Modelling Specialized Knowledge With Conceptual Frames: The TermFrame Approach to a Structured Visual Domain Representation

Špela Vintar, Amanda Saksida, Katarina Vrtovec,

Uroš Stepišnik

Faculty of Arts, University of Ljubljana, Aškerčeva 2, SI – 1000 Ljubljana
E-mail: {spela.vintar, amanda.saksida, katarina.vrtovec, uros.stepisnik} @ff.uni-lj.si

Abstract

We describe an emerging knowledge base for karstology developed in line with the frame-based approach with data for three languages, English, Slovene and Croatian. An annotation framework was developed to identify the definition elements, semantic categories, relations and relation definitors in definitions of karst concepts extracted from specialized corpora. A multi-layered annotation was performed for sets of validated English and Slovene definitions. We present the distribution of semantic categories and typical definition frames for the most prominent semantic categories: surface and underground landforms, hydrological forms and geomes, for English and Slovene. The definition frames specify the typical properties of concepts we expect to be described, and in our case they were initialized by domain experts and then verified through corpus data. The structured domain representation resulting from the annotated corpus allows us to compare knowledge structures between languages, generate ideal definitions and experiment with domain visualisations, graphs and maps of geolocations.

Keywords: frame-based terminology; knowledge modelling; karstology; semantic annotation

1. Introduction

Domain terminologies are often thought of as structured and systematic networks of concepts which allow for efficient and unambiguous communication between experts. Traditional specialized dictionaries proved – through their alphabetic ordering alone – inadequate for representing concepts as abstract units of knowledge, but termbases in digital format can easily accommodate the concept-oriented approach and utilize the terminological entry as the tangible equivalent of the concept residing in the cognitive realm. Indeed, many online multilingual termbases such as IATE[1] or UMLS[2] embody this approach.

[1] https://iate.europa.eu/

[2] https://www.nlm.nih.gov/research/umls/

The frame-based approach to terminology (Faber et al., 2005; Faber, 2009; Faber et al., 2012) has provided a valuable new framework for representing specialized knowledge by combining linguistic information derived from specialized corpora with conceptual structures and by highlighting the fact that the cognitive frames underlying specialized communication are dynamic, context-, language- and culture-dependent (Leitchik & Shelov, 2007; Temmermann & Van Campenhoudt, 2014; Faber & Medina-Rull, 2017). Moreover, the concepts of a specialized domain should not be described in isolation but represented as nodes in an intricate knowledge network illustrating both generic and domain-specific relations between them. A widely known implementation of these principles is the EcoLexicon [3], a multilingual knowledge base for the environmental domain.

The TermFrame project adopts the frame-based rationale, but adapts and extends existing methodologies with the following goals in mind:

- To build a comprehensive structured knowledge base for the domain of karstology in three languages – English, Slovene and Croatian;
- To develop modes of knowledge representation which can be used by linguists, terminologists, experts and data scientists alike, and which adequately show language- and context-dependent differences between knowledge frames;
- To explore new methods of knowledge extraction from specialized texts, so that our results can be generalized and applied to new languages and domains.

This paper focuses on the semantic annotation framework and the resulting resources which can serve both as input for knowledge visualization and as training data for future knowledge extraction tools. It is structured as follows: Section 2 gives a brief overview of related work on terminological definitions and their semantic structure from the Frame Semantics point of view. Section 3 describes the resources built and used in TermFrame, including the tools for term and definition extraction. In Section 4 we give a detailed explanation of our annotation framework and provide examples of annotated definitions, followed by some quantitative data from the annotated corpora and an illustration of the resulting domain representation in Section 5.

2. Definitions and frames

The terminological definition is the most concentrated means of communicating expert knowledge which helps users understand the meaning of a specialized lexical unit (Seppälä & Ruttenberg, 2013: 19). Although its structure was originally defined by Aristotle, the textual reality shows that authors use varying definition styles (Svensen, 1993: 117; Roche et al., 2009), while several attempts have been made to devise a

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

typology of definitions (Blanchon, 1997; Seppälä, 2007; Diki-Kidiri, 2000; Madsen/Thomsen, 2008; Pollak, 2010).

Here we refrain from delving deeper into the definition types and the factors which may influence the author to use a certain defining style over another, although some understanding of this variety is needed for automatic definition extraction, as we show in Section 3. It should be stressed, however, that the choice of semantic elements used to delineate specialized meaning is not arbitrary, and the frame-based approach helps us discern predominant definition templates or frames, or even guide definition formation, as shown for example in San Martin & L'Homme (2014) and Duran-Muñoz (2016). The definition template is usually related to the semantic category of the concept and reflects its role in the domain-specific event.

In our own previous work (Vintar & Grčić Simeunović, 2017), a cross-language analysis of definition frames in karstology revealed interesting differences between English and Croatian. Karst as a core concept is defined in Croatian mostly through its geomorphological features and settings, while in English we found several instances where karst or its subtypes were defined as the geomorphologic or hydrologic functioning of the karst processes. The underlying cognitive frame is in this case clearly language-dependent.

3. TermFrame resources

For the purposes of our research we built three corpora, Slovene, English and Croatian. The corpora contain relevant contemporary works on karstology and are comparable in terms of the domain and text types included. The corpora comprise scientific texts (scientific papers, books, articles, doctoral and master's theses, glossaries and dictionaries) from the field of karstology, which in itself is an interdisciplinary domain partly overlapping with surface and subsurface geomorphology, geology, hydrology and other fields. Table 1 gives basic information about the corpus.

	English	Slovene	Croatian
Tokens	2,721,042	1,208,240	1,229,368
Words	2,195,982	987,801	969,735
Sentences	97,187	51,990	53,017
Documents	57	60	43

Table 1: The TermFrame corpora

Once the corpora were compiled we performed term and definition extraction and other knowledge mining steps described in Pollak et al. (2019). Definition candidates were extracted automatically with the pattern-based setup of ClowdFlows, which according

to previous research performs best (Pollak et al., 2012). At this time the tool yet has to be adapted to Croatian, hence the remainder of this paper reports results for English and Slovene only. Also, in the first stage definition extraction was performed on approximately half of the English corpus. The extracted sentences were manually validated to retain only contexts with valuable explanatory information about the karst concept. Given this relatively broad view many of our definitions do not necessarily comply with the traditional definition structure: in many cases the definiendum appears at the end of the sentence, the genus or hypernym may be missing, and several examples of extensional definitions were found. After validation the yield was 215 and 259 terms for English and Slovene, respectively.

The semantic annotation of definitions was performed in WebAnno, an open source server-based tool which allows users to specify the annotation layers, attributes and tagsets, and perform annotation, curation and monitoring (De Castilho et al., 2014). In our workflow, each definition was annotated by two persons (linguists), then curation was performed by a domain expert. Regular meetings of all annotators and curators were organized to discuss ambiguities and consolidate the annotation procedure.

4. Annotating definitions in TermFrame

4.1 The annotation framework

The development of the annotation framework is an essential step in domain modelling as it attempts to produce a mapping between the cognitive level representing expert knowledge, the textual reality describing this knowledge, and a formal level with structures, categories and relations. The primary purpose of such a mapping is to allow for an accurate and functional representation of the domain. At the same time, a secondary purpose is to provide insight into linguistic features which may be used for automatic knowledge extraction not just in the domain of choice, but potentially also in other domains. Our project team consists of linguists, a cognitive scientist, a karstologist and several experts in NLP, and has developed a framework able to accommodate both these purposes.

The annotation consists of five layers:

1. Definition element. This layer identifies the following elements of the definition: DEFINIENDUM (the term which is being defined), DEFINITOR (the defining phrase of the definition, usually a verbal phrase), GENUS (the hypernym or superordinate term), and SPECIES (the hyponym or subordinate term; relevant in extensional definitions). Though not annotated, the IS_A relation is implicit between DEFINIENDUM and GENUS (sandstone IS_A rock), and SPECIES and DEFINIENDUM (doline IS_A karst depression).

2. Semantic category. This is a hierarchical framework which used the EcoLexicon conceptual hierarchy as a starting point, but was adapted to karstology in collaboration with domain experts. It uses five top-level categories (for details see Figure 1). The concepts represented by the categories were modelled according to the basic karstologic approach (Ford & Williams, 2007; Jennings, 1985) corresponding to surface and subsurface karst landforms (Landform) and a number of related processes (Process). Other categories included typical karst environments (Geome), materials, processes and landforms closely connected to karst environments (Entity/Element/Property) and typical methods and tools in karstology (Instrument/Method).

3. Relation. We use a set of 16 relations, each of which marks a specific property or feature of the definiendum. Relations may span over several words or phrases and do not necessarily overlap with the two previous layers. Thus, in the example sentence in Figure 2 the relation COMPOSED_OF is expressed in the text by *of freshly formed gypsum*.
The following relations were defined by domain experts according to the geomorphologic analytical approach (Pavlopoulos et al., 2009) considering spatial distribution (HAS_LOCATION; HAS_POSITION), morphography (HAS_FORM; CONTAINS), morphometry (HAS_SIZE), morphostructure (OCCURS_IN_MEDIUM; COMPOSED_OF), morphogenesis (HAS_CAUSE), morphodynamics (AFFECTS; HAS_RESULT; HAS_FUNCTION), and morphochronology (OCCURS_IN_TIME). Additional relations were applied for general properties (HAS_ATTRIBUTE; DEFINED_AS), and for research methods (STUDIES; MEASURES).

4. Relation_definitor. This layer was introduced to facilitate potential knowledge extraction experiments, but also for easier access to the concept features expressed by the relations. In the example below, the composition of the definiendum *sandstone* is expressed by the phrase *made of cemented quartz sand*, where *made of* is the relation definitor.

5. Term_canonical. This layer was added primarily for term normalization purposes in elliptic constructions, for example in *water discharge and velocity* we may add water velocity as the canonical or full version of the term.

Typically, the definition has one definiendum, although in our corpus and domain it is not uncommon to list term variants for certain karst phenomena; in such cases (see below) all synonymous term variants were marked as definienda. We may find definitions without a genus, for example extensional or functional definitions. In the case of extensional definitions listing members of a class we mark hyponyms as SPECIES.

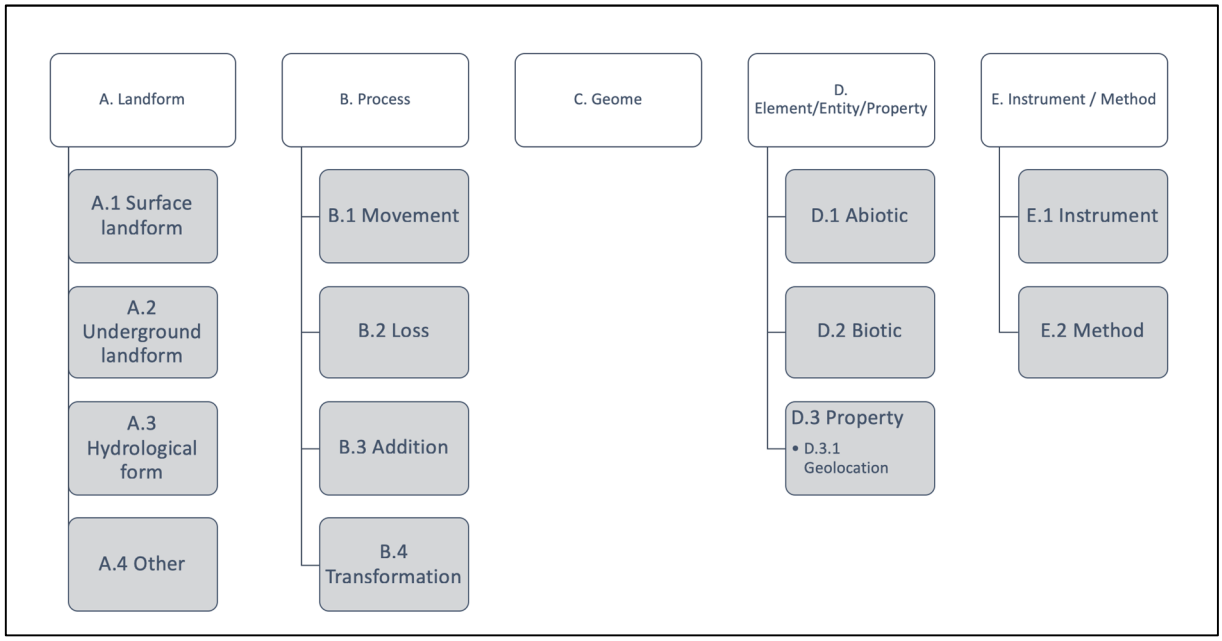

Figure 1: Semantic categories in TermFrame.

Semantic categories are assigned to terms or term-like expressions pertaining to karstology, whereby some categories (e.g. D.1 Abiotic) include terms from the broader domains of geography, geology and chemistry. The definiendum must always be assigned a semantic category, and it is expected that the genus – if present – will share the same category as the definiendum.

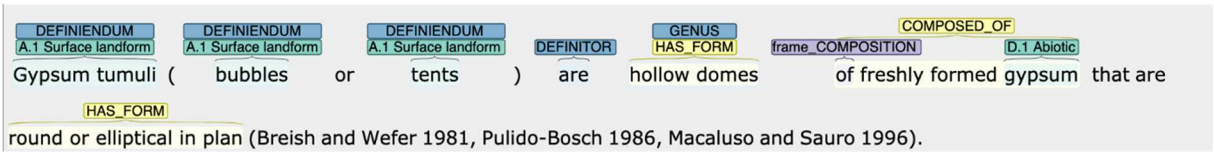

Figure 2: Definition for *gypsum tumuli* with two synonyms.

Relations on the other hand may be assigned to single words, phrases or larger strings, even entire clauses, depending on the context used to explain a particular feature of the definiendum. In addition to the relation itself we annotate the so-called relation definitor, which is the verbal, adjectival or prepositional phrase introducing the relation. For example, the COMPOSED_OF relation might be introduced by *made of, consisting of, of,* HAS_CAUSE by phrases such as *formed by the, driven by, induced by* etc. The relation definitors might help us identify patterns for future experiments with automatic relation identification.

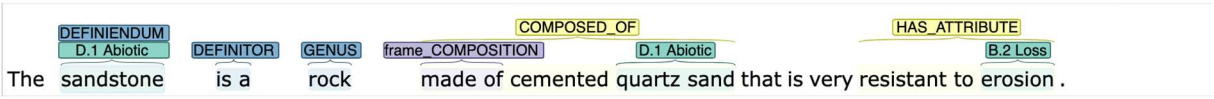

Figure 3: Definition for *sandstone* with two relations.

The choice of relations in a definition is not arbitrary, rather there are certain logical connections between the semantic category and the relations which are used to define it. Such connections can help us predict the relations to be found in a definition. Thus, a surface landform is typically defined using one or several of the following relations: HAS_FORM, HAS_CAUSE, HAS_SIZE and HAS_LOCATION; whereas processes will typically be defined through the HAS_CAUSE, HAS_RESULT, HAS_ATTRIBUTE, OCCURS_IN_TIME and AFFECTS relations. These initial assumptions about definition templates in karstology were formulated by the domain expert prior to the annotation stage. One of the goals of the TermFrame project is to verify such assumptions and compare corpus evidence from three languages with the "ideal" definition template. On the other hand, the ideal template may serve as an aid for generating complete definitions from annotated corpus data.

4.2 Distribution of categories and relations in the English and Slovene TermFrame corpora

In total, 1,061 English and 1,332 Slovene terms were assigned categories, of which 215 English and 286 Slovene terms were definienda. Figure 4 shows the distribution of categories for all annotated terms; we see that in both languages the most frequent category is D.1 Abiotic, followed by surface and underground landforms and geomes. Abiotic elements are frequent categories in definitions because they comprise all kinds of natural entities not specific to karst, such as *bedrock, calcite, deposit, limestone, ridge, sediment*, etc. Amongst the definienda, the most frequent category for both English and Slovene is surface landform (73/119) followed by geomes in Slovene and underground landforms in English.

A geome is a geographical environment or landscape. We find numerous definitions for geomes denoting either types of karst (*cryptokarst, fluviokarst, glacier pseudokarst*) or subsurface environments, usually defined by their hydrologic function (*epikarst,* aquifer, conduit system, subcutaneous zone). Geomes seem more frequent in Slovene, but in fact this is due to numerous definitions for the same concept (e.g. 14 definitions for *kontaktni kras*, six for *kras*).

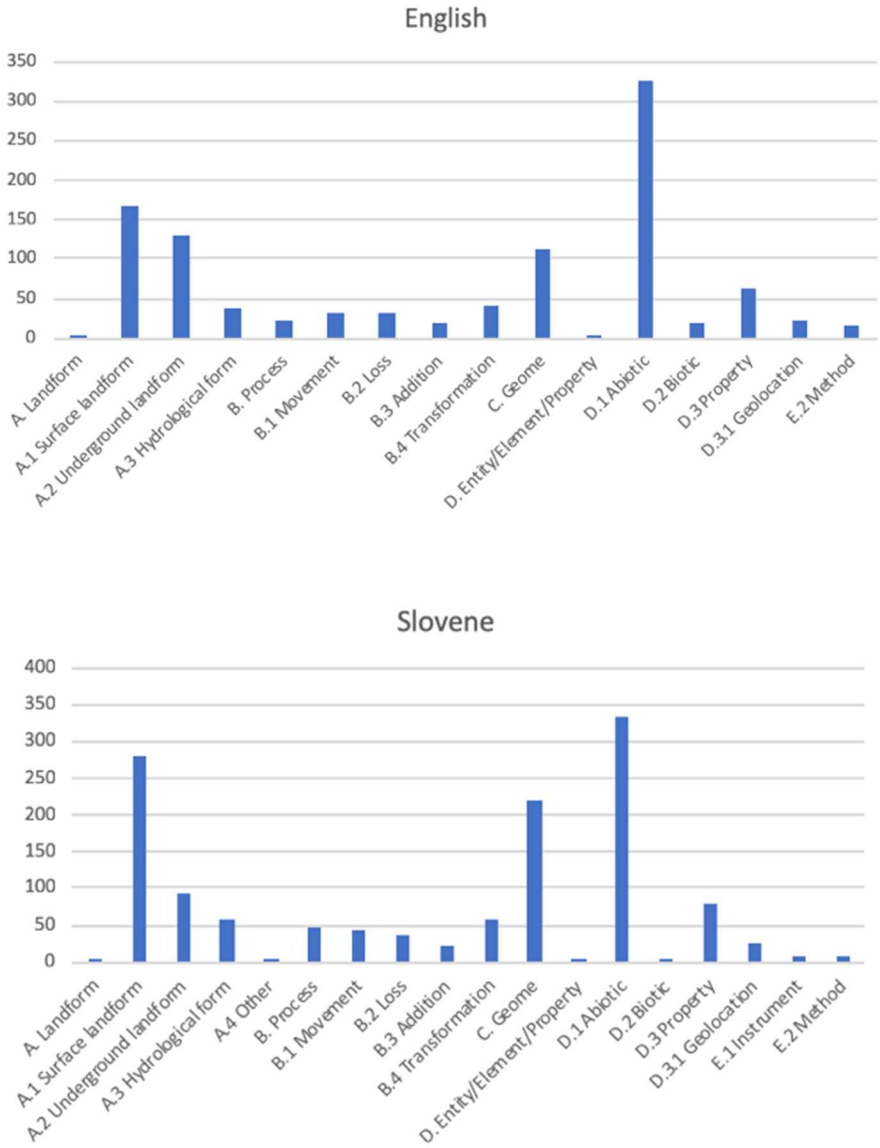

Figure 4: Frequency of categories assigned to English and Slovene terms.

A total of 382 relations were marked in English and 482 in Slovene. In both languages karst concepts are most frequently described through their spatial distribution (HAS_LOCATION), followed by morphography and morphogenesis (HAS_FORM, HAS_CAUSE). This is in accordance with the basic concept of geomorphology (as well as karstology) as a science (Jennings, 1985; White, 1988) that focuses primarily on the shape of landscape features (morphography) and the processes forming them (morphogenesis). Other relations have a similar distribution, apart from the rather general HAS_ATTRIBUTE relation, which appears more frequently in Slovene than in English. Clearly though the frequencies alone do not tell us much about how concepts

are defined in karstology. Looking at the relations occurring with specific semantic categories enables us to discern definition templates.

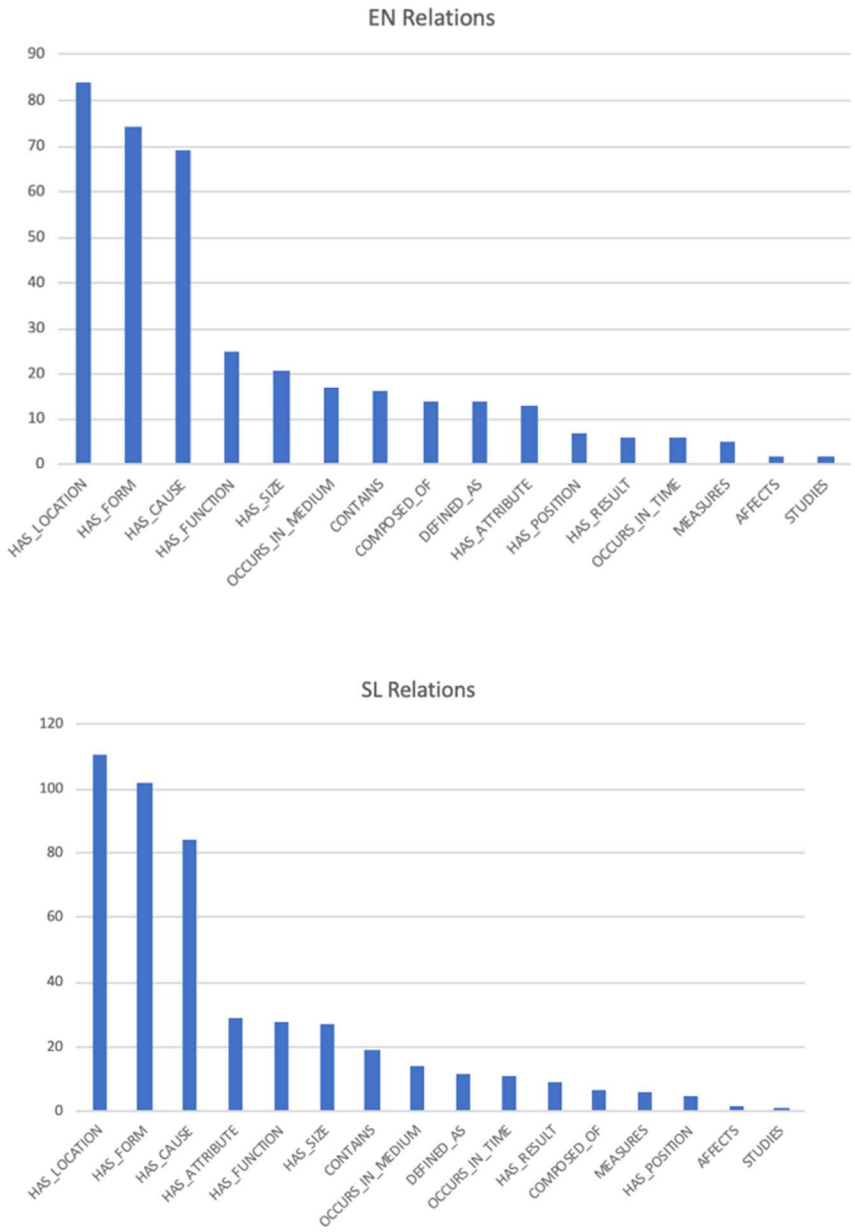

Figure 5: Frequencies of relations found in English and Slovene definitions.

An average definition for a surface landform contains only two relations out of the four typical ones for this category: form, size, location and cause. Sometimes the relation coincides with the genus, as the example in Figure 6 shows. The CONTAINS relation is more frequent with the underground landforms than with other landforms. Thus, *blue holes* contain *tidally influenced waters*, *marginal caves* contain *troglobiotic species*, *vertical shafts* contain *shattered rock and sediment* etc. It is not surprising that

speleothems as subsurface voids have a more pronounced tendency to *contain* something than surface landforms.

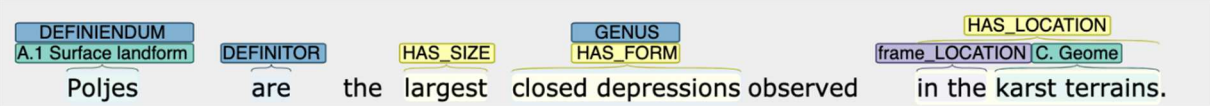

Figure 6: Definition for *polje.*

In contrast to surface and underground forms, hydrological forms are more frequently defined through their function and time pattern. As the examples below illustrate, water is an important agent in karst and hydrological forms are the points in the karst system which may function as storage or transmitters.

Geomes are the second most frequent definiendum category in the Slovene corpus and the third in the English one. We find definitions for environments such as karst and its subtypes (denuded karst, open karst, contact karst, doline karst, epikarst, fluviokarst, hypogene karst, paleokarst, fengcong karst, shilin etc.), but also other large entities and their subparts (aquifer, aquiclude, phreatic zone, zone of vertical circulation etc.). The higher number of geomes in Slovene may be due to the high variability of karst landscapes in Slovenia, which are very actively studied and described by local karstologists.

The most frequent relations used to define geomes in both languages are HAS_CAUSE, HAS_LOCATION, CONTAINS, HAS_ATTRIBUTE, HAS_FORM, HAS_FUNCTION. Interestingly, in English we find three instances where the relation HAS_RESULT is used to define a geome, while no such cases were found for Slovene. The HAS_RESULT relation conceptually requires an agent as subject, in other words a geographical entity would need to instigate some natural activity in order to produce results. In previous work (Vintar & Grčić Simeunović, 2016) we have shown that the cognitive frames underlying definition templates may be language- or culture-dependent, and here we find further evidence for this by defining a geome as a process (see Figure 7). It would appear that English definitions emphasize the morphogenetic aspect, while the Slovene ones prefer the morphodynamic properties of the karst environment as part of the karst system.

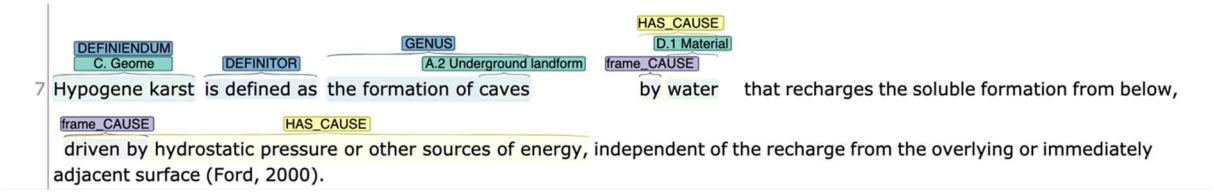

Figure 7: Definition for *hypogene karst.*

5. Towards domain modelling

The TermFrame corpus annotation imposed a rich multi-layered structure onto the previously unstructured content of a large set of documents. The annotation has so far been limited to definitions, although the present annotations can be used for machine learning to extract additional bits of knowledge and the relations among them. The development of a domain representation suited to the needs of experts, researchers, terminologists and lay users remains the primary future task of the project, but several possible directions have already been identified.

For many key concepts in karst we have found several definitions, whereby different authors emphasize different aspects of the definiendum depending on the context, text type and other factors. The identification of the prototypical or ideal definition frame allows us to generate a complete definition from the relations found in different definitions.

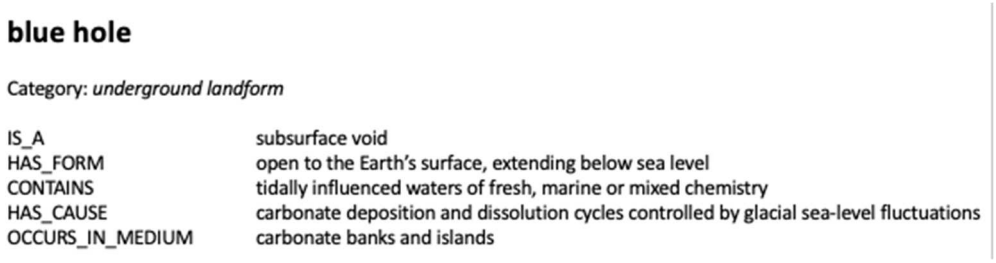

Figure 8: Generating a complete definition frame from several definitions.

Representing the structure of the domain in a graph allows us to see the size of individual concept category hubs, explore nodes and their neighbours, view nodes belonging to several categories and much more (Figure 9). Visualization experiments are underway also for unsupervised detection of communities, see Miljković et al. (2019: 12).

Karstology is essentially a subdomain of geography, and most of the features we explore and represent occur as tangible objects, often sites of interest, in various karst landscapes of the world. Since our corpus contains numerous references to geographical entities, one possibly useful representation is displaying instances of a particular karst feature on a map. Figure 10 presents a map depicting the geolocations of caves extracted from our English corpus using GeoNames.org for co-ordinates.

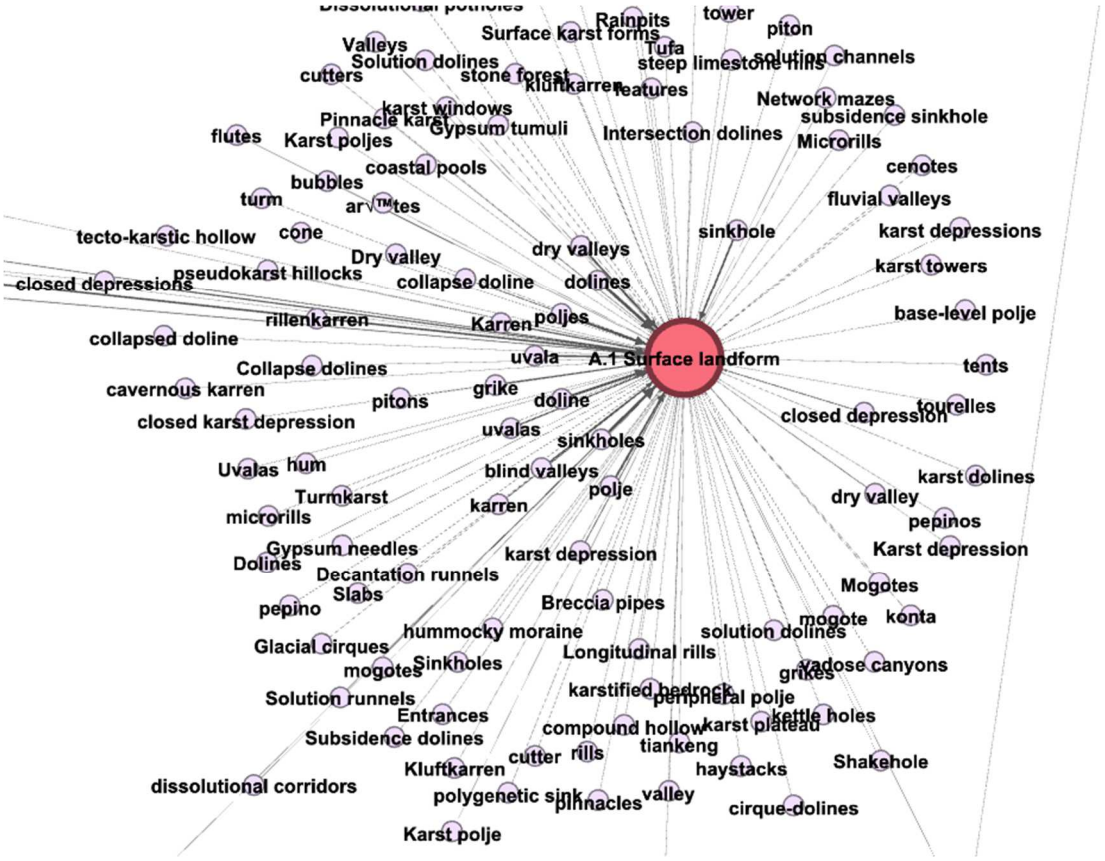

Figure 9: A section of the domain graph representing *surface landforms.*

6. Conclusions

We describe the first stages of the TermFrame project with the construction of a trilingual comparable corpus of karstology and the development of a multi-layer framework for semantic annotation. Analyses of the annotated definitions in English and Slovene allow us to draw conclusions about the cognitive frames underlying knowledge structures in the selected domain, in particular the definition templates for each semantic category. So far these seem similar for both languages, with some differences in frequency distribution and the occurrence of the HAS_RESULT relation to define geomes in English but not Slovene.

Our future plans are to explore the potential of relation definitors in combination with semantic categories to automatically extract or predict relations. Several experiments are underway to extract meaningful knowledge through graph modelling.

7. Acknowledgements

This work is funded by the Slovenian Research Agency grant J6-9372 TermFrame: Terminology and knowledge frames across languages, 2018-2021.

Figure 10: Map of caves mentioned in the TermFrameEN corpus.

8. References

Blanchon, E. (1997). Point de vue en terminologie. *Meta*, *42*(1), pp. 168-173.

De Castilho, R. E., Biemann, C., Gurevych, I. & Yimam, S. M. (2014). WebAnno: a flexible, web-based annotation tool for CLARIN. In *Proceedings of the CLARIN Annual Conference (CAC) 2014*, Soesterberg, Netherlands.

Diki-Kidiri, M. (2000). Une approche culturelle de la terminologie. *Terminologie Nouvelles 21*, Rifal, pp. 58-64.

Duran-Muñoz, I. (2016). Producing frame-based definitions. *Terminology*, 22/2, pp. 223-249.

Faber Benítez, P., Márquez Linares, C., & Vega Expósito, M. (2005). Framing Terminology: A process-oriented approach. *Meta: Journal des traducteurs/Meta: Translators' Journal*, *50*(4).

Faber, P. (2009). The Cognitive Shift in Terminology and Specialized Translation. *MonTI. Monografías de Traducción e Interpretación*. 1: pp. 107-134.

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

Faber, P., León-Araúz, P., & Reimerink, A. (2016). EcoLexicon: new features and challenges. *GLOBALEX workshop, Portorož*, pp. 73-80.

Faber, P. & Medina-Rull, L. (2017) Written in the Wind: Cultural Variation in Terminology. In M. Gryviel (ed.) *Cognitive Approaches to Specialist Languages.* Newcastle-upon-Tyne: Cambridge Scholars, pp. 419-442.

Ford, D. & Williams, P. D. (2007). Karst Hydrogeology and Geomorphology. Wiley, Chichester.

Jennings, J. N. (1985). Karst Geomorphology. Basil Blackwell, Oxford, pp. 293ff.

Leitchik, V. M. & Shelov, S. D. (2007). Commensurability of scientific theories and indeterminacy of terminological concepts. In B.E. Antia (ed.) *Indeterminacy in terminology and LSP: Studies in honour of Heribert Picht.* Amsterdam: John Benjamins, pp. 93-106.

Madsen, B. N., & Thomsen, H. E. (2008). Terminological Principles Used for Ontologies. *Managing Ontologies and Lexical Resources*, pp. 107-122.

Miljković, D., Kralj, J., Stepišnik, U. & Pollak, S. (2019). Communities of related terms in Karst terminology co-occurrence network. *Proceedings of eLex 2019.*

Pavlopoulos, K., Evelpidou, N. & Vassilopoulos, A. (2009). Mapping Geomorphological Environments. Springer, Berlin Heidelberg.

Pollak, S., Vavpetic, A., Kranjc, J., Lavrac, N. & Vintar, Š. (2012). NLP workflow for on-line definition extraction from English and Slovene text corpora. *Proceedings of KONVENS*, pp. 53–60.

Pollak, S., Repar, A., Martinc, M. & Podpečan, V. (2019). Karst exploration: extracting terms and definitions from karst domain corpus. *Proceedings of eLex 2019.*

Roche, C., Calberg-Challot, M., Damas, L., & Rouard, P. (2009). Ontoterminology: A new paradigm for terminology. In *International Conference on Knowledge Engineering and Ontology Development*, pp. 321-326.

San Martin, A. & L'Homme, M.-C. (2014). Definition Patterns for Predicative Terms in Specialized Lexical Resources. In *Proceedings of LREC14*, pp. 3748-3755.

Seppälä, S. (2007). La définition en terminologie: typologies et critères définitoires. In *Terminologie & Ontologies: Théories et Applications* (TOTh 2007), pp. 23-43.

Seppälä, S, & Ruttenberg, A. (2013). Survey on Defining Practices in Ontologies. Concordia University: Montreal. (http://www.webcitation.org/6O2zethfp)

Svensén, B. (1993). *Practical Lexicography: Principles and Methods of Dictionary-Making.* Oxford University Press.

Temmerman, R., & Van Campenhoudt, M. (Eds.). (2014). *Dynamics and Terminology: An interdisciplinary perspective on monolingual and multilingual culture-bound communication* (Vol. 16). Amsterdam: John Benjamins.

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. (2017). Definition frames as language-dependent models of knowledge transfer. *Fachsprache* 1-2/2017, pp. 43-58.

White, W. B., (1988). *Geomorphology and hydrology of karst terrains.* Oxford university press, Oxford.