Communities of Related Terms in a Karst Terminology Co-occurrence Network

Dragana Miljkovic[1], Jan Kralj[1], Uroš Stepišnik[2], Senja Pollak[1,3]

[1] Jožef Stefan Institute, Ljubljana, Slovenia
[2] University of Ljubljana, Ljubljana, Slovenia
[3] University of Edinburgh, UK
E-mail: dragana.miljkovic@ijs.si, jan.kralj@ijs.si, uros.stepisnik@gmail.com, senja.pollak@ijs.si

Abstract

Karst science is an attractive field of interdisciplinary research with rich terminology. This study was performed as part of a project aiming at developing novel approaches to terminology extraction and visualization, in line with the understanding of knowledge, as represented in texts, as conceptually dynamic and linguistically varied. The aim of this paper is to investigate how powerful graph-based methods can be used for visualizing and analysing domain terminology. In order to detect communities in karst terminology, we analyse the frequently co-occurring karst terms in a scientific corpus of karstologic literature. The most frequent co-occurrence pairs, which included ten or more co-occurrences within the whole corpus, are delivered as input to the Louvain community detection algorithm and visualized as a domain graph. The resulting data was evaluated by domain experts who found that the detected term groups are meaningful and correspond to different types of karst phenomena. The results are further discussed in relation to more standard topic modelling approaches, using Latent Dirichlet Allocation and Non-negative Matrix Factorization algorithms.

Keywords: karstology; co-occurrence network; community detection algorithm; network visualization; topic modelling

1. Introduction

Karst science, or karstology, is a well-researched discipline with rich terminology, consisting of many expressions referring to regionally specific phenomena. Contemporary research of the topography that is referred to as a 'karst geomorphologic system' or simply 'karst' includes numerous scientific disciplines that study the karst environments worldwide; however, the earliest research on karst primarily regards Classical Karst, which is located in western Slovenia. Consequently, karstologists use many local Slovenian scientific terms and toponyms for typical geomorphological karst structures not only when writing in Slovene, but also in English and other languages. In this paper, we focus on karts texts in English.

This study was undertaken as part of the TermFrame project[1], which is based on contemporary findings in the field of terminology and cognitive linguistics, and aims to

[1] TermFrame project web site: http://termframe.ff.uni-lj.si/

develop novel methods that can be utilized in the field of terminology research. The focus of these novel methods is on corpus-based approaches to extraction and visualization of terminological knowledge, including text and graph mining and advanced data representation techniques.

Recent attempts in terminological science understand knowledge, as represented in texts, as conceptually dynamic and linguistically varied (Cabré, 1999; Temmerman, 2000; Kageura, 2002). Research advances in cognition have contributed to the Frame-Based Terminology (Faber, 2012; Faber, et al., 2006), which focuses on representing dynamic knowledge and investigating cultural elements in cognitive structures (Rodríguez Redondo, 2004; Grygiel, 2017), while projects such as EcoLexicon[2] attempt to visually represent concept networks. While a limited number of studies have used graph-based approaches in the fields of terminology and lexicography (Meyer & Eppinger, 2018; Krek et al., 2017) and for language comparison (Škrlj & Pollak, 2019), we believe that these methods are still to be fully explored, as they present the potential for novel research of specialized knowledge, as well as for new possibilities of knowledge representation that can be inspiring to contemporary lexicography. We believe that the graph-based method for exploring term co-occurrences can contribute to the needs of frame-based terminology, aiming at facilitating user knowledge acquisition through different types of multimodal and contextualized information (Gil-Berrozpe et al., 2017). This type of graph-based tool also has potential for future data representation in the field of e-lexicography (Granger, 2012), where multimodal data and hybridization between different types of language resources (e.g., dictionaries, encyclopaedias, term banks, lexical databases, translation tools) are commonly observed.

The focus of the present work in the scope of the above-mentioned project is to apply graph-based methods to the terminology of karst research. This has motivated us to explore co-occurrences of the specific karstology terms and visualize the results. Another motivation for the visualization of results is that domain experts are often able to interpret information faster when viewing graphs as opposed to tables (Brewer et al., 2012). More generally, as evident by the rising field of digital humanities, digital content, tools, and methods are transforming the entire field of humanities, changing the paradigms of understanding, asking new research questions and creating new knowledge (Hughes et al., 2015; Hughes, 2012). The work complements the results in karst terminology research presented in Vintar et al. (2019), where frame-based annotation of karst definitions is introduced, and in Pollak et al. (2019), where the authors present the results of term, definition, and triplet extraction from karst literature.

This paper is structured as follows: after presenting the background technologies and related work in Section 2, Section 3 introduces our method, which is based on

[2] http://ecolexicon.ugr.es/en/index.htm

community detection of terms extracted from a karstology corpus and their visualization in the form of a network; along with Section 4, the two sections represent the main contribution of the paper. In Section 5, we discuss the results in relation to a more standard topic modelling methods approach, and we conclude this paper in Section 6.

2. Background technologies and related work

This section presents a brief overview of the state-of-the-art of the fields related to our study methods, including co-occurrence and visualization, community detection algorithms and topic modelling.

2.1 Co-occurrence approach and visualization

Scientific literature in different fields can be explored through a search for the co-occurrences of domain-specific terms and their frequencies. A co-occurrence of two terms means that the terms coexist in the text within a certain window. The idea behind detecting co-occurrences of terms is that closely related terms will appear together more frequently. Moreover, co-occurrences can reveal hidden patterns and interesting features in the texts that are being analysed. For example, the co-occurrence analysis might detect spam messages (Krestel & Chen, 2008) or find meaningful knowledge from biological literature in a systematic and automated way (Al-Aamri et al., 2017). Co-occurrence is also used widely in text classification (Figueiredo et al., 2011) and categorization (Luo & Zincir-Heywood, 2004).

There is a difference between first-order and second-order co-occurrence approaches. For the first-order co-occurrence, one would simply count how many occurrences of one token there are within a specified distance of the particular occurrence of another token and build a vector presentation of the results. A second-order co-occurrence vector would represent some aggregation over the token representations, and in the simplest case this is a sum (Maldonado & Emms, 2012).

Representation of co-occurrence pairs in the form of a network is a common way to aid the domain experts with exploration of research results. Such representations can be used for various purposes, such as word sense disambiguation, which represents a challenge in natural language processing field (Duque et al., 2018). Li et al. (2018) report the discovery of new information in the biomedical domain based on the analysis of the structural characteristics of the co-occurrence network. Additionally, co-occurrence networks are increasingly used when analysing users' behaviour on social media (Correia et al., 2016).

In the field of lexicography, co-occurrence networks have been used with the aim of building a new Slovene thesaurus from data available in a comprehensive English–Slovene dictionary (Krek et al., 2017).

2.2 Community detection algorithms

When co-occurrence networks become too large and complex, their visual inspection becomes difficult. One way to explore complex networks more easily is to use community detection algorithms.

Community detection algorithms can be split into several classes based on the underlying idea that guides the algorithms. It must be noted that a strict split between the different methods is impossible, as these methods are not developed in isolation. For example, many methods that are not strictly classified as modularity-based algorithms still use the concept of modularity in one of their steps.

Divisive algorithms are algorithms that find the community structure of a network by iteratively removing edges from the network. The most widely used algorithm among divisive algorithms is the Girvan Newman algorithm (Girvan & Newman, 2002), which removes the network edges with the largest centrality measure. The reasoning behind this is that edges which are more central to a graph are the edges most likely to cross communities. An alternative algorithm is the Radicchi algorithm, which calculates the edge-clustering coefficient of edges in order to determine which edges must be removed. Here, the reasoning is that edges between communities belong to fewer cycles than edges within communities.

Modularity-based algorithms form the majority of community detection algorithms. While, as mentioned above, the concept of modularity (Newman & Girvan, 2004) is used in almost all algorithms to an extent (especially when attempting to determine the best clustering from a hierarchical clustering of nodes), the algorithms in this class use modularity more centrally than other algorithms. The most prominent modularity-based methods are the Louvain algorithm (Blondel et al., 2008) and the Newman greedy algorithm (Newman & Girvan, 2004). Other methods include variations of the greedy algorithm (Wakita & Tsurumi, 2007), simulated annealing (Guimerà & Amaral, 2005), spectral optimization of modularity via a modularity matrix (Newman, 2006a; Newman, 2006b) or via the graph adjacency matrix (White & Smyth, 2005), and deterministic optimization approaches (Duch & Arenas, 2005).

Spectral algorithms find communities in networks by analysing the eigenvectors of matrices derived from the network. The community structure is extracted either from the eigenvectors of the Laplacian matrix of the network (Donetti & Muñoz, 2004) or from the stochastic matrix of the network (Capocci et al., 2005). In both cases, the idea behind the algorithms is that eigenvectors extracted from the network will have similar values on indices that belong to network vertices in the same community. First, a computation of several eigenvectors belonging to the largest eigenvalues is performed. The resulting eigenvectors form a set of coordinates of points, each belonging to one network vertex, with clustering of these points corresponding to community detection of network vertices.

Another important community detection algorithm is the InfoMap algorithm (Rosvall et al., 2009). This is based on the idea of minimal description length of the walks performed by a random walker traversing the network. The communities in InfoMap are determined by constructing so-called codebooks, which are used to describe walks on the network – corresponding to communities in the network, codebooks yield on average shorter average descriptions of walks. Finally, in the most recent rapid development of network embedding algorithms, some researchers have begun using embedding-based methods for network community detection (Li et al., 2018).

2.3 Topic modelling

In this section, we cover topic modelling, i.e. methods used for discovering various topics that appear in a collection of documents. Topic modelling methods are well-established in the field of text modelling, and can be considered as alternative approaches to co-occurrence community detection. Methods for topic modelling can rely on linear algebra, such as Vector Space Model (VSM) (Becker & Kuropka, 2003) or Matrix Factorization (NMF) (Paatero & Tapper, 1994), while others are based upon statistical distributions, for example Latent Dirichlet Allocation (LDA) (Blei et al., 2003). When using both NMF and LDA for topic modelling, two matrices are constructed from the document-term matrix: the document-topic and topic-term matrices. The topics are derived from the contents of the documents, and the topic-document matrix describes data clusters of related documents. LDA usually performs well when it comes to identifying coherent topics, whereas NMF provides incoherent ones (Stevens et al., 2012). While VSM is based on a similar principle as NMF, it has significant limitations when processing long documents as they have poor similarity values. Because the corpus analysed for the purposes of this paper includes both short and long documents (doctoral dissertations, dictionaries, etc.), this specific method was excluded from consideration.

The aim of this paper is to analyse the communities in karst terminology by analysing the co-occurrence network of frequently co-occurring karst terms in the scientific corpus of karst literature. We defined a co-occurrence of terms as their coexistence in the same sentence, while in order to qualify as frequently co-occurring, a term pair had to occur at least ten times over the span of the entire corpus. We decided to start inspecting karst corpus gathered for the purpose of the TermFrame project with basic first-order co-occurrence vectors and present the results of co-occurrence terms in the form of community network, as it is easily comprehended by domain experts. For our research, we used three leading algorithms in the community detection field: Label propagation, Louvain, and InfoMap. The InfoMap and Label propagation algorithms did not yield meaningful results: both identified one large community and several singletons. For this reason, the Methodology, Results, and Discussion sections all focus exclusively on the results obtained using the Louvain algorithm. We also discuss the results from the community detection experiment in relation to two topic modelling approaches, LDA

and NMF, while the exploration of second-order co-occurrence approaches will be explored in future work.

3. Methodology

First, we tokenized and lemmatized our collection of scientific literature and the corresponding term list. Next, first order co-occurrences of pre-specified terms were identified within the corpus. After this, the Louvain community detection algorithm was used to find the communities of co-occurrence pairs. The schematic of the methodology used in this study is shown in Figure 1, with each step further explained below.

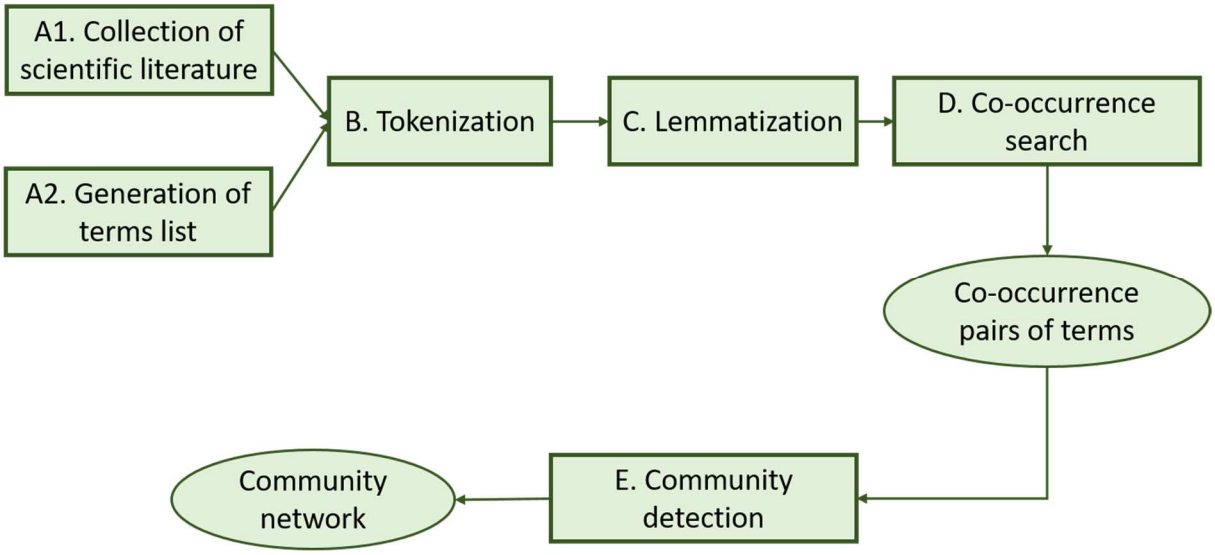

Figure 1: The schematic of the methodology.

A1. Collection of scientific literature represents the compilation of 25 scientific karstology texts, including papers, doctoral dissertations, and the glossary of cave and karst terminology. This corpus was compiled as part of the TermFrame project and is an extended version of earlier work (Vintar & Grčić Simeunović, 2016).[3]

A2. Generation of terms list was performed as a two-phase process. First, relevant terms were automatically extracted from the TermFrame corpus using the LUIZ-CF term extractor (Pollak et al., 2012), which is a variant of LUIZ (Vintar, 2010) refined with scoring and ranking functions. The terms were validated by the domain expert and were used to compile a term list along with the previously acquired terms from the QUIKK termbase[4]. This process of term extraction and evaluation is presented in more detail in Pollak et al. (2019).

[3] We used the corpus version v1.0.

[4] http://islovar.ff.uni-lj.si/karst

B. Tokenization was performed using the NLTK Tokenizer for Python.

C. Lemmatization was performed using the Lemmagen tool (Juršič et al., 2010).

D. Co-occurrence search was performed automatically by the Python script, which stores in a separate file the co-occurring term pairs and the number of their co-occurrences in the whole TermFrame corpus.

E. Community detection was performed using the Louvain algorithm (Blondel et al., 2008), which works by decreasing the modularity of the network, a function that measures the density of links inside communities compared to links between communities. The modularity of a network is defined as:

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

where A_{ij} denotes the weight of the edge between nodes i and j (in our case, the number of co-occurrences), k_i denotes the degree (sum of all adjacent edge weights) of node i, and m denotes the total sum of weights in the network. The term c_i denotes the community to which node i is assigned, meaning the sum above runs over all pairs of i,j where i and j belong to the same community.

4. Results and discussion

For the purposes of this research, we compiled a list of 452 karst terms drawing from a corpus of karstology texts which contained 108,769 sentences in total. Both the list and the literature were tokenized and lemmatized prior to the co-occurrence search, which yielded a list of 10,990 unique co-occurrence pairs using 426 unique lemmatized terms, as well as the data regarding co-occurrence frequency.

The initially obtained co-occurrence pairs would result in a complex network that would be difficult to represent in a comprehensible manner. To simplify the visualization, co-occurrence pairs with frequencies of ten or less were removed from the subsequent analysis. This left us with 1,247 co-occurrence pairs (see Table 1).

	Initial co-occurrence list	Filtered co-occurrence list
Number of co-occurrence pairs	10,990	1,247
Number of unique terms	426	309

Table 1: The summary of the initially obtained co-occurrence list and the filtered version, which contains only the co-occurrence pairs with frequencies of 10 or more.

The 20 most frequent co-occurrence pairs extracted from the karst corpus are listed in Table 2.

ID	Term 1	Term 2	Frequency of appearing	ID	Term 1	Term 2	Frequency of appearing
1	cave	karst	1688	11	limestone	dolomite	368
2	cave	passage	1482	12	cave	karren	349
3	cave	limestone	739	13	solution	karren	319
4	cave	spring	735	14	karren	limestone	311
5	cave	speleothem	664	15	cave	pit	288
6	cave system	cave	597	16	limestone	marble	282
7	cave	gypsum	512	17	karst	spring	270
8	cave	calcite	468	18	karst	term	261
9	karst	limestone	464	19	cave	canyon	261
10	calcite crust	cave	381	20	karst	doline	259

Table 2: The list of common co-occurrence pairs extracted from the karst corpus sorted from most to least frequent.

The filtered co-occurrence pairs served as input for the Louvain algorithm for community detection. Starting with each node in its own community, the algorithm iteratively works in two stages. In the first stage, it searches for the optimum pairs or groups of communities to merge into a larger community and thus increase the modularity of the partition. In the second stage, the algorithm reduces the network to a coarser network based on the discovered communities. The two-stage procedure is then repeated until no increases in modularity can be made. This results in a hierarchy of network node clusters, which can then be cut at any level to produce a clustering of the network nodes. In our case, the algorithm resulted in a three-layer hierarchy. The top level consisted of only two communities and the bottom level of single-node communities. The middle layer was the only layer containing non-trivial information about the structure of the co-occurrence network, and it was therefore subject to further analysis.

The middle layer of the hierarchy, discovered by the Louvain algorithm, consisted of eight communities. Next, we visualized the network using the Barnes-Hut approximation of the force-directed layout to calculate optimal node positions (Jacomy et al., 2014). The discovered communities were then displayed on the network visualization by colouring nodes corresponding to the communities they belong to (see Figure 2).

The karst domain experts analysed the resulting network and found the network visualization particularly interesting, as the communities (listed below) were found to correspond to different types of karst phenomena.

- Community 0: Exokarst landforms ('kamenitza', 'grike', 'stone forest'), which are the result of direct effects of dissolution of bedrock exposed on the surface;

- Community 1: Subsurface landforms, speleogenetic features, and cave environments (e.g. 'passage', 'flowstone deposit', 'cave system'). This community comprises all types of underground voids typical for karst environments regardless of their morphogenesis, including characteristic mechanical and chemical fills within.

- Community 2: Surface karst landforms and environments (e.g. 'uvala', 'doline', 'karst terrain') which are a product of surface and subsurface karst processes, materialising as relief forms or terrain types.

- Community 3: Karst hydrologic processes, environments, and methods (e.g. 'karst recharge', groundwater basin', 'tracer test') incorporate all karst aquifer types, the processes within them, and methods concerning their research.

- Community 4: Karst geology representing terms related to karst lithology (e.g. dolomite), minerals ('calcite') and processes affecting them (e.g. 'dissolution')

- Community 5: Includes only two terms (karrenfield, phreatic-cave), which is not enough to define the topic field.

- Community 6 includes only two terms ('turbulent flow', 'laminar flow'), which is not enough to define the topic field.

- Community 7 includes only two terms ('vadose zone', 'phreatic zone'), which is not enough to define the topic field.

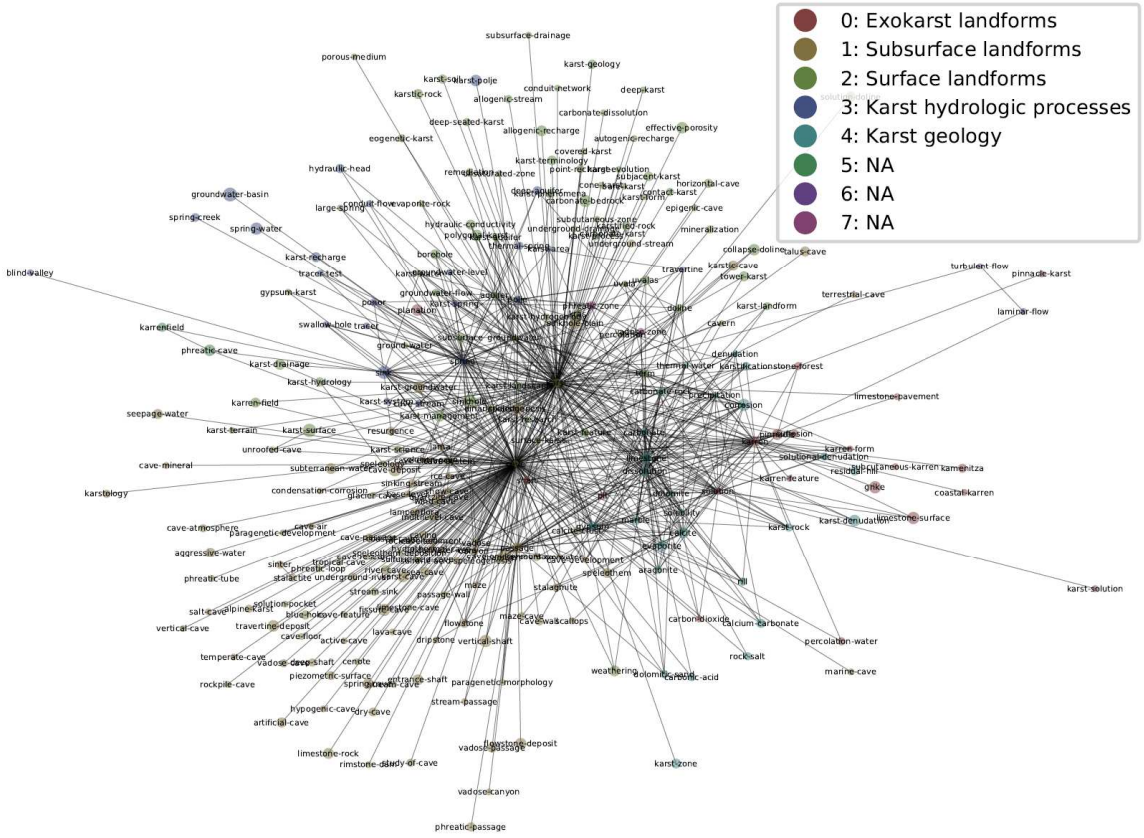

Figure 2: The co-occurrence network, visualized using a force-directed layout, showing the communities discovered within the network. The colours of the nodes correspond to the communities the nodes belong to.

5. Topic modelling experiments

As graph-based modelling is a relatively novel field for harvesting knowledge from specialized corpora, this section discusses our results with respect to more standard topic modelling approaches. For the purpose of this research, we used LDA and NMF algorithms, implemented within a Scikit-learn Python module. The algorithms searched through the complete corpus of 25 documents (described above) containing 108 769 lemmatized sentences, presenting the domain expert with the 25 most important words for each topic. The domain expert subsequently evaluated whether the derived topic words adequately represent specific subfields of karstology. In Table 3, we list the topics and the topic words identified by the NMF and LDA algorithms, which were estimated as meaningful groups by the domain expert. To enable further comparison of results with the community detection experiment, the number of topics was set to eight for both algorithms.

NMF	Topic 0: SPELEOLOGY	cave passage entrance long know study world km large deep map exploration bat sediment mammoth example explore stream important contain river site animal speleothem state
	Topic 1: KARST HYDROLOGY	water flow spring table level aquifer zone high groundwater discharge surface underground stream sea conduit phreatic supply resource fresh mix air rise sink temperature time
	Topic 5: KARST GEOMORPHOLOGY	rock form figure surface limestone large develop small carbonate karren passage process area 10 soil solution high occur dissolution doline lower feature cover sediment deposit
	Topic 6: SPELEOBIOLOGY	species family subterranean know troglobitic habitat include genera number genus group population troglomorphic bat fauna large occur troglobite terrestrial aquatic marine represent small order environment
	Topic 7: GENERAL METHODOLOGY (KARST)	use method data term model technique land date tracer place study time site widely approach human dye analysis test trace map measure determine source work
LDA	Topic 0: SPELEOLOGY	cave sediment passage type channel wall 20 place contain small like 12 width speleothem vertical significant 100 2001 possible figure direction rillenkarren floor stream scale
	Topic 2: KARST GEOLOGY	rock large limestone carbonate cover deposit upper surface gypsum forest dissolution area stone protect calcite earth line layer bed joint various material analysis salt fracture
	Topic 5: KARST HYDROLOGY	water flow spring zone soil deep high aquifer karst surface occur groundwater slope natural condition table value depression low erosion increase result point temperature climate

Table 3: Topic modelling results with Non-negative Matrix Factorization (NMF) and Latent Dirichlet distribution (LDA) applied to karst literature

From a karstologic point of view, the following topics extracted by means of the NMF method describe various aspect of karstology, i.e. different scientific fields regarding karst research:

- Topic 0: Speleology incorporates topic words that are directly referring to cave processes, cave-related landforms, or toponyms regarding to research of caves (i.e. speleology).
- Topic 1: Karst hydrology topic words comprise a variety of terms describing karst aquifers and their study.
- Topic 5: Karst geomorphology topic words correspond to a variety of surface landforms and processes, as well as words labelling their properties.
- Topic 6: Speleobiology topic words are related to cave biota and habitats.
- Topic 7: General karst methodology topic words incorporate a combination of various terms describing research methods from different karst research fields.

LDA identified only three topic groups meaningful to the domain expert, compared to the five identified by NMF:

- Topic 0: Speleology (see NMF Topic 0).
- Topic 2: Karst geology words regarding karst rocks, minerals, and processes concerning them.
- Topic 5: Karst hydrology (see NMF Topic 1).

NMF and community detection experiments have some overlaps in results, such as karst hydrologic processes and karst surface landforms and environments, as well as a partial topic overlap with terms related to speleology.

The results of our proposed community detection methodology have identified several specific topics as evaluated by the expert; however, it can be hard to determine to which extent this is to be attributed to term pre-selection, the community detection algorithm, or to the visualization of results. A detailed study of the role of each component is beyond the scope of this paper, but we believe that graph-based methods coupled with visualization offer great opportunities for investigating terminology as dynamic systems.

An overview of the number of meaningful communities identified by the proposed community detection approach and topic modelling methods (NMF and LDA) is presented in Table 4. All of the topics listed in this paper were manually evaluated by a domain expert. Community detection differs from the topic modelling approaches in that it takes pre-specified terms as input, while topic modelling approaches take as

input all words in the corpus documents. For this reason, a deeper quantitative comparison between these approaches is not feasible.

Number of meaningful topics		
Community detection algorithm	Topic modelling (LDA)	Topic modelling (NMF)
5	3	5

Table 4: Quantitative overview of the discovered topics with topic modelling and graph-based methods.

6. Conclusions and future work

In this work, we used a list of terms extracted from karst scientific literature and then performed a network analysis of karst terminology, wherein the network was constructed from co-occurring karst terms. The community detection algorithms described in this paper grouped specialized terms into semantically related topics, which were also visually presented as coloured nodes in the graphs. In addition, we approached the same corpus from the viewpoint of more standard topic modelling techniques, using LDA and NMF as our main tools.

In future work we plan to include the exploration of second-order co-occurrences, embedding-based topic modelling, and combining graph-based term and community detection methods. In addition, we consider performing a systematic comparison of graph-based community detection and topic modelling approaches, as well as evaluating if term extraction can contribute to these approaches.

Furthermore, we plan to use network representation in the form of triplets {subject, predicate, object}, which can also be a source of identifying novel semantic relations. Within the scope of the TermFrame project, a multi-layer semantic annotation has been performed and the most frequent conceptual frames for specific semantic categories explored. By combining information from manual annotations and the proposed network-based techniques, new knowledge about conceptual frames, semantic relations, and topics could be observed. The potential of graph-based topological analysis lies also in its power to explore structural information, which could reveal potential language and culture-driven differences if, for example, applied to larger comparable corpora of karst texts in different languages.

7. Acknowledgements

This work was financed by Slovenian Research Agency grants J6-9372 (Terminology and Knowledge Frames across Languages - TermFrame) and P2-0103 (Knowledge Technologies). This paper is supported by European Union's Horizon 2020 research

and innovation programme under grant agreement No 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). The results of this work reflects only the authors' views, and the European Commission is not responsible for any use that may be made of the information it contains.

8. References

Al-Aamri, A., Taha, K., Al-Hammadi, Y., Maalouf, M., & Homouz, D. (2017). Constructing Genetic Networks using Biomedical Literature and Rare Event Classification. *Scientific Reports, 7*(1), pp. 2045-2322.

Becker, J., & Kuropka, D. (2003). Topic-based vector space model. *Proceedings of the 6th International Conference on Business Information Systems.* Colorado Springs, USA.

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

Blondel, V., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, P10008.

Brewer, N. T., Gilkey, M. B., Lillie, S. E., Hesse, B. W., & Sheridan, S. L. (2012). Tables or Bar Graphs? Presenting Test Results in Electronic Medical Records. *Medical Decision Making, 32*(4), pp. 545-553.

Cabré, M. T. (1999). *Terminology: Theory, methods, applications.* Amsterdam; Philadelphia: J. Benjamins Publishing Company.

Capocci, A., Servedio, V. D., Caldarelli, G., & Colaiori, F. (2005). Detecting communities in large networks. *Physica A: Statistical Mechanics and its Applications, 352*(2), pp. 669-676.

Correia, R. B., Li, L., & Rocha, L. M. (2016). Monitoring potential drug interactions and reactions via network analysis of Instagram user timeliness. *Pacific Symposium on Biocomputing. 21.* Kohala Coast, Hawaii, USA: World Scientific, pp. 492-503.

Donetti, L., & Muñoz, M. A. (2004). Detecting network communities: a new systematic and efficient algorithm. *Journal of Statistical Mechanics: Theory and Experiment, 2004*(10), P10012.

Duch, J., & Arenas, A. (2005). Community detection in complex networks using extremal optimization. *Physical review E, 72*(2), pp. 027104.

Duque, A., Stevenson, M., Martinez-Romo, J., & Araujo, L. (2018). Co-occurrence graphs for word sense disambiguation in the biomedical domain. *Artificial Intelligence in Medicine, 87*, pp. 9-19.

Eronen, L., & Toivonnen, H. (2012). Biomine: predicting links between biological entities using network models of heterogeneous databases. *BMC Bioinformatics, 13*(1), pp. 119.

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

Faber, P., Montero Martínez, S., Castro Prieto, M. R., Senso Ruiz, J., Prieto Velasco, J. A., León Arauz, P. & Vega Expósito, M. (2006). Process Oriented Terminology Management in the Domain of Coastal Engineering. *Terminology, 12*(2), pp. 136.

Figueiredo, F., Rocha, L., Couto, T., Salles, T., André Gonçalves, M., & Meira Jr, W. (2011). Word co-occurrence features for text classification. *Information Systems, 36*(5), pp. 843-858.

Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences, 99*(12), pp. 7821-7826.

Grygiel, M. (2017). *Cognitive Approaches To Specialist Languages.* (M. Grygiel, Ed.) Cambridge Scholars Publishing.

Guimerà, R. & Amaral, L. A. (2005). Functional cartography of complex metabolic networks. *Nature, 433*(7028), pp. 895-900.

Hughes, L. M. (2012). Using ICT methods and tools in arts and humanities research. In L. M. Hughes (ed.) *Digital Collections: Use, Value and Impact.* London, UK: Facet Publishing, pp. 123-134.

Hughes, L., Constantopoulos, P., & Dallas, C. (2015). Digital Methods in the Humanities: Understanding and Describing their Use across the Disciplines. In S. Schreibman, R. Siemens, & J. Unsworth (eds.) *A New Companion to Digital Humanities.* John Wiley & Sons, pp. 150-170.

Jacomy, M., Venturini, T., Heymann, S., & Bastian, M. (2014). ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS One, 9*(6), e98679.

Juršič, M., Mozetič, I., Erjavec, T., & Lavrač, N. (2010). LemmaGen: multilingual lemmatisation with induced Ripple-Down rules. *Journal of Universal Computer Science, 16*(9), pp. 1190-1214.

Kageura, K. (2002). *The Dynamics of Terminology: A Descriptive Theory of Term Formation and Terminological Growth.* Amsterdam, Philadelphia: John Benjamins Publishing Company.

Krek, S., Laskowski, C., & Robnik-Šikonja, M. (2017). From Translation Equivalents to Synonyms: Creation of a Slovene Thesaurus Using word co-occurrence Network Analysis. In I. Kosem et al. (eds.) *Electronic lexicography in the 21st century: Proceedings of eLex 2017 conference.* Leiden, the Netherlands: Lexical Computing CZ s.r.o., Brno, Czech Republic, pp. 93-109.

Krestel, R., & Chen, L. (2008). Using co-occurrence of tags and resources to identify spammers. *ECML PKDD.* Antwerp: Springer, pp. 38-46.

Li, T., Bai, J., Yang, X., Liu, Q., & Chen, Y. (2018). Co-Occurrence Network of High-Frequency Words in the Bioinformatics Literature: Structural Characteristics and Evolution. *Applied Sciences, 8*, 1994.

Li, Y., Sha, C., Huang, X., & Zhang, Y. (2018). Community detection in attributed graphs: an embedding approach. *Thirty-Second AAAI Conference on Artificial Intelligence.* New Orleans, Louisiana, USA: AAAI .

Liu, Y., McInnes, B. T., Pedersen, T., Melton-Meaux, G., & Pakhomov, S. V. (2012). Semantic relatedness study using second order co-occurrence vectors computed from biomedical corpora, UMLS and WordNet. *2nd ACM SIGHIT International Health Informatics Symposium*. Miami, Florida, USA: ACM, pp. 363-372.

Luo, X., & Zincir-Heywood, A. N. (2004). Combining word based and word co-occurrence based sequence analysis for text categorization. *Machine Learning and Cybernetics.* Shanghai: IEEE, pp. 1580-1585.

Maldonado, A., & Emms, M. (2012). First-order and second-order context representations: geometrical considerations and performance in word-sense disambiguation and discrimination. *JADT 11th International Conference on the Statistical Analysis of Textual Data*. Liege, France, pp. 676-686.

Meyer, P., & Eppinger, M. (2018). fLexiCoGraph: Creating and Managing Curated Graph-Based Lexicographical Data. In J. Čibej, V. Gorjanc, I. Kosem, & S. Krek (eds.) *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts.* Ljubljana, Slovenia: Ljubljana University Press, Faculty of Arts, pp. 1017-1022.

Newman, M. E. (2006a). Finding community structure in networks using the eigenvectors of matrices. *Physical review E, 74*(3), pp. 036104.

Newman, M. E. (2006b). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences, 103*(23), pp. 8577-8582.

Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E, 69*(2), pp. 026113.

Paatero, P., & Tapper, U. (1994). Positive Matrix Factorization: A Non-Negative Factor Model with Optimal Utilization of Error Estimates of Data Values. *Environmetrics, 5*(2), pp. 111-126.

Pollak, S., Repar, A., Martinc, M., & Podpečan, V. (2019). Karst exploration: extracting terms and definitions from karst domain corpus. In I. Kosem et al. (eds.) *Proceedings of eLex 2019.* Sintra, Portugal, pp. 934-956.

Pollak, S., Vavpetič, A., Kranjc, J., Lavrač, N., & Vintar, Š. (2012). NLP workflow for on-line definition extraction from English and Slovene text corpora. *Proceedings of KONVENS 2012*. Vienna, Austria: ÖGAI, pp. 53-60.

Rodríguez Redondo, A. L. (2004). Aspects of cognitive linguistics and neurolinguistics: conceptual structure and category-specific semantic deficits. *Estudios ingleses de la Universidad Complutense, 12*, pp. 43-62.

Rosvall, M., Axelsson, D., & Bergstrom, C. T. (2009). The map equation 178.1. *The European Physical Journal Special Topics, 178*(1), pp. 13-23.

Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012). Exploring Topic Coherence over many models and many topics. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Jeju Island, Korea: Association for Computational Linguistics, pp. 952–961.

Škrlj, B., & Pollak, S. (2019). Language comparison via network topology**.** *Proceedings of the 7th International Conference on Statistical Language and Speech Processing*. LNCS 11816. Springer.

Temmerman, R. (2000). *Towards New Ways of Terminology Description: The Sociocognitive-approach.* Amsterdam, Philadelphia: John Benjamins Publishing Company.

Vintar, Š. (2010). Bilingual Term Recognition Revisited. The Bag-of-Equivalents Term Alignment Approach and Its Evaluation. *Terminology, 16*(2), pp. 141-158.

Vintar, Š., & Grčić Simeunović, L. (2016). Definition frames as language-dependent models of knowledge transfer. *Fachsprache: internationale Zeitschrift für Fachsprachenforschung, - didaktik und Terminologie, 39*(1-2), pp. 43-58.

Vintar, Š., Saksida, A., Stepišnik, U., & Vrtovec, K. (2019). Knowledge frames in karstology: the TermFrame approach to extract knowledge structures from definitions. In *Proceedings of eLex 2019,* Sintra, Portugal.

Wakita, K., & Tsurumi, T. (2007). Finding community structure in mega-scale social networks:[extended abstract]. *16th international conference on World Wide Web*. Banff, Alberta, Canada: ACM, pp. 1275-1276.

White, S., & Smyth, P. (2005). A spectral clustering approach to finding communities in graph. *SIAM International Conference on Data Mining. 5*. Newport Beach, California, USA: SIAM, pp. 76-84.