languages

MDPI

Article

Extraction of Terms Related to Named Rivers

Juan Rojas-Garcia * and Pamela Faber

Department of Translation and Interpreting, University of Granada, 18002 Granada, Spain
* Correspondence: juanrojas@ugr.es

Received: 16 May 2019; Accepted: 15 June 2019; Published: 27 June 2019

Abstract: EcoLexicon 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 for extracting terms associated with named rivers (e.g., *Mississippi River*). Terms were extracted from a specialized corpus, 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.

Keywords: named river; conceptual information extraction; geographic contextualization; text mining; Frame-Based Terminology

1. Introduction

EcoLexicon (http://ecolexicon.ugr.es) is a multilingual, terminological knowledge base (TKB) 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 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 rivers (e.g., *Nile River*) 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 following subsections provide the motivation for the research and the background on distributional semantic models. The rest of this paper is organized as follows. Section 2 explains the materials and methods applied in this study, namely, the automatic identification of named rivers, the selection procedure for terms in distributional semantic models, and the clustering technique for rivers sharing associated terms. Section 3 shows the results obtained. Finally, Section 4 discusses the results and presents the conclusions derived from this work as well as plans for future research.

1.1. Motivations for the Research

Despite the fact that named landforms, among other named entities, are frequently found in specialized texts on the 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[1], GEMET[2], or FAO Term Portal[3]. In contrast, AGROVOC[4] basically contains a list of named landforms with hyponymic information, whereas ENVO[5] 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).

Until now, 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. To cope with this type of situation, TKBs should include the semantic representation of named landforms.

To achieve this aim in EcoLexicon, regarding named rivers, 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 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.

1.2. 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 and 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.

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 Semantic (COALS) (Rohde et al. 2006) is an example of this type of model.

Prediction-based models 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. Work on lexical semantics applying DSMs and clustering techniques includes the identification of semantic relations (Bertels and Speelman 2014), word sense discrimination and disambiguation (Pantel and Lin 2002), automatic metaphor identification (Shutova et al. 2010), and classification of verbs into semantic groups (Gries and Stefanowitsch 2010).

2. Materials and Methods

2.1. Materials

2.1.1. Corpus Data

The terms related to named rivers were extracted from a subcorpus of English texts on coastal engineering, comprising roughly seven million tokens and composed of specialized and

---

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

semi-specialized texts. This subcorpus is part of the English EcoLexicon corpus (23.1 million tokens) (see León-Araúz et al. (2018) for a detailed description).

2.1.2. GeoNames Geographic Database

The automatic detection of the named rivers in the corpus was performed with a GeoNames database dump. GeoNames (http://www.geonames.org) 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.

2.2. Methodology

2.2.1. Pre-Processing

After compilation and cleaning, the corpus texts were tokenized, tagged with parts of speech, lemmatized, and lowercased in R programming language. The multi-word terms in EcoLexicon were then automatically matched in the lemmatized corpus and joined with underscores.

2.2.2. Named River Recognition

Both normalized and alternate names of the rivers in GeoNames were searched in the lemmatized corpus. 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. Because of space constraints, the procedure is not described.

The variants were normalized in the lemmatized corpus and joined with underscores. The 250 rivers with the highest number of mentions in the corpus are shown on the map in Figure 1. Their latitudes and longitudes were retrieved from the GeoNames database dump. This reflects the representativeness of the corpus in reference to river locations.

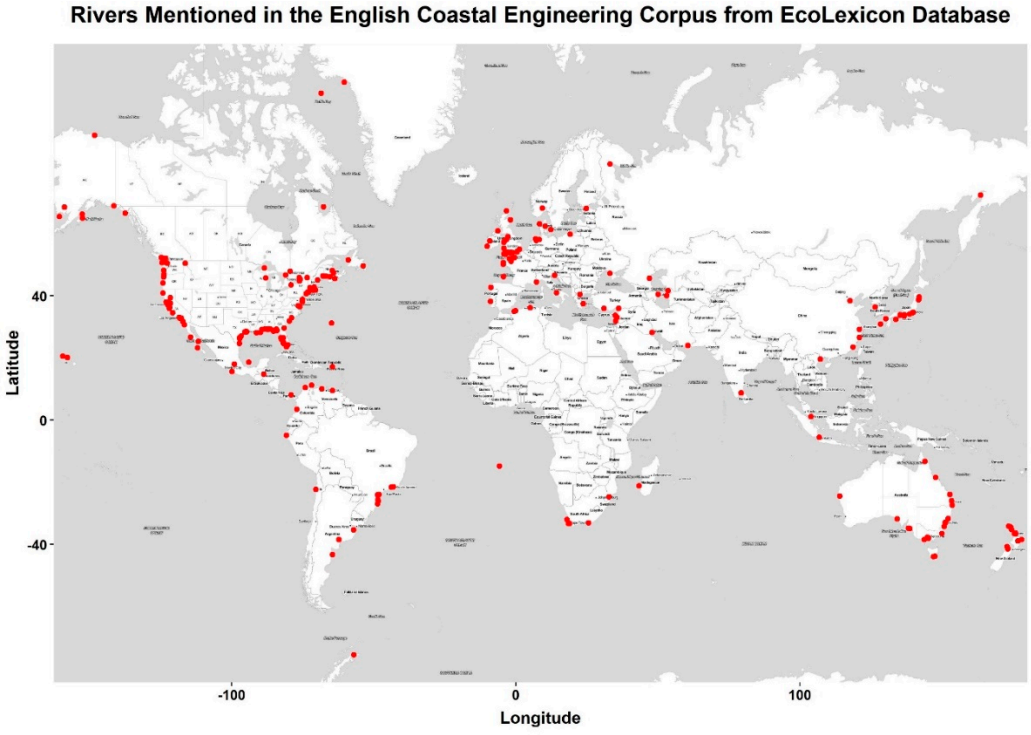

Figure 1. Map with the location and color-coded frequency of the named rivers.

The occurrence frequency of the named rivers ranged from 118 to 1 mention. In our study, only those rivers with a frequency greater than 9 were considered. Figure 2 shows the 55 named rivers that fulfilled this condition, along with their numbers of mentions.

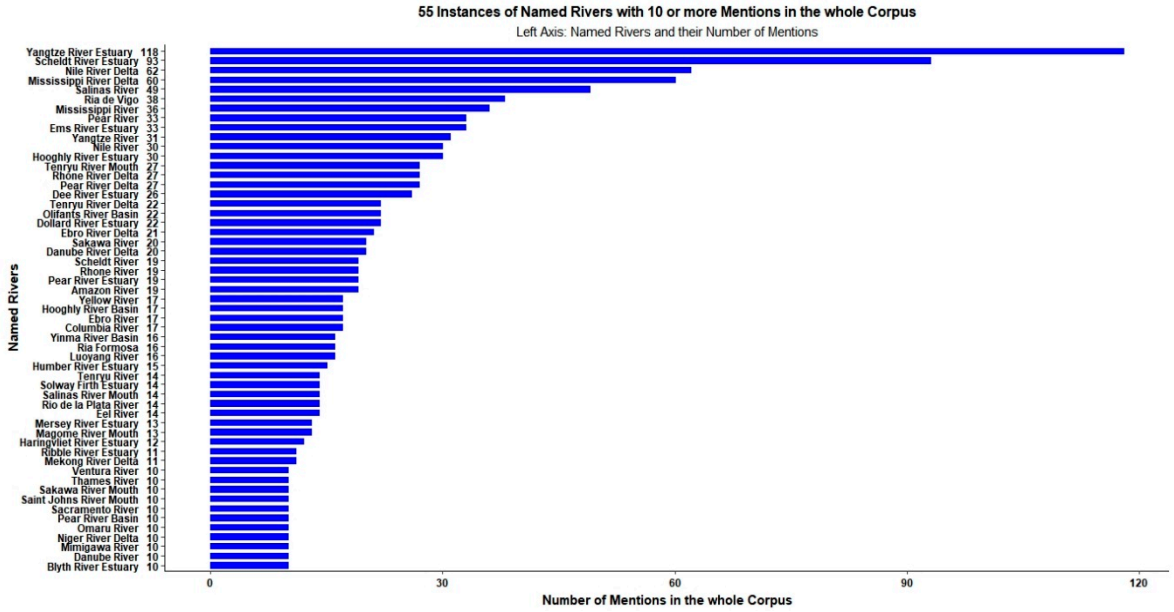

Figure 2. Designations and number of mentions of the 55 named rivers whose occurrence frequency was higher than 9.

2.2.3. 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 and Lenci 2016).

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 2008). 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 and Levy 2007) and capture more semantic relations (Jurafsky and Martin 2017). We followed standard practice and did not use stopwords (i.e., determiners, conjunctions, relative adverbs, and prepositions) as context words (Kiela and 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 matrix, 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.

2.2.4. Term Selection Procedure and Weighting Schemes

Subsequently, a 55×4650 submatrix was extracted, where the rows represented the 55 named rivers, and the columns represented the 4650 terms co-occurring with them. To cluster rivers sharing associated terms, the terms that best discriminated different groups of rivers were selected. This was done by applying Moisl (2015, chp. 3) statistical criteria, whereby only the column vectors with the highest values in raw frequency, variance, variance-to-mean ratio (vmr), and term frequency-inverse document frequency (tf-idf) were retained. Figure 3 shows the co-plot of the four criteria in descending order of magnitude. A threshold of 2000 was set. This meant that only 1858 column terms fulfilled all criteria.

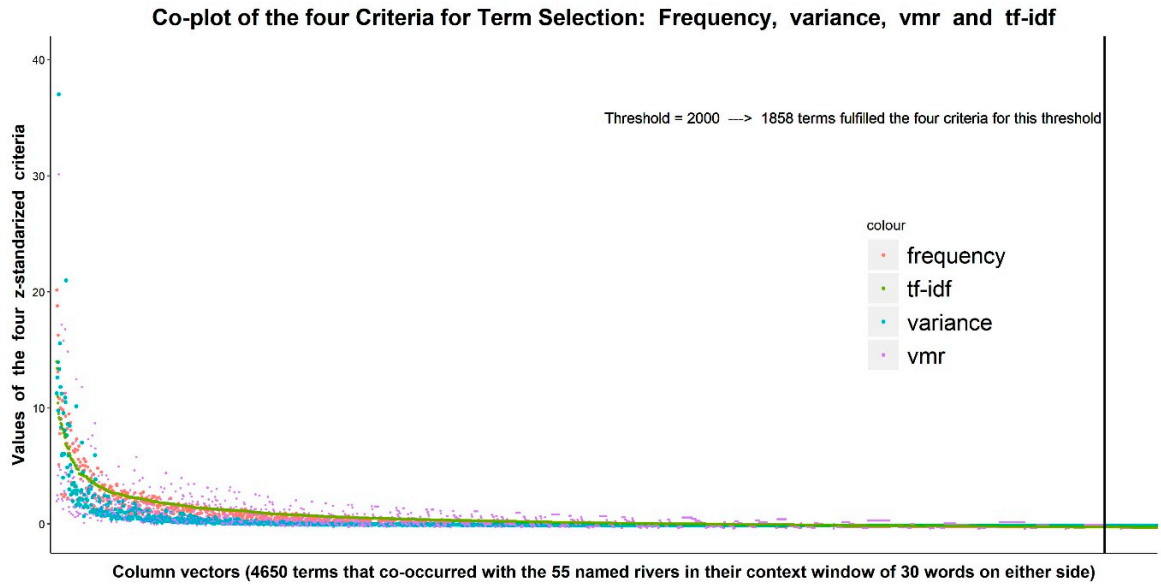

Figure 3. Co-plot of the four criteria for term selection.

Accordingly, a reduced matrix of 1913 × 1913 dimensions (1858 terms plus 55 named rivers) was obtained. The matrix was then subjected to three weighting schemes. First, the statistical log-likelihood measure calculated the association score between all term pairs, since it captures syntagmatic and paradigmatic relations (Bernier-Colborne and Drouin 2016; Lapesa et al. 2014) and achieves better performance for small-sized corpora (Alrabia et al. 2014). Secondly, the scores were transformed by applying logarithms to reduce skewness (Lapesa et al. 2014). Finally, the row vectors were normalized to unit length.

2.2.5. Clustering of Named Rivers

A hierarchical clustering technique was applied to the weighted 55 × 1858 submatrix. The cosine distance was used as the intervector distance measure, and the Ward's method as the clustering algorithm, namely, 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 (Suzuki and Shimodaira 2004) is a method for this in hierarchical clustering, which was implemented in the R package *pvclust* (Suzuki and 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 4.

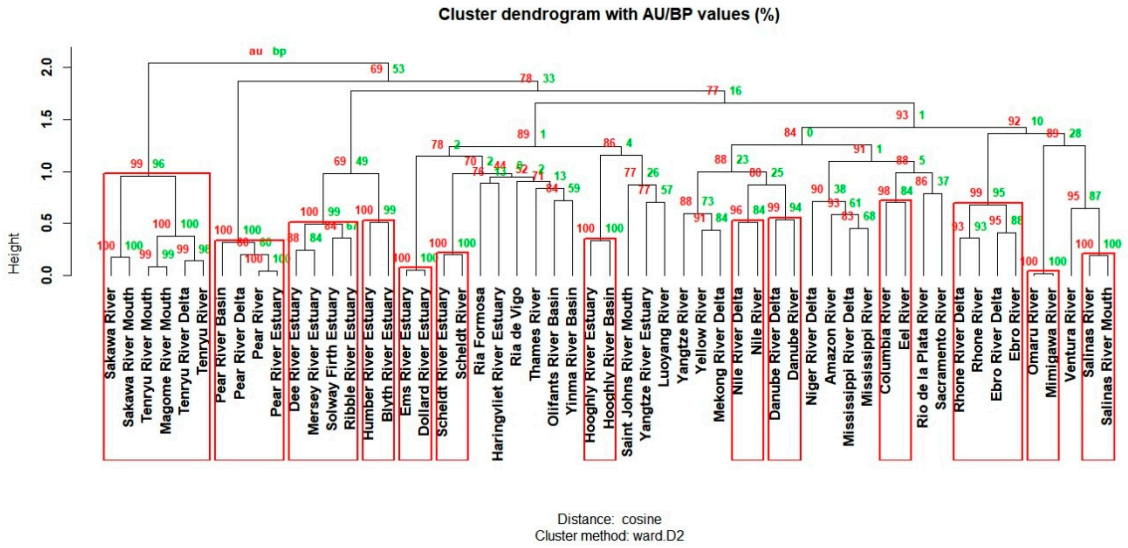

Figure 4. Dendrogram of the hierarchical clustering of the 55 named rivers with 13 clusters.

2.2.6. 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 semantically related to each river, was extracted from the DSM using cosine similarity.
2. For each cluster, the mathematical operation set intersection was applied to the sets of the top 30 terms, most semantically related to the rivers in the same cluster. Only the shared terms with a cosine similarity higher than 0.55 were selected.

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

3. Results

Because of space constraints, only the results for the first and twelfth clusters in Figure 4 (numbering the clusters from left to right) are presented in this paper. As shown in Figure 4, the first cluster is formed by the *Sakawa* and *Tenryu* rivers, *Sakawa*, *Tenryu*, and *Magome* river mouths, and the *Tenryu River* delta, all located in Japan. The *Omaru* and *Mimigawa* rivers, also located in Japan, comprise the twelfth cluster (see Table 1). These clusters were selected because both contain different rivers, and they all flow in Japan. We found it interesting to explore the reasons why different rivers were grouped together, and why there were two groups of Japanese rivers in the dendrogram, rather than only one.

Table 1. Designations and locations of the rivers in the first and twelfth clusters.

Cluster 1 (Japan)	Cluster 12 (Japan)
Sakawa River	Omaru River
Sakawa River Mouth	Mimigawa River
Tenryu River	
Tenryu River Mouth	
Tenryu River Delta	
Magome River Mouth	

For the description of the frames, the semantic relations were manually extracted by querying the corpus in Sketch Engine (Kilgarriff et al. 2004), and analyzing knowledge-rich contexts, namely, "a context indicating at least one item of domain knowledge that could be useful for conceptual analysis"

(Meyer 2001, p. 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*, *prevents*, *accumulates_in*, *inputs*, and *simulates*.

The semantic networks described in the following subsections reflect that most terms related to named rives are complex nominals (e.g., *longshore sand transport*, *beach nourishment*). English complex nominals are multi-word terms (MWTs) with a head noun preceded by a modifying element (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 and 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 and Faber 2018).

3.1. First Cluster: Sakawa, Tenryu and Magome Rivers

After the construction of dams and coastal protection structures (i.e., breakwaters, jetties, etc.), and extensive *riverbed excavation* for sand mining, the sediment supplied from the *Sakawa* and *Tenryu* rivers markedly decreased, resulting in *beach erosion* on both the *Seisho* and *Enshu-nada* coasts, into which the *Sakawa* and *Tenryu* rivers discharge, respectively. Additionally, since *submarine canyons* have developed very close to the shoreline on the *Seisho Coast*, most *river sediment* from the *Sakawa River* sinks into them because of the *fluvial fan* at its mouth, thus causing *sand loss*. Since urgent measures were required to protect both coasts, beach topography changes were predicted. For that reason, the beach modifications were simulated using the *contour-line-change model* considering the following: the variation in grain size of the beach sediments, the *longshore sand transport* through the *submarine canyons*, and the sediment supply from both rivers.

In the case of the *Sakawa River* (see Figure 5), the most favourable result was obtained when nourishment was performed using fine- and coarse-sized materials, known as *mixture materials*, because the *Seisho Coast* advanced, and the *seabed erosion* near the *submarine canyons* was prevented.

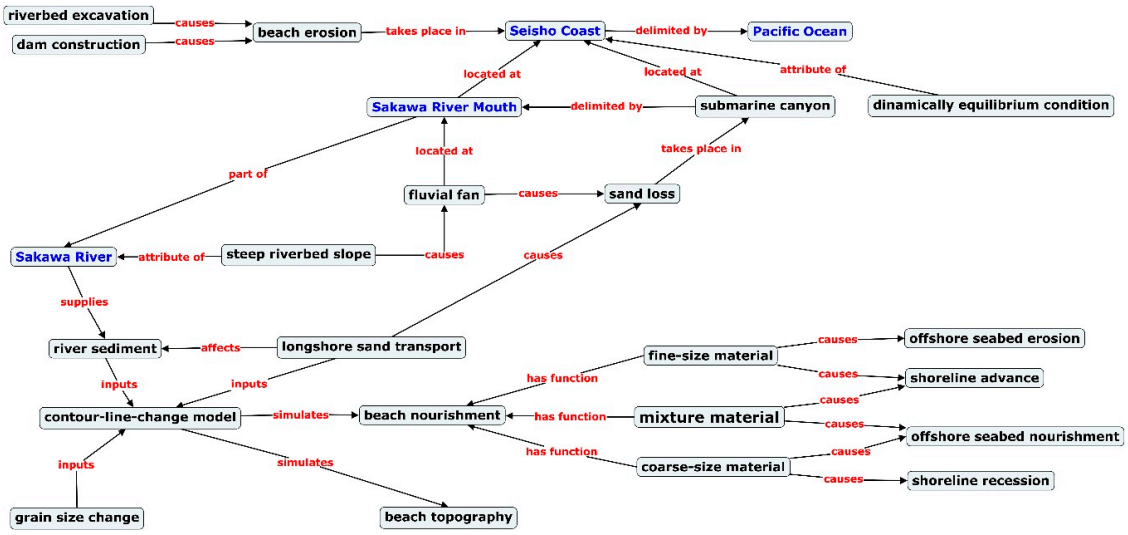

Figure 5. Semantic network of the terms associated with the *Sakawa River*.

In the case of the *Tenryu River* (see Figure 6), *sand bypassing* (i.e., man-induced transfer of sand from a given distance landwards of the coast line, to a beach) at *Sakuma Dam* as a measure against *beach erosion* on the *Enshu-nada Coast* was taken to recover the sandy beach, but *breakwaters*, previously constructed as a measure against *beach erosion*, were a barrier to the movement of sand by *longshore sand transport*.

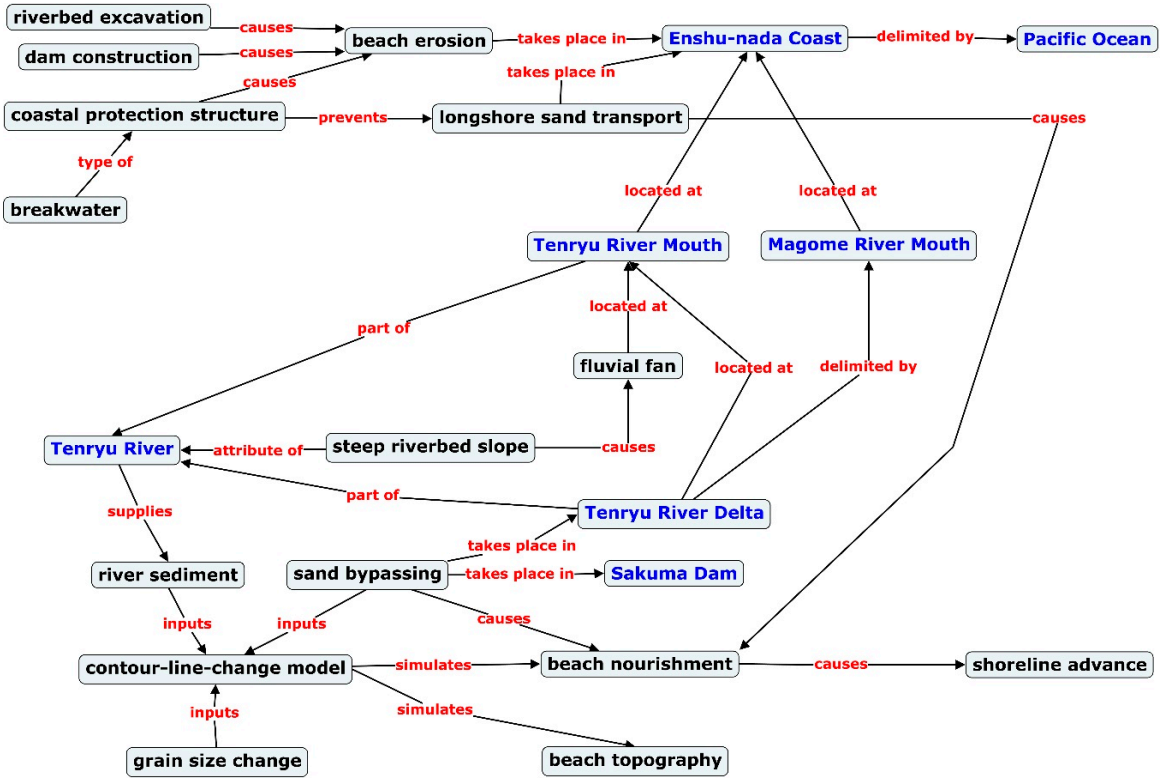

Figure 6. Semantic network of the terms associated with the *Tenryu* and *Magome* rivers.

3.2. Twelfth Cluster: Omaru and Mimigawa Rivers

Owing to the interruption of sediment flow at dams, degradation of the riverbed was observed downstream of the *Omaru*, *Mimigawa*, *Hitotsuse*, and *Ooyodo* rivers. Sediment discharge through these four rivers was thus considered to decrease considerably, causing *coastal erosion* on the *Miyazaki Coast*. The *Sumiyoshi Beach*, located on this coast, is thus a severely eroded beach because of the decrease in sediment supply from the four rivers, and the blocking of *longshore sand transport* by the *breakwater* of the *Miyazaki Port* (see Figure 7).

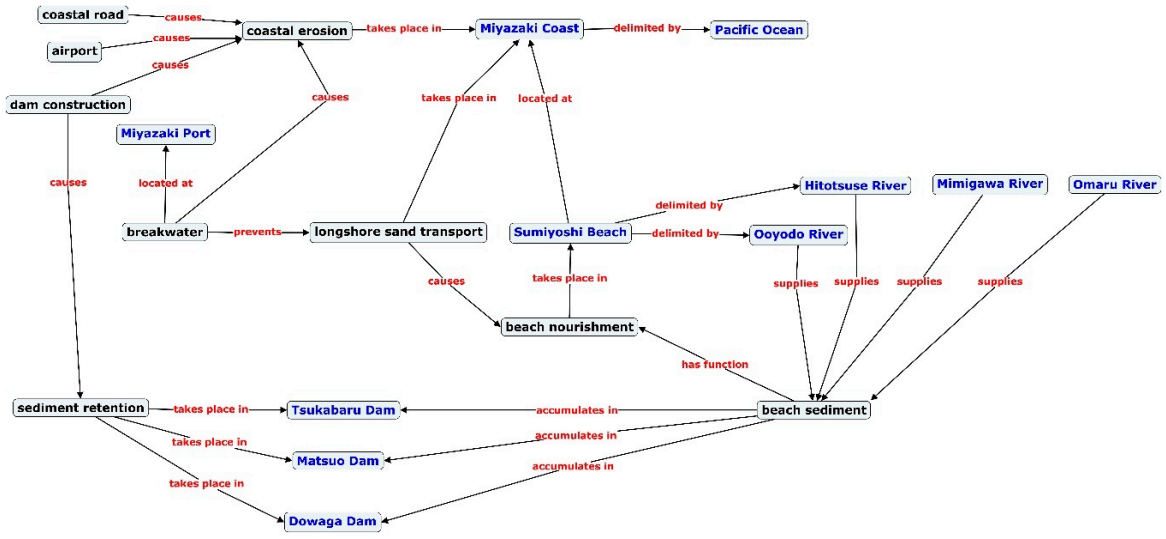

Figure 7. Semantic network of the terms associated with the *Omaru* and *Mimigawa* rivers.

4. Discussion

To extract knowledge for the semantic frames or 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 analyzing 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.* 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, as already done in Rojas-García and Cabezas-García (forthcoming) for other purposes.

The method for the extraction of terms closely associated with named rivers combining, on the one hand, the use of a count-based DSM, weighted by the log-likelihood association measure, and on the other hand, a selection procedure for terms based on four statistical criteria. Although this term selection procedure offered successful results to construct the semantic frames, Topic Modelling (Blei et al. 2003), a domain-specific dimension reduction technique for texts, will be also applied, and a comparison of both methods will be carried out.

The semantic frames presented in the previous section reflect that most terms related to named rives are multi-word terms (MWT) since specialized language units are mostly represented by such compound forms (Nakov 2013). 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. Furthermore, the profusion of MWTs underlines the importance of applying methods (automatic, semi-automatic, or manual) to recognize them for the knowledge representation of a specialized domain.

Finally, the conceptual structures also highlighted that coastal engineering texts attach great importance to the study of the processes that each named river triggers, the processes that affect a certain named river, and the crucial role that a named river plays in preventing coastal erosion. The effective acquisition of this specialized knowledge about named rivers is necessary in communicative situations, such as specialized translation of coastal engineering texts to appropriately render terms into another language (Faber 2012). The semantic networks that underlie the usage of named rivers provide this background knowledge and make the semantic and syntactic behavior of terms explicit by means of the description of conceptual relations and term combinations (Faber 2009).

Author Contributions: Conceptualization, J.R.-G.; methodology, J.R.-G.; validation, J.R.-G. and P.F.; formal analysis, J.R.-G.; investigation, J.R.-G.; resources, P.F.; data curation, J.R.-G. and P.F.; writing—original draft preparation, J.R.-G.; writing—review and editing, P.F.; visualization, J.R.-G.; supervision, P.F.; project administration, P.F.; funding acquisition, J.R.-G. and P.F.

Funding: 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 Juan Rojas-Garcia.

Conflicts of Interest: The authors declare no conflict of interest.

References

Alrabia, Maha, Nawal Alhelewh, AbdulMalik Al-Salman, and Eric Atwell. 2014. An Empirical Study on the Holy Quran Based on A Large Classical Arabic Corpus. *International Journal of Computational Linguistics* 5: 1–13.

Ars, Fatemeh, Jon Willits, and Michael Jones. 2016. Comparing Predictive and Co-occurrence Based Models of Lexical Semantics Trained on Child-directed Speech. Paper presented at 38th Annual Conference of the Cognitive Science Society, CogSci, Austin, TX, USA, August 10–13; pp. 1092–97.

Baroni, Marco, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. Paper presented at 52nd Annual Meeting of the Association for Computational Linguistics, ACL, Baltimore, MD, USA, June 22–27; vol. 1, pp. 238–47.

Bernier-Colborne, Gabriel, and Patrick Drouin. 2016. Evaluation of distributional semantic models: A holistic approach. Paper presented at 5th International Workshop on Computational Terminology, CompuTerm, Osaka, Japan, December 12; pp. 52–61.

Bertels, Ann, and Dirk Speelman. 2014. Clustering for semantic purposes: Exploration of semantic similarity in a technical corpus. *Terminology* 20: 279–303.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3: 993–1022.

Bullinaria, John A., and Joseph P. Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods* 39: 510–26. [CrossRef] [PubMed]

Cabezas-García, Melania, and Pamela Faber. 2018. Phraseology in specialized resources: An approach to complex nominals. *Lexicography* 5: 55–83. [CrossRef]

Evert, Stefan. 2008. Corpora and Collocations. In *Corpus Linguistics. An International Handbook*. Edited by Anke Lüdeling and Merja Kytö. Berlin: Mouton de Gruyter, chp. 58.

Faber, Pamela. 2009. The cognitive shift in terminology and specialized translation. *MonTI. Monografías de Traducción e Interpretación* 1: 107–34. [CrossRef]

Faber, Pamela. 2011. The Dynamics of Specialized Knowledge Representation: Simulational Reconstruction or the Perception action Interface. *Terminology* 17: 9–29.

Faber, Pamela, ed. 2012. *A Cognitive Linguistics View of Terminology and Specialized Language*. Berlin and Boston: De Gruyter Mouton.

Faber, Pamela, Pilar León-Araúz, and Juan Antonio Prieto. 2009. Semantic Relations, Dynamicity, and Terminological Knowledge Bases. *Current Issues in Language Studies* 1: 1–23.

Gries, Stefan, and Anatol Stefanowitsch. 2010. Cluster analysis and the identification of collexeme classes. In *Empirical and Experimental Methods in Cognitive/Functional Research*. Edited by Sally Rice and John Newman. Stanford: CSLI, pp. 73–90.

Jurafsky, Daniel, and James Martin. 2017. Vector Semantics. In *Speech and Language Processing*. Unpublished Draft of August 28.

Kiela, Douwe, and Stephen Clark. 2014. A Systematic Study of Semantic Vector Space Model Parameters. Paper presented at 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC), EACL, Gothenburg, Sweden, April 26–30; pp. 21–30.

Kilgarriff, Adam, Pavel Rychlý, Pavel Smrz, and David Tugwell. 2004. The Sketch Engine. Paper presented at 11th EURALEX International Congress, Lorient, France, July 6–10; pp. 105–15.

Lapesa, Gabriella, Stefan Evert, and Sabine Schulte im Walde. 2014. Contrasting Syntagmatic and Paradigmatic Relations: Insights from Distributional Semantic Models. Paper presented at 3rd Joint Conference on Lexical and Computational Semantics, SEM'2014, Dublin, Ireland, August 23–24; pp. 160–70.

León-Araúz, Pilar, Arianne Reimerink, and Pamela Faber. 2013. Multidimensional and Multimodal Information in EcoLexicon. In *Computational Linguistics*. Edited by Adam Przepiórkowski, Maciej Piasecki, Krzysztof Jassem and Piotr Fuglewicz. Berlin: Springer, pp. 143–61.

León-Araúz, Pilar, Antonio San Martín, and Pamela Faber. 2016. Pattern-based Word Sketches for the Extraction of Semantic Relations. Paper presented at 5th International Workshop on Computational Terminology, CompuTerm, Osaka, Japan, December 12; pp. 73–82.

León-Araúz, Pilar, Antonio San Martín, and Arianne Reimerink. 2018. The EcoLexicon English corpus as an open corpus in Sketch Engine. Paper presented at 18th EURALEX International Congress, Ljubljana, July 17–21; pp. 893–901.

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

Meyer, Ingrid. 2001. Extracting knowledge-rich contexts for terminography: A conceptual and methodological framework. In *Recent Advances in Computational Terminology*. Edited by Didier Bourigault, Chistian Jacquemin and Marie-Claude L'Homme. Amsterdam and Philadelphia: John Benjamins, pp. 279–302.

Meyer, Ingrid, and Kristen Mackintosh. 1996. Refining the terminographer's concept-analysis methods: How can phraseology help? *Terminology* 3: 1–26. [CrossRef]

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. Paper presented at International Conference on Learning Representations, ICLR, Scottsdale, AZ, USA, May 2–4.

Miller, George, and Walter Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes* 6: 1–28. [CrossRef]

Moisl, Hermann. 2015. *Cluster Analysis for Corpus Linguistics*. Berlin: De Gruyter Mouton.

Nakov, Preslav. 2013. On the interpretation of noun compounds: Syntax, semantics, and entailment. *Natural Language Engineering* 19: 291–330. [CrossRef]

Pantel, Patrick, and Dekang Lin. 2002. Discovering Word Senses from Text. Paper presented at ACM Conference on Knowledge Discovery and Data Mining, KDD-02, Edmonton, AB, Canada, July 23–26; pp. 613–19.

Rohde, Douglas, Laura Gonnerman, and David Plaut. 2006. An Improved Model of Semantic Similarity Based on Lexical Co-Occurrence. *Communications of the ACM* 8: 627–33.

Rojas-García, Juan, and Melania Cabezas-García. forthcoming. *Use of Knowledge Patterns for the Evaluation of Semiautomatically-Induced Semantic Clusters*. Serie Forum für Fachsprachen-Forschung; Berlin: Frank & Timme.

Sager, Juan C., David Dungworth, and Peter F. McDonald. 1980. *English Special Languages. Principles and Practice in Science and Technology*. Wiesbaden: Brandstetter Verlag.

Sahlgren, Magnus, and Alessandro Lenci. 2016. The Effects of Data Size and Frequency Range on Distributional Semantic Models. Paper presented at 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, November 1–5; pp. 975–80.

Shutova, Ekaterina, Lin Sun, and Anna Korhonen. 2010. Metaphor identification using verb and noun clustering. Paper presented at 23rd International Conference on Computational Linguistics, COLING, Beijing, China, August 23–27; vol. 2, pp. 1002–10.

Suzuki, Ryota, and Hidetoshi Shimodaira. 2004. An application of multiscale bootstrap resampling to hierarchical clustering of microarray data: How accurate are these clusters? Paper presented at Fifteenth International Conference on Genome Informatics, GIW2004, Yokohama, Japan, December 13–15.

Suzuki, Ryota, and Hidetoshi Shimodaira. 2006. Pvclust: An R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* 22: 1540–42. [CrossRef] [PubMed]