# A Modeling Approach for the Extraction of Semantic Information from a Maritime Corpus

Dieudonné Tsatcha, Eric Saux, Christophe Claramunt

Naval Academy Research Institute, GIS group, Lanvéoc-Poulmic, CC 600, F-29240
Brest Cedex 9, France

**Abstract.** This paper introduces an algorithm for retrieving semantic information from a maritime corpus. The method is based on Natural Language Processing (NPL) and combines a segmentation of large documents with principles of a conceptual vector model (CVM) and synsets of words. This research is applied to the context of intelligent transport systems and maritime navigation. Based on documents regulating maritime traffic, this approach proposes an aid for navigational decision-making while significantly reducing the number of entities and relations required in the modeling process.

**Keywords:** natural language processing, conceptual vector model, semantics, navigational decision aid.

## 1 Introduction

Security for navigation in the maritime context is a significant challenge, which has been the focus of much research and developments even though much work remains to accomplish. Intelligent transport systems (ITS) provide some solutions (e.g., 3-dimensional GIS applied to maritime navigation[1] [10], Automatic Identification System (AIS)) but identification of a modeling approach which takes into account all the environmental components is not easy to achieve. This is mainly due to the fact that the maritime navigation space is complex (e.g., traffic regulation rules, restricted areas) and changes dynamically (tides, currents, winds) this having a direct incidence on navigation. Moreover, legibility (fog, day, night) can influence the perception of the actions to perform. Most ITS try to reduce collisions by improving the visual representation of the environment (Electronic Chart Display Information Systems, Automatic Radar Plotting Aids) [5] or analyzing information coming from sensors (Global Positioning System, Automatic Identification System, Radio Detection And Ranging (RADAR), infrared cameras, Lang Range Identification and Tracking (LRIT)). However and to the best of our knowledge, none of these methods facilitate significantly navigation planning. One objective is to reduce the complexity of representation by interpreting the semantic information generated by one object

[1] http://www.geovs.com/

or several objects available in a maritime map and within a given area at a given time.

The main objective of this research is to build a decision platform for a maritime navigation environment based on some available semantics proposing a safe route to the captain. The assumption made is that the decision platform can significantly improve the captain's cognitive abilities during a high stress or high workload situation. We assume that a route preserves navigation safety if it takes into account the semantics and affordances of all the objects around the ship. An affordance is a quality of an object, or an environment, which allows an individual to perform an action [9]. This implies to take into account restricted areas, the effect of ocean currents, the wind, radio signals and so on. Considering the knowledge that emerges from exterior events and the behavior of the objects located in the vicinity of the ship or those detected from sensors, a sailor should quickly get directional information regarding the route to follow and make the appropriate motion decision. One of the difficulties for the identification of entities is to take into account their salience. For example, a buoy may be expressed differently at night or during the day (cf. figure 1).

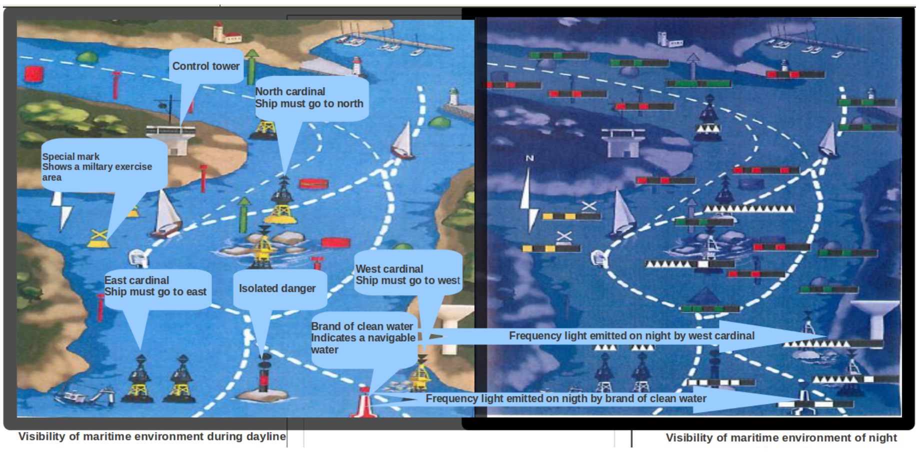
Control tower
North cardinal
Ship must go to north
Special mark
Shows a miltary exercise area
East cardinal
Ship must go to east
Isolated danger
West cardinal
Ship must go to west
Brand of clean water
Indicates a navigable water
Frequency light emitted on night by west cardinal
Frequency light emitted on nigth by brand of clean water
N
Visibility of maritime environment during dayline
Visibility of maritime environment of night

**Fig. 1.** An example of semantic information that influences the route of vessels by day (on the left) and by night (on the right). By day, the objects are recognized based on their shape and colors whereas they are identified by their light signal by night [17].

The development of a decision platform implies modeling a maritime environment and defining a relevant ontology. Maritime knowledge is extracted from documents used to regulate maritime navigation [25,12,18,13,16]. Natural language processing is applied at two levels. The first level permits to extract the terms of the domain. The use of Yatea[2] software [1] coupled with Treetagger [21] extracts the terms with the largest number of occurrences in the documents. The

[2] Yatea is a free piece of software used for lexical disambiguisation of documents.

ones with a small number of occurrences, but which are important in navigation decisions, are also considered. The corpus contains 16 010 sentences defined with 413 076 words (175 578 nouns, 269 22 verbs, 20 703 adjectives, 9 106 adverbs). The second level extracts the semantics of the objects in the finite collection of states set by an expert. This extraction is possible thanks to conceptual vector applied to the sentences that relate to an object and projected in a decision space. By extracting the semantics of the objects, one can find the decision or the future area that the ship should follow. This paper mainly focuses on the second level assuming the first one to be a preprocessing step.

The remainder of the paper is organized as follows. Section 2 introduces the main principles of natural language processing for information retrieval in a general context before introducing the theoretical concepts of conceptual vectors applied to a word and a sentence. Section 3 develops our semantic extraction approach based on the definition of a decision space where sentences are projected within in order to associate the right semantics to a given concept. Section 4 presents a case study and applies our strategy for the extraction of semantics from concepts derived from maritime navigation documents. Finally, section 5 draws some conclusions and outlines further work.

## 2 Theoretical concepts of conceptual vector in Natural Language Processing

Information retrieval (hereafter IR) is an interdisciplinary research domain. Research in IR evolved over time and from early works in the 60's with language indexation experiments [6]. In the 90's, retrieval engines were mainly based on the concept of keyword and without adequate representation of content for both documents and queries [20]. Nowadays, recent progress consists in merging NPL (extracting the lexico-semantic structure of documents) and IR (indexing, matching, etc.) to find the semantic information related to a query [23,8,19]. Information retrieval supports three basic processes [11]: representation of the content of documents, representation of query and comparison of the two previous representations. In order to improve the information retrieval efficiency, documents are transformed into a suitable representation. Becker [2] introduced the different representations that can be used and describes the relations between representations and models. The three most used models in IR research are the vector space model, the probabilistic model and the inference network model [22]. Most systems assign a numeric score to every document and rank it using this score and do not take into account the semantic relatedness between query and sentences which satisfy the query. Tsatsoronis [24] points out the importance of capturing semantics betweeen terms in IR. In this paper, we propose an algorithm developed from the concept of conceptual vector and disambiguisation where the relevance of results depends on this semantic relatedness.

The proposed algorithm is grounded on concepts of conceptual vector of word initially proposed by [15]. Conceptual vectors have been mainly used for information retrieval and for meaning representation in the latent semantic indexing

(LSI) model from latent semantic analysis (LSA) studies in psycholinguistics [20]. Our approach is inspired from [4], which proposes a formalism for the projection of the linguistic notion of semantic field in a vectorial space. A conceptual vector (or vector of concepts) of a word is a set of words in which each word determines a concept where this word can be employed. A conceptual vector of a sentence includes all concepts of the sentence. The latter is based on the direct sum of conceptual vectors of words composing the sentence.

### 2.1 Conceptual vector of word

The definition of a conceptual vector of a word is based on the concept of synset of a word. A synset of a word represents a concept and contains a set of interchangeable words, each of them having the same sense that names the concept [3]. Another sense that names the concept defines another synset of the same initial word. Each word composing the synset that is different from the initial one is called *candidate word*. The definition of the conceptual vector associated to a word is based on a set of synsets and a metric measuring the distance between this word and each candidate word of a synset. The conceptual vector of a word is organized according to grammatical categories (adjective ($a$), adverb ($r$), noun ($n$), verb ($v$)) that the word may belong to and according to decreasing distance values inside a grammatical category. We use the distance defined in RiWordnet[3] [7], developed for creativity support in computation literature proposed by Daniel Howe.

More formally, let $w$ be a word, $S^c = (s_i^c)_{i=1}^{n^c}$ the sets of $n^c$ synsets of $w$ and $C^c = (c_i^c)_{i=1}^{m^c}$ the sets of $m^c$ candidate words of all synsets in category $c$ then the conceptual vector $V(w)$ of the word $w$ is defined as a weighted union of candidate words expressed in each grammatical category:

$$V(w) = \left.\bigcup_{l=1}^{m^a} c_l^a \delta(c_l^a, w)\right)_a \left.\bigcup_{l=1}^{m^r} c_l^r \delta(c_l^r, w)\right)_r \left.\bigcup_{l=1}^{m^n} c_l^n \delta(c_l^n, w)\right)_n \left.\bigcup_{l=1}^{m^v} c_l^v \delta(c_l^v, w)\right)_v \quad (1)$$

For illustration purposes, let us compute the conceptual vector of the word "port". Let $S^a = (s_i^a)_{i=1}^1$, $S^r = \emptyset$, $S^n = (s_i^n)_{i=1}^5$ and $S^v = (s_i^v)_{i=1}^8$ be the different sets of synsets of this word extracted from WordNet where (see Figure 2):

– Adjective (a)
  - $s_1^a$: port, larboard (located on the left side of a ship or aircraft)

– Noun (n)
  - $s_1^n$: port (a place (seaport or airport) where people and merchandise can enter or leave a country)
  - $s_2^n$: port, port wine (sweet dark-red dessert wine originally from Portugal)
  - $s_3^n$: port, embrasure, porthole (an opening (in a wall or ship or armored vehicle) for firing through)

[3] RiWordnet is an API to WordNet that is a lexical database for the English language.

  - $s_4^n$: larboard, port (the left side of a ship or aircraft to someone who is aboard and facing the bow or nose)
  - $s_5^n$: interface, port ((computer science) computer circuit consisting of the hardware and associated circuitry that links one device with another (especially a computer and a hard disk drive or other peripherals))
- Verb (v)
  - $s_1^v$: port (put or turn on the left side, of a ship) "port the helm"
  - $s_2^v$: port (bring to port) "the captain ported the ship at night"
  - $s_3^v$: port (land at or reach a port) "The ship finally ported"
  - $s_4^v$: port (turn or go to the port or left side, of a ship) "The big ship was slowly porting"
  - $s_5^v$: port (carry, bear, convey, or bring) "The small canoe could be ported easily"
  - $s_6^v$: port (carry or hold with both hands diagonally across the body, especially of weapons) "port a rifle"
  - $s_7^v$: port (drink port) "We were porting all in the club after dinner"
  - $s_8^v$: port (modify (software) for use on a different machine or platform)

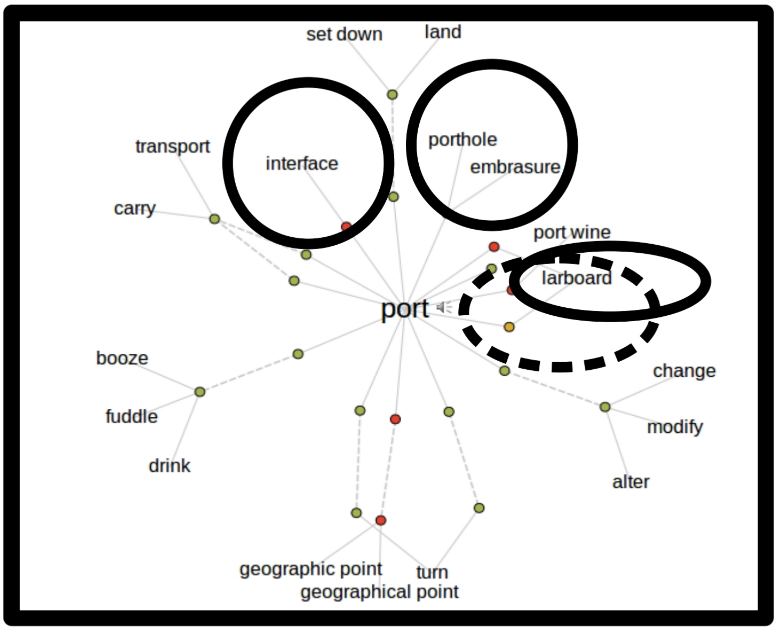

**Fig. 2.** Illustration of the synsets of the word "port" from a visual thesaurus[5]: the circle with continuous line represents the noun synsets while the circle with dashed line represents the adjective synset.

[5] http://www.visualthesaurus.com/landing/

The sets of candidate words for word "port" are defined by $C^a$=(larboard), $C^r = \emptyset$, $C^n$=(embrasure,porthole,larboard,interface) and $C^v = \emptyset$. $C^r = \emptyset$ or $C^v = \emptyset$ means that there is no candidate word, i.e., there is no other sense than the initial one conveyed by the word "port". Finally, the conceptual vector of word "port" is given by:
$V$(port)=(larboard[1.00])$_a$(embrasure[1.00]porthole[1.00]larboard[1.00]interface[1.00])$_n$

The distance $\delta$ between two words $w_1$,$w_2$, or between a word and a candidate word in the context of conceptual vectors, is computed as follow. Let G={adjective (a), adverb (r), noun (n), verb (v)} be the set of grammatical categories in the WordNet dictionary and P={P$^a$,P$^r$,P$^n$,P$^v$} a set of common parents of these words in the WordNet lexical network. $w_1$ and $w_2$ have a common parent if they share some semantic relations (hypernym, hyponym, holonym, troponym etc.). The distance between these two words is defined as:

$$\delta(w1,w2)^6 = \begin{cases} 1 & \text{if } w_1 \text{ and } w_2 \text{ don't have a common parent} \\ \| \left[\frac{min(d(w_1,P^g),d(w_2,P^g))}{min(d(w_1,P^g),d(w_2,P^g))+d(P^g,R)}\right]_{g\in G} \| \end{cases} \quad (2)$$

where $d(w_1,w_2)$ is the number of arcs between nodes $w_1$ and $w_2$, $R$ the root node of the lexical network and $\|\|$ is the infinity norm. The shorter the distance between two words, the higher the semantic proximity between them.

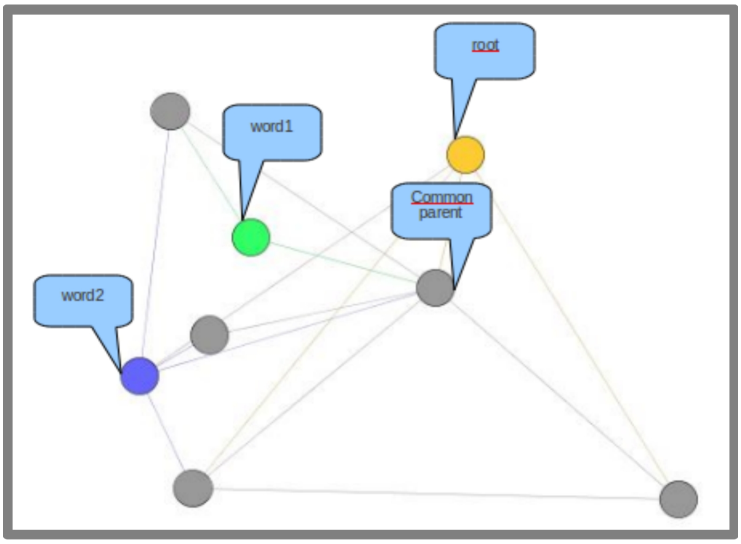

**Fig. 3.** Illustration of the structure of the lexical network used by WordNet.

[6] http://www.rednoise.org/rita/wordnet/documentation/riwordnet_method_getdistance.htm

Figure 3 represents an example of organisation of words in the WordNet dictionary for an arbitrary grammatical category $g = v$. In such a graph, the distance $\delta(word1, word2)$ between the two words is equal to $\parallel \frac{1}{1+1} \parallel$=0.5.

### 2.2 Conceptual vector of sentence

Assuming the principle that a sentence is a collection of polysemic words, we define the conceptual vector of a sentence as the direct sum of conceptual vectors of the words that the sentence contains. For each conceptual vector of a word, we take into account the grammatical category which is the same as the part-of-speech of this word in the sentence. In WordNet, the synsets of a word are computed for four grammatical categories (adjective, adverb, noun, verb). For a word where synset extraction is not possible from its part-of-speech, all the conceptual vectors are determined regardless of its part-of-speech in the sentence. A conceptual vector of a sentence gathers all concepts that the sentence relates. In the definition the conceptual vector does not take into account the type (i.e., declarative, exclamatory, interrogative, imperative) and form (i.e., affirmative/negative, active/passive, neutral/emphatic) of a sentence. Let $s$ be a sentence composed of $n$ words $w_i$, then the conceptual vector of the sentence $s$ is defined by:

$$V(s) = \sum_{i=1}^{n-1} V(w_i) \oplus V(w_{i+1}) \tag{3}$$

The direct sum between two conceptual vectors of a word is defined as a union of candidate words organized by grammatical category, each one weighed by its distance to the word that it represents (cf. equation 2). The direct sum of conceptual vectors is usually used to extend the field of concepts of the working space (i.e., the context).

Let $w_i$ and $w_j$ be two words and $V(w_i)$ and $V(w_j)$ their corresponding conceptual vectors, the direct sum between these two conceptual vectors is expressed by the equation:

$$\begin{aligned} V(w_i) \oplus V(w_j) = \quad & \bigcup_{k=i,j} \bigcup_{l=1}^{m^a_{w_k}} c^a_{l,w_k} \delta(c^a_{l,w_k}, w_k) \Bigg) \Bigg)_a \quad \bigcup_{k=i,j} \bigcup_{l=1}^{m^r_{w_k}} c^r_{l,w_k} \delta(c^r_{l,w_k}, w_k) \Bigg) \Bigg)_r \\ & \bigcup_{k=i,j} \bigcup_{l=1}^{m^n_{w_k}} c^n_{l,w_k} \delta(c^n_{l,w_k}, w_k) \Bigg) \Bigg)_n \quad \bigcup_{k=i,j} \bigcup_{l=1}^{m^v_{w_k}} c^v_{l,w_k} \delta(c^v_{l,w_k}, w_k) \Bigg) \Bigg)_v \end{aligned} \tag{4}$$

In the particular case where $w_i = w_j$, $V(w_i) \oplus V(w_i) = V(w_i)$.

As an example, let us consider the following sentence $s$="The ship is in the port". The initial disambiguisation of this sentence is proposed in table 1:

| **word** | **part-of-speech**[7] | **lemma** |
|---|---|---|
| The | DT | the |
| ship | NN | ship |
| is | VBZ | be |
| in | IN | in |
| the | DT | the |
| port | NN | port |

**Table 1.** Disambiguisation of the sentence

The conceptual vectors of the different words in a sentence $s$ according to their part-of-speech are:

- $V$(the)=$\emptyset$
- $V$(ship)=$\emptyset$
- $V$(is)=(be[0.00]exist[0.00]equal[0.00]constitute[0.00]represent[0.00]comprise[0.00] follow[0.00]embody[0.00]personify[0.00]live[0.00]cost[0.00])$_v$
- $V$(in)=(inwards[1.00]inward[1.00])$_n$(inch[0.00]indium[0.00])$_r$
- $V$(the)=$\emptyset$
- $V$(port)=(embrasure[1.00]porthole[1.00]larboard[1.00]interface[1.00])$_n$

The direct sum between $V$(ship) and $V$(port) is:

$$V(\text{ship}) \oplus V(\text{port}) = (\text{embrasure[1.00]porthole[1.00]larboard[1.00]interface[1.00]})_n$$

The resulting normalised (see section 3.1) conceptual vector is:

$$\begin{aligned} V(\text{“The ship is the in the port”}) &= V(\text{the}) \oplus V(\text{ship}) \oplus V(\text{is}) \oplus V(\text{in}) \oplus V(\text{the}) \oplus V(\text{port}) \\ &= (\text{inch[0.00]indium[0.00]})_r \\ &\quad (\text{inwards[0.38]inward[0.40]embrasure[0.47]} \\ &\quad \text{porthole[0.52]larboard[0.58]interface[0.66]})_n \\ &\quad (\text{be[0.00]exist[0.00]equal[0.00]} \\ &\quad \text{constitute[0.00]represent[0.00]comprise[0.00]follow[0.00]} \\ &\quad \text{embody[0.00]personify[0.00]live[0.00]cost[0.00]})_v \end{aligned}$$

## 3 Decision space for the extraction of semantic information

The goal of this section is to extract the semantic information related to a concept. We introduce a decision space where the different conceptual vectors of

[7] http://en.wikipedia.org/wiki/Brown_Corpus#Partofspeech_tags_used

the sentences which describe this concept must be projected. The decision space contains a list of feasible options identified in the macro-phases of the decision strategy. Jankowski and Nyerges summarized the macro-phase of the decision strategy in three steps [14]: (1) intelligent about the values, objectives and criteria (2) design of a set of feasible options, (3) choice about recommendations. The feasible options should be linked to the objectives and validated by an expert. The conceptual vectors of words that are correlated to the semantics we are searching for define the semantic axes (i.e., the basis) of the decision space.

### 3.1 Projection of a sentence in a decision space

The projection of a sentence $s$ in a decision space corresponds to the projection of the candidate words of the conceptual vector of $s$ (i.e., $V(s)$) in order to valuate the contribution in each semantic direction of the basis. We thus derive the principal direction detailed in the next subsection.

Let us assume that one want to compute the contribution $x_i^c$ of a candidate word of $V(s)$ in the semantic direction $d$ and category $c$ where $n$ is the number of candidate words of the conceptual vector $V(d)$ that defines a semantic axis of the decision space. If $c_i^c$ is the common candidate word of the two conceptual vectors $V(s)$ and $V(d)$ with weights $\delta_s$ and $\delta_d$ respectively then

$$x_i^c = \frac{(1 - \delta_s * \delta_d)}{n}. \tag{5}$$

$x_i^c$ equals to zero if the candidate word of $V(s)$ does not belong to $V(d)$. In equation 5, the value is weighed by the number of candidate words in the semantic axis considering that a candidate word of the conceptual vector of a sentence $V(s)$ has a higher influence if the number of candidate words of $V(d)$ is low. A candidate word of $V(s)$ having a weight equal to 1 (i.e., a poor semantic contribution) may have no contribution in a semantic axis (i.e., $x_i^c = 0$ if $\delta_s = 1$ and $\delta_d = 1$) and is not taken into account in the final decision. To tackle this problem, one can normalise the conceptual vector of a sentence (section 2.2), discarding the case where a weight is equal to 1.

The contribution of a sentence in a category $c$ is the sum of the contribution of the $m_c$ candidate words of this category, i.e., $x^c = \sum_{i=1}^{m_c} x_i^c$. Finally, the contribution of a sentence is the sum of the contributions in each category, i.e., $x = x^a + x^r + x^n + x^v$. The higher the semantic contribution in a direction, the higher the projection value $x$. This process is repeated in all the semantic directions that contribute to the decision space.

Let us illustrate this principle with the following example where one want to find the contribution of the word “stay” in the semantic direction “stop”. We firstly define the conceptual vectors of these two words:

$V(\text{stop})$ = (halt[0.00]block[0.00]check[0.00]
arrest[0.00]blockade[0.12] bar[0.14] end[0.14] finish[0.14]barricade[0.22]
break[0.29]cease[0.67]intercept[0.73]kibosh[1.00]
terminate[1.00]contain[1.00]quit[1.00]discontinue[1.00]$)_v$
(halt[0.00]stoppage[0.00]stopover[0.00]
layover[0.00]arrest[0.00]check[0.00]hitch[0.00]stay[0.00]
occlusive[0.00]plosive[0.00]period[0.00]point[0.00]
diaphragm[0.00]catch[0.00]blockage[0.00]block[0.00]
closure[0.00]occlusion[0.00]$)_n$

$V(\text{stay})$ = (remain[0.00]rest[0.00]stick[0.00]
bide[0.00]abide[0.00]continue[0.00]detain[0.00]delay[0.00]
persist[0.00]outride[0.00]quell[0.00]appease[]$)_v$
(arrest[1.00]check[0.33]halt[0.33]stop[0.33]hitch[0.50]
stoppage[1.00]$)_n$

Secondly, we compute the projection values of the candidate words of $V$(stay) in each category. Projection is always null except for common candidate words $c_1^n$=arrest, $c_2^n$=check, $c_3^n$=halt, $c_4^n$=hitch , $c_5^n$=stoppage. Projections values are computed as follow:

- $x_1^n$=$\frac{1-0.00*1}{35}$=0.029, with $c_1^n$=arrest and $n = 35$, $\delta_{stop}$=0.00 , $\delta_{stay}$=1
- $x_2^n$=$\frac{1-0.00*0.33}{35}$=0.029, with $c_2^n$=check and $n = 35$, $\delta_{stop}$=0.00 , $\delta_{stay}$=0.33
- $x_3^n$=$\frac{1-0.00*0.33}{35}$=0.029, with $c_3^n$=halt and $n = 35$, $\delta_{stop}$=0.00 , $\delta_{stay}$=0.33
- $x_4^n$=$\frac{1-0.00*0.50}{35}$=0.029, with $c_4^n$=hitch and $n = 35$, $\delta_{stop}$=0.00 , $\delta_{stay}$=0.50
- $x_5^n$=$\frac{1-0.00*1}{35}$=0.029, with $c_5^n$=stoppage and $n = 35$, $\delta_{stop}$=0.00 , $\delta_{stay}$=1

For category noun, we deduce that the contribution of word "stay" in the direction "stop" is: $x^n = x_1^n + x_2^n + x_3^n + x_4^n + x_5^n = 0.15$. It results that the final contribution value is: $x = x^a + x^r + x^n + x^v = 0.15$

### 3.2 Principal semantic direction for a concept

This section aims at determining the principal semantics in a decision space that is associated to a concept. The first stage of the process consists in identifying the sentences of the corpus related to this concept and to project each of them in the decision space. The next stage focuses on the computation of the main contribution of these sentences. The contribution of the sentences in one semantic axis of the decision space is the sum of the contributions of the sentences regarding this direction. We apply this principle in all the semantic directions that contribute to the decision space. The semantic direction which warrants the highest trust is the one which has the highest coordinate or score. This semantic direction is called the principal semantic direction associated to the concept.

In the case where at least two directions have the same score, the direction which ensures security of mariner is considered. Regarding experts, these decisions are classified according to the decreasing security order: back, stop, maneuver and continue. As a result in table 2, the decision corresponding to the "low water" concept is "to go back".

| $\frac{Affordance}{Sentence}$ | **maneuver** | **stop** | **continue** | **back** |
|---|---|---|---|---|
| **Total** | 0.12 | **0.12** | 0.02 | 0.12 |

**Table 2.** Projection values of sentences that refer to the concept "low water" in the decision space ($V$(maneuver),$V$(stop),$V$(continue),$V$(back)).

## 4 Case study

Let us illustrate our approach by the analysis of the semantics associated to the concept of "anchorage area". The sentences related to this concept in our corpus are:

$s_1$: "Any vessel anchored outside of the prescribed anchorage limits must move to a prescribed **anchorage area** when space becomes available".

$s_2$: "Whenever, in the opinion of the captain of the port such action may be necessary, he may require any or all vessels in any designated **anchorage area** to moor with two or more anchors".

$s_3$: "Reserved anchors shall be placed well within the anchorage areas, so that no portion of the hull or rigging will at any time extend outside of the **anchorage area**".

$s_4$: "Except in cases where unforeseen circumstances create conditions of imminent peril, or with the permission of the captain of the port, no vessel shall be anchored in baltimore harbor and patapsco river outside of the **anchorage areas** established in this section for more than 24 hours".

$s_5$: "Any vessel anchoring, under great emergency, within this area shall be placed as close to an **anchorage areas** as practicable, and shall move away immediately after the emergency ceases".

In a second stage, one decide to extract the behaviour that a mariner can decide facing to the concept "anchorage area". We restrict our case study to the actions or affordances (continue, stop, go back and maneuver) that identify the four semantic axes of our decision space defined by the basis of conceptual vectors ($V$(maneuver),$V$(stop),$V$(continue),$V$(back)). As regards experts in maritime navigation, these actions describe the different actions a mariner can take in front of an object in the real word. Each of these actions or situations has the following meaning: maneuver indicates to the mariner that he must change his trajectory; stop indicates to the mariner that he must temporarily remain in his navigation area (for example, the vessel enters in an anchorage area or he receives a special signal which requires him to stop the navigation); back denotes that he must turn around because the environment becomes dangerous or impracticable (for example, in the presence of dense fog or strong storm); and continue proposes to mariner he can follow the same trajectory because none unsafe event is detected. The principal affordance we can associate to the concept "anchorage area" is *stop*, because it has the highest coordinate with value $0.06 = 0.01 + 0.00 + 0.00 + 0.01 + 0.04$. The coordinates of the projections of each sentence in the decision space are summarized in table 3 and show a visualisation of this decision space in figure 4.

| $\frac{Affordance}{Sentence}$ | **maneuver** | **stop** | **continue** | **back** |
|---|---|---|---|---|
| $s_1$ | 0.00 | 0.01 | 0.00 | 0.00 |
| $s_2$ | 0.01 | 0.00 | 0.00 | 0.01 |
| $s_3$ | 0.01 | 0.00 | 0.00 | 0.01 |
| $s_4$ | 0.00 | 0.01 | 0.00 | 0.00 |
| $s_5$ | 0.01 | 0.04 | 0.01 | 0.00 |
| **Total** | 0.03 | **0.06** | 0.01 | 0.02 |

**Table 3.** Projection values of sentences that refer to the concept "anchorage area" in the decision space ($V$(maneuver),$V$(stop),$V$(continue),$V$(back)).

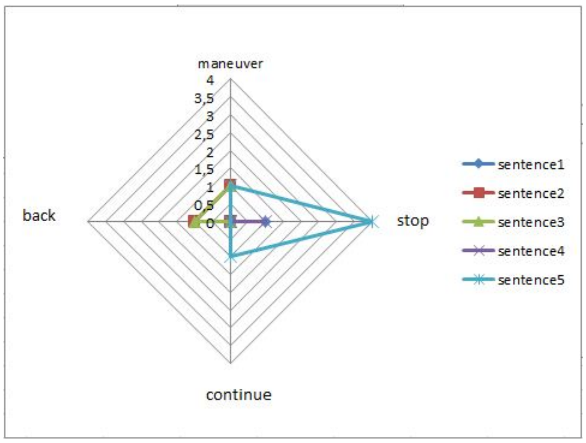

**Fig. 4.** Visualisation of the decision space for the concept "anchorage area".

Let us illustrate our strategy with a second example in which we try to extract the affordance related to the concept "cardinal buoy". Accordingly and using the same documents previously cited, one selects the sentences:

$s_1$: "For example, a particular **cardinal buoy** represented through a symbol on a chart".

$s_2$: "The top marks of **cardinal buoys** consist of the combination of two black cones mounted one above the other on the top of the buoy with the following, combinations:
(a) both cones pointing up = North cardinal,
(b) both pointing down = South cardinal,
(c) one pointing up and the other down with their bases together = East cardinal,
(d) one pointing up and the other pointing down with their points together = West cardinal.

The principal affordance that can be associated to the concept "cardinal buoy" is *maneuver*, because it has the highest coordinate with value 0.04=0.02+0.02. We summarize the coordinates of the projections of each sentence in the decision space in table 4 and show a visualisation of this decision space in figure 4.

| $\frac{Affordance}{Sentence}$ | **maneuver** | **stop** | **continue** | **back** |
|---|---|---|---|---|
| $s_1$ | 0.02 | 0.00 | 0.00 | 0.00 |
| $s_2$ | 0.02 | 0.01 | 0.00 | 0.00 |
| **Total** | **0.04** | 0.01 | 0.00 | 0.01 |

**Table 4.** Projection values of sentences that refer to the concept "cardinal buoy" in the decision space ($V$(maneuver),$V$(stop),$V$(continue),$V$(back)).

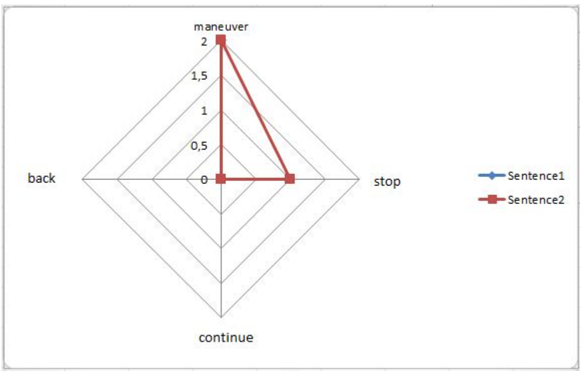

**Fig. 5.** Visualisation of the decision space for the concept "cardinal buoy".

### 4.1 Glosses of the concepts of a sentence

In some cases, the projection of a word or a sentence in a decision space generates a conceptual vector whose euclidean norm equals zero (i.e., each semantic contribution is equal to zero) and no decision emerges. We use the concept of gloss to improve the results. The glosses of a word are the different definitions of it. For example, a gloss of word "port" may be "a place (seaport or airport) where people and merchandise can enter or leave a country" (definition from synset $s_1^n$ of subsection 2.1). As a result when no decision is proposed, the principle of the semantic extraction strategy is to use the definition of a word (i.e., its gloss) to extract the semantic information related to it.

The new coordinate of a sentence whose conceptual vector is null is computed by using the glosses of the each word in this sentence. For each word $w$, we extract the different glosses and project them in our decision space (see subsection 3.1). The most relevant gloss of word $w$ is the gloss having the highest coordinate in the decision space. The infinity norm ($\|\|_\infty$) is applied to find the most contributing gloss of a word. Lastly the coordinates generated by each word of the initial sentence is sumed to get a new coordinate for it.

For example, two sentences in the corpus are related to the concept of "dense fog":

$s_1$: "From April to September there are only a few days with **dense fogs**".

$s_2$: "**Dense fog** is more common offshore and should be expected on unusually warm, humid winter and spring days".

This implies that no decision is taken since the projection of the concept "dense fog" is null in each semantic axis, i.e.:

$s_1$: maneuver[0.00]stop[0.00]continue[0.00]back[0.00]

$s_2$: maneuver[0.00]stop[0.00]continue[0.00]back[0.00]

Consequently, the glosses of the terms of sentences $s_1$ and $s_2$ are used in order to try to find a more accurate decision. The projections of the different glosses in the decision space give the results presented in table 5 and lead to the decision to *go back*:

| $\frac{Affordance}{Sentence}$ | **maneuver** | **stop** | **continue** | **back** |
|---|---|---|---|---|
| $s_1$ | 0.00 | 0.01 | 0.01 | 0.12 |
| $s_2$ | 0.00 | 0.05 | 0.02 | 0.35 |
| **Total** | 0.00 | 0.06 | 0.03 | **0.47** |

**Table 5.** Projection values of sentences that refer to the concept "dense fog" in the decision space ($V$(maneuver),$V$(stop),$V$(continue),$V$(back)) using the concept of gloss.

## 5 Conclusion and further work

This paper introduces a general strategy to extract semantic information from a corpus. We assume that the analysis of documents written by experts in a specific domain gives richer information than the exploitation of usual definitions found in common dictionaries. Accordingly, and in order to propose a vector of concepts of a word or a sentence, we use WordNet a lexical database of English where nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Each component of this conceptual vector called “candidate word” is associated to a value which quantifies the semantic distance between the word or the sentence and its candidate word.

The extracted semantic information identifies the root of the second stage devoted to decision aid. The goal of this second stage is to extract the right decision with respect to a concept. The process applied is to define a decision space made up of different semantic axes in correlation with the application domain in which sentences are projected. The final decision is derived from the analysis of the main contribution observed in the semantic directions. This emphasizes the fact that the semantic richness of the initial corpus is important and influences the success of the strategy more than the choice of the WordNet dictionnary. To improve the results, the initial strategy is extended by considering not only the synsets of a word but also its glosses. The proposed strategy is applied to the extraction of semantic information in the maritime context for navigation aids but the process can easily be applied to other domains.

Further work concerns the development of a real time navigation aid platform which takes into account semantic information generated by objects (lighted buoy, water wayroute, radio signal, ships, etc.) or exterior events (wind, fog, stream, etc.) which appear in the vicinity of the ship. We assume that the descriptions and rules about these objects appear in the initial corpus used for disambiguisation. This platform will be coupled with a spatio-temporal ontology of the maritime environment that will store the initial and the extracted knowledge.

**Acknowledgements**
The authors grateful to the “Établissement Principal du Service Hydrographique et Océanographique de la Marine” (Brest-France) for their nautical documents and for their assistance in this study. Also thank to Thierry Hamon for his supporting in the use of software Yatea. We would also like to thank Mathieu Lafourcade for his introduction to conceptual vectors and Daniel Howe for his help on RiWordnet, and Alicino Ferreira for his careful proofreading of this article.

## References

1. Aubin, S., Hamon, T.: Improving term extraction with terminological resources. In: Salakoski, T., Ginter, F., Pyysalo, S., Pahikkala, T. (eds.) Proceeding of the 5th

International Conference on NLP, FinTAL 2006, Advances in Natural Language Processing. pp. 380–387. No. 4139 in LNAI, Springer (August 2006)

2. Becker, J., Grob, L., Hellingrath, B., Klein, S., Kuchen, H., Müller-Funk, U., Vossen, G.: Advances in Information Systems and Management Science (2004)
3. Beckwith, R., Fellbaum, C., Gross, D., Miller, K.: Introduction to WordNet : An Online Lexical Database. International Journal of Lexicography 3(4), 235–244 (1990)
4. Chauché, J.: Détermination sémantique en analyse structurelle : une expérience basée sur une définition de distance. TAL Information 31(1), 17–2 (1990)
5. Claramunt, C., Fournier, S., Li, X., Peytchev, E.: Real-time Geographical Information for ITS. In: Proceedings of the 5th IEEE International Conference in Intelligent Transportation Systems. pp. 237–242 (2005)
6. Cleverdon, C.W.: Report on Testing and Analysis of an Investigation into the Comparatie Efficiency of Indexing Systems (1962)
7. Daniel, C.H.: *a WordNet Library for Java Processing*, http://www.rednoise.org/rita/wordnet/documentation/index.htm
8. Dumais, S.T., Letsche, T.A., Littman, M.L., Landauer, T.K.: Automatic Cross-Language Retrieval using Latent Semantic Indexing. In: AAAI-97 Spring Symposium Series: Cross-language Text and Speech Retrieval. pp. 18–24 (1997), http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.50.5717
9. Gibson, J.J.: The Theory of Affordances, pp. 67–82. Lawrence Erlbaum (1977)
10. Goralski, R., Gold, C.: Marine GIS: Progress in 3D Visualization for Dynamic GIS. In: Spatial Data Handling. pp. 401–416 (2008)
11. Hiemstra, D.: Using Language Models for Information Retrieval. Ph.D. thesis, Taaluitgeverij Neslia Paniculata (January 2001)
12. International Hydrographic Bureau, MONACO: Recommended ENC Validation Checks (2011)
13. James, M., Vicki, G.: The Handbook of Delaware Boating Laws and Responsabilities (2011)
14. Jankowski, P., Nyerges, T.: Geographic Information Systems for Group Decision Making: Towards a Participatory, Geographic Information Science (2003)
15. Lafourcade, M., Prince, V., Schwab, D.: Vecteurs conceptuels et structuration émergente de terminologie. Traitement Automatiques des Langues 43(1), 43–72 (2002)
16. National Oceanic and Atmospheric Administration, US Department of Commerce: United States Coast Pilot, 44th edn. (2011)
17. Néméta, A.: Code Vagnon Permis Plaisance : Option cotière, Vagnon edn. (2008)
18. Pearson, M.: Mémento Vagnon du Skipper : Moteur et voile (2008)
19. Potthast, M., Stein, B., Anderka, M.: A Wikipedia-based multilingual retrieval model. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) Proceedings of the 30$^{th}$ European Conference on IR Research, ECIR 2008, Advances in Information Retrieval. LNCS, vol. 4956, pp. 522–530. Springer, Berlin/Heidelberg (2008), http://dx.doi.org/10.1007/978-3-540-78646-7_51
20. Salton, G., MacGill, M.: Introduction to Modern Information Retrieval (1983)
21. Schmid, H.: Treetagger - a language independent part-of-speech tagger, http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
22. Singhal, A.: Modern information retrieval: A brief overview. In IEEE Data Engineering Bulletin 24(4), 35–43 (2001)
23. Strzalkowski, T., Carballo, J., Marinescu, M.: Natural language information retrieval: Trec-3-report. In: Proceedings of the 3$^{rd}$ Text Retrieval Conference (1994)

24. Tsatsaronis, G., Panagiotopoulou, V.: A generalized vector space model for text retrieval based on semantic relatedness. The 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-09) (April 2009)
25. U.S. Department of Transportation, United States Coast Guard: Navigation Rules International-InLand (2011)