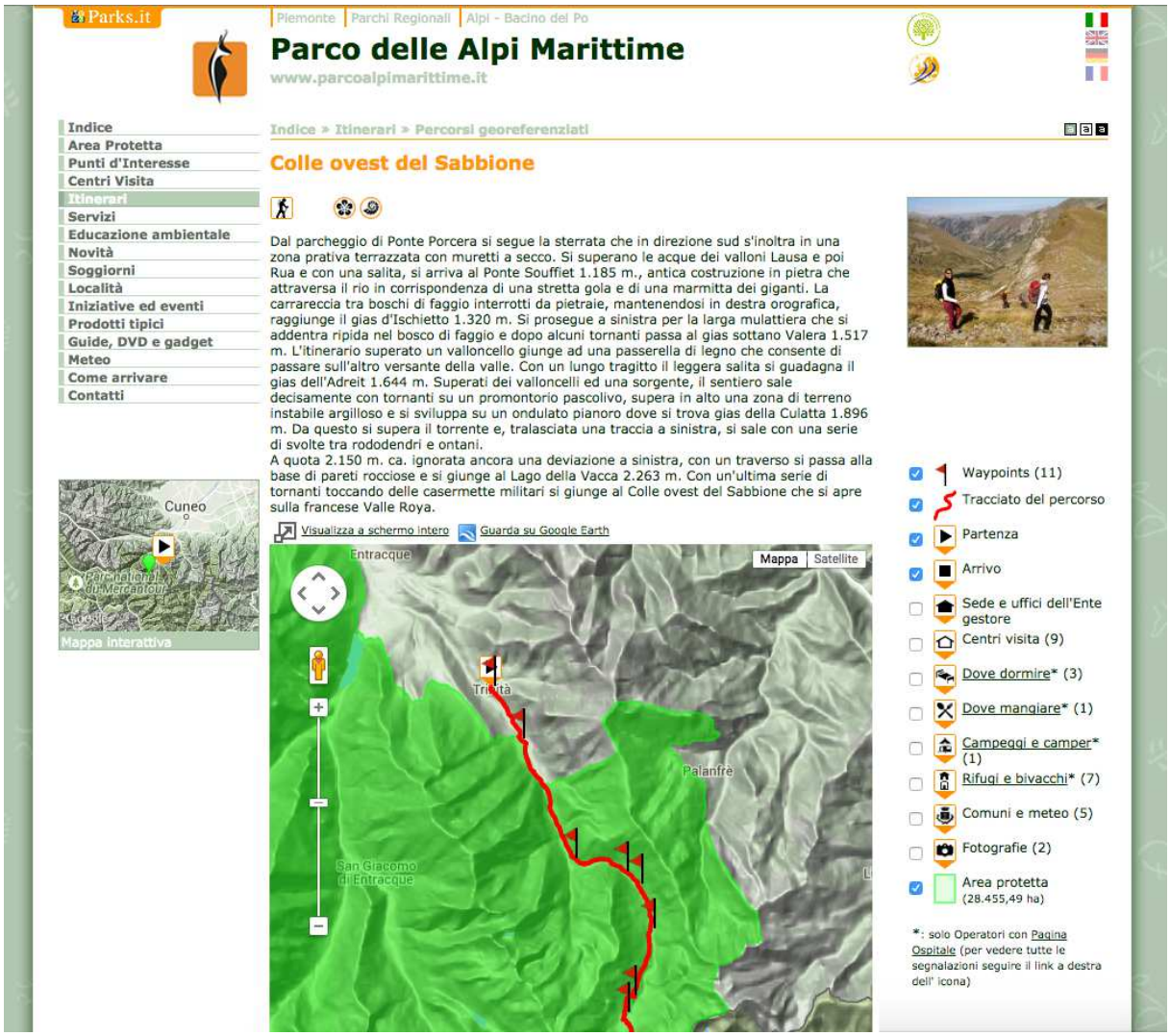


Figure A.4: Colle ovest del Sabbione – **Source:** http://www.parks.it/parco.alpi.marittime

Appendix B

Part-of-Speech tagsets

This appendix lists the tagsets used by TreeTagger (for French, Spanish and Italian languages) and by Talismane (for French).

Tag	Description	Tag	Description
ABR	abreviation	PRP:det	preposition + article
ADJ	adjective	PUN	punctuation
ADV	adverb	PUN:cit	punctuation citation
DET:ART	article	SENT	sentence tag
DET:POS	possessive pronoun	SYM	symbol
INT	interjection	VER:cond	verb conditional
KON	conjunction	VER:futu	verb futur
NAM	proper name	VER:impe	verb imperative
NOM	noun	VER:impf	verb imperfect
PRO	pronoun	VER:infi	verb infinitive
PRO:DEM	demonstrative pronoun	VER:pper	verb past participle
PRO:IND	indefinite pronoun	VER:ppre	verb present participle
PRO:PER	personal pronoun	VER:pres	verb present
PRO:POS	possessive pronoun	VER:simp	verb simple past
PRO:REL	relative pronoun	VER:subi	verb subjunctive imperfect
PRP	preposition	VER:subp	verb subjunctive present

Table B.1: French POS tags used by TreeTagger
Source: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/french-tagset.html

Tag	Description	Tag	Description
ABR	abreviation	PRP:det	preposition plus article
ADJ	adjective	PUN	punctuation
ADV	adverb	PUN:cit	punctuation citation
DET:ART	article	SENT	sentence tag
DET:POS	possessive pronoun	SYM	symbol
INT	interjection	VER:cond	verb conditional
KON	conjunction	VER:futu	verb futur
NAM	proper name	VER:impe	verb imperative
NOM	noun	VER:impf	verb imperfect
PRO	pronoun	VER:infi	verb infinitive
PRO:DEM	demonstrative pronoun	VER:pper	verb past participle
PRO:IND	indefinite pronoun	VER:ppre	verb present participle
PRO:PER	personal pronoun	VER:pres	verb present
PRO:POS	possessive pronoun	VER:simp	verb simple past
PRO:REL	relative pronoun	VER:subi	verb subjunctive imperfect
PRP	preposition	VER:subp	verb subjunctive present

Table B.2: Spanish POS tags used by TreeTagger
Source: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/spanish-tagset.txt

Tag	Description	Tag	Description
ABR	abreviation	PRO:pers	personal pronoun
ADJ	adjective	PRO:poss	possessive pronoun
ADV	adverb	PRO:refl	reflexive pronoun
CON	conjunction	PRO:rela	relative pronoun
DET:def	definite article	SENT	sentence marker
DET:indef	indefinite article	SYM	symbol
FW	foreign word	VER:cimp	verb conjunctive imperfect
INT	interjection	VER:cond	verb conditional
LS	list symbol	VER:cpre	verb conjunctive present
NOM	noun	VER:futu	verb futur
NPR	name	VER:geru	verb gerund
NUM	numeral	VER:impe	verb imperative
PON	punctuation	VER:impf	verb imperfect
PRE	preposition	VER:infi	verb infinitive
PRE:det	preposition + article	VER:pper	verb participle perfect
PRO	pronoun	VER:ppre	verb participle present
PRO:demo	demonstrative pronoun	VER:pres	verb present
PRO:indef	indefinite pronoun	VER:refl:infi	verb reflexive infinitive
PRO:inter	interrogative pronoun	VER:remo	verb simple past

Table B.3: Italian POS tags used by TreeTagger
Source: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/italian-tagset.txt

Tag	Description	Tag	Description
ADJ	adjective	P	preposition
ADV	adverb	P+D	preposition + determinant
ADVWH	interrogative adverb	P+PRO	preposition + pronom
CC	coordinating conjunction	PONCT	punctuation
CLO	clitic (object)	PRO	pronoun
CLR	clitic (reflexive)	PROREL	relative pronoun
CLS	clitic (subject)	PROWH	interrogative pronoun
CS	subordinating conjunction	V	indicative verb
DET	determinent	VIMP	imperative verb
DETWH	interrogative determinant	VINF	infinitive verb
ET	foreign word	VPP	past participle verb
I	interjection	VPR	present participle verb
NC	common noun	VS	subjunctive verb
NPP	proper noun		

Table B.4: French POS tags used by Talismane
Source: http://urieli.github.io/talismane/

Appendix C

Evaluation of the NER task on a French travelogue

Chapter 7 has described the evaluation of our proposed method for Named Entity Recognition and Classification (NERC) over a specific corpus of hiking descriptions. In this appendix, we describe the evaluation of the NER task over the French travelogue called 'Ascension au Pic de Nethou'[108] written by Platon de Tchihatcheff in 1842, who has described his expedition through the Pyrenees. The following text is a short excerpt of this travelogue:

> "La tête des monts Pyrénéens s'élève, fière et peu connue, la chaîne de la Maladetta, dont le point culminant, le pic de Néthou, n'a été encore gravi par personne, et dont les approches sont défendues, presque de toutes parts, par des glaciers formidables. Comme plusieurs tentatives avaient été faites pour atteindre jusqu'à son sommet, sans qu'aucune d'elles ait pu complètement réussir , il était d'un intérêt, pour la science des zones supérieures de ces monts de constater d'une manière exacte ou la possibilité d'y parvenir ou bien la nature des obstacles qui pouvaient s'y opposer. Les expériences de M. de Humboldt, celles de Saussure, si fécondes en résultats pour l'histoire des sciences naturelles, lors de l'ascension de ces savants au Chimboraço et au Mont-Blanc, étaient de beaux exemples à suivre ; et une tentative sérieuse, en faveur de leur rival pyrénéen, devenait en quelque sorte, une oeuvre de conscience. [...] C'est du port de Bénasque que l'on a le premier coup d'oeil de la chaîne de la Maladetta, qui s'étend à l'ouest du pic d'Albe jusqu'au pic de la Fourcanade, à l'est avec sa crête hérissée et son glacier immense, de près de 11.694 mètres de longueur, percé ça et là par des moraines superficielles et des rocs détachés qui s'élèvent sur son dos. Le pic de Néthou, ayant l'air d'un cône obtus et comme voûté, domine toute la chaîne, et, quoique sans rival dans les Pyrénées, il paraît être plutôt, vu du côté nord, le satellite que le chef d'un pic à double pointe qui s'élève à l'ouest de lui, et que l'on appelle abusivement ici le pic de la Maladetta. [...] Vu du port de Bénasque ou de celui de la Picade, ce glacier ne se présente pas comme reposant sur un lit à pente très rapide. Etant presque entièrement couvert de neige, il a plutôt l'air d'une vaste nappe blanche, légèrement jaunie par des débris terrestres, dans quelques parties de sa surface, et soulevée sur plusieurs points en croupes arrondies et fendues. Ce n'est que là qu'on aperçoit cette couleur bleu-verte qui est si caractéristique dans les glaciers des Alpes. Il ne m'est pas arrivé non plus de voir dans les Pyrénées ces belles cavernes transparentes, ce jeu cristallin d'aiguilles et d'arêtes, qui hérissent la surface des glaciers des Bois (vulgairement appelé Mer de Glace) des Bossons, du Rhône, de Grindelwald et de beaucoup d'autres, et qui les font ressembler aux vagues tumultueuses de la mer en courroux, comme surprises par une congélation subite. Ici, on ne trouve rien de pareil, et les glaciers quoiqu'en montrant parfaitement leur profil de vert marin dans la coupe perpendiculaire de leurs crevasses, laissent beaucoup à désirer lorsqu'on veut les comparer à ceux des Alpes."

[108] http://www.amazon.fr/Ascension-au-Pic-Nethou-Aneto/dp/2912233143

Table C.1 shows the total number of words and the distribution of ENEs in this document. Whereas hiking descriptions in which there are only between 1 and 3% of non-spatial ENEs, in this travelogue there are 35% of non-spatial ENEs. Indeed, there are several name of persons (e.g., *colonel Coraboeuf*, *M. Franqueville*, etc.) and several names of plants and trees (e.g., *Pinus sylvestris*, *Rhododendron ferrugineum*, *Lonicera alpina*).

Total # of words	7730	
# of ENEs	220	
- spatial (% of ENEs)	144	(65%)
- non-spatial (% of ENEs)	76	(35%)

Table C.1: Distribution of ENEs

Table C.2 shows the number of ENEs well detected by CasEN and Perdido II without any error (deletion, classification, boundaries or classification and boundaries). The column 'N' shows the number of reference of ENEs in the travelogue. Then, Table C.3 shows the number of slot errors (insertions, deletions, classifications, boundaries and both classifications and boundaries) introduced by CasEN (a) and Perdido II (b).

	N	CasEN		Perdido I	
level 0	100	59	59%	64	64%
level 1	116	33	18%	104	90%
level 2	6	1	17%	5	83%
total	220	93	42%	173	79%

Table C.2: Number of well detected ENEs with CasEN and Perdido II

	N	Insertion (I)		Deletion (D)		Classification (C)		Boundaries (B)		CB	
		(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)
level 0	100	3	21	33	4	2	29	3	1	0	2
level 1	116	0	6	63	0	0	5	9	3	11	4
level 2	6	0	0	2	0	0	1	2	0	0	0
total	220	3	27	98	4	2	35	14	4	11	6

Table C.3: Number of errors with (a) CasEN and (b) Perdido II

Figure C.1 shows the comparison of the percentage of slot errors of the CasEN and Perdido NER tools. Each bar of this chart refers to the percentage of errors, thus, the lower the percentages are, better are the results. Concerning errors of insertion (i.e., false positive), we can notice that CasEN makes very few errors (only 3) and that Perdido makes more insertions (27). Then, we can notice that most of the errors made by CasEN are due to deletion (i.e., non detection) and incorrect boundaries detection. On the contrary, Perdido makes very few deletions and incorrect boundaries detection but most of the errors are due to the classification. Furthermore, this evaluation has shown that PERDIDO obtains comparable results with different types of textual documents (i.e, hiking descriptions and travelogues).

Figure C.2 shows some examples of correct recognition of ENEs of level > 0. This highlights the fact that CasEN succeed to annotate NEs which we consider as being ENEs. Figure C.3 shows some examples of annotation of person entities with the CasEN system. We can notice that CasEN uses external evidences (e.g., colonel, M.,etc.) but do not include these evidences in the annotated entities. Finally, Figure C.4 shows some examples of errors of boundaries detection and classification made by CasEN.

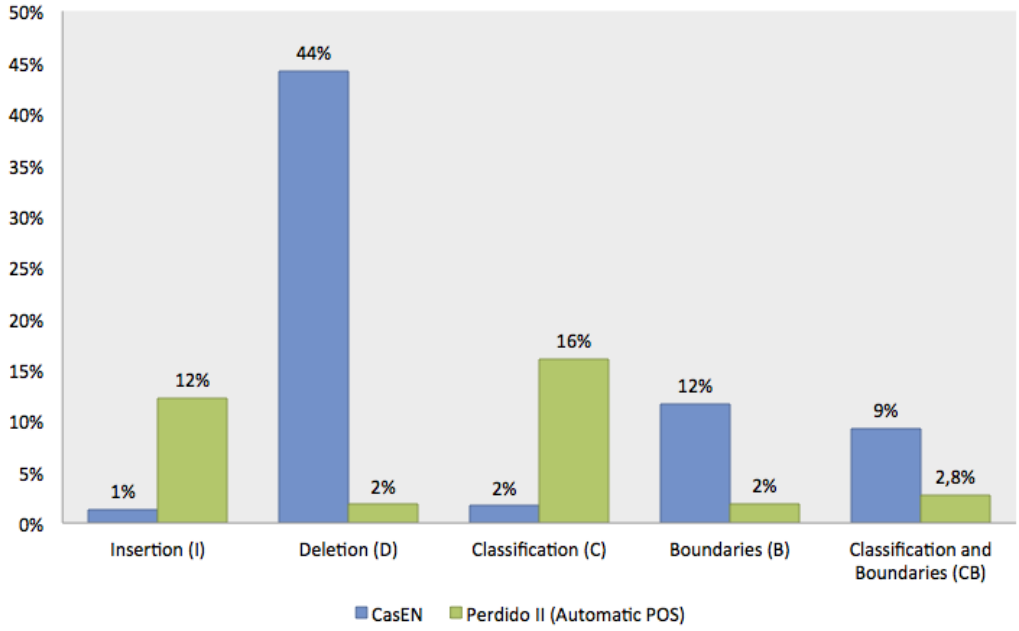


Figure C.1: Comparison of the percentage of slot errors of CasEN and Perdido II

	level 0	level 1	level 2	**total**
SER	38.5%	67.7%	50%	**54%**
Recall	64%	45.7%	50%	**54%**
Precision	95.52%	100%	100%	**97.6%**
Precision classification	92.54%	79.2%	100%	**87%**
Precision boundaries	91%	62.3%	33.3%	**77.3%**

(a) CasEN

	level 0	level 1	level 2	**total**
SER	42%	12.1%	8.3%	**25.5%**
Recall	96%	100%	100%	**98.2%**
Precision	82.1%	95.1%	100%	**89%**
Precision classification	55.6%	87.7%	83.3%	**72.2%**
Precision boundaries	79.5%	89.3%	100%	**84.9%**

(b) Perdido II

Table C.4: Evaluation of the NERC task

```
<geogName>pic de Néthou</geogName>
<geogName>sommet du Néthou</geogName>
<placeName>port de la Picade</placeName>
<address>chemin du port de la Picade</address>
<placeName>port de Bénasque</placeName>
<geogName>plaines de la Catalogne</geogName>
<geogName>lac de Grigueno</geogName>
<geogName>mont Saint-Bernard</geogName>
<geogName>bassin de Corunes</geogName>
<geogName>terres de France</geogName>
<placeName>port de Grigueno</placeName>
<orgName>territoire de l'Aragon</orgName>
<geogName>vallée du Plan</geogName>
<geogName>vallée de l'Essera</geogName>
```

Figure C.2: Examples of correct recognition of ENEs of level > 0 with CasEN

```
colonel <persName>Coraboeuf</persName>
M. <persName>Fontan</persName>
M. le docteur <persName>Fontan</persName>
```

Figure C.3: Examples of boundaries detection errors done by CasEN

```
maire de <placeName>Luchon</placeName>
le baromètre Gay-<placeName>Lussac</placeName>
directeur de l' <orgName>Observatoire</orgName> de <placeName>Toulouse</placeName>
la crête-mère <persName>de la Maladetta</persName>
mon guide <persName>de Luz</persName>
aux mines de <placeName>Pasco</placeName>
<geogName>pics des Monts</geogName> Maudits
les Cordillères des <geogName>Andes</geogName>
```

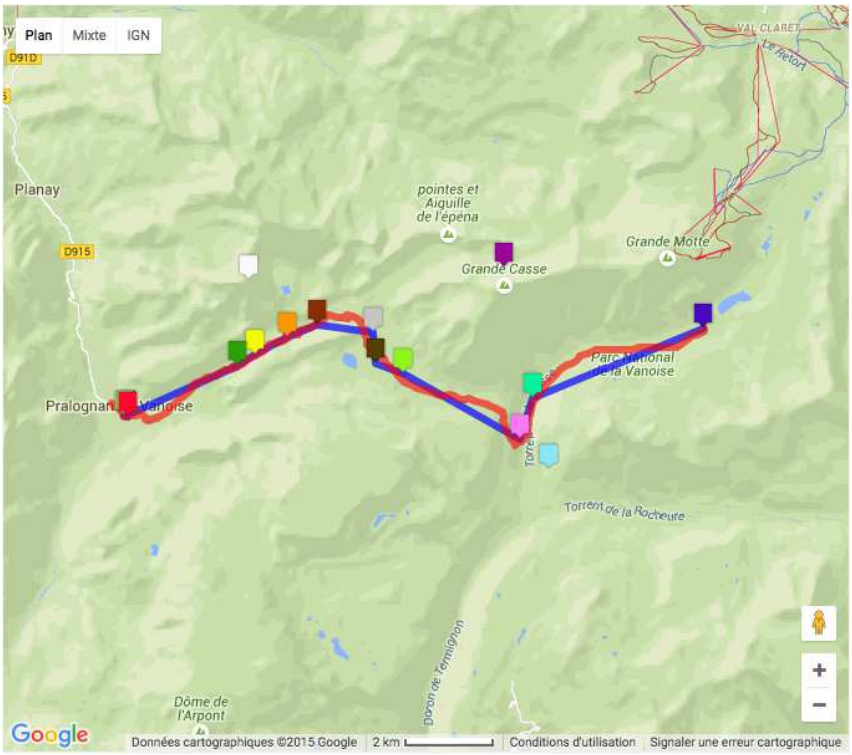
Figure C.4: Examples of errors done by CasEN

Appendix D

Examples of results of the PERDIDO processing chain

A Pralognan suivre la route entre l' hôtel de la Vanoise et celui du Petit Mont Blanc et continuer tout droit Passer **les bourgs** de Barioz et de Bieux On tombe alors sur le GR55 à suivre jusqu'au hameau des Fontanettes Le sentier suit par la gauche un télésiège jusqu'au refuge des Barmettes Poursuivre par le pont de la Glière Arrivé près_des chalets de la Glière ne_pas prendre le sentier de droite qui mène au Moriond mais filer tout **droit** Plus loin passer sur le pont du Chanton Peu après on parvient au lac des Vaches que l'on traverse sur des **rochers** Le dénivelé durant cette partie est assez conséquent de l' **ordre** de 1100m mais rien que **le passage** du lac des Vaches et la vue sur la Grande Casse méritent le détour Au bout de ce lac le chemin revient un_peu vers la gauche A un petit carrefour ne_pas partir vers la Pointe du Creux Noir mais bifurquer vers la droite en **direction** du lac Long que l'on contournera également par la droite On parvient ensuite au refuge du Col de la Vanoise Continuer sur le GR55 Un bon tronçon de sentier avec peu de dénivelé contourne le Lac Rond par sa droite et passe plus loin devant une croix Plus tard une bonne descente de 300m assez raide se profile On peut apercevoir tout en bas le **pont** de Croe-Vie Descendre tout en bas pour arriver à ce dernier et le traverser A sa hauteur au carrefour ne_pas prendre à droite vers le refuge d' Entre-Deux-Eaux mais partir à gauche et rester sur le GR55 Suivre alors le torrent de la Leisse par la droite pendant plusieurs kilomètres pour arriver à l' **étape du jour** Cette partie se fait dans une vallée sauvage et encaissée de toute beauté avec ses moraines sur un chemin en pente douce mais régulière Au loin surgit le refuge de la Leisse perché sur une butte qui sera un_peu dure à gravir en cette fin de **journée** Voilà vous êtes arrivés

(a) Annotated text

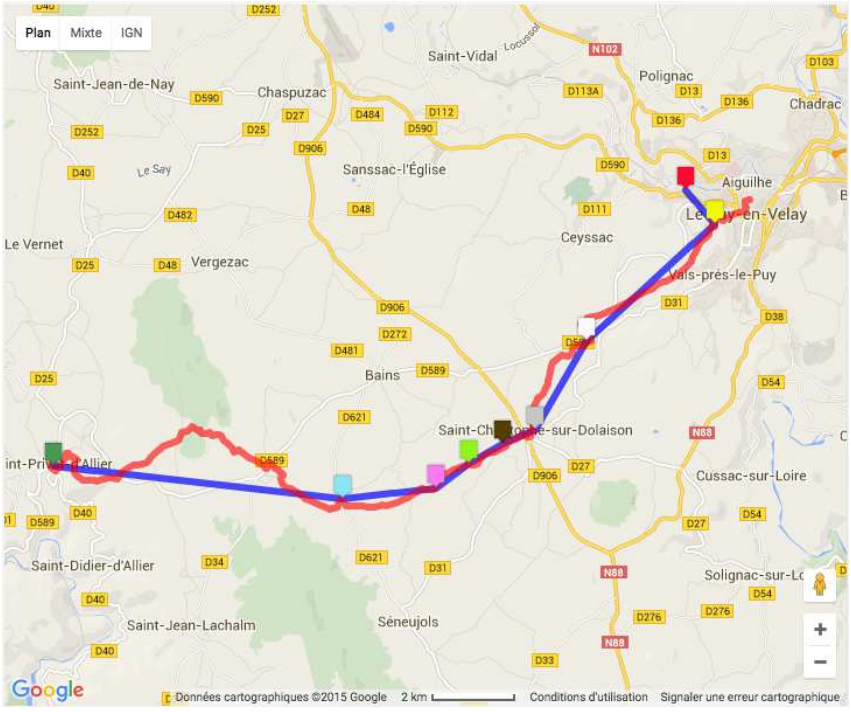


(b) Map-based representation

Figure D.1: Results of the PERDIDO processing chain

Départ des marches de la cathédrale pour descendre **la rue des tables** des clous directionnels ont été posés au sol Juste après **la fontaine** du Choriste prendre tout à gauche la rue Raphaëlle pour atteindre la place du Plot et sa **fontaine** situés en centre-ville De cette place emprunter successivement rue Saint-Jacques rue des Capucins et rue de Compostelle A_partir_de la rue des Capucins la route goudronnée monte en **pente douce** Après l' entrée de l' usine Fontanille dans le virage prendre à gauche le large chemin Continuer tout droit en passant devant **la croix** dite de Pouvignac Au virage suivant ignorer à gauche le chemin d' accès au sentier des chibottes pour poursuivre sur **le même chemin caillouteux** Passez devant un bâtiment agricole Rester en poursuivant en droite **ligne** sur **ce chemin** jusqu'à atteindre la D589 La couper pour prendre le sentier caillouteux en face L' itinéraire est balisé par les traits blanc rouge du GR qu' il faut repérer sur les pierres et les arbres et suivre jusqu'à Saint-Privat Après avoir une nouvelle fois coupé la D589 atteindre La Roche A la sortie près_d' une table de pique-nique partir sur la gauche en suivant **le sentier** Au croisement suivant bifurquer à gauche ignorer à droite la variante également balisée par les traits blanc rouge du GR Déboucher au village de Saint-Christophe-sur-Dolaison et faite le tour de l' église pour observer son remarquable clocher à peigne Poursuivre en quittant **le bourg** prendre à droite pour passer devant des **tables de pique-nique** Puis ressortir de l' autre côté de la départementale en empruntant un petit tunnel Face_à la croix prendre à gauche Traverser successivement les hameaux de Tallode Liac et Lic en suivant **le balisage blanc** rouge du GR Vous arrivez ensuite au village de Ramourouscle Au centre bifurquer à gauche Suivre la route jusqu'à la chapelle Saint-Roch Traverser Monbonnet pour rejoindre la D589 La suivre à gauche sur 150m avant_de la traverser pour s' engager à droite sur un chemin montant entre les maisons Traverser un plateau qui à un **moment** se scinde en deux Prendre le chemin de gauche qui grimpe et monte vers **la crête sombre de sapins** Arriver à un **embranchement le chemin** continue vers la droite Pénétrer dans la **forêt** en suivant **le balisage blanc** rouge du GR Vous atteignez le lac de l' Oeuf que vous longez par la gauche Déboucher sur une **route** La suivre à gauche sur 150m avant_de prendre rapidement un chemin qui s' abaisse à droite signalisation Continuer tout droit en restant sur ce chemin dans les bois Il se poursuit par une petite route à **découvert** bordée de **champs cultivés** Couper la D589 pour prendre en face la route menant au village de Chier Traverser **la place** centrale en passant devant **la mairie** Dévaler ensuite un chemin herbeux puis un sentier pierreux avant_de franchir un **ruisseau** à proximité duquel se trouve un moulin Déboucher sur **la route** goudronnée à l' entrée de Saint-Privat-d'Allier Dirigez -vous alors vers le centre-bourg pour suivre **la rue montant** en **direction** du garage Jobert pour atteindre **la mairie** en **face** de laquelle vous pouvez monter dans **la navette spéciale** départ à h pile pour retourner en **bus** au Puy en Velay.

(a) Annotated text

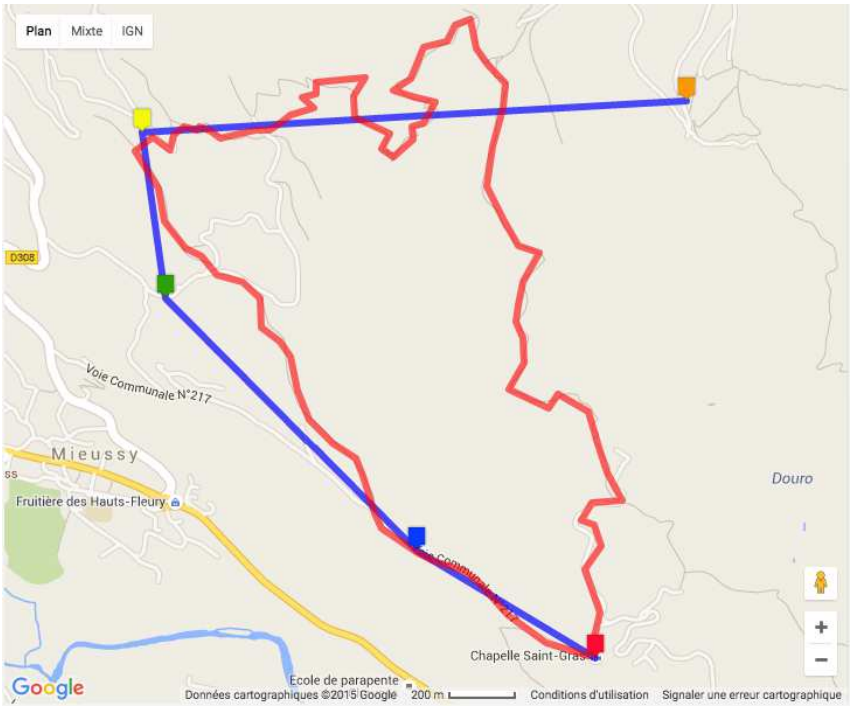


(b) Map-based representation

Figure D.2: Results of the PERDIDO processing chain

Dans le prolongement de la route **:direction** Le Jourdy **:**prenez la piste qui descend et vous amène à la Chapelle Saint-Gras dédiée au protecteur des cultures et du bétail Continuez tout droit sur le chemin carrossable après être passé au hameau de Guillard prenez à droite **:direction** Roche Pallud **:**0h15 Le chemin monte de **façon** assez raide puis se radoucit Suivez **le chemin principal** balises et pictogrammes aux croisements qui par des séries de montées et petites descentes vous conduit à Cloiset puis au Jourdy Rejoignez **la route** prenez à droite et suivez -la jusqu' à son **extrémité** 0h40 Poursuivez dans l' axe de la route par une **piste raide** Au niveau d' une coupe forestière où la vue sur **:les massifs** des Aravis **:**du Bargy et du Môle se dégage vous allez croiser une autre piste Poursuivez tout **droit** et prenez tout de suite à gauche un sentier bien marqué balise C' est le début du chemin de Croix qui vous mènera par une **montée** soutenue à **:la grotte** du Jourdy **:**1h25 **:La statue Notre Dame** du Bon Secours **:**protège les maisons des éboulements Le torrent qui y coule vient directement **:**du **:**Plateau de Sommand**:** Le réseau souterrain se développe sur environ 400m de longueur Prenez à droite Le sentier d'abord à plat descend ensuite puis remonte sur la gauche parallèle au pied de la falaise Il rejoint **un ressaut rocheux** câbles puis serpente pour atteindre **un nouveau ressaut** câbles Le sentier est indiqué par moment par des poteaux fléchés suivez **ces indications** Ce passage assez raide nécessite un effort important et une grande attention Poursuivez au-dessus balise par le sentier en forêt qui débouche au niveau du téléski En face de vous se trouve **:le site de décollage de parapentes** de Pertuiset **:**Prenez à droite la piste en herbe qui vous amène au **:croisement** de Roche Pallud **:**2h25 Partez à droite et suivez **la piste** descente soutenue qui vous ramènera au parking de départ Vous remarquerez par endroit les traces de l' ancien chemin pavé qui pendant longtemps a été le principal moyen d' accès aux **:alpages** de Sommand **:**Attention par temps humide certaines zones sont particulièrement glissantes

(a) Annotated text

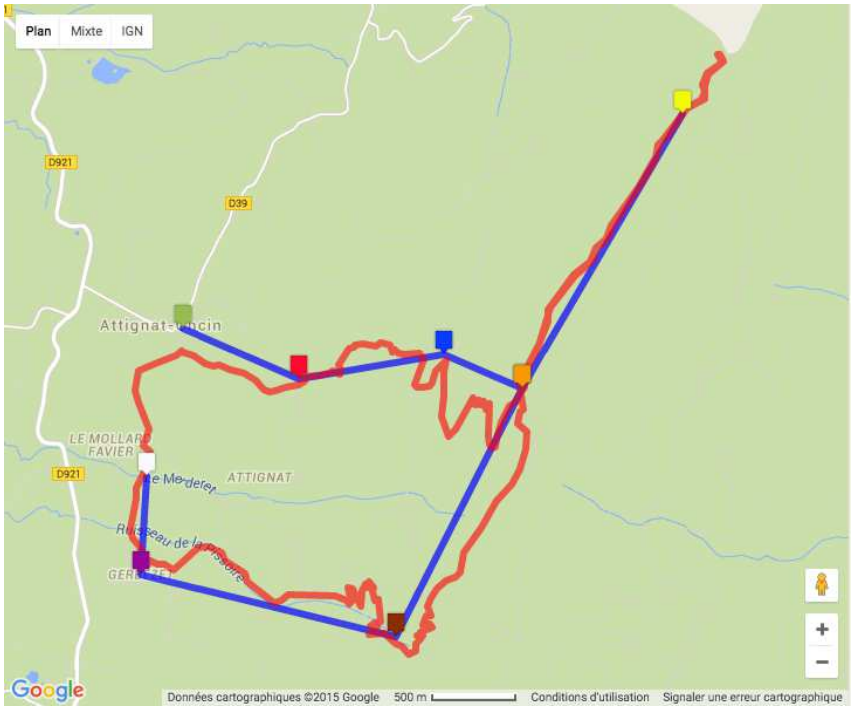


(b) Map-based representation

Figure D.3: Results of the PERDIDO processing chain

Se rendre dans la commune d' Attignat Oncin et stationner au parking du cimetière A partir du parking prendre la route goudronnée qui monte au Hameau de La Fauchère le traverser et gagner les réservoirs bien visibles au dessus du hameau Trouver le sentier balisé par un panneau en bois le sentier monte rudement jusqu' à la Rousse 960m et 45 minutes Traverser la piste et reprendre légèrement plus haut le sentier En suivant toujours le balisage jaune on va de sentiers en portions de pistes parvenir à l' altitude 1360m Sommet de l' Archelle en 1h45 depuis le départ et 750m de dénivelée La table d' orientation ayant disparu on va se diriger par le sentier de crête vers le belvédère du Mont Grêle 1410m ce sentier de 18km était défoncé sur le de sa longueur par les engins de débardage Le belvédère est ainsi atteint en 40 minutes on y trouve quelques explications sur la formation du massif de l' Epine et du lac d' Aiguebelette joli panorama sur la Chartreuse et les deux lacs Revenir à l' Archelle puis parcourir le pâturage en friche et profiter du panorama sur les sommets de Chartreuse en particulier Grand Som et Petit Som qui semblent tout proches Le sentier de descente par le Chemin de la Pissoire est bien balisé il indique deux heures et 59km pour rejoindre Attignat Oncin Ce chemin herbeux devient rapidement plus difficile raviné d' une part et défoncé par des 4X4 d' autre part A l' altitude 1170m Chemin de la Pissoire prendre la direction la Pissoire c' est un sentier pentu et glissant qui va longer et traverser le ruisseau de la Pissoire au niveau d' une petite cascade On atteint une route forestière que l' on quitte trois lacets plus bas pour reprendre le sentier d'abord et le chemin ensuite tous deux suivent le cours de la Pissoire ils vont nous amener jusqu' à Gerbezet 630m à partir de là on suivra le GR9 en suivant la route d'abord et un chemin qui va longer et traverser des prés Il faudra traverser le ruisseau du Merderet et peu après on atteint une petite route qu' on descendra à gauche puis emprunter la route plus importante à droite jusqu' au village

(a) Annotated text



(b) Map-based representation

Figure D.4: Results of the PERDIDO processing chain

Appendix E

External links

This appendix lists the resources provided during the PERDIDO project. These resources have been described in this dissertation and are available online.

- Specification files of the PERDIDO markup language described in Chapter 5:
  - ODD: http://erig.univ-pau.fr/PERDIDO/ns/Perdido_odd.xml
  - XML Schema: http://erig.univ-pau.fr/PERDIDO/ns/Perdido.xsd
  - DTD: http://erig.univ-pau.fr/PERDIDO/ns/Perdido.dtd
- Online PERDIDO demonstration tool described in Section 6.4:
  - http://erig.univ-pau.fr/PERDIDO/
- Video demonstrating the PERDIDO application described in Section 6.4:
  - http://erig.univ-pau.fr/PERDIDO/demo/Perdido.mp4
- Video demonstrating the manual annotation tool described in Chapter 7:
  - http://erig.univ-pau.fr/PERDIDO/demo/TextTagging.mp4

Acronyms

ACE Automatic Content Extraction.

ACH Association of Computers in the Humanities.

AHP Analytic Hierarchy Process.

AI Artificial Intelligence.

ALLC Association of Literary and Linguistic Computing.

CRF Conditional Random Fields.

DAG Directed Acyclic Graph.

DBMS Database Management Systems.

DBSCAN Density-Based Spatial Clustering of Applications with Noise.

DCC Double Cross Calculus.

DFS Depth First Search.

EGN EuroGeoNames.

ENE Expanded Named Entity.

ESNE Expanded Spatial Named Entity.

GIR Geographical Information Retrieval.

GIS Geographic Information System.

GML Geography Markup Language.

GPS Global Positioning System.

GPX GPS Exchange Format.

GUM Generalized Upper Model.

HMM Hidden Markov Models.

HSS Holistic Spatial Semantic.

HTML Hypertext Markup Language.

HTTP HyperText Transfer Protocol.

IE Information Extraction.

INSPIRE INfrastructure for SPatial InfoRmation in Europe.

IR Information Retrieval.

KML Keyhole Markup Language.

ME Maximum Entropy Models.

MUC Message Understanding Conference.

NE Named Entity.

NER Named Entity Recognition.

NERC Named Entity Recognition and Classification.

NLP Natural Language Processing.

NMCA National Mapping and Cadastral Agency.

ODD One Document Does it all.

OGC Open Geospatial Consortium.

OSM OpenStreeMap.

POI Point of Interest.

POS Part-of-speech.

RCC Region Connection Calculus.

RDF Resource Description Framework.

REST Representational State Transfer.

SABE Seamless Administrative Boundaries of Europe.

SER Slot Error Rate.

SGML Standard Generalized Markup Language.

SpRL Spatial Role Labeling.

SSH Spatial Semantic Hierarchy.

STML Spatio-Temporal Markup Language.

SVM Support Vector Machines.

TEI Text Encoding Initiative.

TGML Temporal Geographical Markup Language.

TGN Thesaurus of Geographic Names.

TRML Toponym Resolution Markup Language.

UNECE United Nations Economic Commission for Europe.

UNGEGN United Nations Group of Experts on Geographical Names.

URI Uniform Resource Identifier.

W3C World Wide Web Consortium.

WCS Web Coverage Service.

WFS Web Feature Service.

WMS Web Map Service.

WMTS Web Map Tile Service.

WSD Word Sense Disambiguation.

WSM Weighted Sum Model.

XML Extensible Markup Language.

Bibliography

Abadie, N. and Mustière, S. (2010). Constitution et exploitation d'une taxonomie géographique à partir des spécifications de bases de données. *Revue Internationale de Géomatique*, 20(2):145–177.

Abeillé, A., Clément, L., and Toussenel, F. (2003). Building a treebank for french. In Abeillé, A., editor, *Treebanks*, number 20 in Text, Speech and Language Technology, pages 165–187. Springer Netherlands.

Abney, S. (1996). Partial parsing via finite-state cascades. *Natural Language Engineering*, 2(04):337–344.

Agrawal, R. J. and Shanahan, J. G. (2010). Location disambiguation in local searches using gradient boosted decision trees. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '10, pages 129–136.

Ahern, S., Naaman, M., Nair, R., and Yang, J. H.-I. (2007). World Explorer: Visualizing Aggregate Data from Unstructured Text in Geo-referenced Collections. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '07, pages 1–10, New York, NY, USA. ACM.

Aji, A., Sun, X., Vo, H., Liu, Q., Lee, R., Zhang, X., Saltz, J. H., and Wang, F. (2013). Demonstration of Hadoop-GIS: a spatial data warehousing system over MapReduce. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 518–521.

Allen, G. L. (1997). From Knowledge to Words to Wayfinding: Issues in the Production and Comprehension of Route Directions. In *Proceedings of the International Conference on Spatial Information Theory: A Theoretical Basis for GIS*, COSIT '97, pages 363–372, London, UK, UK. Springer-Verlag.

Allen, J. F. (1983). Maintaining Knowledge About Temporal Intervals. *Commun. ACM*, 26(11):832–843.

Amitay, E., Har'El, N., Sivan, R., and Soffer, A. (2004). Web-a-where: Geotagging Web Content. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pages 273–280, New York, NY, USA. ACM.

Anders, K.-H. and Sester, M. (2000). Parameter-free cluster detection in spatial databases and its application to typification. *International Archives of Photogrammetry and Remote Sensing*, 33(B4/1; PART 4):75–83.

Asher, N., Muller, P., and Gaio, M. (2008). Spatial entities are temporal entities too: the case of motion verbs. In *Spatial entities are temporal entities too: the case of motion verbs*, pages 16–21, Marrakech, Maroc.

Aurnague, M. (2011). How motion verbs are spatial: The spatial foundations of intransitive motion verbs in french. *Lingvisticae Investigationes*, 34(1):1–34.

Aurnague, M. and Vieu, L. (2015). Function vs. regions in spatial language: a fundamental distinction. In Astésano, C. and Jucla, M., editors, *Neuropsycholinguistic perspectives on language cognition. Essays in honour of Jean-Luc Nespoulous*, volume 4 of *Explorations in Cognitive Psychology*, pages 31–45. Psychology Press.

Aurnague, M., Vieu, L., and Borillo, A. (2010). *Langage et cognition spatiale, Sciences Cognitives*, chapter La représentation formelle des concepts spatiaux dans la langue, pages 69–102. Denis, M.

Barba-Romero, S. (2001). The Spanish Government Uses a Discrete Multicriteria DSS to Determine Data-Processing Acquisitions. *Interfaces*, 31(4):123–131.

Bartha, G. and Kocsis, S. (2011). Standardization of geographic data: The european inspire directive. *European Journal of Geography*, 2(2):79–89.

Bateman, J. A., Hois, J., Ross, R., and Tenbrink, T. (2010). A linguistic ontology of space for natural language processing. *Artificial Intelligence*, 174(14):1027–1071.

Batista, D. S., Ferreira, J. D., Couto, F. M., and Silva, M. J. (2012). Toponym disambiguation using ontology-based semantic similarity. In Caseli, H., Villavicencio, A., Teixeira, A., and Perdigão, F., editors, *Computational Processing of the Portuguese Language*, number 7243 in Lecture Notes in Computer Science, pages 179–185. Springer Berlin Heidelberg.

Béchet, N., Chauché, J., Prince, V., and Roche, M. (2014). How to combine text-mining methods to validate induced verb-object relations? *Comput. Sci. Inf. Syst.*, 11(1):133–155.

Belouaer, L., Brosset, D., and Claramunt, C. (2013). Modeling spatial knowledge from verbal descriptions. In Tenbrink, T., Stell, J., Galton, A., and Wood, Z., editors, *Spatial Information Theory*, number 8116 in Lecture Notes in Computer Science, pages 338–357. Springer International Publishing.

Bensalem (2010). Toponym disambiguation by arborescent relationships. *Journal of Computer Science*, 6(6):653–659.

Berretti, S., Del Bimbo, A., and Vicario, E. (2003). Weighted walkthroughs between extended entities for retrieval by spatial arrangement. *IEEE Transactions on Multimedia*, 5(1):52–70.

Berthele, R. (2004). The typology of motion and posture verbs: A variationist account. *TRENDS IN LINGUISTICS STUDIES AND MONOGRAPHS*, 153:93–126.

Bikel, D. M., Miller, S., Schwartz, R., and Weischedel, R. (1997). Nymble: A High-performance Learning Name-finder. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, ANLC '97, pages 194–201, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bizer, C., Tom, H., and Berners-Lee, T. (2009). Linked data: the story so far. *International Journal on Semantic Web and Information System (IJSWIS)*, 5(3):1–22.

Bloom, P. (1994). *Language and space*. MIT press.

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

Borillo, A. (1990). A propos de la localisation spatiale. *Langue française*, 86(1):75–84.

Borillo, A. (1998). *L'espace et son expression en français, L'essentiel*. Orphrys.

Borillo, A. (2004). Quand les adverbiaux de localisation spatiale constituent des facteurs d'enchaînement spatio-temporel dans le discours. In *Colloque Chronos 6*, pages 123–138, Genève.

Borthwick, A. (1998). *A maximum entropy approach to named entity recognition*. PhD thesis, New York University.

Breier, M. (2013). The way is the goal–modelling of historical roads.

Brosset, D., Claramunt, C., and Saux, E. (2008). Wayfinding in Natural and Urban Environments: A Comparative Study. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 43(1):21–30.

Burggraf, D. S. (2006). Geography markup language. *Data Science Journal*, 5:178–204.

Buscaldi, D. (2011). Approaches to disambiguating toponyms. *SIGSPATIAL Special*, 3(2):16–19.

Buscaldi, D. and Magnini, B. (2010). Grounding toponyms in an italian local news corpus. In *Proceedings of the 6th Workshop on Geographic Information Retrieval*, GIR '10, pages 15:1–15:5, New York, NY, USA. ACM.

Buscaldi, D. and Rosso, P. (2008a). A conceptual density-based approach for the disambiguation of toponyms. *Int. J. Geogr. Inf. Sci.*, 22(3):301–313.

Buscaldi, D. and Rosso, P. (2008b). Map-based vs. knowledge-based toponym disambiguation. In *Proceedings of the 2nd international workshop on Geographic information retrieval*, GIR '08, pages 19–22, New York, NY, USA. ACM.

Béchet, F., Sagot, B., and Stern, R. (2011). Coopération de méthodes statistiques et symboliques pour l'adaptation non-supervisée d'un système d'étiquetage en entités nommées. In *TALN'2011 - Traitement Automatique des Langues Naturelles*.

Carreras, X. and Màrquez, L. (2005). Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, CONLL '05, pages 152–164, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chomsky, N. (1965). *Aspects of the theory of syntax*. M.I.T. Press, Oxford, England.

Claramunt, C., Parent, C., Spaccapietra, S., and Thériault, M. (1999). Database Modelling for Environmental and Land Use Changes. In Stillwell, D. J., Geertman, P. S., and Openshaw, D. S., editors, *Geographical Information and Planning*, Advances in Spatial Science, pages 181–202. Springer Berlin Heidelberg.

Clementini, E. (2009). *A Conceptual Framework for Modelling Spatial Relations*. PhD thesis, Institut National des Sciences Appliquées de Lyon, Lyon.

Clementini, E. and Cohn, A. (2014). Rcc*-9 and cbm*. In Duckham, M., Pebesma, E., Stewart, K., and Frank, A., editors, *Geographic Information Science*, volume 8728 of *Lecture Notes in Computer Science*, pages 349–365. Springer International Publishing.

Clementini, E., Felice, P. D., and Hernández, D. (1997). Qualitative representation of positional information. *Artificial Intelligence*, 95(2):317–356.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537.

Constant, M. (2003). *Grammaires locales pour l'analyse automatique de textes : méthodes de construction et outils de gestion*. PhD thesis, Université Paris-Est.

De Felice, G., Fogliaroni, P., and Wallgrün, J. O. (2011). A Hybrid Geometric-qualitative Spatial Reasoning System and Its Application in GIS. In *Proceedings of the 10th International Conference on Spatial Information Theory*, COSIT'11, pages 188–209, Berlin, Heidelberg. Springer-Verlag.

Delling, D., Sanders, P., Schultes, D., and Wagner, D. (2009). Engineering Route Planning Algorithms. In Lerner, J., Wagner, D., and Zweig, K. A., editors, *Algorithmics of Large and Complex Networks*, pages 117–139. Springer-Verlag, Berlin, Heidelberg.

DeLozier, G., Baldridge, J., and London, L. (2015). Gazetteer-independent toponym resolution using geographic word profiles. In *Proceedings of AAAI 2015*, Austin, Texas, USA.

Denis, M. (1997). The description of routes: a cognitive approach to the production of spatial discourse. *Cahiers de Psychologie Cognitive*, 16:409–458.

Derungs, C. and Purves, R. S. (2014). From text to landscape: locating, identifying and mapping the use of landscape features in a swiss alpine corpus. *International Journal of Geographical Information Science*, 28(6):1272–1293.

Dutoit, T. (1997). *An Introduction to Text-to-Speech Synthesis*. Springer Science & Business Media.

Egan, T. (2015). Manner and Path: evidence from a multilingual corpus. *CogniTextes. Revue de l'Association française de linguistique cognitive*, (Volume 12).

Egenhofer, M. and Franzosa, R. (1991). Point-set topological spatial relations. *International journal for Geographical Information Systems*, 5(2):161–174.

Egenhofer, M. J. (1991). Reasoning About Binary Topological Relations. In *Proceedings of the Second International Symposium on Advances in Spatial Databases*, SSD '91, pages 143–160, London, UK, UK. Springer-Verlag.

Egenhofer, M. J. (2002). Toward the Semantic Geospatial Web. In *Proceedings of the 10th ACM International Symposium on Advances in Geographic Information Systems*, GIS '02, pages 1–4, New York, NY, USA. ACM.

Egenhofer, M. J. and Shariff, A. R. B. M. (1998). Metric Details for Natural-language Spatial Relations. *ACM Trans. Inf. Syst.*, 16(4):295–321.

Ehrmann, M. (2008). *Les entités nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation*. PhD thesis, Paris 7 - Denis Diderot.

Eldawy, A., Li, Y., Mokbel, M. F., and Janardan, R. (2013). Cg hadoop: computational geometry in mapreduce. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 284–293.

Ester, M., Kriegel, H.-P., Sander, J., Wimmer, M., and Xu, X. (1998). Incremental clustering for mining in a data warehousing environment. In *VLDB*, volume 98, pages 323–333.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pages 226–231.

Fellbaum, C. (2012). *WordNet*. Blackwell Publishing Ltd.

Feuerhake, U. and Sester, M. (2013). Mining group movement patterns. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 510–513.

Florczyk, A. J., Lopez-Pellicer, F. J., Muro-Medrano, P. R., Nogueras-Iso, J., and Zarazaga-Soria, F. J. (2010). Semantic selection of georeferencing services for urban management. *Journal of Information Technology in Construction*, 15 (Special Issue Bringing urban ontologies into practice):111–121.

Fogliaroni, P. and Clementini, E. (2015). Modeling visibility in 3d space: A qualitative frame of reference. In *3D Geoinformation Science*, pages 243–258. Springer.

Frank, A. U. (1991). Qualitative spatial reasoning with cardinal directions. In *Proceedings of the Seventh Austrian Conference on Artificial Intelligence*, pages 157–167. Springer.

Frank, A. U. (1998). Formal Models for Cognition - Taxonomy of Spatial Location Description and Frames of Reference. In *Spatial Cognition, An Interdisciplinary Approach to Representing and Processing Spatial Knowledge*, pages 293–312, London, UK, UK. Springer-Verlag.

Frank, A. U. and Mark, D. M. (1991). *Language Issues for Geographical Information Systems*.

Freksa, C. (1992). Using orientation information for qualitative spatial reasoning. In Frank, A. U., Campari, I., and Formentini, U., editors, *Theories and methods of spatio-temporal reasoning in geographic space*, volume 639 of *LNCS*, pages 162–178, Berlin. Springer.

Friburger, N. (2002). *Reconnaissance automatique des noms propres: application à la classification automatique de textes journalistiques*. Thèse doctorat, Université François-Rabelais, Tours, France.

Friburger, N. and Maurel, D. (2004). Finite-state transducer cascades to extract named entities in texts. *Theoretical Computer Science*, 313(1):93–104.

Fu, G., Jones, C. B., and Abdelmoty, A. I. (2005). Building a Geographical Ontology for Intelligent Spatial Search on the Web. ACTA Press.

Gaio, M. and Madelaine, J. (1996). Un modèle de l'illusion perceptive : simulation versus expérimentation. *Intellectica*, (22):67–91.

Gaio, M., Sallaberry, C., Etcheverry, P., Marquesuzaà, C., and Lesbegueries, J. (2008). A global Process to Access Documents' Contents from a Geographical Point of View. *Journal of Visual Languages & Computing*, 19(1):03–23.

Gaio, M., Sallaberry, C., and Nguyen, V. T. (2012). Typage de noms toponymiques à des fins d'indexation géographique. *TAL*, 53(2):1–35.

Galliano, S., Gravier, G., and Chaubard, L. (2009). The ester 2 evaluation campaign for the rich transcription of french radio broadcasts. In *In In: Proceddings of Interspeech, Brighton (United Kingdom*.

Garbin, E. and Mani, I. (2005). Disambiguating toponyms in news. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 363–370, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gildea, D. and Jurafsky, D. (2002). Automatic Labeling of Semantic Roles. *Comput. Linguist.*, 28(3):245–288.

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221.

Gregory, I., Donaldson, C., Murrieta-Flores, P., and Rayson, P. (2015). Geoparsing, GIS, and Textual Analysis: Current Developments in Spatial Humanities Research. *International Journal of Humanities and Arts Computing*, 9(1):1–14.

Grishman, R. and Sundheim, B. (1996). Message Understanding Conference-6: A Brief History. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 1*, COLING '96, pages 466–471, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gross, M. (1997). The construction of local grammars. *Finite-State Language Processing*, pages 329–354.

Guerrero Nieto, M., García Rodríguez, M. J., Urrutia Zambrana, A., Vaca, S., Libardo, W., Poveda, B., and Angel, M. (2010). Incorporating timeml into a gis. *International Journal of Computational Linguistics and Applications*, 1(1-2):269–283.

Götze, J. and Boye, J. (2015). "Turn Left" Versus "Walk Towards the Café": When Relative Directions Work Better Than Landmarks. In Bacao, F., Santos, M. Y., and Painho, M., editors, *AGILE 2015*, Lecture Notes in Geoinformation and Cartography, pages 253–267. Springer International Publishing.

Güting, R. H., de Almeida, T., and Ding, Z. (2006). Modeling and Querying Moving Objects in Networks. *The VLDB Journal*, 15(2):165–190.

Habib, M. and Van Keulen, M. (2012). Improving toponym disambiguation by iteratively enhancing certainty of extraction. In *KDIR 2012 - Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*.

Hahn, U. and Mani, I. (2000). The challenges of automatic summarization. *Computer*, 33(11):29–36.

Haklay, M. (2010). How good is volunteered geographical information? a comparative study of openstreetmap and ordnance survey datasets. *Environment and Planning B Planning and Design*, (37):682–703.

Hao, Q., Cai, R., Wang, C., Xiao, R., Yang, J.-M., Pang, Y., and Zhang, L. (2010). Equip tourists with knowledge mined from travelogues. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 401–410, New York, NY, USA. ACM.

Hernandez, D. (1994). *Qualitative Representation of Spatial Knowledge*. Springer-Verlag New York, Inc.

Hernández, D. (1993). Maintaining Qualitative Spatial Knowledge. In Frank, A. U. and Campari, I., editors, *COSIT'93*, volume 761 of *LNCS*, pages 33–53. Springer-Verlag.

Hill, L. L. (2000). Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints. In *Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL '00, pages 280–290, London, UK, UK. Springer-Verlag.

Hirschman, L. and Gaizauskas, R. (2001). Natural Language Question Answering: The View from Here. *Nat. Lang. Eng.*, 7(4):275–300.

Hollenstein, L. and Purves, R. (2010). Exploring place through user-generated content: using flickr todescribe city cores. *Journal of Spatial Information Science*, (1).

Hornsby, K. and Egenhofer, M. J. (2002). Modeling Moving Objects over Multiple Granularities. *Annals of Mathematics and Artificial Intelligence*, 36(1-2):177–194.

Iacobini, C. (2009). The role of dialects in the emergence of Italian phrasal verbs. *Morphology*, 19(1):15–44.

INSPIRE (2014). INSPIRE Data Specification on Geographical Names — Technical Guidelines. Technical Report D2.8.1.3, European Commission.

Intagorn, S. and Lerman, K. (2011). Learning boundaries of vague places from noisy annotations. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 425–428. ACM.

Ioannis, G., Peter, K., Martin, R., Kai-Florian, R., and Tyler, T. (2014). Wayfinding decision situations: A conceptual model and evaluation. In *Eighth International Conference on Geographic Information Science (GIScience 2014)*, Geographic Information Science, pages 221–234, Vienna, Austria.

Ireson, N. and Ciravegna, F. (2010). Toponym resolution in social media. In *The Semantic Web–ISWC 2010*, pages 370–385. Springer.

ISO (2003). Geographic information — Spatial referencing by geographic identifiers. ISO 19112::2003, International Organization for Standardization, Geneva, Switzerland.

Jackendoff, R. (2012). Language as a source of evidence for theories of spatial representation. *Perception*, 41(9):1128–1152.

Jardine, N. and van Rijsbergen, C. J. (1971). The Use of Hierarchic Clustering in Information Retrieval. *Information Storage and Retrieval*, 7(5):217–240.

Jonasson, K. (1994). *Le nom propre*. Duculot, Belgique, Louvain-la-Neuve.

Jones, C. B. and Purves, R. S. (2008). Geographical information retrieval. *International Journal of Geographical Information Science*, 22(3):219–228.

Jones, C. B., Purves, R. S., Clough, P. D., and Joho, H. (2008). Modelling vague places with knowledge from the web. *Int. J. Geogr. Inf. Sci.*, 22(10):1045–1065.

Kao, A. and Poteet, S. R. (2006). *Natural Language Processing and Text Mining*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Karger, D., Motwani, R., and Ramkumar, G. D. S. (1997). On approximating the longest path in a graph. *Algorithmica*, 18(1):82–98.

Kemmerer, D. (2005). The spatial and temporal meanings of English prepositions can be independently impaired. *Neuropsychologia*, 43(5):797–806.

Kim, J., Sridhara, V., and Bohacek, S. (2009). Realistic mobility simulation of urban mesh networks. *Ad Hoc Networks*, 7(2):411–430.

Kitchin, R. (2013). Big data and human geography Opportunities, challenges and risks. *Dialogues in Human Geography*, 3(3):262–267.

Kokashvili, S. (2012). *Syntaxe et sémantique des verbes de déplacement, de mouvement et de position en français et en géorgien modernes.* Paris 4.

Kordjamshidi, P., Bethard, S., and Moens, M.-F. (2012). SemEval-2012 task 3: Spatial role labeling. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, pages 365–373, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kordjamshidi, P., Van Otterlo, M., and Moens, M.-F. (2011). Spatial role labeling: Towards extraction of spatial relations from natural language. *ACM Trans. Speech Lang. Process.*, 8(3):4:1–4:36.

Krieg-Brückner, B. and Shi, H. (2006). Orientation Calculi and Route Graphs: Towards Semantic Representations for Route Descriptions. In Raubal, Miller, H. J., Frank, A. U., and Goodchild, M. F., editors, *Proceedings of the 4th International Conference on Geographic Information Science*, GIScience'06, pages 234–250, Berlin, Heidelberg. Springer-Verlag.

Kuipers, B. (2000). The Spatial Semantic Hierarchy. *Artificial Intelligence*, 119(1–2):191–233.

Landau, B. and Jackendoff, R. (1993). "What" and "where" in spatial language and spatial cognition. *Behavioral and Brain Sciences*, 16(02):217–238.

Langran, G. (1992). *Time in geographic information systems*. CRC Press.

Laube, P., Imfeld, S., and Weibel, R. (2005). Discovering relative motion patterns in groups of moving point objects. *International Journal of Geographical Information Science*, 19(6):639–668.

Laur, D. (1991). *Sémantique du déplacement et de la localisation en français: une étude des verbes, des prépositions et de leurs relations dans la phrase simple*. PhD thesis, A.N.R.T, Lille.

Laur, D. (1993). La relation entre le verbe et la préposition dans la sémantique du déplacement. *Langages*, 27(110):47–67.

Lauritzen, S. L. and Wermuth, N. (1989). Graphical Models for Associations between Variables, some of which are Qualitative and some Quantitative. *The Annals of Statistics*, 17(1):31–57.

Lehto, L., Latvala, P., and Kähkönen, J. (2013). An Implementation of the OGC's WFS Gazetteer Service Application Profile. pages 11–14.

Leidner, J. L. (2004). *Towards a Reference Corpus for Automatic Toponym Resolution Evaluation*.

Leidner, J. L. (2006). An evaluation dataset for the toponym resolution task. *Computers, Environment and Urban Systems*, 30(4):400–417.

Leidner, J. L. (2007). *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. Universal-Publishers.

Leidner, J. L. and Lieberman, M. D. (2011). Detecting geographical references in the form of place names and associated spatial natural language. *SIGSPATIAL Special*, 3(2):5–11.

Lejeune, S., Pierre, P., Grelot, J.-P., Patrice, F., and Mathurin, R. (2003). *CHARTE DE TOPONYME - Toponymie du territoire français*. Institut National de l'Information Géographique et Forestière (IGN).

Lesbegueries, J. (2007). *Plate-forme pour l'indexation spatiale multi-niveaux d'un corpus territorialisé*. PhD thesis, Université de Pau et des Payse de l'Adour, Pau.

Lesbegueries, J., Gaio, M., and Loustau, P. (2006). Geographical information access for non-structured data. In *Proceedings of the 2006 ACM Symposium on Applied Computing (SAC)*, pages 83–89, Dijon, France. ACM.

Levinson, S. C. (1996). Language and space. *Annual Review of Anthropology*, 25(1):353–382.

Levinson, S. C. (2003). *Space in language and cognition: explorations in cognitive diversity*. Number 5 in Language, culture, and cognition. Cambridge University Press, Cambridge ; New York.

Lewis, D. D. and Jones, K. S. (1996). Natural language processing for information retrieval. *Commun. ACM*, 39(1):92–101.

Li, R., Fuest, S., and Schwering, A. (2014). The effects of different verbal route instructions on spatial orientation. In Huerta Guijarro, J., Schade, S., and Granell Canut, C., editors, *the 17th AGILE conference on geographic information science, Castellon, Spain*. Springer.

Li, Y., Moffat, A., Stokes, N., and Cavedon, L. (2006). Exploring probabilistic toponym resolution for geographical information retrieval. In *3rd Workshop on Geographic Information Retrieval (GIR*, pages 17–22.

Lieberman, M. D. and Samet, H. (2012). Adaptive context features for toponym resolution in streaming news. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 731–740. ACM.

Lieberman, M. D., Samet, H., and Sankaranarayanan, J. (2010). Geotagging with local lexicons to build indexes for textually-specified spatial data. In *Data Engineering (ICDE), 2010 IEEE 26th International Conference on*, pages 201–212. IEEE.

Ligozat, G. (1998). Reasoning about Cardinal Directions. *J. Vis. Lang. Comput.*, 9(1):23–44.

Lopez-Pellicer, F. J., Lacasta, J., Florczyk, A., Nogueras-Iso, J., and Zarazaga-Soria, F. J. (2012). An ontology for the representation of spatiotemporal jurisdictional domains in information retrieval systems. *International Journal of Geographical Information Science*, 26(4):579–597.

Loustau, P. (2008). *Interprétation automatique d'itinéraires dans des récits de voyages.* PhD thesis, Université de Pau et des Pays de l'Adour.

Loustau, P., Gaio, M., and Nodenot, T. (2008). Interprétation automatique d'itinéraires à partir d'un corpus de récits de voyages pilotée par un usage pédagogique. *RNTI*, E(13):177, 206.

Lynch, K. (1960). *The image of the city*, volume 11. MIT press.

MacMahon, M., Stankiewicz, B., and Kuipers, B. (2006). Walk the Talk: Connecting Language, Knowledge, and Action in Route Instructions. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*, AAAI'06, pages 1475–1482, Boston, Massachusetts. AAAI Press.

Makhoul, J., Kubala, F., Schwartz, R., and Weischedel, R. (1999). Performance measures for information extraction. In *In Proceedings of DARPA Broadcast News Workshop*, pages 249–252.

Mani, I. (1999). *Advances in Automatic Text Summarization*. MIT Press, Cambridge, MA, USA.

Mani, I., Hitzeman, J., Richer, J., Harris, D., Quimby, R., and Wellner, B. (2008). Spatialml: Annotation scheme, corpora, and tools. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), European Language Resources Association (ELRA)*, Marrakech, Morocco.

Mani, I. and MacMillan, T. R. (1996). Identifying unknown proper names in newswire text. In Boguraev, B. and Pustejovsky, J., editors, *Corpus Processing for Lexical Acquisition*, pages 41–59. MIT Press, Cambridge, MA, USA.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Martins, B., Anastácio, I., and Calado, P. (2010). A Machine Learning Approach for Resolving Place References in Text. In Painho, M., Santos, M. Y., and Pundt, H., editors, *Geospatial Thinking*, number 0 in Lecture Notes in Geoinformation and Cartography, pages 221–236. Springer Berlin Heidelberg.

Martins, B., Silva, M. J., and Chaves, M. S. (2005). Challenges and Resources for Evaluating Geographical IR. In *Proceedings of the 2005 Workshop on Geographic Information Retrieval*, GIR '05, pages 65–69, New York, NY, USA. ACM.

Maurel, D., Friburger, N., Antoine, J.-Y., Eshkol-Taravella, I., and Nouvel, D. (2011). Cascades de transducteurs autour de la reconnaissance des entités nommées. *TAL*, 52(1):69–96.

Maurel, D., Spędzia-Baron, M., Bouchou-Markhoff, B., and Vitas, D. (2014). Prolexbase. A Multilingual Relational Database of Proper Names. *Cahiers de Linguistique.*, 40(2):49–71.

McCallum, A. and Li, W. (2003). Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-enhanced Lexicons. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 188–191, Stroudsburg, PA, USA. Association for Computational Linguistics.

McDonald, D. D. (1996). Corpus Processing for Lexical Acquisition. pages 21–39. MIT Press, Cambridge, MA, USA.

Michon, P.-E. and Denis, M. (2001). When and Why Are Visual Landmarks Used in Giving Directions? In Montello, D. R., editor, *Proceedings of the International Conference on Spatial Information Theory: Foundations of Geographic Information Science*, COSIT 2001, pages 292–305, London, UK, UK. Springer-Verlag.

Mikheev, A., Grover, C., and Moens, M. (1998). Description of the LTG system used for MUC-7. In *In Proceedings of 7th Message Understanding Conference (MUC-7*.

Mikheev, A., Moens, M., and Grover, C. (1999). Named entity recognition without gazetteers. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, EACL '99, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Miller, G. A. and Johnson-Laird, P. N. (1976). *Language and perception*, volume viii. Belknap Press, Cambridge, MA, England.

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill series in computer science. McGraw-Hill, New York, NY [u.a.], international ed., [reprint.] edition.

Moncla, L., Gaio, M., and Mustière, S. (2014a). Automatic itinerary reconstruction from texts. In *Eighth International Conference on Geographic Information Science (GIScience 2014)*, Geographic Information Science, pages 253–267, Vienna, Austria.

Moncla, L., Gaio, M., Nogueras-Iso, J., and Mustière, S. (2015). Reconstruction of itineraries from annotated text with an informed spanning tree algorithm. *International Journal of Geographical Information Science*. (Accepted for publication).

Moncla, L., Renteria-Agualimpia, W., Nogueras-Iso, J., and Gaio, M. (2014b). Geocoding for texts with fine-grain toponyms: An experiment on a geoparsed hiking descriptions corpus. In *Proceedings of the 22Nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '14, pages 183–192, New York, NY, USA. ACM.

Montello, D. R. (2005). *Navigation*, pages 257–294. Cambridge University Press, Cambridge.

Morrow, D. G. and Clark, H. H. (1988). Interpreting words in spatial descriptions. *Language and Cognitive Processes*, 3(4):275–291.

Mouna Snoussi, Jérôme Gensel, and Paule-Annick Davoine (2012). Extending TimeML and SpatialML languages to handle imperfect spatio-temporal information in the context of natural hazards studies. In *Proceedings of the AGILE'2012 International Conference on Geographic Information Science, Avignon*, Avignon.

Muller, P. (1998). A qualitative theory of motion based on spatio-temporal primitives. In *Proceedings of the Sixth International Conference on Knowledge Representation and Reasoning (KR98)*, pages 131–141.

Muller, P. and Tannier, X. (2004). Annotating and measuring temporal relations in texts. In *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics.

Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Navigli, R. (2009). Word Sense Disambiguation: A Survey. *ACM Comput. Surv.*, 41(2):10:1–10:69.

Nedas, K. A., Egenhofer, M. J., and Wilmsen, D. (2007). Metric details of topological line–line relations. *International Journal of Geographical Information Science*, 21(1):21–48.

Nguyen, V. T. (2012). *Méthode d'extraction d'informations géographiques à des fins d'enrichissement d'une ontologie de domaine, Geographical information extraction method in order to enrich a domain ontology.* thèse de doctorat en informatique, Université de Pau et des Pays de l'Adour, Pau.

Nguyen, V. T., Gaio, M., and Moncla, L. (2013). Topographic subtyping of place named entities: a linguistic approach. In Danny Vandenbroucke, Bénédicte Bucher, J. C., editor, *The 15th AGILE International Conference on Geographic Information Science*, pages 1–5, Louvain. Springer.

Nouvel, D., Antoine, J.-Y., and Friburger, N. (2014). Pattern mining for named entity recognition. In Vetulani, Z. and Mariani, J., editors, *Human Language Technology Challenges for Computer Science and Linguistics*, Lecture Notes in Computer Science, pages 226–237. Springer International Publishing.

Nouvel, D., Antoine, J.-Y., Friburger, N., and Soulet, A. (2012). Coupling knowledge-based and data-driven systems for named entity recognition. In *Innovative hybrid approaches to the processing of textual data (HYBRID'12, EACL Workshop, poster).*

Overell, S. and Rüger, S. (2008). Using co-occurrence models for placename disambiguation. *Int. J. Geogr. Inf. Sci.*, 22(3):265–287.

O'Keefe, J. (1996). The spatial prepositions in english, vector grammar, and the cognitive map theory. *Language and space*, pages 277–316.

Padró, L. and Stanilovsky, E. (2012). Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey. ELRA.

Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In Chair), N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Palacio, D. (2010). *Combinaison de critères par contraintes pour la Recherche dÍnformation Géographique*. PhD thesis, Université de Pau et des Pays de l'Adour.

Palmer, M., Gildea, D., and Xue, N. (2010). Semantic Role Labeling. *Synthesis Lectures on Human Language Technologies*, 3(1):1–103.

Pang, B. and Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135.

Paumier, S. (2003). *De la reconnaissance de formes linguistiques à l'analyse syntaxique*. thèse de doctorat en informatique, Université de Marne-la-Vallée.

Pautasso, C., Wilde, E., and Alarcon, R., editors (2014). *REST: Advanced Research Topics and Practical Applications*. Springer New York, New York, NY.

Poibeau, T. (2003). Extraction automatique d'information: du texte brut au web sémantique. In *Extraction automatique d'information: du texte brut au web sémantique*. Hermès Lavoisier.

Poibeau, T. (2011). *Traitement automatique du contenu textuel*. Lavoisier.

Popescu, A., Grefenstette, G., and Moëllic, P. A. (2008). Gazetiki: Automatic Creation of a Geographical Gazetteer. In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '08, pages 85–93, New York, NY, USA. ACM.

Pourcel, S. and Kopecka, A. (2005). Motion expression in French: typological diversity. *Durham & Newcastle working papers in linguistics*, 11:139–153.

Pradhan, S., Hacioglu, K., Ward, W., Martin, J. H., and Jurafsky, D. (2003). Semantic Role Parsing: Adding Semantic Structure to Unstructured Text. In *Proceedings of the Third IEEE International Conference on Data Mining*, ICDM '03, pages 629–632, Washington, DC, USA. IEEE Computer Society.

Prim, R. C. (1957). Shortest connection networks and some generalizations. *Bell system technical journal*, 36(6):1389–1401.

Pustejovsky, J., Knippen, R., Littman, J., and Saurí, R. (2005). Temporal and event information in natural language text. *Language Resources and Evaluation*, 39(2):123–164.

Pustejovsky, J. and Moszkowicz, J. L. (2008). Integrating motion predicate classes with spatial and temporal annotations. In *Proceedings of COLING*, pages 95–98.

Pustejovsky, J., Moszkowicz, J. L., and Verhagen, M. (2012). A linguistically grounded annotation language for spatial information. *TAL*, 53(2):87–113.

Pustejovsky, J. and Yocum, Z. (2013). Capturing motion in ISO-SpaceBank. *Proceedings of the 9th Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pages 25–34.

Qin, T., Xiao, R., Fang, L., Xie, X., and Zhang, L. (2010). An efficient location extraction algorithm by leveraging web contextual information. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '10, pages 53–60.

Rabiner, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Randell, D. A., Cui, Z., and Cohn, A. G. (1992). A spatial logic based on regions and connection. In *3rd International Conference on Knowledge Representation and Reasoning*, pages 165–176, San Mateo. Morgan Kaufmann.

Ratinov, L. and Roth, D. (2009). Design challenges and misconceptions in named entity recognition. In *CoNLL*.

Rattenbury, T., Good, N., and Naaman, M. (2007). Towards Automatic Extraction of Event and Place Semantics from Flickr Tags. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 103–110, New York, NY, USA. ACM.

Rattenbury, T. and Naaman, M. (2009). Methods for extracting place semantics from flickr tags. *ACM Transactions on the Web (TWEB)*, 3(1):1.

Rauch, E., Bukatin, M., and Baker, K. (2003). A Confidence-based Framework for Disambiguating Geographic Terms. In *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References - Volume 1*, HLT-NAACL-GEOREF '03, pages 50–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

Renz, J. (2002). *Qualitative Spatial Reasoning with Topological Information*. Springer-Verlag, Berlin, Heidelberg.

Roberts, K., Adrian Bejan, C., and Harabagiu, S. (2010). Toponym disambiguation using events. In *Proceedings of the Twenty-Third International Florida Artificial Intelligence Research Society Conference (FLAIRS 2010)*, pages 271–276.

Rothenberg, J. (2000). Preserving authentic digital information. *Authenticity in a digital environment*.

Saaty, T. L. (1999). *Decision making for leaders: the analytic hierarchy process for decisions in a complex world*, volume 2. RWS publications.

Santorini, B. (1990). Part-of-speech tagging guidelines for the penn treebank project (3rd revision). Technical report, Department of Computer and Information Science, University of Pennsylvania.

Sarda, L. (2001). L'expression du déplacement dans la construction transitive directe. *Syntaxe et sémantique*, 2(1):121–135.

Sarjakoski, L. T., Kettunen, P., Flink, H.-M., Laakso, M., Rönneberg, M., and Sarjakoski, T. (2011). Analysis of verbal route descriptions and landmarks for hiking. *Personal and Ubiquitous Computing*, 16(8):1001–1011.

Scheider, S. and Purves, R. (2013). Semantic place localization from narratives. In *Proceedings of The First ACM SIGSPATIAL International Workshop on Computational Models of Place*, COMP '13, pages 16:16–16:19, New York, NY, USA. ACM.

Schilder, F. and Habel, C. (2001). From Temporal Expressions to Temporal Information: Semantic Tagging of News Messages. In *Proceedings of the Workshop on Temporal and Spatial Information Processing - Volume 13*, TASIP '01, pages 9:1–9:8, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Sekine, S., Sudo, K., and Nobata, C. (2002). Extended named entity hierarchy. In *Proceedings of the Third International Conference on Language Resources and Evaluation, LREC 2002, May 29-31, 2002, Las Palmas, Canary Islands, Spain*.

Serdyukov, P., Murdock, V., and Van Zwol, R. (2009). Placing flickr photos on a map. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 484–491. ACM.

Sinnott, R. (1984). Virtues of the haversine. *Sky and Telescope*, 68(2):159.

Slobin, D. I. (1996). Two ways to travel: Verbs of motion in english and spanish. *Grammatical constructions: Their form and meaning*, pages 195–219.

Smart, P. D., Jones, C., and Twaroch, F. (2010). Multi-source toponym data integration and mediation for a meta-gazetteer service. In Fabrikant, S., Reichenbacher, T., Kreveld, M., and Schlieder, C., editors, *Geographic Information Science*, volume 6292 of *Lecture Notes in Computer Science*, pages 234–248. Springer Berlin Heidelberg.

Smith, D. A. and Crane, G. (2001). Disambiguating Geographic Names in a Historical Digital Library. In *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL '01, pages 127–136, London, UK, UK. Springer-Verlag.

Smith, D. A. and Mann, G. S. (2003). Bootstrapping toponym classifiers. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references - Volume 1*, HLT-NAACL-GEOREF '03, pages 45–49, Stroudsburg, PA, USA. Association for Computational Linguistics.

Spaccapietra, S., Parent, C., Damiani, M. L., de Macedo, J. A., Porto, F., and Vangenot, C. (2008). A conceptual view on trajectories. *Data & Knowledge Engineering*, 65(1):126–146.

Speriosu, M. and Baldridge, J. (2013). Text-driven toponym resolution using indirect supervision. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1466—-1476, Sofia, Bulgaria. ACL.

Sui, D. and Goodchild, M. (2011). The convergence of GIS and social media: challenges for GIScience. *International Journal of Geographical Information Science*, 25(11):1737–1748.

Szarvas, G., Farkas, R., and Kocsor, A. (2006). A Multilingual Named Entity Recognition System Using Boosting and C4.5 Decision Tree Learning Algorithms. In Todorovski, L., Lavrač, N., and Jantke, K. P., editors, *Discovery Science*, number 4265 in Lecture Notes in Computer Science, pages 267–278. Springer Berlin Heidelberg.

Takeuchi, K. and Collier, N. (2002). Use of Support Vector Machines in Extended Named Entity Recognition. In *Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, COLING-02, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.

Talmy, L. (1983). *How language structures space*. Number 4 in Berkeley cognitive science report. Cognitive Science Program, Institute of Cognitive Studies, University of California at Berkeley, Berkeley, CA, Etats-Unis.

Talmy, L. (1985). *Lexicalization patterns: Semantic structure in lexical forms. Language typology and syntactic description, vol. 3, Grammatical categories and the lexicon, ed. by Timothy Shopen, 57–149.* Cambridge: Cambridge University Press.

Talmy, L. (2000). *Toward a Cognitive Semantics*. The MIT Press.

Tarjan, R. (1972). Depth-First Search and Linear Graph Algorithms. *SIAM Journal on Computing*, 1(2):146–160.

Tarquini, F., Felice, G., Fogliaroni, P., and Clementini, E. (2007). A Qualitative Model for Visibility Relations. In *Proceedings of the 30th Annual German Conference on Advances in Artificial Intelligence*, KI '07, pages 510–513, Berlin, Heidelberg. Springer-Verlag.

TEI P5 (2014). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. TEI Consortium. http://www.tei-c.org/Guidelines/P5/ (accessed July 2015). P5, version 2.7.0. Last updated on 16th September 2014.

Tom, A. and Denis, M. (2004). Language and spatial cognition: comparing the roles of landmarks and street names in route instructions. *Applied Cognitive Psychology*, 18(9):1213–1230.

Tran, M. (2006). *Prolexbase : un dictionnaire relationnel multilingue de noms propre : conception, implémentation et gestion en ligne*. Thèse doctorat d'informatique, Université François Rabelais Tours.

Tran, M. and Maurel, D. (2006). Prolexbase : Un dictionnaire relationnel multilingue de noms propres. *Traitement Automatique des Langues*, 47(3):115–139.

Triantaphyllou, E. (2000). *Multi-criteria Decision Making Methods: A Comparative Study*, volume 44 of *Applied Optimization*. Springer US, Boston, MA.

Tversky, B. and Lee, P. U. (1998). How Space Structures Language. In *Spatial Cognition, An Interdisciplinary Approach to Representing and Processing Spatial Knowledge*, pages 157–176, London, UK, UK. Springer-Verlag.

Urieli, A. (2013). *Robust French syntax analysis : reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. phdthesis, Université Toulouse le Mirail - Toulouse II.

Vandeloise, C. (1986). *L'Espace en français. Sémantique des prépositions spatiales*. Editions du Seuil.

Vasardani, M., Timpf, S., Winter, S., and Tomko, M. (2013). From Descriptions to Depictions: A Conceptual Framework. In *Proc. 11th Intl. conf on Spatial Information Theory, COSIT*, pages 299–319.

Wacholder, N., Ravin, Y., and Choi, M. (1997). Disambiguation of Proper Names in Text. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, ANLC '97, pages 202–208, Stroudsburg, PA, USA. Association for Computational Linguistics.

Werner, S., Krieg-Brückner, B., and Herrmann, T. (2000). Modelling Navigational Knowledge by Route Graphs. In Freksa, C., Habel, C., Brauer, W., and Wender, K. F., editors, *Spatial Cognition II, Integrating Abstract Theories, Empirical Studies, Formal Methods, and Practical Applications*, pages 295–316, London, UK, UK. Springer-Verlag.

Winter, S. and Raubal, M. (2006). Time Geography for Ad-Hoc Shared-Ride Trip Planning. In *7th International Conference on Mobile Data Management, 2006. MDM 2006*, pages 6–6. IEEE.

Woodruff, A. G. and Plaunt, C. (1994). GIPSY: Automated Geographic Indexing of Text Documents. *J. Am. Soc. Inf. Sci.*, 45(9):645–655.

WäLchli, B. (2001). A typology of displacement (with special reference to Latvian). *STUF - Language Typology and Universals*, 54(3).

Yahiaoui, S., Josselin, D., Marchand-Lagier, C., and Douvinet, J. (2014). Vérification et (re)construction automatiques des limites des bureaux de vote par l'étude des textes juridiques. In *SAGEO 2014*, Grenoble, France.

Yu, L. (2011). Linked Open Data. In *A Developer's Guide to the Semantic Web*, pages 409–466. Springer Berlin Heidelberg.

Yuan, Y. and Raubal, M. (2012). Extracting dynamic urban mobility patterns from mobile phone data. In Xiao, N., Kwan, M.-P., Goodchild, M. F., and Shekhar, S., editors, *Geographic Information Science*, number 7478 in Lecture Notes in Computer Science, pages 354–367. Springer Berlin Heidelberg.

Zhang, X., Mitra, P., Klippel, A., and MacEachren, A. (2010). Automatic extraction of destinations, origins and route parts from human generated route directions. In Fabrikant, S. I., Reichenbacher, T., Kreveld, M. v., and Schlieder, C., editors, *Geographic Information Science*, number 6292 in Lecture Notes in Computer Science, pages 279–294. Springer Berlin Heidelberg.

Zhang, X., Qiu, B., Mitra, P., Xu, S., Klippel, A., and MacEachren, A. M. (2012). Disambiguating road names in text route descriptions using exact-all-hop shortest path algorithm. In *ECAI'12*, pages 876–881.

Zhao, J., Jin, P., Zhang, Q., and Wen, R. (2014). Exploiting location information for web search. *Computers in Human Behavior*, 30:378–388.

Zhou, G. and Su, J. (2002). Named Entity Recognition Using an HMM-based Chunk Tagger. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 473–480, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zimmermann, K. (1993). Enhancing qualitative spatial reasoning — Combining orientation and distance. In Frank, A. U. and Campari, I., editors, *Spatial Information Theory A Theoretical Basis for GIS*, number 716 in Lecture Notes in Computer Science, pages 69–76. Springer Berlin Heidelberg.

Zimmermann, K. and Freksa, C. (1996). Qualitative spatial reasoning using orientation, distance, and path knowledge. *Applied Intelligence*, 6(1):49–58.

Zipf, A. and Krüger, S. (2001). TGML - extending GML by temporal constructs - a proposal for a spatiotemporal framework in XML. In *In: Proceedings of ACM GIS 2001. Atlanta USA*, pages 117–148.

Zlatev, J. (2010). Spatial semantics. In Cuyckens, H. and Geeraerts, D., editors, *The Oxford Handbook of Cognitive Linguistics*, pages 318–350. Oxford University Press.

ÉCOLE DOCTORALE :
École doctorale des sciences exactes et leurs applications
Escuela de Doctorado de la Universidad de Zaragoza

LABORATOIRE :
Laboratoire d'Informatique de l'Université de Pau et des Pays de l'Adour - EA 3000
Departamento de Informática e Ingeniería de Sistemas, Universidad de Zaragoza
Laboratoire COGIT, IGN, Université Paris-Est

Ludovic Moncla

prenom.nom@univ-pau.fr

Université de Pau et des Pays de l'Adour
Avenue de l'Université, BP 576
64012 Pau Cedex (France)

Universidad de Zaragoza
C/ Pedro Cerbuna 12
50009 - Zaragoza (España)

UNIVERSITÉ DE PAU ET DES PAYS DE L'ADOUR

1542 Universidad Zaragoza