Argentinian Journal of Applied Linguistics
Vol. 7, No. 1, May 2019, 27-57

Extraction of terms for the construction of semantic frames for named bays

Juan Rojas-Garcia* - University of Granada (Spain)
Pamela Faber - University of Granada (Spain)

(Received 9/01/19; final version received 23/03/19)

ABSTRACT

EcoLexicon (http://ecolexicon.ugr.es) is a terminological knowledge base on environmental science, whose design permits the geographic contextualization of data. For the geographic contextualization of LANDFORM concepts, this paper presents a semi-automatic method of extracting terms associated with named bays (i.e., *Greenwich Bay*).

Terms were extracted from a specialized corpus, where named bays were automatically identified. Statistical procedures were applied for selecting both terms and bays in distributional semantic models to construct the conceptual structures underlying the usage of named bays. The bays sharing associated terms were also clustered and represented in the same conceptual network.

The results showed that the method successfully described the semantic frames of named bays with explanatory adequacy, according to the premises of Frame-based Terminology.

Keywords: Named bay, Conceptual information extraction, Geographical contextualization, Text mining, Frame-based Terminology.

RESUMEN

EcoLexicon (http://ecolexicon.ugr.es) es una base de conocimiento terminológica sobre ciencias medioambientales, cuyo diseño permite la contextualización geográfica de conceptos de la categoría ACCIDENTE GEOGRÁFICO. Para tal fin, este artículo presenta un método semiautomático para extraer términos asociados con bahías con nombre propio (e.gr., *Bahía de Pensacola*).

Los términos se extrajeron de un corpus especializado, donde las designaciones de bahías se identificaron automáticamente. Se aplicaron procedimientos estadísticos para seleccionar bahías y términos, que se proyectaron en espacios semánticos vectoriales, y se emplearon para construir las estructuras conceptuales que subyacían en el uso de la bahías.

Los resultados muestran que el método es apropiado para describir los marcos semánticos que evocan las bahías, según las premisas de la Terminología basada en Marcos.

Palabras clave: Bahía con nombre propio, Extracción de información conceptual, Contextualización geográfica, Minería de textos, Terminología basada en Marcos.

* Corresponding author, e-mail: juanrojas@ugr.es

THE ELECTRONIC RESOURCE EcoLexicon is a multilingual, terminological knowledge base on environmental science (http://ecolexicon.ugr.es) that is the practical application of Frame-based Terminology (Faber, 2012). Since most concepts designated by environmental terms are multidimensional (Faber, 2011), the flexible design of EcoLexicon permits the contextualization of data so that they are more relevant to specific subdomains, communicative situations, and geographic areas (León-Araúz, Reimerink & Faber, 2013). However, the geographic contextualization of LANDFORM concepts depends on knowing which terms are semantically related to each landform, and how these terms are related to each other.

This paper presents a semi-automatic method of extracting terms associated with named bays (i.e., *Escambia Bay*) as a type of landform from a corpus of English Coastal Engineering texts. The aim is to represent that knowledge in a semantic network in EcoLexicon according to the theoretical premises of Frame-based Terminology.

The rest of this paper is organized as follows. Section 2 provides motivations for the research, and background on distributional semantic models and clustering techniques. Section 3 explains the materials and methods applied in this study, namely, the automatic identification of named bays, the selection procedures for terms and bays in distributional semantic models, and the clustering technique for bays sharing associated terms. Section 4 shows the results obtained. Finally, Section 5 discusses the results and presents the conclusions derived from this work as well as plans for future research.

Background and Literature Review

Motivations for the Research

Despite the fact that named landforms, among other named entities, are frequently found in specialized texts on environment, their representation and inclusion in knowledge resources has received little research attention, as evidenced by the lack of named landforms in terminological resources for the environment such as DicoEnviro[2], GEMET[3] or FAO Term Portal[4]. In contrast, AGROVOC[5] contains basically a list of named landforms with hyponymic information, whereas ENVO[6] provides descriptions of the named landforms with only geographic details, and minimal semantic information consisting of the relation *located_in* (and *tributary_of* in the case of named rivers and bays).

So far, knowledge resources have limited themselves to representing concepts such as BAY, RIVER or BEACH, on the assumption that the concepts linked to each of them are applicable, respectively, to all named bays, rivers and beaches in the real world. This issue is evident in the following description of forcing mechanisms acting on suspended sediment concentrations (SSC) in bays and rivers.

According to Moskalski and Torres (2012), temporal variations in the SSC of bays and rivers are the result of a variety of forcing mechanisms. River discharge is a primary controlling factor, as well as tides, meteorological forcing (i.e., wind-wave resuspension,

offshore winds, storm and precipitation), and human activities. Several of these mechanisms tend to act simultaneously. Nonetheless, the specific mix of active mechanisms is different in each bay and river. For example, SSC in San Francisco Bay is controlled by spring-neap tidal variability, winds, freshwater runoff, and longitudinal salinity differences, whereas precipitation and river discharge are the mechanisms in Suisun Bay. In Yangtze River, SSC is controlled by tides and wind forcing, whereas river discharge, tides, circulation, and stratification are the active forcing mechanisms in York River.

Consequently, in a knowledge resource, a list of forcing mechanism concepts semantically linked to BAY and RIVER concepts would not represent the knowledge really transmitted in specialized texts. To cope with this type of situation, terminological knowledge bases should include the semantic representation of named landforms.

To achieve that aim in EcoLexicon regarding named bays, the knowledge should be represented in a semantic network according to the theoretical premises of Frame-based Terminology, which propose knowledge representations with explanatory adequacy for enhanced knowledge acquisition (Faber, 2009). Hence, each named bay should appear in the context of a specialized semantic frame that highlights both its relation to other terms and the relations between those terms. The construction of these semantic networks and the semi-automatic extraction of terms from a specialized corpus are described in this paper. As far as we know, this framework has not been studied in the context of specialized lexicography, which constitutes an original aspect of this work.

Distributional Semantic Models

Distributional semantic models (DSMs) represent the meaning of a term as a vector, based on its statistical co-occurrence with other terms in the corpus. According to the distributional hypothesis, semantically similar terms tend to have similar contextual distributions (Miller & Charles, 1991). The semantic relatedness of two terms is estimated by calculating a similarity measure of their vectors, such as Euclidean distance or cosine similarity (Salton & Lesk, 1968), *inter alia*.

Existing DSMs can be classified, based on two criteria, namely, the leveraged distributional information (Sahlgren, 2008), and the underlying language model (Baroni, Dinu & Kruszewski, 2014). According to the former criterion, models can be syntagmatic or paradigmatic.

Syntagmatic models capture combinatorial relations between terms, namely, non-hierarchical relations such as the effect of an entity on a process (e.g., ... *the Bay of Fundy, because of its basin geometry, amplifies tides*); where a process takes place (e.g., *Wind system changes affect also relative sea level as observed, for example, in the Hudson Bay*); or the location of an entity (e.g., *Many of the beaches along eastern Hudson Bay are characterized by boulder-strewn tidal flats*). Such syntagmatic relations are reflected in terms that co-occur within the same text region, either sentence, paragraph, or document

(Manning, Raghavan & Schütze, 2008). Latent Sematic Analysis (LSA) (Deerwester, Dumais, Furnas, Landauer & Harshman, 1990) is an example of a syntagmatic model, whereby a term-document matrix of co-occurrences is first built to collect the normalized frequency of a term in a document, and the Singular Value Decomposition (Jolliffe, 2002) is then applied to reduce the number of columns to a few orthogonal latent dimensions.

Paradigmatic models are based on taxonomic relations such as hyponymy (e.g., *The Bay of Fundy is a low wave-energy environment that is dominated by tidal processes*) and meronymy (e.g., *Debris litters the bay floor along parts of the developed western shoreline of Greenwich Bay*). In these methods, a term-term matrix of co-occurrences indicates how many times context terms co-occur with a target term within a sliding context window, which spans a certain number of terms on either side of the target term. Hyperspace Analogue to Language (HAL) (Lund, Burges & Atchley, 1995) is an example of a paradigmatic model.

According to the second classification criterion, DSMs are either count-based or prediction-based. *Count-based models* calculate the frequency of the terms that occur within a term's context (i.e, a sentence, paragraph, document or context window of a certain size). LSA, HAL, Global Vectors (GloVe) (Pennington, Socher & Manning, 2014), and Correlated Occurrence Analogue to Lexical Semantic (COALS) (Rohde, Gonnerman & Plaut, 2006) are examples of this type of model. *Prediction-based models* exploit neural probabilistic language models, which represent terms by predicting the next term based on previous terms. Examples of predictive models include the continuous bag-of-words (CBOW) and skip-gram models (Mikolov, Chen, Corrado & Dean, 2013), Parallel Document Context (Sun, Guo, Lan, Xu, & Cheng, 2015), and Collobert and Weston model (Collobert & Weston, 2008).

The applications of DSMs in lexical and computational semantics include the following:

- Identification of semantic relations. DSMs are useful tools for Terminology, since they can help identify semantic relations between terms based on corpus data (Bertels & Speelman, 2014; Bernier-Colborne & L'Homme, 2015; Reimerink & León-Araúz, 2017). In addition, knowledge of a few seed terms and their relationships can help to infer analogous relationships for other similar terms that are nearby in the DSM (Hearst & Schütze, 1993; Widdows, 2003; Thompson, Batista-Navarro, Kontonatsios, Carter, Toon, McNaught, Timmermann, Worboys & Ananiadou, 2015).
- Information retrieval. Search engines can locate documents based on synonyms and related terms as well as matching keywords (Deerwester et al. 1990; Nguyen, Soto, Kontonatsios, Batista-Navarro & Ananiadou, 2017).
- Word sense discrimination and disambiguation. The vectors for each of the occurrences of the same term in a corpus (called context vectors) can be clustered, and the centroids of these clusters can be treated as word senses. An occurrence of the

same ambiguous term can then be mapped to one of these word senses, with a confidence level derived from the similarity between the context vector for this occurrence and the nearest centroids (Schütze, 1997 and 1998; Pantel & Lin, 2002).

- Use of word vectors as features for automatic recognition of named entities in text corpora (Turian, Ratinov, Bengio & Roth, 2009; El bazi & Laachfoubi, 2016), and for representation of proper names (Herbelot, 2015).

Clustering Analysis

Clustering is one of the most important unsupervised learning techniques in data analysis (Everitt, Landau & Leese, 2001). It classifies objects into groups (clusters) based on shared features. In hierarchical clustering, objects are successively integrated in inclusive clusters, depicted in dendrograms (Xu & Wunsch, 2009). Clustering techniques are used in many disciplines for purposes of Information Retrieval (Manning et al., 2008) and Text Mining (Feldman & Sanger, 2007), and, increasingly, in Corpus Linguistics (Moisl, 2009).

Work in lexical semantics that applies clustering techniques includes, *inter alia*, analysis of word distribution data in text to derive syntactic and semantic lexical categories (Bullinaria, 2008; Katrenko & Adriaans, 2008; Kiss, 1973; Miller, 1971); automatic induction of verb classes from verb selectional preferences extracted from corpus data (Sun & Korhonen, 2009); automatic metaphor identification in unrestricted text (Shutova, Sun & Korhonen, 2010); and classification of verbs into semantic groups based upon the relationship between words and grammatical constructions (Gries & Stefanowitsch, 2010).

Materials and Methods

Corpus Data

The terms related to named bays were extracted from a subcorpus of English texts on Coastal Engineering. This subcorpus, which comprises roughly 7 million tokens, is composed of specialized and semi-specialized texts, and is an integral part of the EcoLexicon English Corpus (23.1 million tokens) (see León-Araúz, San Martín and Reimerink [2018] for a detailed description).

GeoNames Geographical Database

The automatic detection of the named bays in the corpus was performed with a GeoNames database dump. GeoNames (http://www.geonames.org) has over 10 million proper names for 645 different geographical entities, such as bays, beaches, rivers, mountains, etc. For each entity, information about their normalized designations, alternate designations, latitude, longitude, and location name is stored. A daily GeoNames database dump is publicly available as a worldwide text file.

Pre-processing

After their compilation and cleaning, the corpus texts were tokenized, tagged with parts of speech, lemmatized, and lowercased with the Stanford *CoreNLP* package for R programming language. The multiword terms stored in EcoLexicon were then automatically matched in the lemmatized corpus and joined with underscores.

Named Bays Recognition

Both normalized and alternate names of the bays in GeoNames were searched in the lemmatized corpus. A total of 306 designations were recognized and listed. Nevertheless, since various designations can refer to the same bay because of syntactic variation (e.g., *Bay of Fundy* and *Fundy Bay*) and orthographic variation (e.g., *Choctaw*[*h*]*atchee Bay*), a procedure was created to identify variants and give them a single designation in the corpus.

In the case of syntactic variations without *of*, the preposition was automatically added to the names without it (e.g., *Fundy Bay* was converted to *Bay of Fundy*) and matched in the list of recognized designations. This was only problematic when the variants referred to different bays, such as the case of *Naples Bay* (USA) and *Bay of Naples* (Italy).

Orthographic variations were identified with a matrix of the Levenshtein edit distances between the 306 designations. The Levenshtein distance between two strings is the number of deletions, insertions, or substitutions required to transform the first string into the second one. As such, the pairs of strings with an edit distance of 1 or 2 were manually inspected to discover the orthographic changes.

Once the variants were normalized (Table 1) in the lemmatized corpus and joined with underscores, the number of named bays was 294. They are shown on the map in Figure 1, with color-coded rectangles that depict their frequency in the corpus. Their latitudes and longitudes were retrieved from the GeoNames database dump. This reflects the representativeness of the corpus in reference to bay locations and their number of mentions. As shown in Figure 1, most of the named bays are located in the USA.

Variant	Normalized designation
Paranague Bay	Paranagua Bay
Paranaguo Bay	Paranagua Bay
Choctawatchee Bay	Choctawhatchee Bay
Fundy Bay	Bay of Fundy
Funday Bay	Bay of Fundy
Ingleses Bay	Bay of Ingleses
Josiah's Bay	Josias Bay
Josiah Bay	Josias Bay
Westernport Bay	Western Port Bay
Port Phillip	Port Phillip Bay
Greenwich cove	Greenwich Bay
Halfmoon bay	Half Moon Bay

Table 1. Variants referring to the same bay and their normalized designation.

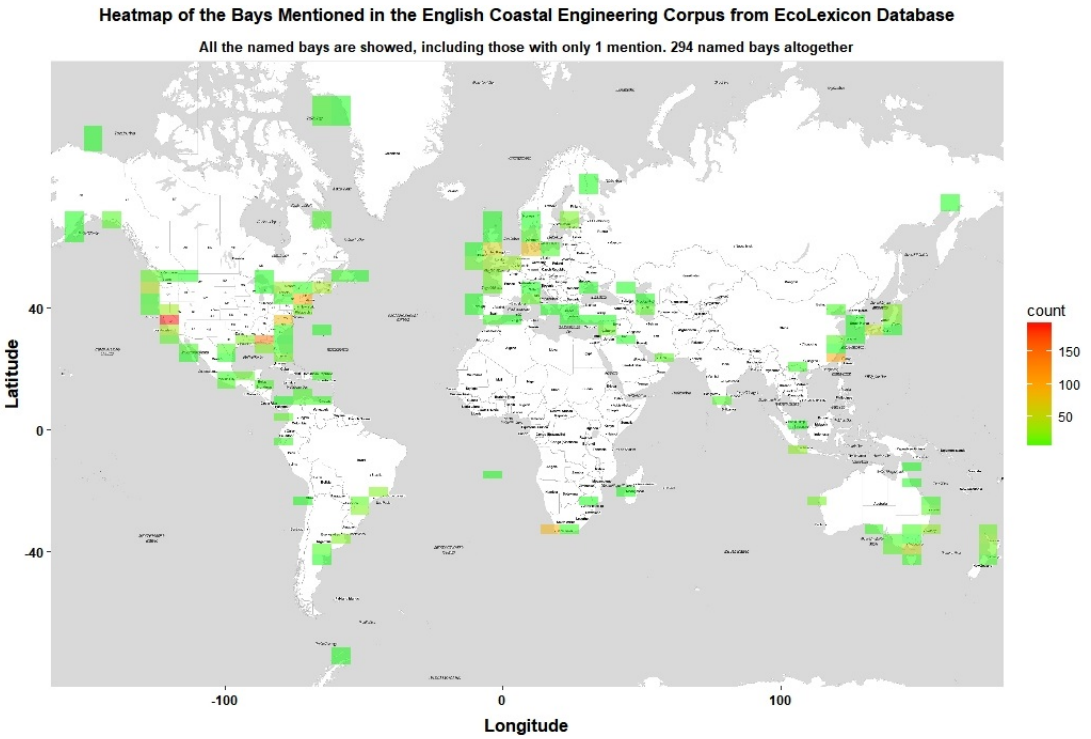

Figure 1. Map with the location and color-coded frequency of the 294 named bays.

A critical issue was the retrieval of the geographical coordinates of the bays. Although latitudes and longitudes could be retrieved from the GeoNames database dump, occasionally, the same designation referred to bays in different countries. For instance, the corpus only located *False Bay* in South Africa. However, GeoNames indicated that bays with the same name also existed in India, Yemen, the USA, Canada, and Australia. Such cases had to be resolved by corpus queries.

With regard to the occurrence frequency of the named bays in the corpus, the values ranged from 127 (*Monterey Bay*) to only one mention (150 of the 294 named bays). In our study, only those bays with an occurrence frequency greater than 5 were considered, since DSMs perform poorly with low-frequency terms (Luhn, 1957). Table 2 shows the 55 named bays that fulfilled this condition, whereas Figure 2 shows their number of mentions.

Country	Named Bays
United States (20)	**California State:** San Francisco Bay, Suisun Bay, Monterey Bay, San Diego Bay, Morro Bay, Back Bay. **Florida State:** Escambia Bay, Pensacola Bay, Tampa Bay, Saint Joseph Bay, Florida Bay. **State of New York:** Long Island Sound, Naples Bay. **State of Rhode Island:** Greenwich Bay, Narragansett Bay. **Other States:** Chesapeake Bay (Virginia), Siletz Bay (Oregon), Mobile Bay (Alabama), Delaware Bay (Delaware)
Australia (5)	**Victoria:** Port Phillip Bay, Western Port Bay, Apollo Bay.

	Other States: Botany Bay (New South Wales), Shark Bay (Western Australia)
United Kingdom (4)	**England:** Pevensey Bay, Start Bay, Morecambe Bay, Liverpool Bay
Japan (3)	Tosa Bay (Shikoku Island), Tokyo Bay (Kanagawa), Kamaishi Bay (Iwate)
Brazil (2)	Sepetiba Bay (Rio de Janeiro), Imbituba Bay (Santa Catarina)
Canada (2)	Bay of Fundy (Nova Scotia), Hudson Bay (Ontario)
France (2)	Baie des Anges (Provence-Alpes-Côte d'Azur), Baie des Veys (Normandy)
Mexico (2)	Bay of Campeche (Campeche), Todos Santos Bay (Baja California)
New Zealand (2)	**Auckland:** Bay of Plenty, Tauranga Harbor
South Africa (2)	**Western Cape:** False Bay, Gordons Bay
The Netherlands (2)	**Wadden Sea:** Ley Bay, Dollard Bay
Argentina (1)	Samborombon Bay (Buenos Aires)
China (1)	Quanzhou Bay (Fujian)
Colombia (1)	Buenaventura Bay (Valle del Cauca)
Denmark (1)	Kogo Bay (Zealand)
Estonia (1)	Tallinn Bay (Harjumaa)
Indonesia (1)	Jakarta Bay (Jakarta)
Iran (1)	Chabahar Bay (Sistan and Baluchestan)
Ireland (1)	Dingle Bay (Munster)
Spain (1)	Bay of Biscay (Basque Country)

Table 2. Designations and locations of the 55 named bays whose occurrence frequency was higher than 5.

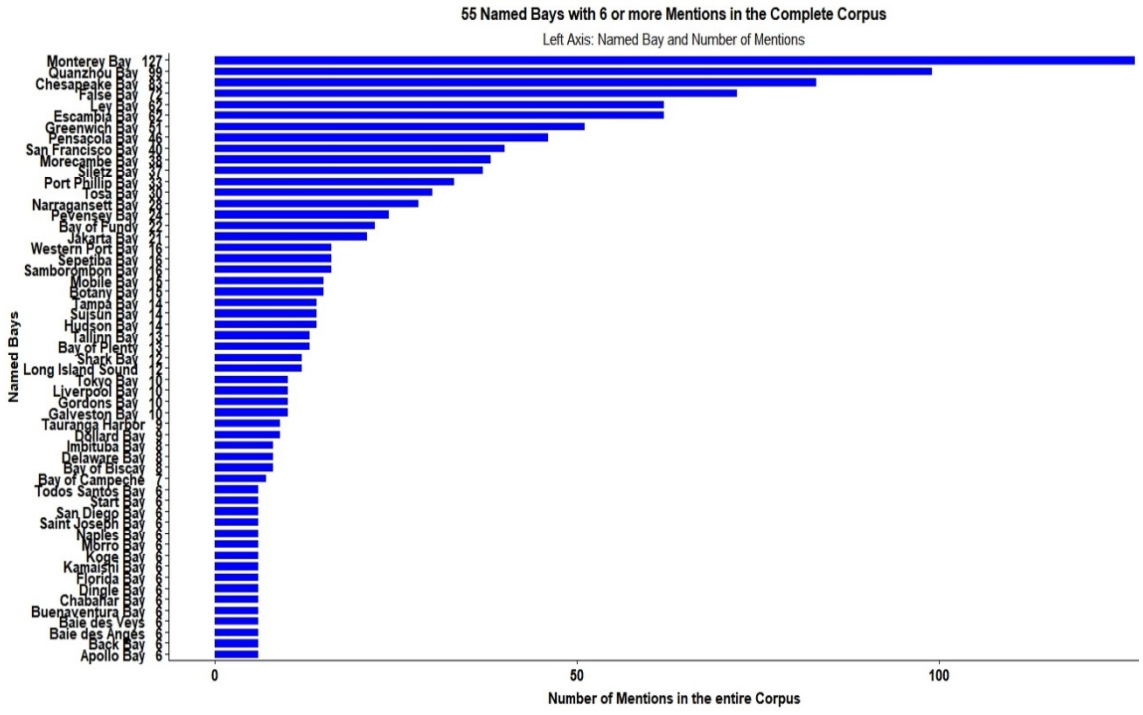

Figure 2. Designations and number of mentions of the 55 named bays whose occurrence frequency was higher than 5.

Term-term matrix construction

After the 294 named bays were joined with underscores in the lemmatized corpus, a count-based DSM was built with the R package *quanteda* for text mining. A count-based DSM was selected to obtain term vectors since this type of DSM outperforms prediction-based ones on small-sized corpora of under 10 million tokens (Ars, Willits & Jones, 2016; Sahlgren & Lenci, 2016).

In the DSM, only terms larger than 2 characters were considered, and numbers and punctuation marks were removed. Additionally, the minimal occurrence frequency was set to 5 so that the co-occurrences were statistically reliable (Evert, 2007). A sliding context window was set up to span 20 terms on either side of the target term because for small corpora, large windows lead to larger counts and greater statistical reliability (Rohde et al., 2006, p. 31; Bullinaria & Levy, 2007, p. 522). Furthermore, when the window is larger, the relations in the DSM will be more semantic than syntactic (Jurafsky & Martin, 2017, p. 5). Since closed-class words are often considered too uninformative to be suitable context words (Kiela & Clark, 2014), stopwords, adjectives and adverbs were not used as context words.

The resulting DSM was a 4,431×4,431 frequency matrix *A*, whose row vectors represented the 55 named bays plus the 4,376 different terms inside the context windows of 20 terms on either side of those bays.

Selection of bays and terms for clustering purposes

Subsequently, a 55×4,376 submatrix *B* was extracted from *A*, where the rows represented the 55 named bays, and the columns represented the 4,376 terms co-occurring with the bays. To cluster the bays of *B* sharing the same associated terms, it was necessary to select both the bays and the terms that best discriminated different groups of bays. This was done by removing the bays and the terms that could act as random noise and adversely affect the clustering results (Kaufman & Rousseeuw, 1990). The remainder of this section explains the selection method of bays and terms for clustering purposes.

An issue often highlighted in the literature on the clustering of rows in a frequency matrix abstracted from corpus data is that variation in document length will affect the clustering results. These documents are thus clustered in accordance with relative length rather than with a more interesting latent structure in the data (Moisl, Maguire & Allen, 2006; Rojas-Garcia, Faber & Batista-Navarro, 2018; Thabet, 2005). The conventional solution to the problem is to normalize the values in the frequency matrix to mitigate the effect of length variation. Normalization by mean document length (Spärck, Walker & Robertson, 2000) is widely used in Information Retrieval literature.

Nevertheless, as stated by Moils (2011), there is a limit to the effectiveness of normalization, and it has to do with the probabilities with which the terms in the column vectors occur in the corpus. Some documents in the matrix rows might be too short to give

accurate population probability estimates for the terms, and since length normalization methods accentuate such inaccuracies, the result is that analysis based on the normalized data inaccurately clusters the rows. One solution consists in statistically ascertaining which documents are too short to provide good estimates and to remove the corresponding rows from the matrix.

For that aim, Moisl (2011, pp. 42-45) proposes a formula that calculates the document length necessary to estimate the probability of each term in the column vectors with a 95% confidence level. Therefore, the formula can be applied to establish a minimum length threshold for the documents and to eliminate any documents under that threshold.

In our case, a document was considered to be the set of all context windows where a certain named bay appeared, and thus corresponded to a row of matrix *B*. As such, we had 55 named-bay documents. Similarly, the length of a document was considered to be the total number of words appearing in the set of all context windows of a certain named bay. The document lengths ranged from 4,950 words (for *Monterey Bay*) to 232 words (for *Kamaishi Bay*). Moisl's (2011) method was then applied to matrix *B* to determine: (1) which of the 55 named bays should be eliminated from our analysis; and (2) which terms helped to distinguish different groups of the retained bays.

Table 3 shows the length for named-bay documents needed by each of the 4,376 terms in the columns of matrix *B* so that their population probabilities could be estimated with a 95% confidence level, according to Moisl's (2011) formula. The terms in Table 3 were sorted in ascending order of the required document length.

Index	Term	Length needed for named-bay documents
1	**beach**	**416**
2	sea_surface_temperature	475
3	island	530
4	river	574
5	bay	597
6	wave	644
7	hurricane	655
[…]	[…]	[…]
325	la_niña	4,927
326	natural_area	4,942
327	criterion	4,944
328	**canal**	**4,944**
329	spring	4,952
330	pass	4,957
331	season	4,968
332	organic_material	4,975
[…]	[…]	[…]
4,371	swash_flow	522,884

4,372	morphologic_change	522,884
4,373	locally_generated_wave	522,884
4,374	counter-circulation	522,884
4,375	rip-opposite_megacusp	522,884
4,376	rip_current_experiment	522,884

Table 3. Length needed for named-bay documents (mostright column) associated with each of the 4,376 terms (middle column) co-occurring with the bays, according to Moisl's (2011) formula.

Since the lowest document-length value needed by the terms was 416 words (for the term *beach* in the first row of Table 3), those bays whose document length was smaller than the minimum length threshold 416 were eliminated from the analysis. This meant that only 29 bays of 55 were retained. As expected, the 29 named bays selected by Moisl's (2011) method were those with the highest number of mentions in the corpus (Figure 2), from Monterey Bay (127 mentions) to Shark Bay (12 mentions).

Regarding the selection of terms, since the maximum length of our named-bay documents was 4,950 words, only the first 328 terms in Table 3 were retained for clustering purposes because their needed document lengths were less than 4,950 words. These results are plotted in Figure 3, where the 4,376 terms co-occurring with the 55 bays are on the horizontal axis (sorted in ascending order of the needed document length), and their required document lengths are on the vertical axis. The red horizontal line indicates the maximum length of the named-bay documents (4,950 words), and the green vertical line marks the 328 terms whose needed document lengths were equal to or less than the maximum named-bay document length.

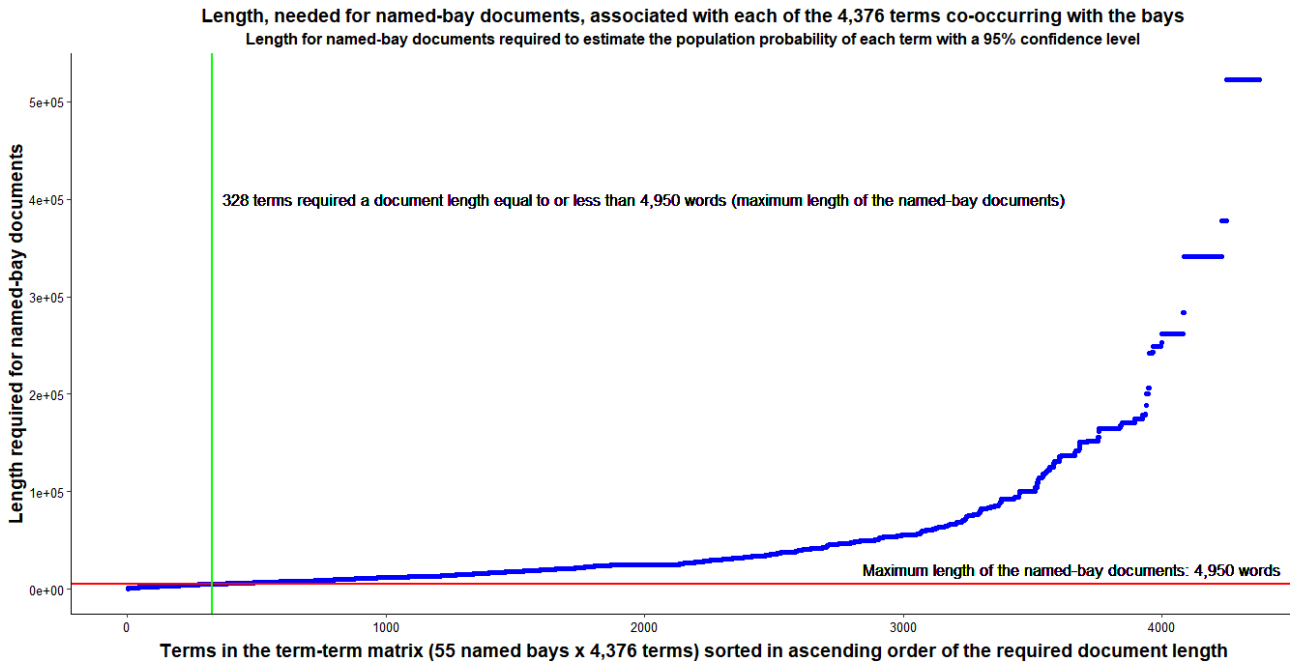

Figure 3. The required document lengths (vertical axis) associated with each of the 4,376 terms (horizontal axis) co-occurring with the 55 named bays.

Of the 328 terms selected by Moisl's (2011) method, only 310 terms co-occurred with the 29 retained bays. Therefore, a 29×310 submatrix *C* was extracted from *B* to group the bay vectors. A visualization of the 310-dimensional bay vectors in a 2-dimensional space is shown in Figure 4. This was accomplished by first weighting the frequency matrix *C* by using the *log-likelihood* association measure (see following section), reducing the number of dimensions via Singular Value Decomposition (Jolliffe, 2002), and plotting the data points according to the first two principal-component coordinates.

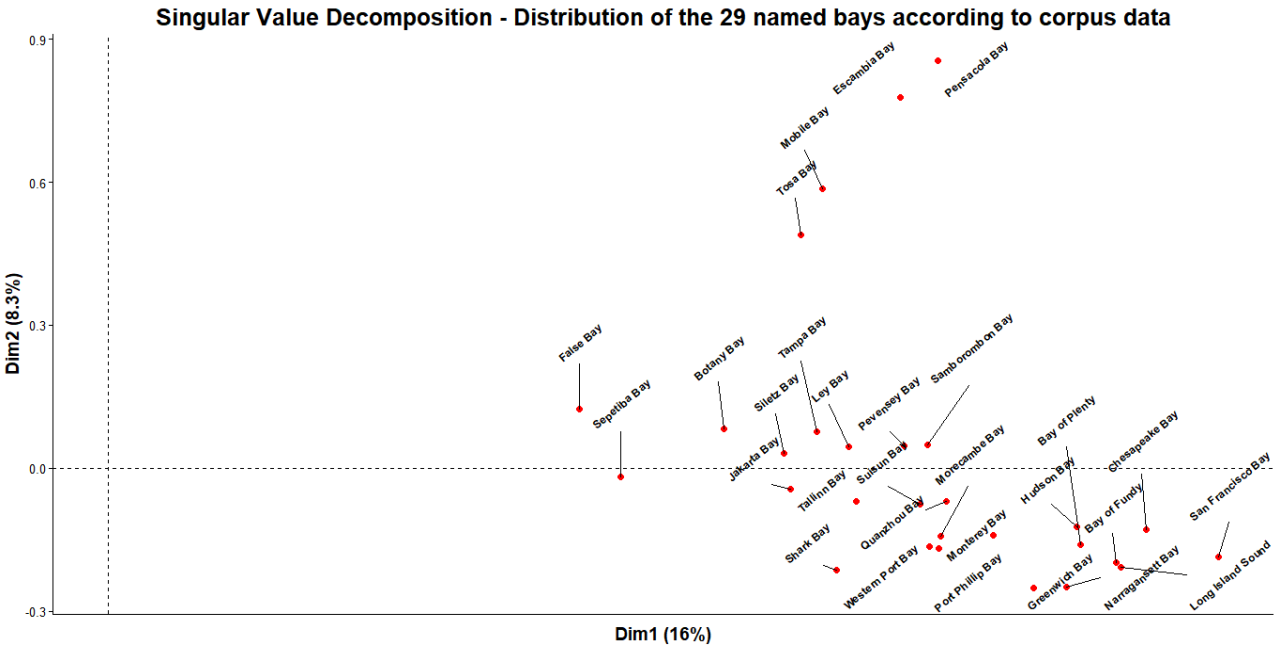

Figure 4. Visualization of the 29 bay vectors in a 2-dimensional space.

Clustering of named bays and weighting schemes

According to Moisl (2011, pp. 30-31), the 29×310 frequency matrix *C* was first normalized by mean document length. Next, we applied a hierarchical clustering technique, using the squared Euclidean distance as the intervector distance measure and Ward's Method as the clustering algorithm (Xu & Wunsch, 2009).

Since it is not clear how strong a cluster is supported by data (Suzuki & Shimodaira, 2004), a means for assessing the certainty of the existence of a cluster in corpus data was devised. Multiscale bootstrap resampling (Shimodaira, 2004) is a method for this in hierarchical clustering, which is implemented in the R package *pvclust* (Suzuki & Shimodaira, 2006). For each cluster, this method produces a number ranging from zero to one. This number is the approximately unbiased probability value (AU *p*-value), which represents the possibility that the cluster is a true cluster. The greater the AU *p*-value, the greater the probability that the cluster is a true cluster supported by corpus data. An AU

p-value equal to or greater than 95% significance level is most commonly adopted in research.

In the clustering results, 2 groups of bays, with AU *p*-values equal to or greater than 95%, were considered (Figure 5). Unfortunately, the existence of only 2 groups with such a large number of bays inside each of the clusters was not conducive to appropriately describing the semantic frames of the bays. As such, the normalization by mean document length was disregarded because it led to unreliable clustering. Consequently, other weighting schemes were tested instead.

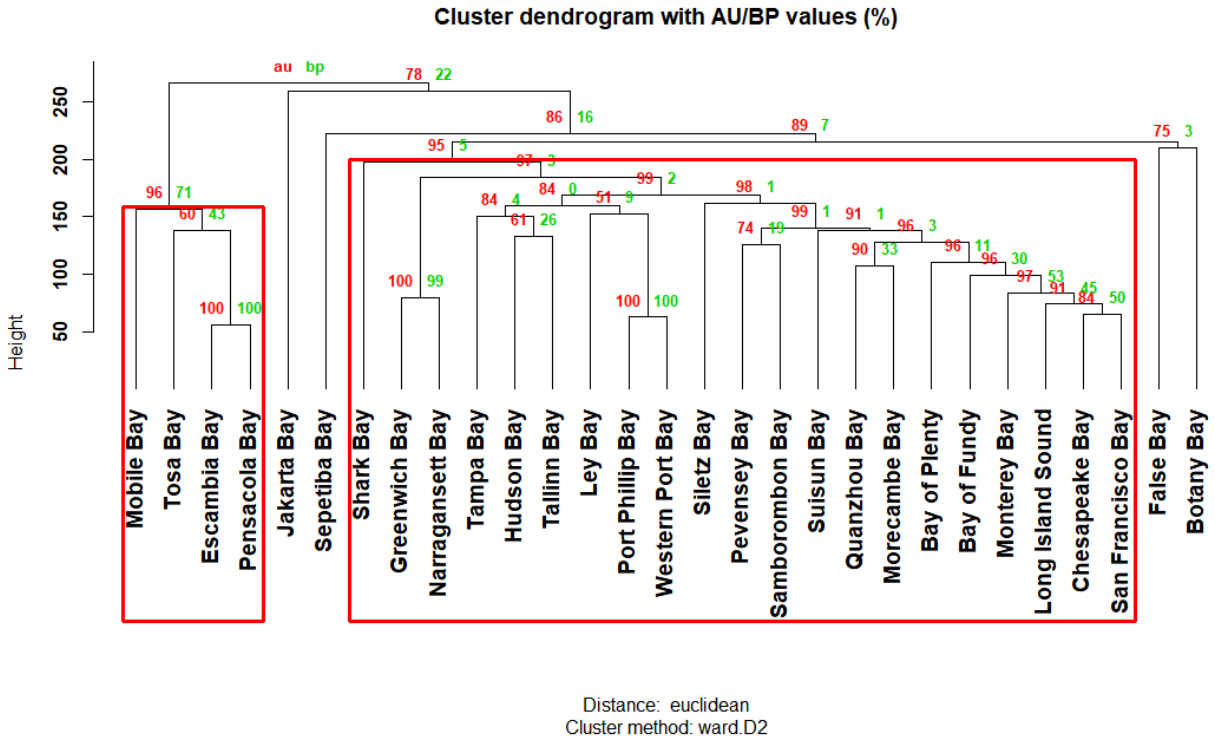

Figure 5. Dendrogram of the hierarchical clustering of the 29 named bays, along with 2 red rectangles indicating clusters with red-colored AU *p*-values ≥ 95% (red values at branches). The matrix was previously normalized by mean document length, which led to unreliable clustering results.

The frequency matrix *C* was subjected to three weighting schemes. First, the statistical *log-likelihood* measure (Dunning, 1993) was applied to calculate the association score between all term pairs, including the named bays (Evert, 2007, pp. 24-30). Research on computational linguistics reveals that *log-likelihood* is able to capture syntagmatic and paradigmatic relations (Bernier-Colborne & Drouin, 2016, p. 58; Lapesa et al., 2014, p. 168) and to achieve better performance for medium-to-low-frequency data than other association measures (Alrabia et al., 2014, p. 4; Krenn, 2000). However, the calculation of the *log-likelihood* scores was modified to cope with these critical situations:

- When the observed frequency was less than the expected one, the score was set to 0, as recommended by Evert (2007, p. 22). Otherwise, the score would have been

negative showing repulsion between terms, whereas our interest was in the stronger attraction to each other.

- When a term pair did not co-occur (i.e., its observed frequency was 0), the score was set to 0. Otherwise, the score would have obtained a low value, indicating a certain attraction between the pair of terms despite the absence of co-occurrence in corpus data.
- When a term co-occurred with only one bay, the corresponding addend in the *log-likelihood* formula (i.e., the addend where the observed frequency O_{21} takes part, according to Evert [2007, p. 25]) was set to 0. Otherwise, the score would have tended to minus infinity, and its value would have been undetermined.

Secondly, the association scores were transformed by adding 1 and calculating the natural logarithmic to reduce skewness (Lapesa et al., 2014). Finally, the row vectors were normalized to unit length to minimize the negative effects of extreme values on the Euclidean distance-based clustering technique.

The hierarchical clustering technique was then applied to the weighted matrix *C*. As a result, 5 groups of bays with AU *p*-values equal to or greater than 99% were considered to be strongly supported by corpus data (Figure 6). Two bays comprised each of the 5 clusters, which provided evidence that the clustering results with a *log-likelihood* measure was more reliable than those with mean document length. Accordingly, this paper focuses on the 10 named bays inside the 5 clusters shown in Figure 6.

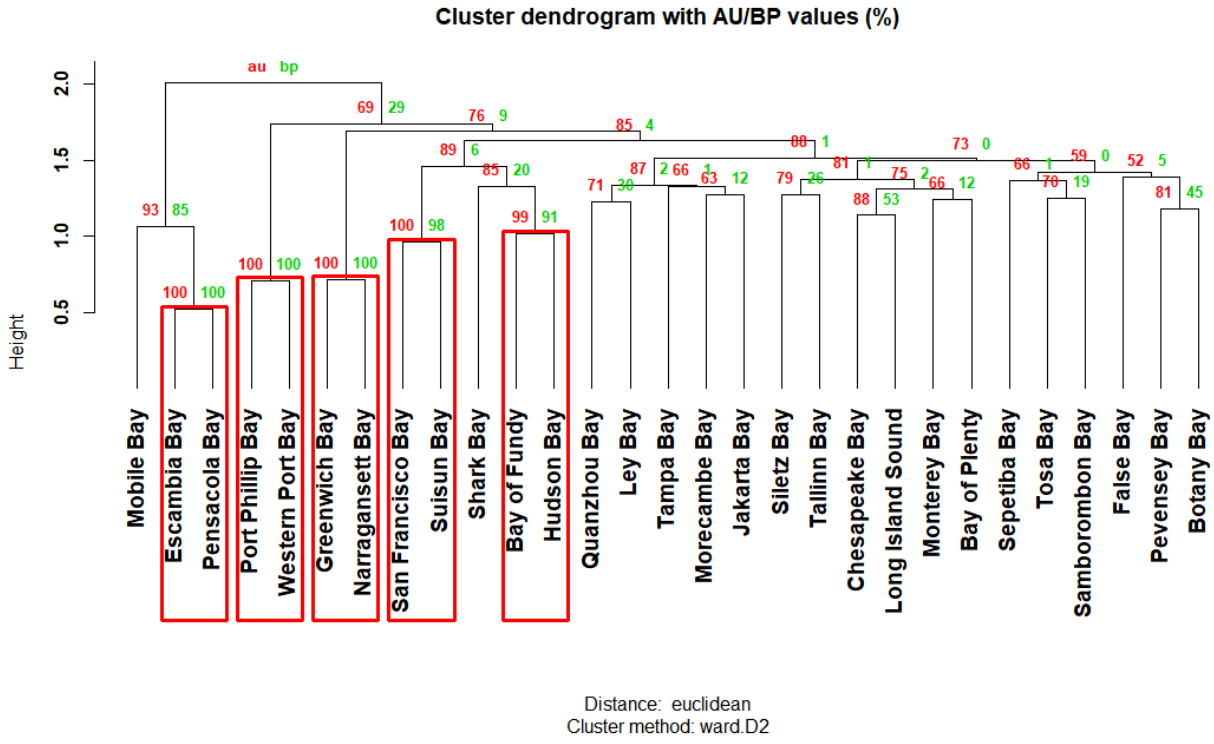

Figure 6. Dendrogram of the hierarchical clustering of the 29 named bays, along with 5 red rectangles indicating clusters that are strongly supported by corpus data (red-colored AU *p*-values ≥ 99%). The matrix was previously weighted by *log-likelihood* measure, which led to reliable clustering results.

Figure 7 shows a scatter plot of these 5 clusters via Singular Value Decomposition. The weighted matrix *C* was also used to visualize the bay vectors in Figure 4.

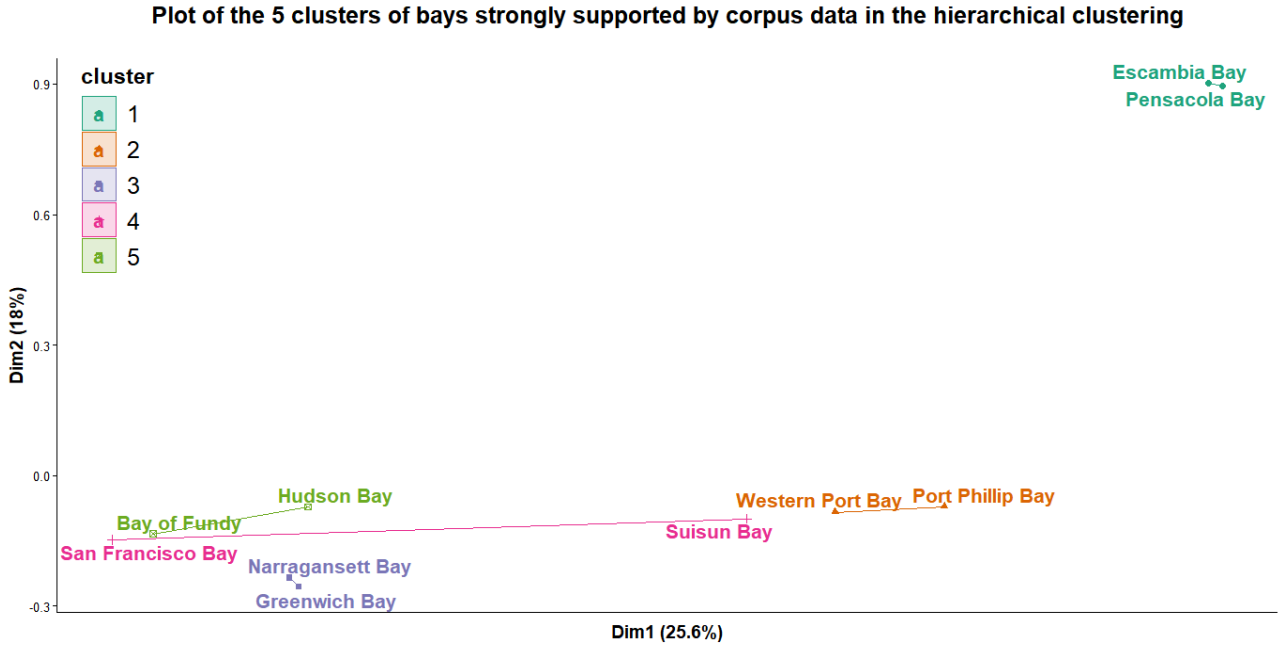

Figure 7. Scatter plot of the 5 clusters of bays, strongly supported by corpus data, via Singular Value Decomposition.

Selection of terms for semantic network construction

With a view to evaluating the procedure for term selection that best captured the terms related to the 29 named bays for the construction of semantic networks, 5 methods were devised.

Method One.

A 339×339 squared frequency matrix *D1* was built, whose rows represented the 29 named bays plus the 310 terms selected by Moisl's (2011) method. The columns also represented the same bays and terms co-occurring with the target words in the rows. *D1* was weighted by the *log-likelihood* measure. Then, the scores were transformed by adding 1 and calculating the natural logarithmic to reduce skewness (Lapesa et al., 2014).

The matrix *D1* tested whether the 310 terms selected by Moisl's (2011) method were sufficient to understand and represent the semantic frames in which the 29 named bays appeared.

Method Two.

A 3.867×3.867 frequency matrix *D2* was built, whose rows represented the 29 named bays plus the 3.838 terms co-occurring with them. *D2* was weighted in the same way as *D1*. *D2* tested whether no term selection method could optimally describe the semantic frames of the bays.

Method Three.

The number of columns in the weighted matrix *D2* was reduced to only two by applying the innovative dimensionality reduction technique UMAP (Uniform Manifold Approximation and Projection) (McInnes & Healy, 2018). It eliminates information redundancy among column variables and helps to identify local latent structures in corpus data. As a result, a 3.867×2 matrix *D3* was obtained.

D3 was tested whether such an innovative dimensionality reduction technique applied to all the terms co-occurring with the 29 bays was an improvement over *D2*.

Method Four.

In the same way as Moisl's (2011) method was used to select bays and terms, another statistical method was employed to select the terms that best described the 29 bays, based on Moisl (2015, pp. 77-93). In Corpus Linguistics, Moisl (2015) suggests retaining the term columns with the highest values in four statistical criteria: *raw frequency*, *variance*, *variance-to-mean ratio* (vmr) and *term frequency-inverse document frequency* (tf-idf).

Moisl's (2015) method was applied to a 29×3.838 frequency matrix, whose rows represented the 29 named bays. The columns represented all the terms co-occurring with them (excluding the bays). Figure 8 shows the co-plot of the four criteria, *z*-standardized for comparability reasons, and sorted in descending order of magnitude. A threshold of up to 1000 was set. This meant that only 847 terms fulfilled all criteria.

We estimated that between 25 and 30 terms would be necessary for a named bay to describe its semantic frame. A total number of terms ranging from 725 to 870 would thus be required for the description of the 29 bays. The threshold was set accordingly, so that the number of selected terms was within the interval 725-870 terms.

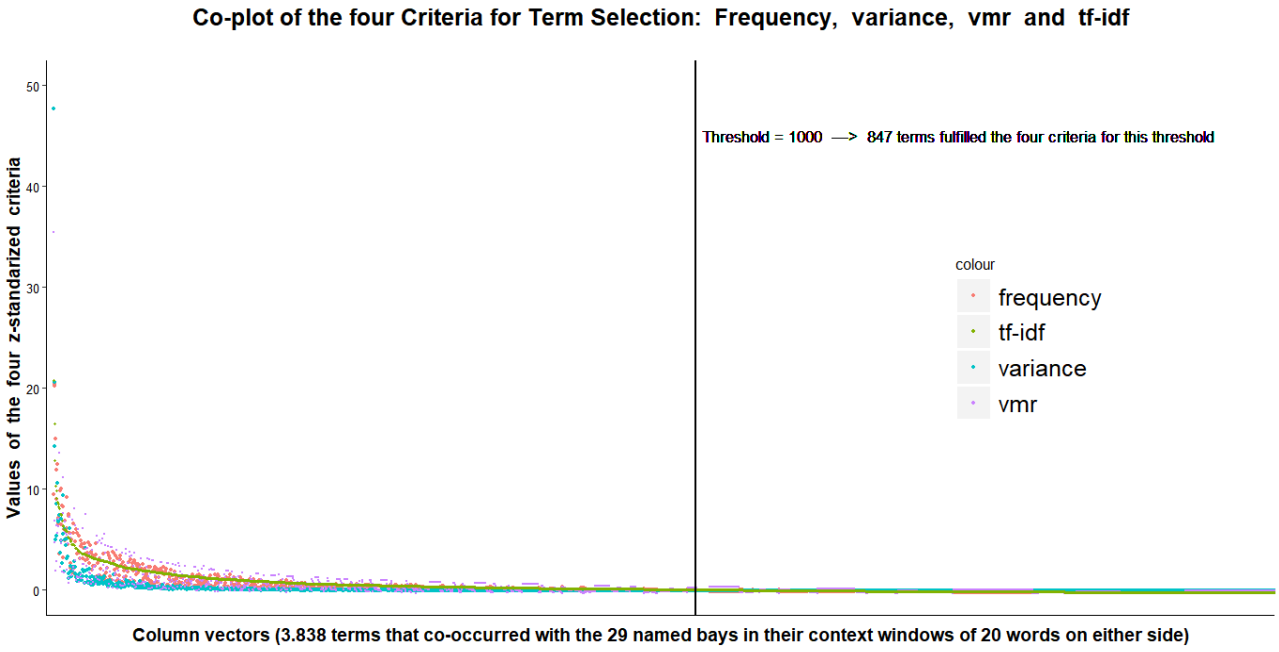

Figure 8. Co-plot of the 4 criteria for term selection: Frequency, variance, vmr, and tf-idf.

An 876×876 frequency matrix *D4* was obtained, where the rows represented the 29 named bays plus the 847 terms selected by Moisl's (2015) four statistical criteria. *D4* was weighted in the same way as *D1*.

D4 tested whether a term selection method was needed to appropriately describe the semantic frame of a named bay.

Method Five.

Finally, the 876 columns in the weighted matrix *D4* were reduced to only two columns by applying the UMAP technique. As a result, an 876×2 matrix *D5* was obtained. *D5* tested whether dimensionality reduction by UMAP, applied to selected terms, was an improvement over *D4*.

Terms characterizing each cluster

To ascertain the terms closely associated with each of the 5 clusters for semantic network construction, the following procedure was used:

1. For each of the 10 named bays in the clusters (Figure 6), a set of the top-30 terms, most semantically related to each bay according to their cosine similarities, was extracted from the corresponding DSM.
2. For each cluster, the mathematical operation *set intersection* was applied to the sets of the top-30 terms most semantically related to both bays in the same cluster. Only the shared terms with a cosine similarity higher than 0.4 were selected.

A reduced set of terms was thus obtained for each cluster to describe the named bays, based on shared associated terms.

Results

Analysis of the term selection methods

For each of the clusters, the term selection methods produced 5 sets of terms, which characterized them. Those term sets were qualitatively compared to gold standard sets of terms, manually extracted from the context windows of the 10 bays clustered in Figure 6, which best described each of the clusters for semantic network construction. For space constraints, only the main results of the comparisons are highlighted:

- The method that systematically produced the best sets of terms characterizing each cluster for semantic network construction was the Method Four, consisting in term selection based on Moisl's (2015) four statistical criteria.
- Method One, which consisted of bay and term selection according to Moisl (2011), produced sets of terms that could be used to infer the scientific topic in which the bays of a cluster were involved. However, because of the small number of terms selected (310 terms for 29 bays), the term sets were not conducive to understandable knowledge representations because, most of the time, the number of terms was not sufficient to derive a clear semantic relation between them.
- Method Two, with all the 3,838 terms co-occurring with the bays, produced meaningful sets of terms, which were suitable for semantic network construction. Nonetheless, surprisingly, the number of terms in the sets was lower than that in the sets obtained with Method Four. The reason was that the number of shared terms in each cluster with a cosine similarity higher than 0.4 was lower with Method Two, and higher with Method Four. In addition, for most clusters, some terms in the sets obtained with Method Two were not relevant for frame construction.
- Method Three and Method Five, whereby the number of columns was reduced to two with the UMAP technique, produced unreliable term sets. Firstly, both methods selected some terms with low occurrence frequency that could be disregarded for frame description. Secondly, they selected certain terms that were related to only one of the bays in a cluster. Thirdly, both methods selected some terms that were not related to any of the bays in a cluster. Those unrelated terms were associated with some of the terms that were directed related to the bays in a cluster, but in a thematic context different from that in which the bays were involved.

For the construction of the semantic frames presented in the next section, Method Four was thus applied.

Semantic frames describing the bay clusters

Interestingly, the five clusters in Figures 6 and 7 contained bay pairs located in the same geographical areas, as shown in Table 4.

Cluster 1 (USA)	**Cluster 2** (Australia)	**Cluster 3** (USA)	**Cluster 4** (USA)	**Cluster 5** (Canada)
Escambia Bay (Florida)	**Port Phillip Bay** (Victoria)	**Greenwich Bay** (Rhode Island)	**San Francisco Bay** (California)	**Bay of Fundy** (Nova Scotia)
Pensacola Bay (Florida)	**Western Port Bay** (Victoria)	**Narragansett Bay** (Rhode Island)	**Suisun Bay** (California)	**Hudson Bay** (Ontario)

Table 4. Designations and locations of the bays in the 5 clusters.

For the description of the frames, the semantic relations were manually extracted by querying the corpus in Sketch Engine (Kilgarriff, Rychly, Smrz & Tugwell, 2004), and analysing knowledge-rich contexts (Meyer, 2001). The query results were concordances of any elements between the bays in a cluster and related terms in a ±30 span. The semantic relations were those in EcoLexicon (Faber, León-Araúz & Prieto, 2009), with the addition of *does_not_affect*, *not_located_at*, *increases*, *decreases*, *belongs_to*, *uses*, *simulates*, and *becomes*.

In the first cluster, *Escambia* and *Pensacola* bays are thematically related by numerical parameter studies that simulate: (1) hurricane-induced storm surges, waves and winds, and the land dissipation effect on wind; (2) the effects of these features and inlet-bay configuration on open-coast storm-surge hydrographs. To validate simulation results, researchers employ historical data of the effects of *Hurricane Ivan* on both bays. Figure 9 shows the terms highly associated with the bays and their semantic relations.

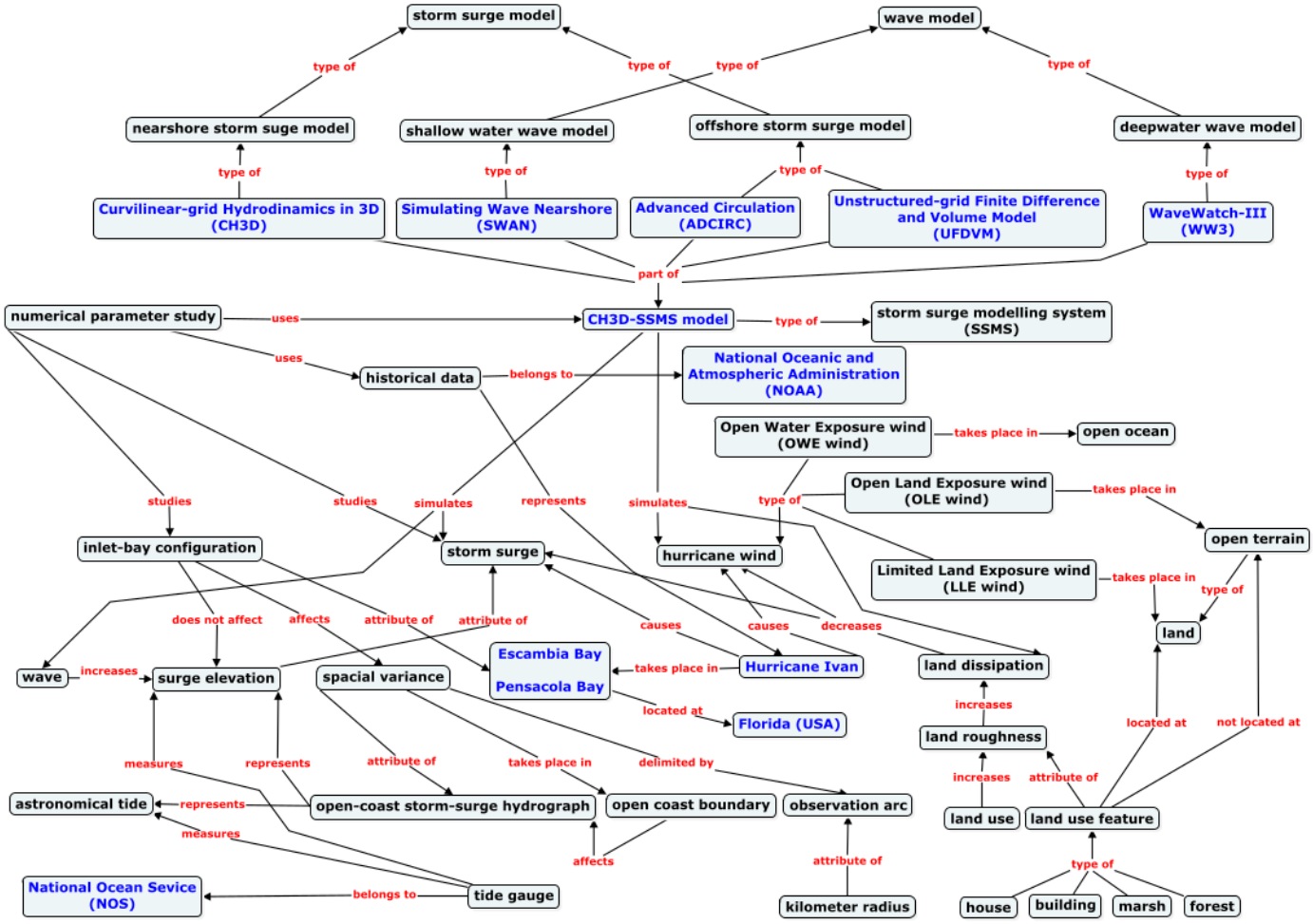

Figure 9. Semantic network of the terms associated with the *Escambia* and *Pensacola* bays.

The bays in the second cluster are involved in the topic of Integrated Coastal Management (ICM). Since the environmental condition of the Victorian coast (Australia) has not improved despite thirty years of ICM, case studies have been carried out in different coastal environments located on the *Port Phillip* and *Western Port* bays: a coastal headland (Point Nepean), a coastal lakes system (Gippsland Lakes), and an urbanising coastal region (Geelong region). These environments were examined to develop an approach that incorporates ICM in a Sustainable Coastal Planning, which responds to the pressures of urban growth, tourism, decline in water quality, climate change on coasts, coastal planning, and environmental protection (Figure 10).

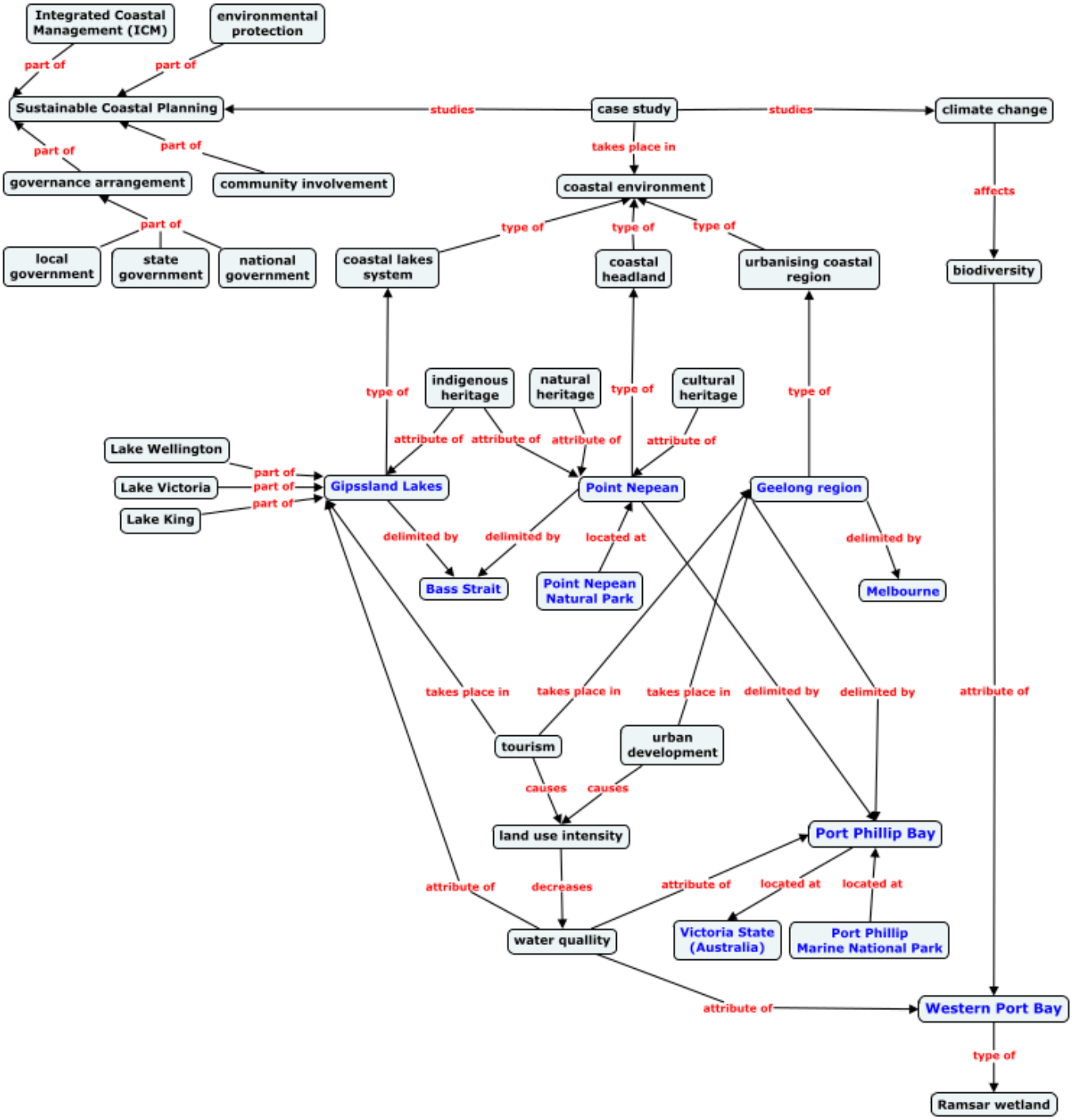

Figure 10. Semantic network of the terms associated with the *Port Phillip* and *Western Port* bays.

In the third cluster, *Greenwich* and *Narragansett* bays are sites for the study of benthic geologic habitats, namely, spatially recognizable areas in bay floors with special geologic and biologic characteristics. These habitats are identified by using imagery, and then classified according to criteria such as sediment particle size (Figure 11).

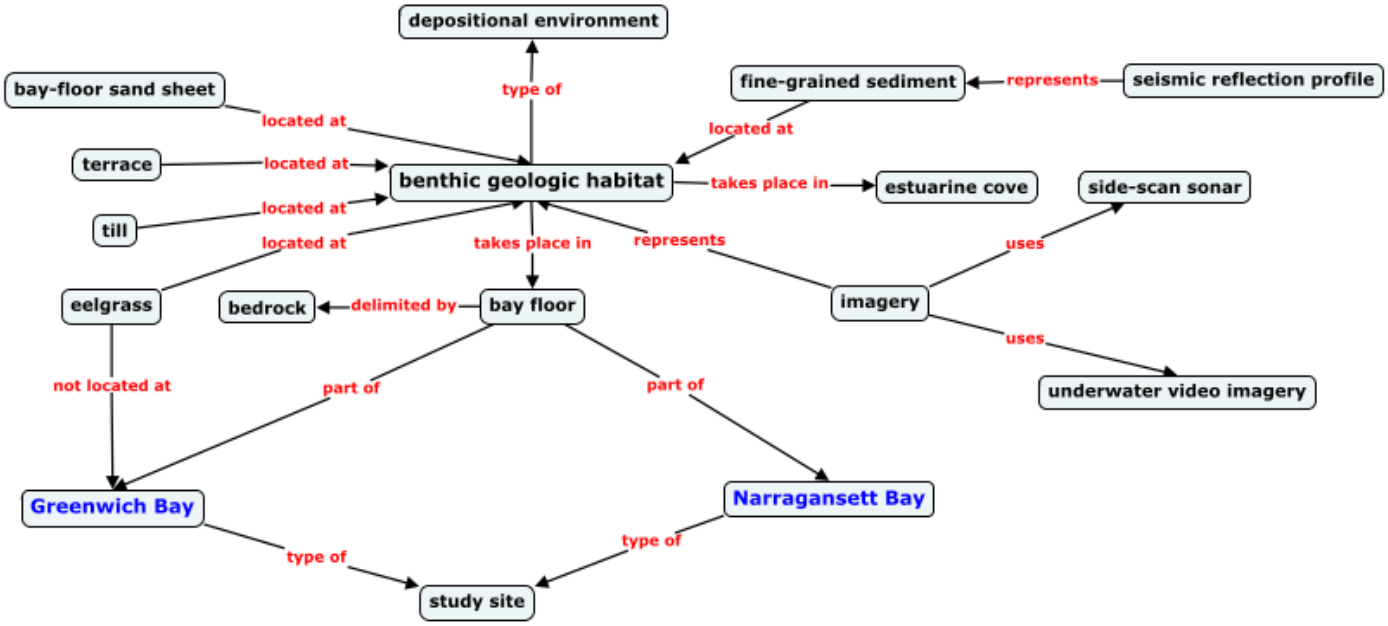

Figure 11. Semantic network of the terms associated with the *Greenwich* and *Narragansett* bays.

In the fourth cluster, *San Francisco* and *Suisun* bays are involved in research studies to determine whether the timescale dependence of forcing mechanisms on suspended sediment concentration (SSC) is typical in estuaries, based on SSC data. Of the forcing mechanisms, several tend to be concurrently active in estuaries, rather than only one. Multiple active forcing mechanisms have been observed in estuaries, but the specific mix of active mechanisms is different in each (Figure 12).

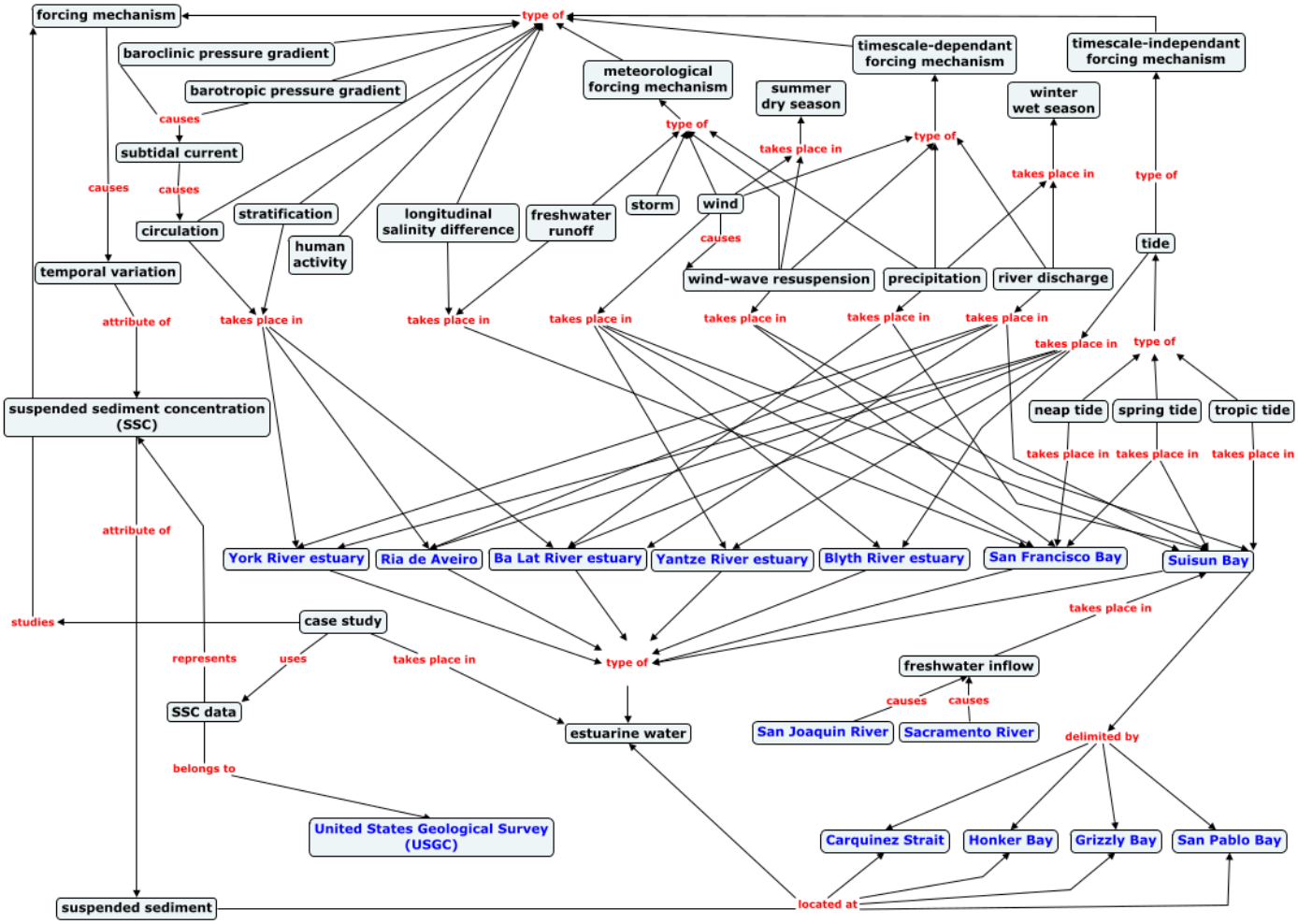

Figure 12. Semantic network of the terms associated with the *San Francisco* and *Suisun* bays.

Finally, in the fifth cluster, *Bay of Fundy* and *Hudson Bay* are low wave-energy environments with large sedimentation rates and tidal ranges, which originate tidal flats and tidal marshes. The *Bay of Fundy* is a vertically mixed estuary. With limited freshwater inputs and the largest tidal ranges in the world (over 15 meters), it is used to generate electricity, thanks to a Straflo turbine. These strong tides also erode joint planes (vertical cracks) of cliffs on the bay. As a result, joint planes enlarge and become caves, which erode further and form arches. When the roof of these arches collapses, the stacks on the bay are formed (Figure 13).

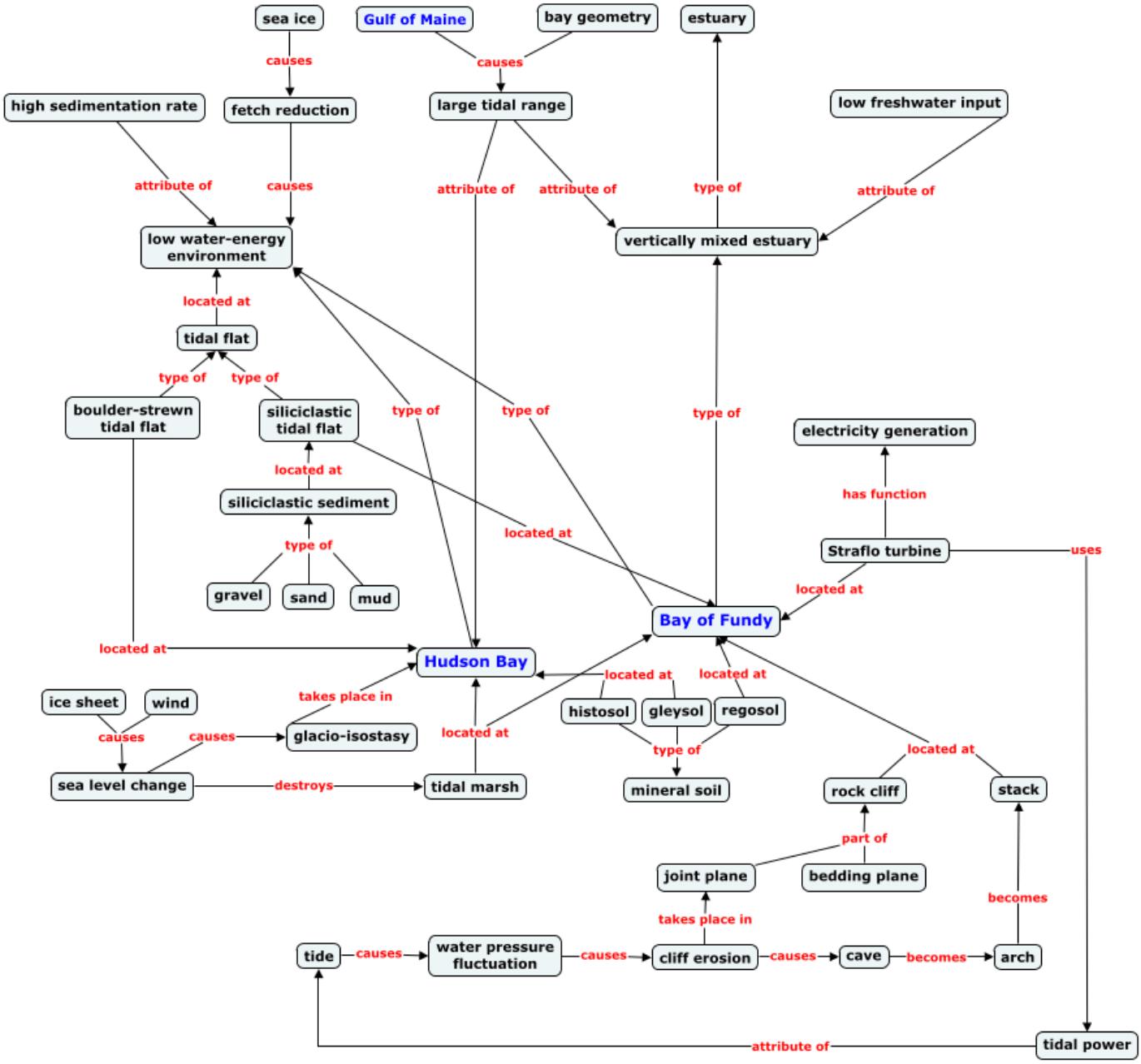

Figure 13. Semantic network of the terms associated with *Bay of Fundy* and *Hudson Bay*.

Conclusions

To extract knowledge for the semantic frames or conceptual structures (Faber, 2012) that underlie the usage of named bays in Coastal Engineering texts, a semi-automated method for the extraction of terms and semantic relations was devised. The semantic relations linking concepts in the semantic frames were manually extracted, based on the corpus analysis of knowledge-rich contexts (Meyer, 2001), a time-consuming task that is essential for the explanatory adequacy of frames (Faber, 2009). In future research, the knowledge patterns by León-Araúz, San Martín & Faber (2016) for the automatic extraction of semantic relations will be tested.

The method for the extraction of terms closely associated with named bays combined selection procedures for both terms and bays, with the use of a count-based DSM, weighted by a *log-likelihood* association measure. The selection of 29 named bays from an initial set of 55 bays with an occurrence frequency greater than 5 was performed by using Moisl's (2011) statistical method. It consisted in determining which bays had suitable document lengths for accurate estimation purposes. This bay selection procedure, along with a matrix normalization by *log-likelihood* measure, yielded reliable clustering results when the bays were automatically grouped based on their shared terms. Surprisingly, the normalization by mean document length, widely used in Information Retrieval, and suggested by Moisl (2011) because of its intuitive simplicity, did not achieve the desired clustering results. This reinforces the view that the performance of conventional procedures used in Natural Language Processing (NLP) largely depends on the nature of the task.

Regarding the term selection procedures, of the five methods tested, that of Moisl (2015, pp. 77-93), based on four statistical criteria, obtained the best performance for semantic network construction when qualitatively compared with gold standard sets of terms. Nonetheless, for reliable bay clustering, the best term selection procedure was that of Moisl (2011). This finding reveals that the best set of terms characterizing named bays is different, depending on whether the ultimate goal is clustering or frame description.

The two methods for term selection including dimensionality reduction by UMAP produced poor results. Since the reduction to two dimensions was probably insufficient, a larger number of dimensions will be tested in the future. Moreover, Topic Modelling (Blei et al., 2003), a domain-specific dimension reduction technique for texts, will be also applied.

Finally, the semantic frames in the previous section reflect that most terms related to named bays are multiword terms (MWT) since specialized language units are mostly represented by such compound forms (Nakov, 2013). The MWT extraction was possible because they were previously matched and joined with underscored in the lemmatized corpus, thanks to the list of MWTs stored in EcoLexicon. This implies that EcoLexicon is a valuable resource for any NLP tasks related to specialized corpora on environmental science.

Acknowledgements

This research was carried out as part of project FFI2017-89127-P, Translation-Oriented Terminology Tools for Environmental Texts (TOTEM), funded by the Spanish Ministry of Economy and Competitiveness. Funding was also provided by an FPU grant given by the Spanish Ministry of Education to the first author.

Footnotes

2. http://olst.ling.umontreal.ca/cgi-bin/dicoenviro/search_enviro.cgi
3. https://www.eionet.europa.eu/gemet/en/themes/
4. http://www.fao.org/faoterm/en/

5. http://aims.fao.org/en/agrovoc
6. http://www.environmentontology.org/Browse-EnvO

References

Alrabia, M., Alhelewh, N., Al-Salman, A. & Atwell, E. (2014). An Empirical Study On The Holy Quran Based On A Large Classical Arabic Corpus. *International Journal of Computational Linguistics*, 5(1), 1-13.

Asr, F., Willits, J. & Jones, M. (2016). Comparing Predictive and Co-occurrence Based Models of Lexical Semantics Trained on Child-directed Speech. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 1092-1097). Philadelphia (Pennsylvania): CogSci.

Baroni, M., Dinu, G. & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (pp. 238-247). Baltimore: ACL, vol. 1.

Bernier-Colborne, G. & Drouin, P. (2016). Evaluation of distributional semantic models: a holistic approach. In *Proceedings of the 5th International Workshop on Computational Terminology* (pp. 52-61). Osaka: Computerm.

Bernier-Colborne, G. & L'Homme, M.C. (2015). Using a distributional neighbourhood graph to enrich semantic frames in the field of the environment. In *Proceedings of the Conference Terminology and Artificial Intelligence* (pp. 9-16). Granada (Spain): TIA.

Bertels, A. & Speelman, D. (2014). Clustering for semantic purposes: Exploration of semantic similarity in a technical corpus. *Terminology*, 20(2), 279-303.

Blei, D.M.; Ng, A.Y. & Jordan, M.I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993-1022.

Bullinaria, J.A. & Levy, J.P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39(3), 510-526.

Bullinaria, J.A. (2008). Semantic Categorization Using Simple Word Co-occurrence Statistics. In M. Baroni, S. Evert & A. Lenci (Eds.), *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics* (pp. 1-8). Hamburg: ESSLLI.

Collobert, R. & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning* (pp. 160-167). New York: ICML.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T. & Harshman, R. (1990). Indexing by latent semantic analysis". *Journal of the American Society for Information Science*, 41(6), 391-407.

Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1), 61-74.

El bazi1, I. & Laachfoubi, N. (2016). Arabic Named Entity Recognition using Word Representations. *International Journal of Computer Science and Information Security*, 14(8), 956-965.

Everitt, B., Landau, S. & Leese, M. (2001). *Cluster Analysis* (4th ed.). London: Arnold.

Evert, S. (2007). Corpora and Collocations. Extended Manuscript of Chapter 58 of A. Lüdeling & M. Kytö (Eds.) (2008), *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter. Retrieved from http://www.stefan-evert.de/PUB/Evert2007HSK_extended_manuscript.pdf (last access: 2018-12-20).

Faber, P. (2009). The cognitive shift in terminology and specialized translation. *MonTI. Monografías de Traducción e Interpretación*, 1, 107-134.

Faber, P. (2011). The Dynamics of Specialized Knowledge Representation: Simulational Reconstruction or the Perception-action Interface. *Terminology*, 17 (1), 9-29.

Faber, P. (Ed.) (2012). *A Cognitive Linguistics View of Terminology and Specialized Language*. Berlin/Boston: De Gruyter Mouton.

Faber, P., León-Araúz, P. & Prieto, J.A. (2009). Semantic Relations, Dynamicity, and Terminological Knowledge Bases. *Current Issues in Language Studies*, 1, 1-23.

Feldman, R. & Sanger, J. (2007). *The Text Mining Handbook*. Cambridge: Cambridge University Press.

Gries, S. & Stefanowitsch, A. (2010). Cluster analysis and the identification of collexeme classes. In S. Rice & J. Newman (Eds.), *Empirical and experimental methods in cognitive/functional research* (pp. 73-90). Stanford (California): CSLI.

Hearst, M. & Schütze, H. (1993). Customizing a lexicon to better suit a computational task. In B. Boguraev & J. Pustejovsky (Eds.), *Proceedings of a Workshop Sponsored by the Special Interest Group on the Lexicon of the Association for Computational Linguistics* (pp. 55-69). Columbus (Ohio): ACL SIGLEX.

Herbelot, A. (2015). Mr Darcy and Mr Toad, gentlemen: distributional names and their kinds. In *Proceedings of the 11th International Conference on Computational Semantics* (pp. 151-161). London: ACL.

Hermann, M. (2011). Finding the Minimum Document Length for Reliable Clustering of Multi-Document Natural Language Corpora. *Journal of Quantitative Linguistics*, 18(1), 23-52.

Jolliffe, I. (2002). *Principal Component Analysis* (2nd ed.). New York: Springer.

Jurafsky, D. & Martin, J. (2017). Vector Semantics. In *Speech and Language Processing*. Draft of August 7, 2017. Retrieved from https://web.stanford.edu/~jurafsky/slp3/15.pdf (last access: 2018-12-20).

Katrenko, S. & Adriaans, P. (2008). Qualia Structures and their Impact on the Concrete Noun Categorization Task. In M. Baroni, S. Evert & A. Lenci (Eds.), *Proceedings of*

the ESSLLI Workshop on Distributional Lexical Semantics (pp. 17-24). Hamburg: ESSLLI.

Kaufman, L. & Rousseeuw, P. (1990). *Finding Groups in Data*. Hoboken (New Jersey): Wiley-Interscience.

Kazama, J., De Saeger, S., Kuroda, K., Murata, M. & Torisawa, K. (2010). A Bayesian Method for Robust Estimation of Distributional Similarities. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 247-256). Uppsala (Sweden): ACL.

Kiela, D. & Clark, S. (2014). A Systematic Study of Semantic Vector Space Model Parameters. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality* (pp. 21-30). Gothenburg (Sweden): EACL.

Kilgarriff, A., Rychly, P., Smrz, P. & Tugwell, D. (2004). The Sketch Engine. In G. Williams & S. Vessier (Eds.), *Proceedings of the 11th EURALEX International Congress* (pp. 105-116). Lorient: EURALEX.

Kiss, G. (1973). Grammatical Word Classes: A Learning Process and Its Simulation. *Psychology of Learning and Motivation*, 7.

Krenn, B. (2000). *The Usual Suspects: Data-Oriented Models for the Identification and Representation of Lexical Collocations*. Saarbrücken: DFKI & Universität des Saarlandes, vol. 7, Saarbrücken Dissertations in Computational Linguistics and Language Technology.

Lapesa, G., Evert, S. & Schulte im Walde, S. (2014). Contrasting Syntagmatic and Paradigmatic Relations: Insights from Distributional Semantic Models. In *Proceedings of the 3rd Joint Conference on Lexical and Computational Semantics* (pp. 160-170). Dublin: SEM.

León-Araúz, P., Reimerink, A. & Faber, P. (2013). Multidimensional and Multimodal Information in EcoLexicon. In A. Przepiórkowski, M. Piasecki, K. Jassem & P. Fuglewicz (Eds.), *Computational Linguistics* (pp. 143-161). Berlin: Springer.

León-Araúz, P., San Martín, A. & Faber, P. (2016). Pattern-based Word Sketches for the Extraction of Semantic Relations". In *Proceedings of the 5th International Workshop on Computational Terminology* (pp. 73–82). Osaka (Japan): Computerm.

León-Araúz, P., San Martín, A. & Reimerink, A. (2018). The EcoLexicon English Corpus as an open corpus in Sketch Engine. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (Eds.), *Proceedings of the 18th EURALEX International Congress* (pp. 893-901). Ljubljana: Euralex.

Luhn, H. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1(4), 309-317.

Lund, K., Burges, C. & Atchley, R.A. (1995). Semantic and associative priming in a high-dimensional semantic space. In J.D. Moore & J.F. Lehman (Eds.), *Proceedings*

of the 17th Annual Conference of the Cognitive Science Society (pp. 660-665). Pittsbugh (USA): University of Pittsburgh.

Manning, C.D., Raghavan, P. & Schütze, H. (1998). *Introduction to Information Retrieval*. Cambridge (England): Cambridge University Press.

McInnes, L & Healy, J. (2018). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. ArXiv e-prints 1802.03426. Retrieved from https://arxiv.org/pdf/1802.03426.pdf (last access: 2018-12-20).

Meyer, I. (2001). Extracting knowledge-rich contexts for terminography: A conceptual and methodogical framework. In D. Bourigault, C. Jacquemin & M.C. L'Homme, (Eds), *Recent Advances in Computational Terminology* (279-302). Amsterdam/Philadelphia: John Benjamins.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient estimation of word representations in vector space. In *Workshop Proceedings of International Conference on Learning Representations*. Scottsdale (Arizona): ICLR.

Miller, G.A. & Charles, W.G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1), 1-28.

Miller, G.A. (1971). Empirical methods in the study of semantics. In D. Steinberg & L. Jakobovits (Eds.), *Semantics: An Interdisciplinary Reader*. Cambridge: Cambridge University Press; 569–585.

Moisl, H. (2009). Exploratory multivariate analysis. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics. An International Handbook* (vol. 2, pp. 874-899). Berlin: Walter de Gruyter.

Moisl, H. (2015). *Cluster Analysis for Corpus Linguistics* (pp. 77-93). Berlin/Munich/Boston: De Gruyter Mouton.

Moisl, H., Maguire, W. & Allen, W. (2006). Phonetic variation in Tyneside: Exploratory multivariate analysis of the Newcastle Electronic Corpus of Tyneside English. In F. Hinskens (Ed.), *Language Variation – European Perspectives* (pp. 127-141). Amsterdam: John Benjamins.

Moskalski, S. & Torres, R. (2012). Influences of tides, weather, and discharge on suspended sediment concentration. *Continental Shelf Research*, 37, 36-45. Retrieved from https://www.sciencedirect.com/science/article/pii/S0278434312000180 (last access: 2018-12-20).

Nakov, P. (2013). On the interpretation of noun compounds: Syntax, semantics, and entailment. *Natural Language Engineering*, 19(3), 291-330.

Nguyen, N.T.H., Soto, A.J., Kontonatsios, G., Batista-Navarro, R. & Ananiadou, S. (2017). Constructing a biodiversity terminological inventory. *PLoS ONE*, 12(4), e0175277.

Pantel, P. & Lin, D. (2002). Discovering Word Senses from Text. In *Proceedings of ACM Conference on Knowledge Discovery and Data Mining* (pp. 613-619). Edmonton (Canada): KDD-02.

Pennington, J., Socher, R. & Manning, C.D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods for Natural Language Processing* (pp. 1532-1543). Doha (Qatar): EMNLP.

Reimerink, A. & León-Araúz, P. (2017). Predicate-Argument Analysis to Build a Phraseology Module and to Increase Conceptual Relation Expressiveness. In R. Mitkov (Ed.), *Computational and Corpus-Based Phraseology. Second International Conference, Europhras 2017, Proceedings* (pp. 176-190). London: Springer.

Rohde, D., Gonnerman, L. & Plaut, D. (2006). An Improved Model of Semantic Similarity Based on Lexical Co-Occurrence. *Communications of the ACM*, 8, 627-633.

Rojas-Garcia, J., Faber, P. & Batista-Navarro, R. (2018). Conceptual information extraction for named bays from a specialized corpus. In T. Read, S. Montaner & B. Sedano (Eds.), *Technological Innovation for Specialized Linguistic Domains* (pp. 97-113). Beau Bassin (Mauricio): Éditions Universitaires Européenes.

Sahlgren, M. & Lenci, A. (2016). The Effects of Data Size and Frequency Range on Distributional Semantic Models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 975-980). Austin (Texas): ACL.

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Linguistics*, 20(1), 33-54.

Salton, G. & Lesk, M.E. (1968). Computer evaluation of indexing and text processing. *Journal of the ACM*, 15(1), 8-36.

Schütze, H. (1997). *Ambiguity Resolution in Language Learning: Computational and Cognitive Models*. Stanford (California): Cambridge University Press, Center for the Study of Language and Information Publication, vol. 71, Lecture Notes.

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, 24(1), 97-124.

Shutova, E., Sun, L. & Korhonen, A. (2010). Metaphor identification using verb and noun clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics* (pp.1002-1010). Beijing (China): COLING, vol. 2.

Spärck J.K., Walker, S. & Robertson, S. (2000). A probabilistic model of information retrieval: development and comparative experiments, part 2. In *Information Processing and Management*, 36, 809-840.

Sun, F., Guo, J., Lan, Y., Xu, J. & Cheng, X. (2015). Learning Word Representations by Jointly Modeling Syntagmatic and Paradigmatic Relations. In *Proceedings of the 53rd Annual Meeting of the ASL and the 7th International Joint Conference on Natural Language Processing* (pp. 136-145). Beijing (China): ACL.

Sun, L. & Korhonen, A. (2009). Improving verb clustering with automatically acquired selectional preferences. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (pp. 638-647). Singapore: EMNLP.

Suzuki, R. & Shimodaira, H. (2004). An application of multiscale bootstrap resampling to hierarchical clustering of microarray data: How accurate are these clusters? In *Proceedings of the Fifteenth International Conference on Genome Informatics*, P034.

Suzuki, R. & Shimodaira, H. (2006). Pvclust: An R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, 22(12), 1540-1542.

Thabet, N. (2005). Understanding the thematic structure of the Qur'an: an exploratory multivariate approach. In *Proceedings of the ACL Student Research Workshop* (pp. 7-12). Michigan: ACL.

Thompson, P., Batista-Navarro, R., Kontonatsios, G., Carter, J., Toon, E., McNaught, J., Timmermann, C., Worboys, M. & Ananiadou, S. (2015). Text Mining the History of Medicine. *PLoS ONE*, 11(1), e0144717.

Turian, J., Ratinov, L., Bengio, Y. & Roth, D. (2009). A preliminary evaluation of word representations for named-entity recognition. In *Proceedings of NIPS Workshop on Grammar Induction, Representation of Language and Language Learning*. Whistler (Canada): Neural Information Processing Systems Foundation.

Widdows, D. (2003). Unsupervised methods for developing taxonomies by combining syntactic and statistical information. In *Proceedings of Human Language Technology* (pp. 276-283). Edmonton (Canada): HLT/NAACL.

Xu, R. & Wunsch, D. (2009). *Clustering*. Hoboken (New Jersey): IEEE Press/Wiley.