RAEL: Revista Electrónica de Lingüística Aplicada
Vol./Núm.: 19.1
Enero-diciembre 2020
Páginas: 97-119
Artículo recibido: 04/02/2020
Artículo aceptado: 06/05/2020
Artículo publicado 29/06/2020
Url: http://www.aesla.org.es/ojs/index.php/RAEL/article/view/381

Application of Topic Modelling for the Construction of Semantic Frames for Named Rivers

Construcción de marcos semánticos para ríos con nombre propio mediante *Topic Modelling*

JUAN ROJAS-GARCIA[1]
UNIVERSITY OF GRANADA

EcoLexicon is a terminological knowledge base on environmental science whose design permits the geographic contextualization of data. For the geographic contextualization of concepts related to named landforms, this paper presents a semi-automatic method of extracting terms associated with named rivers (e.g., *Salinas River*). Terms were extracted from a specialized corpus on Coastal Engineering, where named rivers were automatically identified. Statistical procedures were applied for selecting both terms and rivers in distributional semantic models to construct the conceptual structures underlying the usage of named rivers. The rivers 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 rivers with explanatory adequacy, according to the premises of Frame-based Terminology. Furthermore, the semantic networks unveiled that the named rivers were thematically related to sediment concentration in rivers, sediment discharge into bays, the negative effects of sediment supply decrease on coastal erosion, and national shoreline management plans for managing risks due to flooding and erosion.

Keywords: *named river*; *Frame-based Terminology*; *conceptual information extraction*; *topic modelling*; *text mining*

EcoLexicon es una base de conocimiento terminológica sobre ciencias medioambientales cuyo diseño permite la contextualización geográfica de conceptos relacionados con accidentes geográficos. Para tal fin, este artículo presenta un método semiautomático para extraer términos asociados con ríos con nombre propio (v.gr., *Río Salinas*). Los términos se extrajeron de un corpus especializado en Ingeniería de Costas, donde las designaciones de ríos se identificaron automáticamente. Se aplicaron procedimientos estadísticos para seleccionar ríos y términos, que se proyectaron en espacios semánticos vectoriales, y se emplearon para construir las estructuras conceptuales que subyacían en el uso de los ríos. Los resultados muestran que el método es apropiado para describir los marcos semánticos que evocan los ríos, según las premisas de la Terminología basada en Marcos. Además, las redes semánticas revelaron que los ríos estaban relacionados temáticamente con la concentración de sedimentos, su descarga en las bahías, los efectos perniciosos de su reducción para la erosión costera, y planes nacionales de mantenimiento de costas para gestionar los riesgos de inundación y erosión.

Palabras clave: *río con nombre propio*; *Terminología basada en Marcos*; *extracción de información conceptual*; *topic modelling*; *minería de textos*

[1] Email: juanrojas@ugr.es.

1. INTRODUCTION

EcoLexicon[2] is a multilingual, terminological knowledge base on environmental science 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 et al., 2013). However, the geographic contextualization of named landforms (e.g., *Pearl River*, *Monterey Bay*, *Sunset Beach*) is barely tackled in terminological resources because of two reasons in our opinion: a) they are considered mere instances of concepts such as RIVER, BAY, or BEACH, and their specific relational behaviour with other concepts in a specialized knowledge domain is thus neglected and not semantically described; b) their semantic representation depends on knowing which terms are related to each named landform, and how these terms are related to each other, a time-consuming task taking into account that terminologists do not often resort to natural language processing systems beyond corpus tools such as Sketch Engine (Kilgarriff et al., 2004).

Consequently, this paper presents a semi-automatic method of extracting terms associated with named rivers (e.g., *Omaru River*) as types of landform from a corpus of English Coastal Engineering texts. The aim is to represent that knowledge in semantic networks in EcoLexicon according to the theoretical premises of Frame-based Terminology. Hence, on the hypothesis that named rivers should be considered concepts rather than instances in the Coastal Engineering domain, each named river should appear in the context of a specialized semantic frame that highlights both its relation to other terms and the relations between those terms.

These semantic frames, such as that shown in Figure 1 underlying the linguistic usage of *Escambia* and *Pensacola* bays in Coastal Engineering texts, provide the background knowledge about named rivers necessary in communicative situations, such as specialized translation to appropriately render terms into another language (Faber, 2012). Moreover, they make the semantic and syntactic behavior of terms explicit by means of the description of conceptual relations and term combinations (Faber, 2009).

[2] http://ecolexicon.ugr.es.

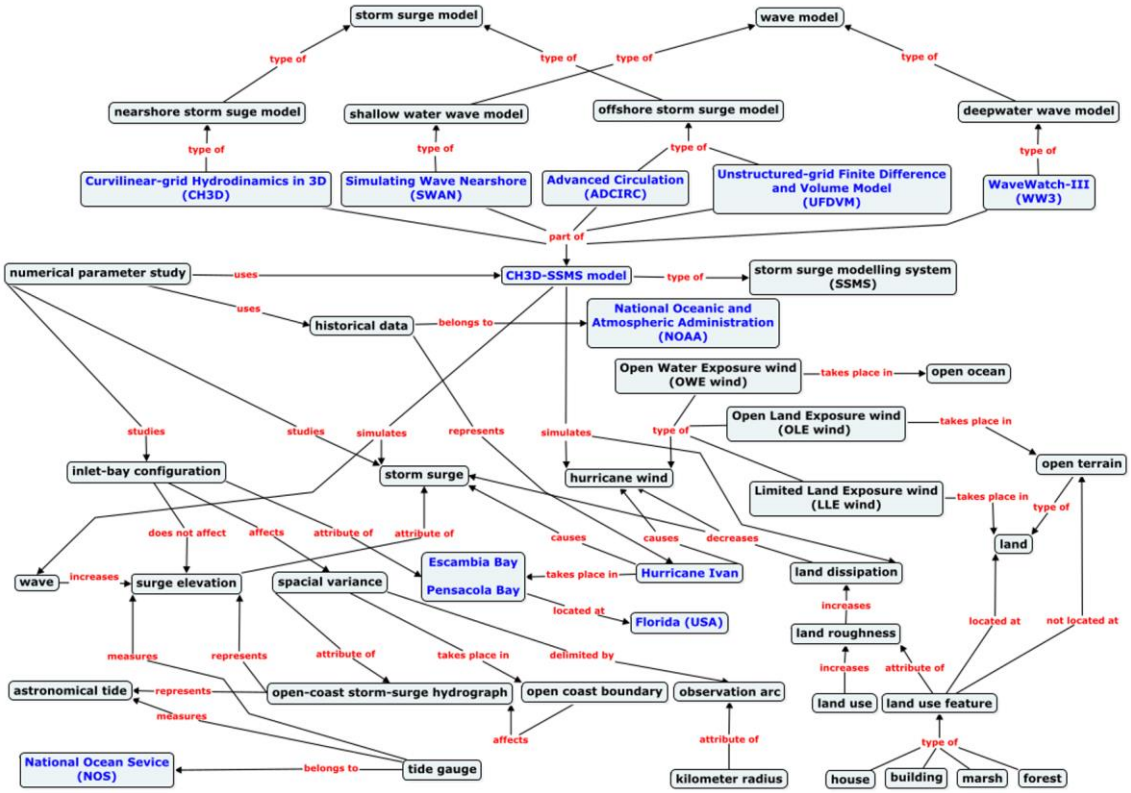

Figure 1: ***Semantic network of the terms associated with* Escambia *and* Pensacola *bays in Coastal Engineering texts***

The rest of this paper is organized as follows. Section 2 provides motivations for the research, and background on distributional semantic models and topic modelling. Section 3 explains the materials and methods applied in this study, namely, the automatic identification of named rivers, the selection procedures for terms and rivers in distributional semantic models, the clustering technique for rivers sharing associated terms, and the topic model for both the extraction of terms associated with each named river and the construction of its corresponding specialized semantic frame. 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.

2. THEORETICAL FRAMEWORK

2.1 *Motivations for the research*

Although 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[3], GEMET[4] or FAO Term Portal[5]. In contrast, AGROVOC[6] basically contains a list of named landforms with hyponymic information, whereas ENVO[7] provides descriptions of the named landforms with only

[3] http://olst.ling.umontreal.ca/cgi-bin/dicoenviro/search_enviro.cgi.
[4] https://www.eionet.europa.eu/gemet/en/themes/.
[5] http://www.fao.org/faoterm/en/.
[6] http://aims.fao.org/en/agrovoc.
[7] http://www.environmentontology.org/Browse-EnvO.

geographic details, and minimal semantic information consisting of the relation *located_in* (and *tributary_of* in the case of named rivers and bays).

Up to the present, 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 also appropriate, respectively, to all instances of 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 mechanisms 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 rivers, the knowledge should be represented in a semantic network according to the theoretical premises of Frame-based Terminology (Faber, 2012), which propose knowledge representations with explanatory adequacy for enhanced knowledge acquisition in communicative situations such as specialized translation (Faber, 2009). Hence, on the hypothesis that named rivers should be considered concepts rather than instances, each named river 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 is an innovative aspect of this work. Needless to say that the extraction and description of named landforms from text corpora have been applied in the field of Geographic Information Retrieval (Derungs & Puves, 2014; Derungs & Samardžić, 2018; Wartmann et al., 2018), but not with the purposes of the Frame-based Terminology.

2.2 *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*.

Depending on the language model (Baroni et al., 2014), DSMs are either count-based or prediction-based. Count-based DSMs calculate the frequency of terms within a term's context (i.e., a sentence, paragraph, document, or a sliding context window spanning a given number of terms on either side of the target term). The Correlated Occurrence Analogue to Lexical Semantics (COALS) (Rohde et al., 2006) is an example of this type of model.

Prediction-based models mostly exploit neural probabilistic language models, which represent terms by predicting the next term on the basis of previous terms. Examples of predictive models include the continuous bag-of-words (CBOW) and skip-gram (SG) models (Mikolov et al., 2013).

DSMs have been used in combination with clustering (i.e., automatic classification of objects into groups based on shared features). Work on lexical semantics applying DSMs and clustering techniques includes the identification of semantic relations (Bertels & Speelman, 2014), word sense discrimination and disambiguation (Pantel & Lin, 2002), automatic metaphor identification (Shutova et al., 2010), and classification of verbs into semantic groups (Gries & Stefanowitsch, 2010).

2.3 *Topic modelling for text mining*

Probabilistic topic modelling is a machine learning technique that automatically identifies themes or topics in a given corpus (Blei, 2012). This digital technology allows to explore documents based on the topics that run through them, rather than on keywords search alone. Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is the approach to topic modelling that has been most frequently employed. The following explanation of topic models describes LDA and is largely based on Griffiths and Steyvers (2004), Murakami et al. (2017), and Spies (2018).

In topic modelling, each term in each corpus document is assigned to one topic. A document thus consists of multiple topics of different probability (e.g., 20 percent Topic A, 10 percent Topic B, 5 percent Topic C, and so forth), approximately following the proportion of terms in the document that are assigned to each topic. All documents in a corpus share the same set of topics, but with different proportions. Therefore, a document can deal with multiple topics, and the terms that appear in that document reflect the particular set of topics it addresses.

In natural language processing, the way of modelling the contributions of different topics to a document is to treat each topic as a probability distribution over terms, viewing a document as a probabilistic mixture of these topics. Starting from observed data in a corpus (i.e., occurrence frequency of the terms in the documents), LDA is able to infer a latent structure from the corpus, consisting of a set of topics.

The content of a topic is thus reflected in the terms to which it assigns high probability. For example, high probabilities for «woods», «hill» and «stream» would suggest that a topic refers to the countryside, whereas high probabilities for «check», «bank» and «credit» would suggest that a topic refers to finance.

According to Spies (2018), from a cognitive view, topic modelling can be related to human capabilities to categorize documents. Psychological research found strong empirical evidence supporting cognitive adequacy of LDA, in comparison to semantic spaces such as Latent Semantic Analysis (Landauer et al., 2011).

As DSMs, a topic model provides a form of semantic representation, a computational analogue of how human might form semantic representations through their linguistic experience. Accordingly, the association between terms can be estimated. Since the term vectors are probability distributions over topics, the relatedness is quantified by means of information-theoretic measures for probability distributions such as Hellinger distance (Csiszár & Shields, 2004), or Jensen-Shannon divergence (Lee, 1999), *inter alia*.

In digital linguistics, topic models have previously been used for a variety of applications, including metaphor identification (Navarro Colorado & Tomás, 2015), thematic

exploration of specialized corpora (Murakami et al., 2017) and literary corpora (Jockers & Mimno, 2013), and selectional preferences for predicate arguments (Ritter et al., 2010).

3. MATERIALS AND METHODS

3.1 *Materials*

3.1.1 *Corpus data*

The terms related to named rivers were extracted from a subcorpus of English texts on Coastal Engineering, comprising roughly 7 million tokens and composed of specialized texts (scientific articles, technical reports, and PhD dissertations), and semi-specialized texts (textbooks and encyclopedias on Coastal Engineering). This subcorpus is part of the English EcoLexicon Corpus (23.1 million tokens) (see León-Araúz et al. (2018) for a detailed description).

3.1.2 *GeoNames geographic database*

The automatic detection of the named rivers in the corpus was performed with a GeoNames database dump. GeoNames[8] has over 10 million proper names for 645 different geographic entities, such as bays, beaches, rivers, and mountains. 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.

3.2 *Methodology*

3.2.1 *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 multi-word terms in EcoLexicon were then automatically matched in the lemmatized corpus and joined with underscores.

3.2.2 *Recognition of named rivers*

Both normalized and alternate names of the rivers in GeoNames were searched in the lemmatized corpus. A total of 783 designations were recognized and listed. Since various designations can refer to the same river because of syntactic variation (e.g., *Nile River* and *River Nile*), and orthographic variation (e.g., *Yangtze* and *Yangtse River*), a procedure was created to identify variants and give them a single designation in the corpus.

In the case of syntactic variations, all the designations with the word *River* in the last position were automatically transformed to the variant with *River* in the first position (e.g., *Dee River* was converted to *River Dee*) and matched in the list of recognized designations.

Orthographic variations were identified with a matrix of the Levenshtein edit distances between the 783 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 in the lemmatized corpus and joined with underscores, the number of named rivers was 674. The 250 rivers with the highest number of mentions in the corpus are shown on the map in Figure 2. Their latitudes and longitudes were

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

retrieved from the GeoNames database dump. This reflects the representativeness of the corpus in reference to river locations.

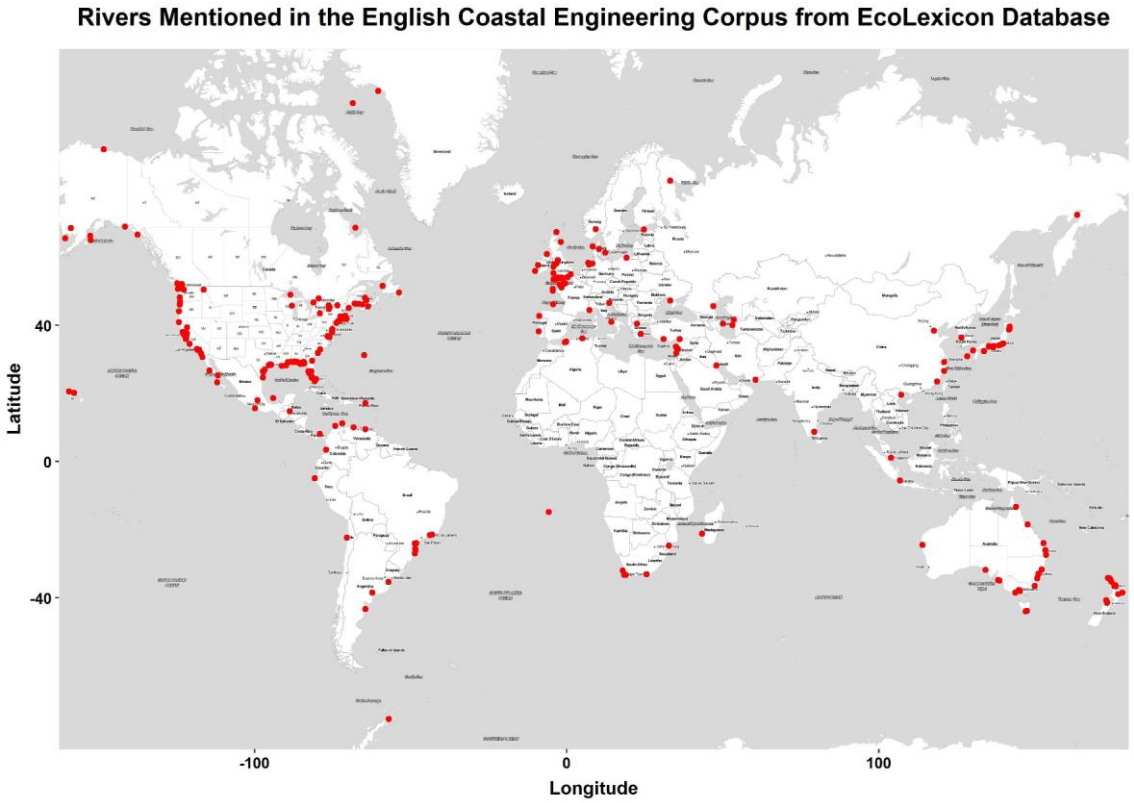

Figure 2: ***Map with the location of the named rivers***

A critical issue was the retrieval of the geographical coordinates of the rivers. Although latitudes and longitudes could be retrieved from the GeoNames database dump, occasionally, the same designation referred to rivers in different countries. For instance, the corpus only located *Yellow River* in China. However, GeoNames indicated that rivers with the same name also existed in the USA, Canada, Ireland, and Papua New Guinea. Such cases had to be resolved by corpus queries.

The occurrence frequency of the named rivers ranged from 118 (*Yantze River Estuary*) to only one mention (349 out of 674 named rivers). In our study, only those rivers with a frequency greater than 9 were considered, since DSMs perform poorly with low-frequency terms (Luhn, 1957). Figure 3 shows the 55 named rivers that fulfilled this condition, along with their number of mentions.

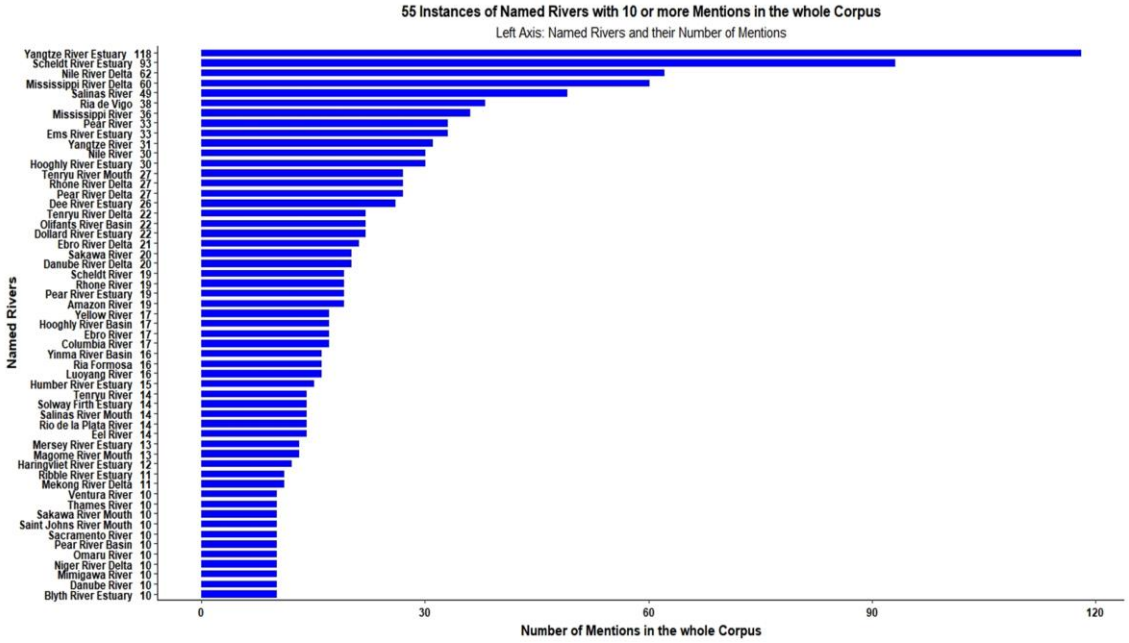

Figure 3: ***Designations and number of mentions of the 55 named rivers whose occurrence frequency was higher than 9***

3.2.3. *Term-term matrix construction*

A count-based DSM was selected to obtain term vectors since this type of DSM outperforms prediction-based ones on small-sized corpora (Ars et al., 2016; Sahlgren & Lenci, 2016). The DSM was built with the R package *quanteda* for text mining.

For the construction of the DSM, terms with fewer than three characters, numbers, and punctuation marks were removed. Additionally, the minimal occurrence frequency was set to 5 (Evert, 2007). The sliding context window spanned 30 terms on either side of the target term because large windows improve the DSM performance for small corpora (Rohde et al. 2006; Bullinaria & Levy, 2007) and capture more semantic relations (Jurafsky & Martin, 2019). We followed standard practice and did not use stopwords (i.e., determiners, conjunctions, relative adverbs, and prepositions) as context words (Kiela & Clark, 2014). Since only nouns are represented in the semantic networks, adjectives, adverbs, and verbs were also disregarded as context words.

The resulting DSM was a 4705 × 4705 frequency matrix *A*, whose row vectors represented the 55 named rivers plus the 4650 terms inside the context windows of 30 terms on either side of those rivers.

3.2.4 *Selection of rivers and terms for clustering purposes*

Subsequently, a 55 × 4650 submatrix *B* was extracted from *A*, where the rows represented the 55 named rivers, and the columns represented the 4650 terms co-occurring with the rivers. To cluster the rivers of *B* sharing the same associated terms, it was necessary to select both the rivers and the terms that best discriminated different groups of rivers. This was done by removing the rivers 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 rivers 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 (Thabet, 2005; Moisl et al., 2006). 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 et al., 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: 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 river appeared, and thus corresponded to a row of matrix *B*. As such, we had 55 named-river 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 river. The document lengths ranged from 6507 words (for *Yantze River Estuary*) to 563 words (for *Blyth River Estuary*). Moisl's (2011) method was then applied to matrix *B* to determine: a) which of the 55 named rivers should be eliminated from our analysis, if any; and b) which terms helped to distinguish different groups of the retained rivers.

Table 1 shows the length for named-river documents needed by each of the 4650 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 1 were sorted in ascending order of the required document length.

Table 1: ***Length needed for named-river documents (mostright column) associated with each of the 4650 terms (middle column) co-occurring with the rivers, according to Moisl's (2011) formula***

Index	Term	Length needed for named-river documents
1	**shoreline**	**391**
2	sediment_load	587
3	dam	648
4	sediment	677
5	reservoir	744
[…]	[…]	[…]
428	aquifer	6435
429	clay	6438
430	gaoyao_station	6438
431	**morphology**	**6505**
432	sand_transfer	6516
[…]	[…]	[…]
4649	turbulent_viscosity	687372
4650	specific_gravity	687372

Since the lowest document-length value needed by the terms was 391 words (for the term *shoreline* in the first row of Table 1), those rivers whose document length were smaller than the minimum length threshold 391 would have to be eliminated from the analysis. This meant that the 55 rivers were retained because they all had a document length larger than 391.

Regarding the selection of terms, since the maximum length of our named-river documents was 6507 words, only the first 431 terms in Table 1 were retained for clustering purposes because their needed document lengths were less than 6507 words. These results are plotted in Figure 4, where the 4650 terms co-occurring with the 55 rivers 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-river documents (6507 words), and the green vertical line marks the 431 terms whose needed document lengths were equal to or less than the maximum named-river document length. Therefore, a 55 × 431 submatrix *C* was extracted from *B* to group the river vectors.

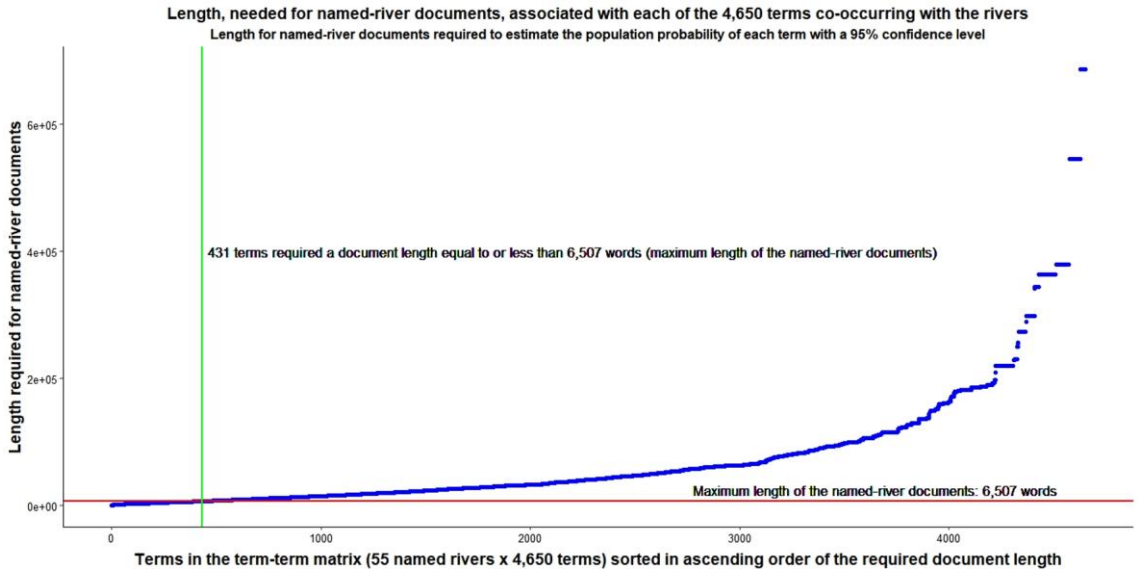

Figure 4: ***The required document lengths (vertical axis) associated with each of the 4650 terms (horizontal axis) co-occurring with the 55 named rivers***

3.2.5. *Clustering of named rivers and weighting schemes*

The 55 × 431 frequency submatrix *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 rivers (Evert, 2007: 24-30). Research on computational linguistics reveals that log-likelihood is able to capture syntagmatic and paradigmatic relations (Bernier-Colborne & Drouin, 2016; Lapesa et al., 2014) and to achieve better performance for medium-to-low-frequency data than other association measures (Alrabia et al., 2014; Krenn, 2000). However, the calculation of the log-likelihood scores was modified to cope with these critical situations:

1) When the observed frequency was less than the expected one, the score was set to 0, as recommended by Evert (2007: 22). Otherwise, the score would have been negative showing repulsion between terms, whereas our interest was in the stronger attraction to each other.
2) 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 two terms despite the absence of co-occurrence in corpus data.
3) When a term co-occurred with only one river, the corresponding addend in the log-likelihood formula (i.e., the addend where the observed frequency O_{21} takes part, according to Evert (2007: 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 logarithm 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.

A hierarchical clustering technique was then applied to the weighted 55×431 submatrix *C*. The cosine distance was used as the intervector distance measure, and Ward's method as the clustering algorithm (i.e., a criterion for choosing the pair of clusters to merge at each step, based on the minimum increase in total within-cluster variance).

Since it was not clear how strongly a cluster was supported by data, a means for assessing the certainty of the existence of a cluster in corpus data was devised. Multiscale bootstrap resampling is a method for this in hierarchical clustering, which was 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.

Thirteen groups of rivers with *p*-values higher than 95% were strongly supported by corpus data, as marked by the red rectangles in the dendrogram in Figure 5.

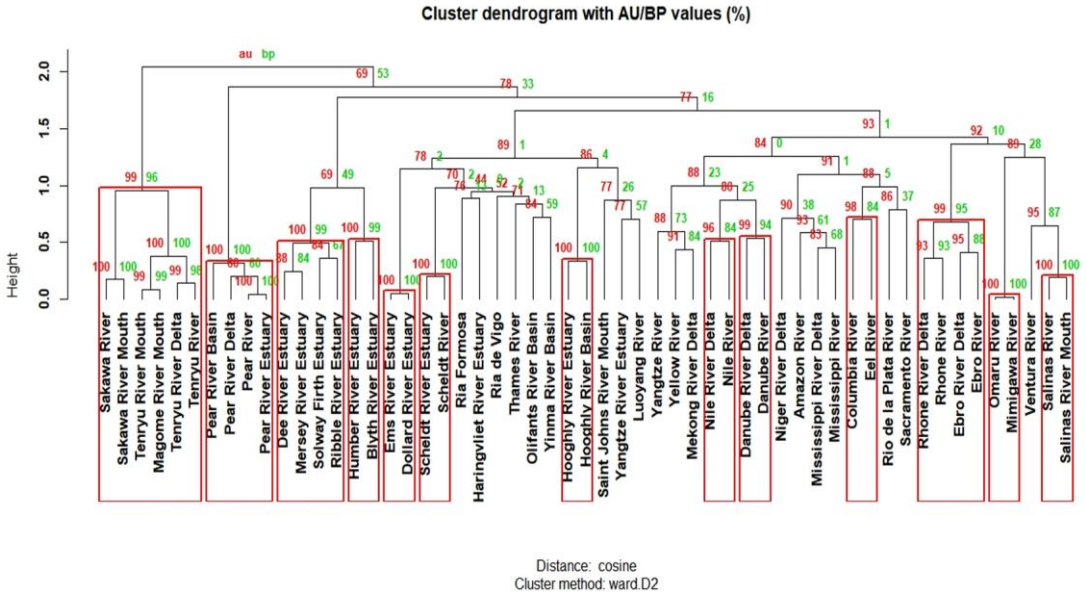

Figure 5: ***Dendrogram of the hierarchical clustering of the 55 named rivers with 13 clusters***

3.2.6. *Selection of terms for semantic network construction*

Since the 431 terms of the submatrix *C* were not sufficient to straightforwardly construct the semantic networks for the 13 clusters of rivers, another statistical method was employed to select the terms that best described the 55 rivers. In Corpus Linguistics, Moisl (2015: 77-93) 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 the 55×4650 frequency submatrix *B*, whose rows represented the 55 named rivers. The columns represented all the terms co-occurring with them (excluding the rivers) inside their context windows of 30 terms on either side.

Figure 6 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 2000 was set. This meant that only 1858 terms fulfilled all criteria for the construction of the semantic networks.

We estimated that 30 terms would be necessary for a named river to describe its semantic frame. A total number of terms around 1650 would thus be required for the description of the 55 rivers. The threshold was set accordingly, so that the number of selected terms was about 1650 terms (1858 selected terms in our study).

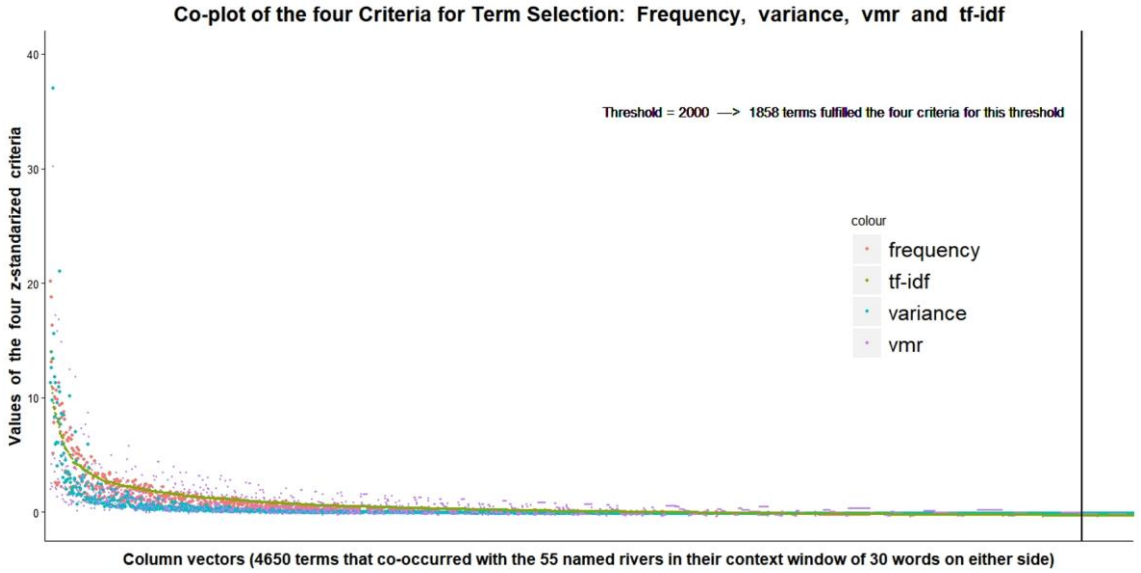

Figure 6: ***Co-plot of the four criteria for term selection: Frequency, variance, vmr, and tf-idf***

3.2.7. *Topic modelling for the extraction of terms associated with named rivers*

Once 1858 terms were selected for the semantic description of the 55 rivers, the relatedness of each river to those terms was estimated by means of a topic model. The Biterm Topic Model (BTM) (Yan et al., 2013), based on LDA, was applied to the lemmatized corpus, but containing only the occurrences of the 55 rivers and the 1858 terms selected.

In our case, BTM was chosen for the following reasons. First, it was found that BTM outperforms LDA for small corpora and short texts (note that the corpus contained only 1913 term types, namely, 55 rivers plus 1858 terms) because it helps to alleviate the data sparsity problem of LDA (i.e., the low co-occurrence frequency of term pairs reduces the semantic coherence of the topics) (*ibidem*). Furthermore, BTM explicitly models the term co-occurrences in local context windows rather than in the document level, thus capturing the short-range dependencies between terms.

For BTM, a context window size of 30 terms was set, the same value as that in the DSMs used for the clustering of the rivers and the selection of the terms for the construction of semantic frames. However, the appropriate number of topics needs to be found by experimentation, calculating the harmonic mean of the document log-likelihood estimated by different models.

The harmonic mean of the document log-likelihood is a traditional measure used to select the topic model with the best generalization capability, namely, the ability of the model to identify the topics treated in unseen document, based on the analysis of the topics appearing in the documents of a training corpus. The greater the harmonic mean of the document log-likelihood, the better.

The *BTM* package for R programming language was applied to the corpus, and 39 models were computed for the topic numbers ranging from 2 to 40. The estimated harmonic

mean of the document log-likelihood for each model is shown in Figure 7, where the optimal number of topics was found to be 40.

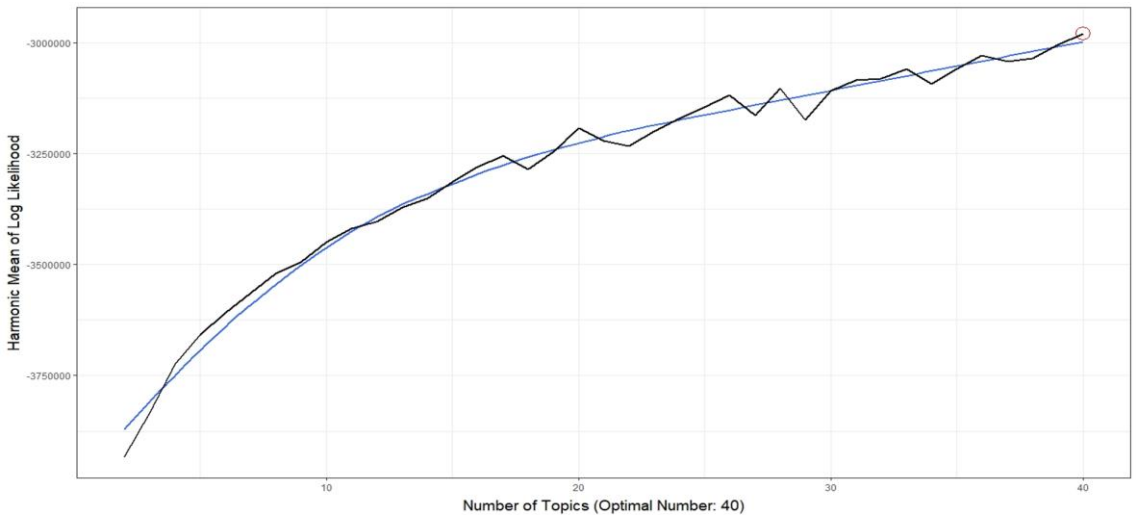

Figure 7: ***Estimated harmonic mean of the document log-likelihood of 39 topic models, with a number of topics ranging from 2 to 40, respectively. The optimal number of topics was 40***

A 1913×40 matrix D was thus extracted from the topic model, where the rows represented the 1858 terms plus the 55 rivers, and the columns the 40 inferred topics. Since each cell contained the probability that a term or river belonged to a topic, the matrix D is called term-topic matrix in the literature.

The 40 topics of the model are represented in Figure 8 by means of the R package *LDAvis*. Figure 8 illustrates the relation between the topics (left), and the top-30 most relevant terms for the topic number 13, related to sediment transport in rivers, bays, and beaches (right).

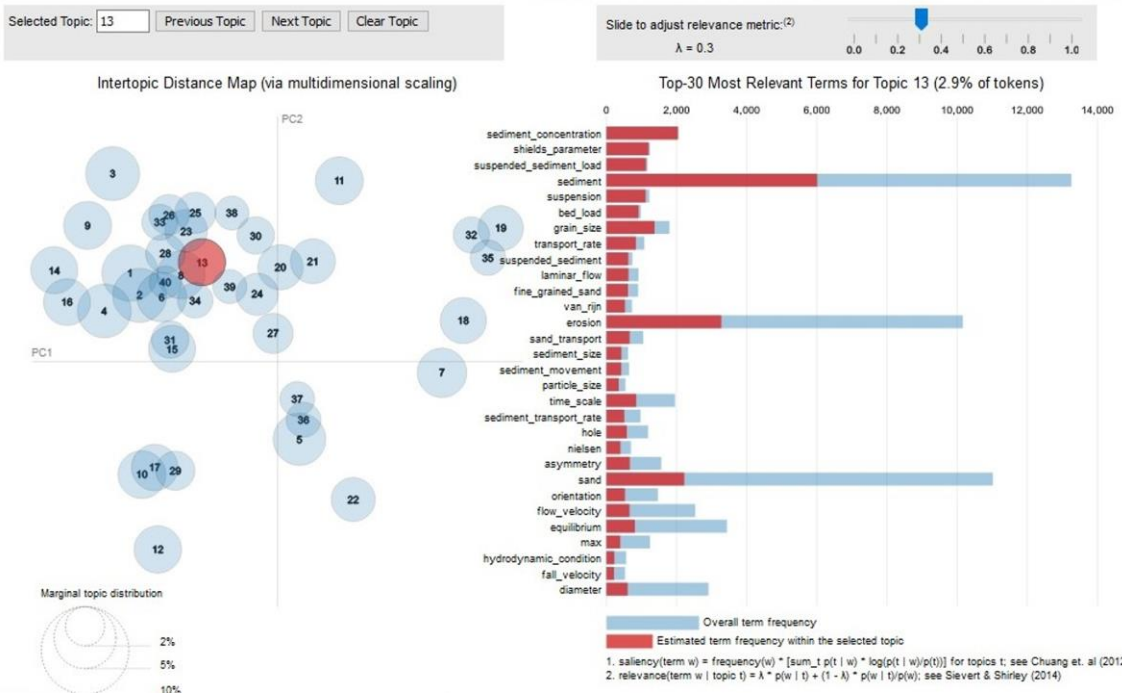

Figure 8: ***Left: Representation of the 40 topics of the model. Right: The top-30 most relevant terms for the topic number 13, related to sediment transport in rivers, bays, and beaches***

3.2.8. *Terms characterizing each cluster*

To ascertain the terms strongly associated with each of the 13 clusters, the following procedure was used:

1) For each of the named rivers in the 13 clusters, a set of the top-30 terms, most associated with each river, was extracted from the term-topic matrix *D* using Hellinger similarity, namely, the inverse magnitude of Hellinger distance. Hellinger similarity ranges from zero to one. The greater the Hellinger similarity between two term vectors, the stronger the relatedness of the terms.
2) For each cluster, the mathematical operation set intersection was applied to the sets of the top-30 terms, most associated with the rivers in the same cluster. Only the shared terms with a Hellinger similarity higher than 0.4 were selected.

A reduced set of terms was thus obtained for each cluster to describe the named rivers.

4. RESULTS

4.1 *Qualitative evaluation of the term extraction method*

For each of the clusters, the term selection method, described in the above section, produced a set of terms characterizing the named rivers. Those term sets were qualitatively compared to gold standard sets of terms, manually extracted from the context windows of the 36 rivers clustered in the 13 groups in Figure 5, which best described, in our opinion, each of the clusters for semantic network construction.

The comparison of the 13 sets of terms, obtained by the term selection method, with the corresponding gold standard term sets unveiled that most terms in the gold standard sets were collected by the selection method. Therefore, we assessed the reliability of the method to be highly enough to ensure that the selected terms could adequately describe the river clusters. For the construction of the semantic frames presented in the next subsection, the term selection method was thus applied.

4.2 *Semantic frames describing the river clusters*

Because of space constraints, only the results for some clusters are provided. Numbering the clusters in Figure 5 from left to right, the clusters number two, three, and thirteen are described. As shown in Figure 5, the second cluster is formed by the basin, delta, and estuary of the *Pearl River*, and the river itself, located in China. The *Salinas River* and its mouth, placed in California (the USA), comprise the last cluster. Both clusters were selected because the named rivers, despite being located in different world areas, are related to the same topics, namely, those of sediment concentration and sediment transport in rivers, sediment discharge into bays and seas, and the negative effects of sediment supply decrease on coastal erosion because of human activities. The third cluster consists of the *Dee*, *Mersey*, *Ribble*, and *Solway Firth* estuaries, all located in Great Britain and involved in national shoreline management plans for managing risks due to flood and erosion in coasts and rivers.

For the description of the frames, the semantic relations were manually extracted by querying the corpus in Sketch Engine (Kilgarriff et al. 2004), and by analysing knowledge-rich contexts, namely, "a context indicating at least one item of domain knowledge that could be useful for conceptual analysis" (Meyer, 2001: 281). The query results were concordances of any elements between the river in a cluster and related terms in a ±40 span. The semantic relations were those in EcoLexicon (Faber et al. 2009), with the

addition of *supplies*, *accumulates_in*, *increases*, *decreases*, *tributary_of*, *erodes*, *loses_into*, *discharges_into*, *located_near*, *develops*, *applies_to*, *monitors*, *monitored_in*, *uses*, and *simulates*, necessary for the explanatory adequacy of the frames (Faber, 2009). Furthermore, the semantic frames shown in the following were validated by a Coastal Engineering expert from the University of Granada (Spain).

4.2.1 *Second cluster in Figure 5:* Pearl River

Predicting *sediment load* (kg/year) in a river system has long been a goal of earth scientists for numerous reasons, including alternation of fish habitats, changes in the load from anthropogenic effects, and the evolution of deltas, estuaries, and coastal environments. Hence, hydrologists have made efforts in applying *sediment rating curves* that can empirically describe the relationship between *suspended sediment concentration* (g/km^3) and *water discharge* (m^3/s) for a certain location. In *sediment rating curves*, *sediment rating parameters* also intervene, which are often associated with *river bed morphology* and *soil erodibility*. Engineers use *sediment rating curves* for predicting the life span of a *dam* on a river, and earth scientists use them to study the erosional and depositional environments.

Dam and *reservoir construction* are regarded as the main cause of the decline in *sediment load*. For that reason, the issue of *sediment load* in the *Pearl River Delta* was studied. Attention was paid to the *sediment rating parameters* of the *sediment rating curves*. The parameters reflected a temporal relationship between *water discharge* and *suspended sediment concentration* due to *human activities*, such as *land use* and *reservoir construction*. These activities are causing a decrease in sediment supply from the *Pearl River*, with grave consequences on the coast (see Figure 9).

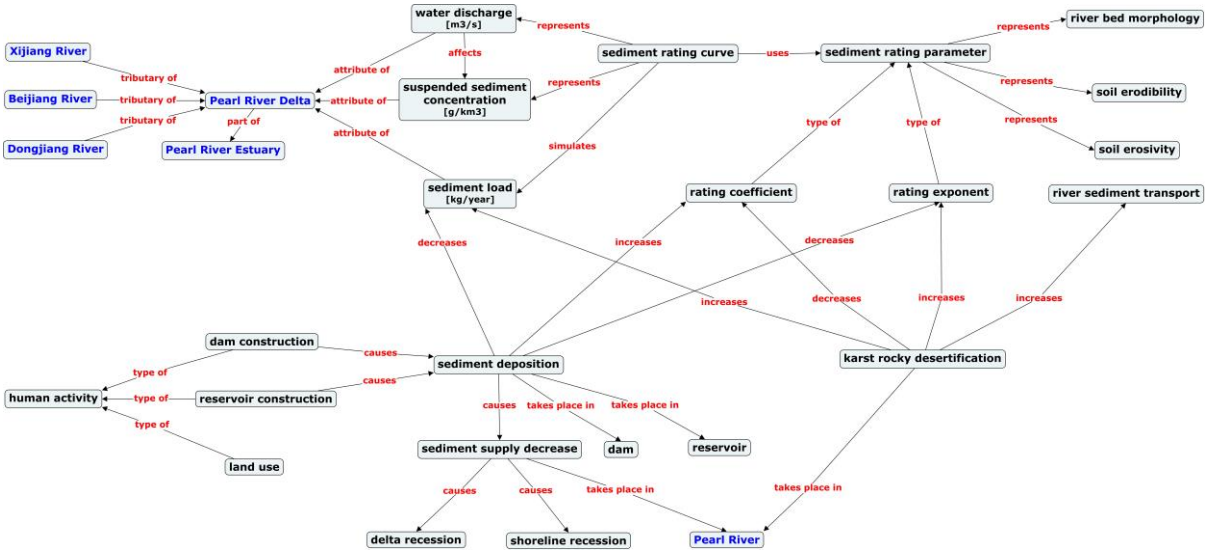

Figure 9: ***Semantic network of the terms associated with the*** **Pearl River**

4.2.2 *Thirteenth cluster in Figure 5:* Salinas River

Sediment is a resource essential to the economic and environmental vitality of *Monterey Bay* beaches, and the mitigation of *shoreline erosion*. Sources of *sand* to the southern *Monterey Bay* are from discharge of the *Salinas River*, and erosion of the *beaches* and *coastal dunes*. However, human activities and natural processes are changing the sand availability, namely, the *dams* constructed along the *Salinas River* have decreased its *sand supply*; most sediment from the river is driven north and potentially lost into *Monterey Submarine Canyon*; and *beach sand mining* and *sea level rise* cause *dune erosion* to progress at a higher rate.

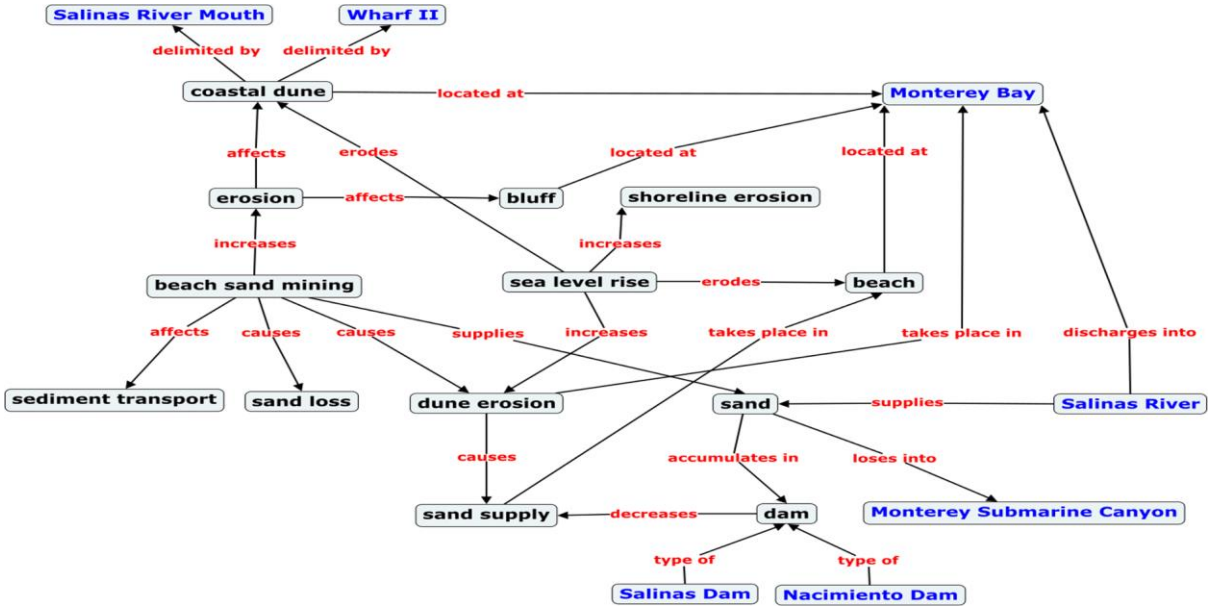

Figure 10: ***Semantic network of the terms associated with the*** **Salinas River**

4.2.3 *Third cluster in Figure 5:* Dee, Mersey, Ribble, *and* Solway Firth *estuaries*

In Great Britain, the *Department for Environment, Food and Rural Affairs*, and the *Welsh Assembly Government* have required to produce *Shoreline Management Plans* (SMPs) for the length of coastline which stretches from *Great Orme's Head* in Wales to the *Scottish Border* on the *Solway Firth Estuary*, including the *Dee*, *Mersey*, and *Ribble* estuaries.

The overall aim of SMP is the *flood* and *erosion risk management* along the coast. Hence, SMP sets out policies for managing the coastline to reduce those risks to *urban areas*, *industrial* and *commercial activities*, and natural environments such as *Marine Protection Areas*. One of those policies is the *managed realignment* (MR), namely, removing *coastal defenses* or building new ones further inland to allow an area to become flooded by the sea. MR, usually pursued in estuarine areas, permits: the restoration of *accommodation space* containing *sediment sinks* for sediments mobilized by erosion; *habitat creation*, such as salt marshes and mud flats; and the long-term *coastal defense resilience*. However, in areas where there are benefits in reverting to natural processes through MR, there may be an increase in *tidal flooding* or *erosion risk* with associated negative impacts on *historic assets*.

Other plans, incorporated into SMP, have been developed to co-ordinate works for flood and erosion risk management, such as *Catchment Flood Management Plans*, which predominantly consider *fluvial flood risks*. SMP also includes a *monitoring programme* to check *shoreline features* and *wetland bird surveys*, among others, and *strategic studies*, for instance, for the *extreme water level prediction* in the *Dee estuary* (see Figure 11).

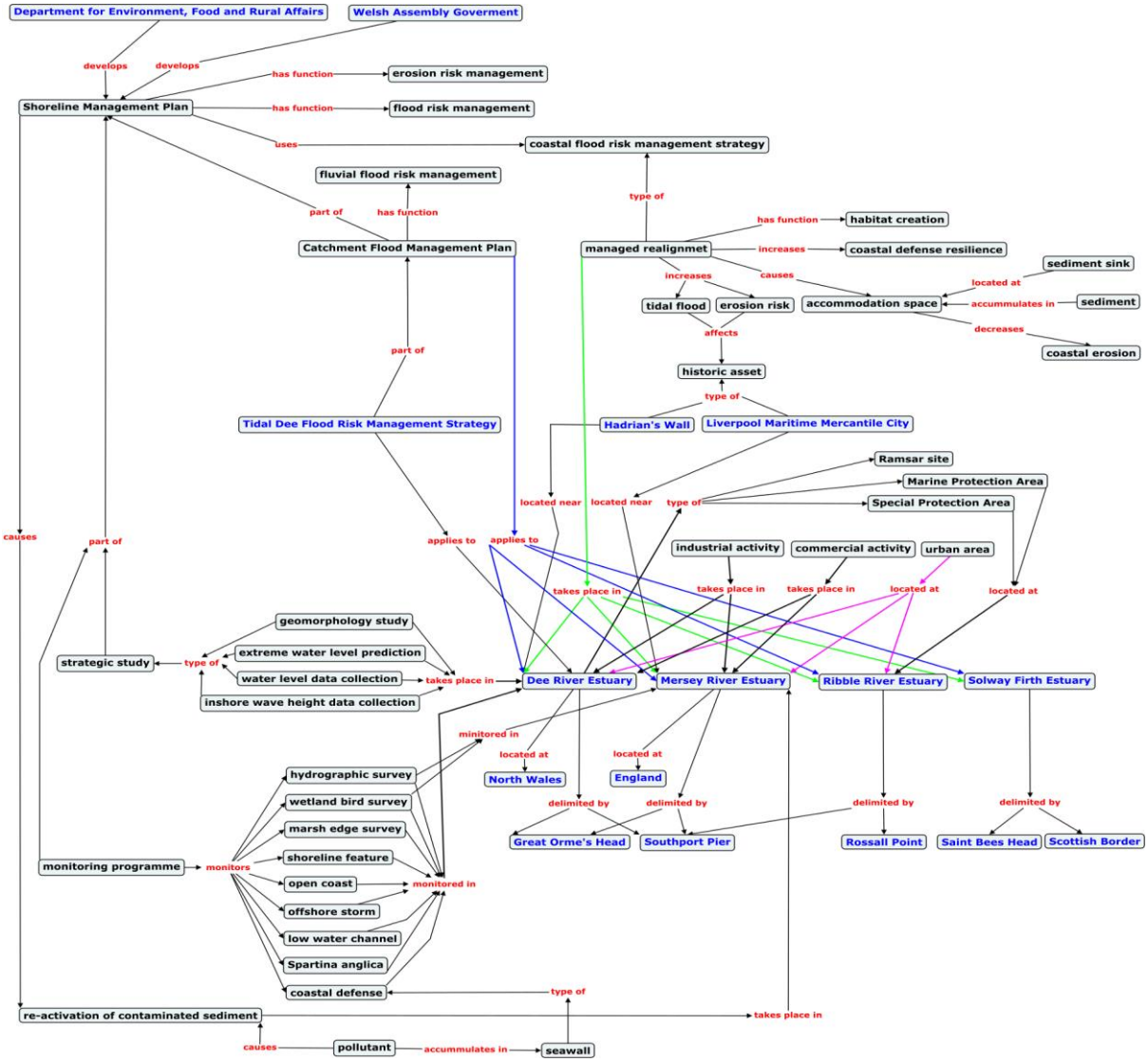

Figure 11: ***Semantic network of the terms associated with the*** **Dee, Mersey, Ribble,** ***and*** **Solway Firth** ***estuaries***

5. CONCLUSIONS

To extract knowledge for the conceptual structures (Faber, 2012) that underlie the usage of named rivers 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 by querying the corpus in Sketch Engine, and by analysing knowledge-rich contexts. The query results were concordances of any elements between the river in a cluster and related terms in a ±40 span. It was a time-consuming task, although essential for the explanatory adequacy of frames (Faber, 2009). In future research, the automatic extraction of semantic relations for named rivers by means of knowledge patterns (KPs) (Meyer, 2001) will be tested. KPs are lexico-syntactic markers that generally convey semantic relations in real texts. For instance, examples of generic-specific KPs are *such as*, *is a kind of*, *and other*, and so on. In León-Araúz et al. (2016), a KP-based sketch grammar for Sketch Engine was developed, which automatically provides a list of terms that hold a specific semantic relation with a target term. In future work, these KPs will be applied to our corpus.

The method for the extraction of terms closely associated with named rivers offered successful results to construct semantic frames with explanatory adequacy, according to the premises of Frame-based Terminology. It combined, on the one hand, a count-based DSM, weighted by the log-likelihood association measure, to cluster rivers, and selection procedures for both rivers and terms based on statistical criteria. On the other hand, a topic model was employed to extract the terms related to each named river.

The semantic frames in the previous section reflected that most terms related to named rivers are complex nominals (e.g., *sediment rating curves*, *suspended sediment concentration*, *beach sand mining*). English complex nominals are multi-word terms (MWTs) with a head noun preceded by modifying elements (i.e., nouns or adjectives) (Levi, 1978). The abundance of MWTs is due to, at least, three reasons: specialized language units are mostly represented by such compound forms (Nakov, 2013); complex nominals provide relevant information for the conceptual structuring of a specialized domain (Meyer & Mackintosh, 1996); and they are frequently used to designate specialized concepts in English (Sager et al. 1980). For these reasons, complex nominals should be included in the semantic networks and in TKBs such as EcoLexicon (Cabezas-García & Faber, 2018).

Nevertheless, MWT extraction was possible because they were previously matched and joined by means of underscoring in the lemmatized corpus, thanks to the list of MWTs stored in EcoLexicon. This confirms that EcoLexicon is a valuable resource for any natural language processing tasks related to specialized corpora on environmental science. Therefore, the profusion of MWTs underlines the importance of applying either automatic, semi-automatic, or manual methods to recognize them for the knowledge representation of a specialized domain.

According to the premises of Frame-based Terminology, the semantic frames presented underlie the usage of named rivers and their associated terms in Coastal Engineering texts, and provide the background knowledge about them, necessary in communicative situations such as specialized translation to appropriately render terms into another language (Faber, 2012). Moreover, the frames make the specific semantic behavior of named rivers in Coastal Engineering domain explicit by means of the description of semantic relations and term combinations (Faber, 2009).

Finally, the conceptual structures also highlighted that, in Coastal Engineering texts, each named river is found to be semantically related to different, specific topics and concepts (e.g., one of the topics is the prediction of suspended sediment concentration by applying sediment rating curves; other topics are the sediment supply decrease due to dam construction, and the negative effects of beach sand mining on dune erosion). On the evidence of these findings supporting our working hypothesis, we thus defend that named rivers in Coastal Engineering domain should be semantically represented in terminological resources. Consequently, it would be more appropriate for named rivers to be deemed concepts for themselves, rather than mere instances of the RIVER concept.

ACKNOWLEDGMENTS

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 author.

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*, vol. 1 (pp. 238-247). Baltimore: ACL.

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.

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.

Blei, D.M. (2012). Probabilistic topic models. *Communications of the ACM*, 55 (4), 77-84.

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.

Cabezas-García, M., & Faber, P. (2018). Phraseology in specialized resources: An approach to complex nominals. *Lexicography*, 5(1), 55-83.

Csiszár, I., & Shields, P.C. (2004). Information theory and statistics: A tutorial. *Foundations and Trends in Communications and Information Theory*, 1(4), 417-528.

Derungs, C. & Purves, R.S. (2014). From text to landscape: Locating, identifying and mapping the use of landscape features in a Swiss Alpine corpus. *International Journal of Geographical Information Science*, 28(6), 1272-1293.

Derungs, C. & Samardžić, T. (2018). Are prominent mountains frequently mentioned in text? Exploring the spatial expressiveness of text frequency. *International Journal of Geographical Information Science*, 32(5), 856-873.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1), 61-74.

Evert, S. (2007). Corpora and collocations. Extended manuscript of chapter 58 in 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: 2020-02-03).

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 and 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.

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.

Griffiths, T.L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(1), 5228-5235.

Jockers, M.L., & Mimno, D. (2013). Significant themes in 19-century literature. *Poetics*, 41(6), 750-769.

Jurafsky, D., & Martin, J.H. (2019). Vector semantics and embeddings. In *Speech and Language Processing*. Draft of October 2, 2019. Retrieved from https://web.stanford.edu/~jurafsky/slp3/6.pdf (last access: 2020-02-03).

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

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.

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

Landauer, T.K., McNamara, D.S., Dennis, S. & Kintsch, W. (Eds.). (2011). *Handbook of Latent Semantic Analysis*. New York: Routledge.

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.

Lee, L. (1999). Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics* (pp. 25–32). College Park (Maryland): ACL.

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.

Levi, J. (1978). *The Syntax and Semantics of Complex Nominals*. New York: Academic Press.

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

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

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 and Philadelphia: John Benjamins.

Meyer, I., & Mackintosh, K. (1996). Refining the terminographer's concept-analysis methods: How can phraseology help? *Terminology*, 3, 1-26.

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.

Moisl, H. (2011). Finding the minimum document length for reliable clustering of multi-document natural language corpora. *Journal of Quantitative Linguistics*, 18 (1), 23-52.

Moisl, H. (2015). *Cluster Analysis for Corpus Linguistics* (pp. 77-93). Berlin, Munich and 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: 2020-02-03).

Murakami, A., Thompson, P., Hunston, S., & Vajn, D. (2017). 'What is this corpus about?': Using topic modelling to explore a specialised corpus. *Corpora*, 12(2), 243-277.

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

Navarro Colorado, B., & Tomás, D. (2015). A fully unsupervised topic modeling approach to metaphor identification. In *Actas del XXXI Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural* (without pagination). Alicante (Spain): SEPLN.

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.

Ritter, A., Mausam, & Etzioni, O. (2010). A latent Dirichlet allocation method for selectional preferences. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 424-434). Uppsala (Sweden): ACL.

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.

Sager, J.C., Dungworth, D., & McDonald, P.F. (1980). *English Special Languages. Principles and Practice in Science and Technology.* Wiesbaden: Brandstetter Verlag.

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.

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

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

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.

Spies, M. (2018). Probabilistic topic models for small corpora – An empirical study. In C. Roche (Ed.), *TOTh 2017 – Terminologie & Ontologie: Théories et Applications* (pp. 137-160). Chambéry (France): Éditions de l'Université Savoie Mont Blanc.

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.

Wartmann, F.M., Acheson, E. & Purves, R.S. (2018). Describing and comparing landscapes using tags, texts, and free lists: An interdisciplinary approach. *International Journal of Geographical Information Science*, 32(8), 1572-1592.

Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013). A biterm topic model for short texts. In *Proceedings of the 22nd International Conference on World Wide Web* (pp. 1445-1456). Rio de Janeiro (Brazil): WWW.