Extraction de connaissances pour la description de l'environnement maritime côtier à partir de textes d'aide à la navigation

Léa Lamotte*, Nathalie Abadie*, Éric Saux**, Éric Kergosien***

* Univ. Paris-Est, LASTIG STRUDEL, IGN, ENSG, F-94160 Saint-Mandé
lea.lamotte@ign.fr, nathalie-f.abadie@ign.fr
http://recherche.ign.fr
**IRENav, Lanvéoc-Poulmic, CC 600, F-29240 Brest Cedex 9
eric.saux@ecole-navale.fr
http://www.ecole-navale.fr
***GERiiCO, Univ. de Lille, F-59000 Villeneuve d'Ascq
eric.kergosien@univ-lille.fr
https://geriico-recherche.univ-lille3.fr

Résumé. Les référentiels de données géoréférencées sont de plus en plus utilisés pour permettre l'annotation spatiale de documents textuels et ainsi faciliter l'accès à leur contenu, voire son analyse spatiale. En revanche, peu de travaux se sont intéressés à l'extraction d'information géographique à partir de textes pour alimenter de tels référentiels. Le travail présenté dans cet article explore les potentialités de l'extraction d'information spatiale indirecte (noms de lieux, relations spatiales, etc.) dans les textes des Instructions Nautiques produites par le Service Hydrographique et Océanographique de la Marine (SHOM). La méthode proposée combine une approche lexicale et une approche à base de patrons linguistiques puis est comparée aux principales approches d'extraction d'information géographique en français.

1 Introduction

Les Instructions Nautiques (IN) sont des documents textuels publiés par le Service Hydrographique et Océanographique de la Marine (SHOM), décrivant les amers et dangers pour la navigation côtière, les courants, les ports et mouillages, leurs chenaux d'accès, leurs équipements et les services proposés aux navigateurs, etc. L'extrait suivant présente une description d'un paysage typique des IN : *"[...]La Pointe de Pléneuf est débordée vers le NW par l'îlot Le Verdelet, puis jusqu'à 1,5 M au large par le Plateau des Jaunes, découvrant ; celui-ci, ainsi que les hauts-fonds rocheux qui s'étendent dans l'Ouest du Verdelet, sont couverts par le secteur rouge (146°-196°) du feu de la tourelle La Petite Muette située à l'entrée de Dahouët.[...]"*.

Dans cet article, nous proposons une approche qui vise à extraire des Entités Spatiales Nommées (ESN) ou non et leurs relations spatiales à partir des IN en vue de les représenter au sein d'une ontologie. L'intérêt de ce modèle de représentation de connaissances est double.

D'une part, sa structure de graphe permet d'intégrer facilement de nouvelles propriétés pour décrire les entités (Kim et al., 2015). D'autre part, ESN et relations spatiales constituent les composantes de base des systèmes de références spatiales par identificateurs géographiques (ISO, 2019) - ou systèmes de références spatiales indirects. Ceux-ci permettent de localiser une position de l'espace décrite par une référence spatiale comportant une relation spatiale - inclusion, distance, adjacence, etc. - et une ou plusieurs entités géographiques de localisations connues. Lesbegueries et al. (2006) parlent d'entités spatiales relatives (ESR), comme par exemple "près du port Saint-Goustan". En permettant de formaliser les références spatiales indirectes, ce modèle permet une médiation entre descriptions textuelles du terrain et bases de données géographiques vectorielles. Ce travail s'inscrit ainsi dans la continuité de travaux destinés à faciliter la mise à jour des IN par un langage contrôlé (Haralambous et al., 2017).

Cet article est organisé de la façon suivante. En section 2, nous présentons un état de l'art des approches d'extraction d'entités et de relations spatiales. La section 3 décrit l'approche que nous proposons. En section 4, nous détaillons les résultats obtenus sur notre corpus avant de discuter des conclusions et perspectives de ce travail en section 5.

2 Extraction d'information géographique à partir de textes

De nombreuses méthodes permettent de reconnaître les EN en général et les Entités Spatiales (ES) en particulier (Nadeau et Sekine, 2007). Parmi les méthodes d'extraction d'informations, les approches statistiques étudient généralement les termes co-occurrents par analyse de leur distribution dans un corpus (Agirre et al., 2000) ou par des mesures calculant la probabilité d'occurrence d'un ensemble de termes (Velardi et al., 2001). Ces méthodes ne permettent pas toujours de qualifier des termes comme étant des EN, notamment dans le cas des ES ou organisations. Des méthodes de fouille de données fondées sur l'extraction de motifs permettent de déterminer des règles de transduction afin de repérer les EN (Maurel et al., 2011). Ces règles utilisent des informations syntaxiques propres aux phrases. D'autres approches combinent un ensemble de règles à un ensemble de dictionnaires (Mansouri et al., 2008). Ce type d'approches donne des résultats intéressants lorsqu'elles sont appliquées à un corpus de domaine spécialisé, comme c'est le cas pour (Moncla, 2015). D'autres combinent des règles à une approche de fouille de textes pour l'identification et la désambiguïsation des ES et des organisations (Tahrat et al., 2013). Les relations peuvent être identifiées par des calculs de similarité entre leurs contextes syntaxiques (Grefenstette, 1994), par prédiction à l'aide de réseaux bayésiens (Weissenbacher et Nazarenko, 2007), par des techniques de fouille de textes (Grčar et al., 2009) ou encore par inférence de connaissances à l'aide d'algorithmes d'apprentissage (Giuliano et al., 2006). Ces méthodes sont efficaces, mais elles n'identifient pas toujours la sémantique de la relation, et ne permettent pas de les identifier de façon précise. Pour la reconnaissance des classes d'EN, de nombreuses approches s'appuient sur des méthodes d'apprentissage supervisé comme les SVM (Joachims, 1998) ou encore les champs aléatoires conditionnels (Zidouni et al., 2009). Ces algorithmes s'appuient sur des textes préalablement étiquetés et sur des descripteurs de tokens (positions des termes, étiquettes grammaticales, informations lexicales, casse, etc.) (Carreras et al., 2003) pour apprendre des modèles de classification d'EN.

Disposant d'un corpus extrêmement spécialisé et dépourvu d'annotations préalables, nous proposons dans un premier temps une approche combinant un ensemble de règles à des dictionnaires de type gazetiers.

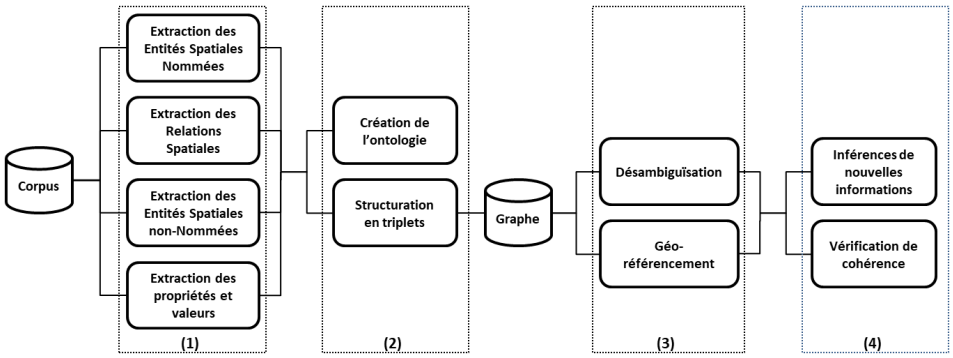

FIG. 1 – *De l'extraction d'information géographique à partir de textes à son exploitation au sein d'une base de connaissances.*

3 Référencement indirect d'entités spatiales

L'approche que nous proposons constitue la première étape de la chaîne de traitement, présentée en figure 1. Celle-ci inclut le marquage des ES et de leurs références spatiales indirectes (1), la génération d'une base de connaissances (2), le géoréférencement des ES qu'elle contient (3) et son axiomatisation pour vérifier sa cohérence et inférer des connaissances non explicitées dans les textes (4). En lien avec le modèle proposé par Chen et al. (2018) et les différents référentiels spatiaux existants (relatif, absolu, intrinsèque) (Shusterman et Li, 2016), nous définissons le 5-tuple <Locatum,Relatum,Relation,Origine,Orientation> afin de caractériser une relation spatiale entre deux objets.

3.1 Extraction des entités spatiales

Pour l'extraction des ESN, nous proposons une approche à base de règles et de gazetiers. Pour pallier le manque d'exhaustivité éventuel des gazetiers, nous introduisons un lexique des descripteurs de lieux. Associé à des critères sur la casse, ce dernier nous permet d'accroître l'extraction des ESN mais aussi d'enrichir les gazetiers initiaux afin de permettre l'identification de nouveaux toponymes (sans descripteurs). Les types des ESN extraites seront nécessaires pour permettre la création des concepts de l'ontologie, et les ES et les toponymes pour peupler ces concepts. Nos patrons sont schématisés sous la forme [1] suivante :

(<descr> <DP>*)* (<maj> <DP>*)* <topo> (<DP>* <maj>)*
<descr> (<DP> <descr>)* <DP>* <maj> (<DP>* <maj>)*

Cependant, les ES contenant un toponyme ne constituent pas la totalité des ES utilisées dans les relations. Il est courant que celles-ci soient composées d'un syntagme nominal ou pronominal (p. ex. : "Ils sont situés à 55 Milles à l'Ouest de Bishop Rock"). Pour les repérer, nous proposons d'utiliser le patron [2] suivant :

1. <descr> désigne un descripteur, <DP> représente un ensemble de déterminant et de prépositions, <maj> désigne un mot commençant par une majuscule, et <topo> est un toponyme connu.

2. <DP> désigne un sous-ensemble des déterminants et prépositions du français, <Car> représente les différentes combinaisons de points cardinaux et <Mod> symbolise un ensemble de modifieurs sélectionnés.

<DET>* <NOM>* <Car>* <DP>* <ADJ>* <descripteur> <Mod>* (<DET> <WORD>)*

3.2 Extraction des relations spatiales

La tâche d'extraction des relations spatiales est divisée en deux étapes : marquage et classification des expressions incluses dans le texte et résolution de leurs arguments.

L'étape de ***marquage et de classification***, effectuée sur notre corpus de développement, dresse une liste des expressions de relations spatiales à extraire ainsi que leur types : descriptions quantitatives et qualitatives de distance (p. ex. : "à 8 M de", "près de"), directions (absolues, relatives et intrinsèques), "entre", "à l'intérieur de", "contient", "traverse" et adjacence. Parmi ces relations seules les relations de direction et de distance pourraient nécessiter l'extraction d'informations concernant le référentiel utilisé pour être correctement interprétées. Cet inventaire exhaustif des différentes formulations des types de relations spatiales nous a servi de base lors de la constitution des patrons lexicaux qui alimentent notre système d'extraction. Il s'agit de patrons très simples, essentiellement fondés sur la détection de verbes particuliers [3], que nous avons identifiés comme caractéristiques d'une relation spatiale donnée.

Pour l'étape de ***résolution des arguments***, nous proposons d'utiliser une représentation simplifiée du texte, qui n'en conserve que la séquentialité, les ponctuations fortes, les conjonctions de coordinations, les entités spatiales et les relations identifiées. Nous avons ensuite formalisé un ensemble de règles visant à déterminer pour chaque relation, quelles ES environnantes sont les plus susceptibles de constituer le locatum et le relatum. Par exemple, une relation précédée par une ponctuation forte, et suivie de deux entités spatiales juxtaposées, prend la première comme relatum, et la seconde comme locatum.

"la Grande Passe, entre l'Île Tristan et le môle de la Pointe Biron"

["'Grande Passe', '**entre**', 'Île Tristan', 'et', 'môle de la Pointe Biron'"]

Dans la partie introductive à la section 3, nous avons défini une relation spatiale à partir d'un 5-tuple. L'orientation et l'origine peuvent être omises (référentiel absolu), partiellement confondues avec d'autres éléments du repère (référentiel intrinsèque) ou encore totalement indépendantes (référentiel relatif). Cependant, même dans ce dernier cas, l'origine est souvent sous-spécifiée et incluse dans l'orientation (ex : "En venant du Nord"). Nous regroupons alors ces deux informations et marquons dans le texte tous les segments exprimant une orientation. Par exemple, "en passant par" suivi d'une ESN ou d'un groupe nominal nous indique une origine, à partir de laquelle nous allons pouvoir déduire un vecteur d'orientation <origine, relatum>. De la même façon, une forme conjuguée du verbe "orienter" suivie d'un point cardinal, donne l'orientation intrinsèque de l'objet auquel se rapporte l'expression.

4 Application aux Instruction Nautiques et évaluation

Notre corpus est constitué du volume des IN décrivant les côtes de la frontière belge à la pointe de Penmarc'h (SHOM, 2018). Nous disposons d'une version au format XML, dotée de balises correspondant à la structure de l'ouvrage : chapitre, sous-chapitres, paragraphes, etc. Nous avons découpé notre corpus en trois : un corpus d'entraînement, dédié à l'élaboration de notre approche d'extraction, un corpus de développement, destiné à l'évaluation intermédiaire

3. Les listes exhaustives des verbes recherchés pour chaque type de relation sont disponibles en ligne (https ://github.com/LeaLamotte/RelationsSpatiales).

et l'amélioration de cette dernière et un corpus de test, utilisé pour l'évaluation finale. Ce découpage a été réalisé suivant deux hypothèses. D'une part, chaque portion de côte est décrite par un même locuteur : pour se doter de corpus non biaisés par le style d'écriture d'une personne, un découpage non géographique du texte est donc préférable. D'autre part, la résolution des co-références nécessite de conserver des portions de texte cohérentes. Nous avons donc découpé le fichier au niveau des balises "paragraphe" et les avons numérotées, selon l'ordre du texte. Les paragraphes ayant reçu un numéro *n* multiple de 5, ont été assignés au corpus de test (soit 36 paragraphes, comportant 40 981 tokens) et ceux avec *n + 1* multiple de 5, au corpus de développement (soit 36 paragraphes, comportant 34 963 tokens). Les paragraphes restants constituent le corpus d'entraînement (108 paragraphes, 146 971 tokens).

Nous avons développé notre chaîne d'extraction avec l'outil Unitex [4]. Nous nous sommes appuyés sur les cascades de transducteurs, en reprenant le fonctionnement en deux cascades, "analyse" et "synthèse", tel que développé dans (Maurel et al., 2011).

4.1 Extraction des entités spatiales

Pour construire nos lexiques, nous avons utilisé les bases de données géographiques "BD NYME®" et "Toponymes", respectivement produites par l'Institut National de l'Information Géographique et Forestière et le SHOM. Pour limiter le bruit, nous n'avons conservé que les toponymes figurant dans un rayon de 10 km autour des côtes décrites dans le corpus et avons extrait leurs descripteurs associés pour créer le lexique des types d'entités géographiques.

Nous avons comparé les résultats de notre système, avec ceux obtenus par les outils Perdido [5] et CasEN [6] (voir tableau 1). Pour CasEN, les faibles scores s'expliquent probablement par le fait qu'il semble détecter les points cardinaux et les articles définis au pluriel placés en début de phrase comme des ESN. De plus, et comme Perdido, ses lexiques ne comportent probablement que peu de termes propres au domaine maritime côtier, ce qui pénalise leur rappel.

	Précision	Rappel	F-1
CasEN	0.16	0.19	0.17
Perdido	0.73	0.16	0.26
Notre approche	0.84	0.61	0.7

TAB. 1 – *Résultats de l'extraction d'ENS réalisée sur le corpus de développement.*

Sur notre corpus de test, notre approche produit une précision de 79%, un rappel de 72% pour une F-mesure de 75%. Nous ne cherchons à extraire les ES non nommées que lorsqu'elles participent à une relation spatiale. Avec notre approche à base de lexiques, nous obtenons donc des résultats relativement bruités avec une précision de 32%. De plus, nous avons encore des difficultés de segmentation, qui pénalisent le rappel de façon significative (57%).

4. développé par le Laboratoire d'Informatique Gaspard-Monge `https://unitexgramlab.org/fr`
5. `http://erig.univ-pau.fr/PERDIDO/api.jsp`
6. `http://tln.li.univ-tours.fr/Tln_Ortolang.html`

4.2 Extraction des relations spatiales

Notre corpus de test comprend 1618 relations spatiales, dont 528 directions, 317 distances, 85 adjacences, 11 "traverse", 195 “contient”, 245 “à l'intérieur” et 115 "entre". Notre système présente des performances qui varient d'un type de relation à l'autre (voir tableau 2). Les distances et les directions sont globalement bien reconnues. Les relations "traverse" et "contient" sont également bien traitées, car il existe peu de façons de les exprimer. En revanche, les relations d'adjacence, moins évidentes, sont difficilement reconnues, avec un rappel inférieur à 40%. Les relations de type "à l'intérieur de" sont également difficiles à cerner, principalement à cause des très nombreux usages de la préposition "dans" en français. Les performances de la catégorie "entre" souffrent également de la fréquence à laquelle la préposition "de" est utilisé près d'un "à", sans pour autant constituer une structure de type "de Saint-Brieuc à Saint-Malo". Cependant, sur la globalité du corpus de test, nous obtenons des résultats prometteurs, tant sur la tâche de marquage des relations (76% de micro-moyenne de la F-mesure), que sur la tâche de marquage et classification (72% de micro-moyenne de la F-mesure).

	Relations (marquage)	Relations (classification)	Distance	Direction	Adjacence	Traverse	Contient	À l'intérieur	Entre	Triplet (marquage manuel)	Triplet (marquage automatique)
Précision	0.78	0.74	0.78	0.82	0.67	1.0	0.72	0.56	0.57	0.69	0.28
Rappel	0.74	0.71	0.76	0.93	0.39	0.6	0.71	0.42	0.76	0.63	0.19
F-1	0.76	0.72	0.77	0.87	0.49	0.75	0.72	0.48	0.65	0.66	0.23

TAB. 2 – *Résultats de l'extraction de relations spatiales.*

En revanche, notre méthode de résolution des arguments de ces relations s'avère peu performante. Elle souffre en effet des imprécisions de tout le marquage précédent. Mais même sur un corpus parfaitement annoté, nous obtenons des résultats modestes, avec une précision de 69% et un rappel de 63%. En effet, les IN constituent un texte long, avec des phrases complexes présentant des structures qu'une approche linéaire comme la nôtre ne suffit pas à traiter. À titre d'exemple les phrases (1) et (2) vont toutes les deux nous donner une structure du type ['entité 1', 'relation 1', 'entité 2', 'relation 2', 'entité 3'], mais des relations 2 avec des sujets différents : dans l'exemple (a) il s'agit de l'entité 1 ; dans l'exemple (b), l'entité 2.

(a) "le port est établi en amont de l'écluse **sur** la rive droite"
{le port, **sur**, la rive droite}
(b) "un port à l'abri d'une jetée enracinée **au pied de** Mont Orgueil Castle"
{une jetée, **au pied de**, Mont Orgueil Castle}

Enfin, dans les IN, la plupart des descriptions utilisant des repères relatifs et intrinsèques sont des descriptions d'itinéraires. Nous avons par conséquent un déplacement mental dans le paysage, qui s'accompagne d'un déplacement implicite du repère, soit d'un objet localisé par rapport au suivant, soit le long du trajet. Les origines et orientations du repère sont rarement fournies explicitement. À ce stade de notre travail, nous obtenons pour l'extraction des indications d'orientation une précision de 74%, un rappel de 40% et une F-mesure de 52%.

5 Discussion et perspectives

Les descriptions d'environnement naturel présentes dans les textes des IN du SHOM sont très riches et complémentaires à celles contenues dans les cartes marines. En effet, celles-ci sont soumises à l'interprétation de leur utilisateur ce qui peut être insuffisant dans certaines situations. Celle-ci doit alors être confortée et complétée par des descriptions et directives présentes dans les IN. Ainsi, il nous paraît important d'extraire et de lier les connaissances présentes dans les IN à celles présentes dans les cartes électroniques de navigation en vue d'une exploitation commune au sein d'un même système d'information.

Nous avons abordé exclusivement la question de l'extraction des ES nommées et non-nommées ainsi que celle de leurs relations spatiales et des référentiels associés à des fins de géoréférencement. Comparée aux outils CasEN et Perdido, l'approche proposée pour l'extraction des ESN offre des résultats supérieurs, notamment en raison des particularités de notre corpus qui mettent en défaut les approches usuelles. Les résultats obtenus pour la tâche d'extraction des relations spatiales sont très encourageants, même si les performances varient d'une catégorie de relations spatiales à l'autre. Pour la résolution des arguments des relations spatiales, les tests ont mis en avant la difficulté à déterminer le locatum et le relatum du fait de la grande diversité des structures de phrases de notre corpus. Cette étape constitue donc une piste d'amélioration à court terme. Bien que la solution proposée doive être améliorée, l'étape suivante vise à proposer une modélisation ontologique des connaissances extraites de ces documents textuels et à les lier aux connaissances cartographiques modélisées au format S-57 [7].

Références

Agirre, E., O. Ansa, E.-H. Hovy, et D. Martìnez (2000). Enriching very large ontologies using the www. In *ECAI Workshop on Ontology Learning*.

Carreras, X., L. S. M. Arque, et L. S. PadrO (2003). A simple named entity extractor using adaboost. In *Proceedings of CoNLL-2003*, pp. 152–155.

Chen, H., M. Vasardani, S. Winter, et M. Tomko (2018). A graph database model for knowledge extracted from place descriptions. *International Journal of Geo-Information 7*(6).

Giuliano, C., A. Lavelli, et L. Romano (2006). Exploiting shallow linguistic information for relation extraction from biomedical literature. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)*.

Grčar, M., E. Klien, et B. Novak (2009). Using term-matching algorithms for the annotation of geo-services. *Knowledge Discovery Enhanced with Semantic and Social Information 220*(8), 127–143.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Norwell, MA, USA : Kluwer Academic Publishers.

Haralambous, Y., J. Sauvage-Vincent, et J. Puentes (2017). A hybrid (visual/natural) controlled language. *Language Resources and Evaluation 51*(1), 93–129.

ISO (2019). Iso 19112:2019 geographic information — spatial referencing by geographic identifiers. Technical report, International Organisation for Standardization.

7. www.s-57.com

Joachims, T. (1998). Text categorization with suport vector machines: Learning with many relevant features. In *ECML*, pp. 137–142.

Kim, J., M. Vasardani, et S. Winter (2015). Harvesting large corpora for generating place graphs. In *International workshop on cognitive engineering for spatial information processes (CESIP), in conjunction with COSIT*, Volume 12.

Lesbegueries, J., M. Gaio, et P. Loustau (2006). Geographical information access for non-structured data. In *Proceedings of the 2006 ACM symposium on Applied computing*, SAC'06, pp. 83–89. New York, NY, USA: ACM.

Mansouri, A., L. S. Affendey, et A. Mamat (2008). Named entity recognition approaches. *TAL 52*(1), 339–344.

Maurel, D., N. Friburger, J.-Y. Antoine, I. Eshkol, et D. Nouvel (2011). Cascades de transducteurs autour de la reconnaissance des entités nommées. *Traitement Automatique des Langues 52*(1), 69–96.

Moncla, L. (2015). *Automatic reconstruction of itineraries from descriptive texts*. Ph. D. thesis, Pau.

Nadeau, D. et S. Sekine (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes 30*(1), 3—26.

SHOM (2018). *Instructions nautiques, France (côtes Nord et Ouest), Du cap de La Hague à la pointe de Penmarc'h (Version à jour au 19 décembre 2018)*. SHOM.

Shusterman, A. et P. Li (2016). Frames of reference in spatial language acquisition. *Cognitive Psychology 88*, 115–161.

Tahrat, S., E. Kergosien, S. Bringay, M. Roche, et M. Teisseire (2013). Text2geo: from textual data to geospatial information. In *WIMS: Web Intelligence, Mining and Semantics*.

Velardi, P., P. Fabriani, et M. Missikoff (2001). Using text processing techniques to automatically enrich a domain ontology. In *FOIS*, pp. 270–284.

Weissenbacher, D. et A. Nazarenko (2007). Identifier les pronoms anaphoriques et trouver leurs antécédents : l'intérêt de la classification bayésienne. In *Proceedings of Traitement Automatique des Langues Naturelles, ATALA*, pp. 145–155.

Zidouni, A., M. Quafafou, et H. Glotin (2009). Structured named entity retrieval in audio broadcast news. *S. D. Kollias et Y. S. Avrithis (Eds.), CBMI*, 126–131.

Summary

Frames of reference are increasingly being used to enable the spatial annotation of textual documents and thus facilitate access to their content and even their spatial analysis. Nevertheless, little work had been done on the extraction of geographical information from texts to feed such frames. The work presented in this article explores the potentialities of indirect spatial information extraction (place names, spatial relationships, etc.) in the texts of the Nautical Instructions produced by the French Marine Hydrographic and Oceanographic Service. The proposed method combines a lexical approach and a linguistic pattern-based approach and is then compared to the main approaches to extracting geographic information in French.