Karst Exploration: Extracting Terms and Definitions from Karst Domain Corpus

Senja Pollak[1,2], Andraž Repar[1], Matej Martinc[1], Vid Podpečan[1]

[1]Jožef Stefan Institute, Ljubljana, Slovenia
[2] Usher Institute of Population Health Sciences and Informatics,
Edinburgh Medical School, Edinburgh, UK
E-mail: senja.pollak@ijs.si, repar.andraz@gmail.com, matej.martinc@ijs.si,
vid.podpecan@ijs.si

Abstract

In this paper, we present the extraction of specialized knowledge from a corpus of karstology literature. Domain terms are extracted by comparing the domain corpus to a reference corpus, and several heuristics to improve the extraction process are proposed (filtering based on nested terms, stopwords and fuzzy matching). We also use a word embedding model to extend the list of terms, and evaluate the potential of the approach from a term extraction perspective, as well as in terms of semantic relatedness. This step is followed by an automated term alignment and analysis of the Slovene and English karst terminology in terms of cognates. Finally, the corpus is used for extracting domain definitions, as well as triplets, where the latter can be considered as a potential resource for complementary knowledge-rich context extraction and visualization.

Keywords: karstology; term extraction; term embeddings; term alignment; definition extraction; triplets; specialized corpora

1. Introduction

The totality of means of expression in a language can be divided into general language and specialized language. Even if there is no distinct boundary between the two, it can be said that general language defines the sum of the means of linguistic expression encountered by most speakers of a given language, whereas specialized language goes beyond the general vocabulary based on the socio-linguistic or the subject-related aspect. The latter arises as a consequence of constant development and specialization in the fields of science, technology, and sociology (Svensen, 1993: 48-49). Similar to the definition of technical language by Svensen, in the context of terminology, specialized language, also called language for special purposes, is defined as a "language used in a subject field and characterized by the use of specific linguistic means of expression" (ISO 1087-1:2000).

If lexicologists and lexicographers mainly focus on words or lexemes, terminologists focus on terms, i.e., the words with a protected status when used in special subject domains (Pearson, 1998: 7). In contemporary approaches, the dichotomy 'word-term' no longer exists. For Kageura (2002) terms are functional variants of words. Cabré

Castellví (2003: 189) claims that all terms are words by nature and notes that "we recognize the terminological units from their meaning in a subject field, their internal structure and their lexical meaning". According to Myking (2007: 86), the traditional terminology is concept-based and the new directions are lexeme-based.

A definition is a characterization of the meaning of the lexeme (Jackson, 2002: 93). It is "a representation of a concept by a descriptive statement which serves to differentiate it from related concepts" (ISO 12620:2009). The concept to be defined is called a *definiendum*, the part defining its meaning *definiens*, and the optional element (usually a verb) connecting the two parts in a sentence is called a hinge.

Granger (2012) highlights the six most significant innovations of electronic lexicography in comparison to the traditional methods: a) corpus integration, meaning the inclusion of authentic texts in the dictionaries; b) more and better data, since there are no more space limitations and one has the possibility to add multimedia data; c) efficiency of access (quick search and different possibilities of database organization); d) customization, meaning that the content can be adapted to the user's needs; e) hybridization, denoting that the limits between different types of language resources—e.g., dictionaries, encyclopaedias, term banks, lexical databases, translation tools—are breaking down; and f) user input, since collaborative or community-based input is integrated. Similar can be claimed for terminological work, where recent approaches in terminology science consider knowledge (represented in texts) as conceptually dynamic and linguistically varied (Cabré, 1999; Kageura, 2002), and where novel methods in data acquisition, organization and representation, are being constantly developed. Knowledge can be extracted from specialized resources automatically, benefiting from the advances in the field of natural language processing. Moreover, attempts in dynamic, visual representation of domain knowledge have been proposed in recent years, e.g., EcoLexicon[1] (Faber et al., 2016).

In this work, we present the extraction of specialized knowledge from a corpus of karstology, i.e. an interdisciplinary domain at the intersection of geology, hydrology, and speleology. The domain is of high interest, as karst is possibly the most prominent geographical feature of Slovenia (with karst formations being some of popular tourist and natural attractions in the country). It is also an interesting example of how terminology is dynamically evolving in a cross-linguistic context. The literature published in English contains many local Slovenian scientific terms and toponyms for typical geomorphological karst structures, which makes it appropriate for research and identification of cognates, as well as homonym terms, with possible differences in meaning across cultures.

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

Within the TermFrame[2] project, we focus on the specialized knowledge of karst science, and plan to develop methods that allow for context- and language-dependent investigation into a domain, relying on semi-automated tools. In this paper, we apply some of the methods that we have previously developed to a new domain, resulting in a repository of karst term and definition candidates in Slovene and English, contributing to the karstology terminological science. Next, we propose a word embedding based term list extension and triplet extraction method that can be used for visualization. These are novel components, contributing to terminological domain modelling.

This paper is structured as follows. After presenting the related work in automated specialized knowledge extraction in Section 2, we present the resources used (Section 3), methods (Section 4), results (Section 5) and conclude the paper with a discussion and plans for future work (Section 6).

2. Related work

Terminological work has undergone a significant change with the emergence of computational approaches resulting in semi-automated extraction of terms, definitions and other knowledge structures from raw text. Automatic terminology extraction has been implemented for various languages, including English (e.g., Sclano & Velardi, 2007; Frantzi & Ananiadou, 1999; Drouin, 2003) and Slovene (e.g., Vintar, 2010; Pollak et al., 2012), which are the languages in our corpus. In the last few years, word embeddings (Mikolov et al., 2013) have become a very popular natural language processing technique, and several attempts have already been made to utilize word embeddings for terminology extraction (e.g., Amjadian et al., 2016; Zhang et al., 2017). We use word embeddings techniques for extending term lists.

Numerous approaches have also been proposed in bilingual term extraction and alignment, including Gaussier (1998), Kupiec (1993), Lefever et al. (2009), Vintar (2010), Baisa et al. (2015), as well as Aker et al. (2013), who treat bilingual term alignment as a binary classification task. The modified version of the latter approach described in Repar et al. (2018), is also used in this paper.

Automated definition extraction approaches have been developed for several languages, including English (e.g., Navigli & Velardi, 2010), Slovene (e.g., Fišer et al., 2010) and multilingual methods (e.g., Faralli & Navigli, 2013). In our work we use a pattern-based definition extraction method for English and Slovene (Pollak et al., 2012).

In addition to definitions, authors have focused on extracting different types of semantic relations. Pattern-based approaches (Hearst, 1992; Roller et al., 2018), and machine learning techniques have also been proposed (cf. Nastase et al., 2013). In contrast to

[2] http://termframe.ff.uni-lj.si/

extracting predefined semantic relations, the Open Information Extraction (OIE) paradigm considers relations as expressed by parts of speech (Fader et al., 2011), paths in a syntactic parse tree (Ciaramita et al., 2005), or sequences of high-frequency words (Davidov & Rappoport, 2006). In our experiments we use the ReVerb triplet extractor by Etzioni et al. (2011).

This study presents the knowledge extraction steps within the TermFrame project, complementing previous work in karstology modelling presented in Vintar and Grčić-Simeunović (2017), and contributing to the emerging karstology knowledge base. The extracted knowledge was used in the frame-based annotation approach, identifying the semantic categories, relations and relation definitors in definitions of karst concepts, as presented in Vintar et al. (2019), as well as in topic modelling using term co-occurrence network presented in Miljković et al. (2019). The work is also closely related to Faber et al. (2016), a multilingual visual thesaurus of environmental science, which was developed following a frame-based, cognitively-oriented approach to terminology.

3. Resources

The corpus of karstology was constructed within the TermFrame project; it consists of Slovene, Croatian and English texts. We focus on the Slovene and English parts of the TermFrame corpus (v1.0). The English subcorpus contains cca. 1.6 M words and the Slovene one cca. 1 M words (see Table 1 for details).

	English	Slovene
Vocabulary size	64,079	73,813
Documents	24	60
Sentences	103,322	57,575
Words	1,673,132	1,041,475
Tokens	1,972,320	1,231,039
Type-to-token ratio	0.032	0.060

Table 1: Statistics for English and Slovenian subcorpora.

In addition, we are using a short gold standard list of Karst domain terms, called the QUIKK term base[3]. The QUIKK term base consists of terms in four languages, but for the purposes of our experiments, the Slovene and English term lists are used, containing 57 and 185 terms, respectively.

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

4. Methods

4.1 Term candidate extraction

First, we present the procedure of extracting terms by comparing the words in the noun phrases in the domain and reference corpora, and next we present a method using word embeddings to extend the list of terms.

4.1.1 Statistical term extraction

For extracting domain terms we use the LUIZ-CF term extractor (Pollak et al., 2012), which is a variant of LUIZ (Vintar, 2010) refined with scoring and ranking functions. The term extraction uses part-of-speech patterns for detecting noun phrases and compares the frequencies of words (lemmas) in the noun phrase in the domain corpus of karstology and a reference corpus.

The output is a list of term candidates in Slovene and English, above a selected frequency[4] and/or termhood threshold. In addition, we applied the following filtering and term merging procedures:

- *Nested term filtering*: Nested terms are the terms that appear within other longer terms and may or may not appear by themselves in the corpus (Frantzi et al., 2000). As in Repar et al. (2019), the difference between a term and its nested term is defined by a frequency difference threshold: if a term in a corpus appears predominantly within a longer string, only the longer term is returned. If not (if a shorter term appears independently of a longer term more frequently than the set parameter), both terms are included in the final output.[5]

- *Stop word filtering*: If a term candidate is found on the stop word list, the term is excluded from the final list.[6]

- *Term merging by fuzzy matching:* Frequently, we can find terms that are extracted as separate terms but are in fact duplicates because they are written in different variants. This can be due to spelling variations (e.g., British and American English, using hyphenation or not), typos (which are relatively

[4] We set minimum frequency to 15.

[5] In our experiments, the parameter is set to 15 to match minimum frequency.

[6] General stop words are not problematic, as they are frequent also in a reference corpus, and therefore not identified as terms by LUIZ-CF. However, the words specific to the academic discourse, are not frequent in general language and therefore often appear as extracted term candidates. To exclude them, we use the following short stop word list: *example, use, source, method, approach, table, figure, percentage, et, al., km.*

frequent when we deal with large text collections), errors due to pdf-to-text conversions etc. The proposed term merging is based on Levenshtein edit distance (Levenshtein, 1966): if two terms are nearly identical (default threshold is 95%), they will be merged and mapped to a common identifier. In addition, a rule which handles the case when two terms have a different prefix but the same tail and should not be recognized as duplicates can be applied.

4.1.2 Extending term lists with word embeddings

Word embeddings are vector representations of words, where each word is assigned a multidimensional vector of real numbers, characterizing the word based on the lexical context in which it appears. When vectors are computed on very large corpora, and especially with recent advances in models using neural networks, these representations have seen a huge success within various natural language processing tasks.

The embeddings capture certain degree of semantics, as words that are similar or semantically related are closer together in the vector space. Previous research conducted by Diaz et al. (2016) showed that embeddings can be successfully used for expanding queries on topic specific texts. In this research, we test if word embeddings can be used for a similar task of extending the gold standard term lists to find more domain terms. According to the research conducted by Diaz et al. (2016), embeddings trained only on small topic specific corpora outperform non-topic specific general embeddings trained on very large general corpora for the task of query expansion due to strong language use variation in specialized corpora. Therefore, we use the same approach for extending the term list and train custom embeddings on the specialized corpus instead of using pretrained embeddings.

In our experiments, we have trained FastText embeddings (Bojanowski et al., 2017) on the Slovenian and English karst subcorpora and use them to find the twenty closest words (according to cosine distance between embeddings) for the first fifty terms in the QUIKK term base[7]. These related words are sorted according to their proximity to the term and the first, second, tenth and twentieth ranked words are used in manual evaluation. Embeddings for multi-word terms are generated by averaging the word embeddings for each word in the term.[8]

[7] To be exact, 50 English terms, and 47 Slovene terms, since only 47 Slovenian terms from the QUIKK term base appear in the Slovenian corpus.

[8] There are several possible multi-word term aggregation approaches, such as summation of component word vectors, averaging of component word vectors, creating multi-word term vectors, etc. As comparing different techniques is beyond the scope of this study, we decided for the simple averaging technique, as previous research on this topic conducted on the medical domain (Henry et al., 2018) found no statistically significant difference between any multi-word term aggregation method.

4.2 Cognates detection and term alignment

English terms are mapped to Slovene equivalents using a data mining approach by Aker et al. (2013) reimplemented in Repar et al. (2018). Bilingual term alignment is treated as a binary classification, with a support vector machine classifier trained on various dictionary and cognate-based features that express correspondences between the words (composing a term) in the target and source languages. The first take advantage of dictionaries (Giza++) created from large parallel corpora, and the latter exploit string-based word similarity between languages (cf. Gaizauskas et al., 2012). In addition, the cognate-based features (see Table 2) allow users to identify cognate term pairs, which are interesting as karst terms in different languages clearly share their origin, but there exist also well-known examples of non-equivalent cognates (e.g., Slovene "dolina" vs. English "doline").

Feature	Description
Longest Common Subsequence Ratio	Measures the longest common non-consecutive sequence of characters between two strings
Longest Common Substring Ratio	Measures the longest common consecutive string (LCST) of characters that two strings have in common
Dice similarity	2*LCST / (len(source) + len(target))
Normalized Levensthein distance (LD)	1 - LD / max(len(source), len(target))

Table 2: Cognate-based features used for term alignment.

4.3 Definition candidates extraction

We use the pattern-based module of the definition extractor (Pollak et al., 2012), which is available online.[9] The soft pattern matching is used to extract sentences of forms NP is NP, NP refers to NP, NP denotes NP, etc., and the parameters contain language (EN, SL), as well as the position of the term in Slovene (if the term must be at the beginning of the sentence, after a larger set of predefined start patterns (our choice) or anywhere in a sentence).

4.4 Triplet extraction

As predefined definition patterns (cf. Section 4.3) were designed for extracting specific knowledge contexts, we complement the approach by open-relation extraction (this experiment is conducted only for English, as for Slovene the tools are not available).

[9] http://clowdflows.org/workflow/8165/

We use ReVerb (Fader et al., 2011), which extracts relation phrases and their arguments and results in triplets of form:

<argument1, relation phrase, argument2>

We believe that in the case that argument1 and argument2 match domain terms, the triplets can be exploited as a method for extraction of knowledge-rich contexts (an alternative to definitions). They are also a useful input for visualization of terminological knowledge and can meet the needs of frame-based terminology, aiming at facilitating user knowledge acquisition through different types of multimodal and contextualized information, in order to respond to cognitive, communicative, and linguistic needs (Gil-Berrozpe et al., 2017). Previously, triplets have been used in other domains, e.g., in systems biology for building networks from domain literature (Miljković et al., 2012).

5. Evaluation setting and results

5.1 Term candidate extraction

5.1.1 Statistical term extraction

We extracted 4,397 English term candidates and 2,946 Slovene term candidates. A domain expert and a linguist specialized in terminology with high domain understanding manually evaluated all term candidates for Slovene and the top 1,823 (above a selected threshold)[10] term candidates for English. The following categories were used:

- Not a term (label: 0)
- Karst term (label: 1)
- Broader domain terms (label: 2)
- Named entity (label: 3)

To distinguish between karst and broader domain terms, the following criterion is used. While karstology is in itself an interdisciplinary field, in TermFrame the focus is on karst geomorphology entailing surface and underground landforms, and karst hydrology

[10] The reason for the discrepancy in the number of evaluated terms is that the evaluation for Slovene yielded a much lower number of terms (categories 1 or 2) in Slovene than in English. Since we need a large number of terms for additional steps, i.e. term alignment, we instructed the evaluators to process the full list of term candidates for Slovene. If we took the same number of top terms for Slovene as for English (top 1,823), we get the following results (cf. Table 3): Not a term: 1,187, Karst term: 140, Domain term: 174, Named entity: 220, Precision: 0.293.

with its typical forms and processes. Terms from neighbouring domains (geography, biology, geochemistry, etc.) which are not exclusive to karst are considered broader domain terms. In case of disagreement, the two annotators achieved consensus on the final category. As presented in Table 3, the resulting list of terms contains 351 karst terms for English and 158 for Slovene. The newly extracted karst terms, such as *cave, uvala, doline, denudation* describing landforms, processes, environment, etc., can serve for the extension of the manual QUIKK karstology term base, while for example the term candidate *karst region* is not considered a term because it is too generic and compositional, denoting a different underlying semantic relation (a region which contains karst).

The precision of term extraction is 0.516 for English and 0.235 for Slovene. For examples of terms in each category, see Table 4, while top terms sorted by termhood score for English and Slovene are presented in Tables 5 and 6, respectively.

Lang	Evaluated terms	Not a term	Karst term	Broader domain term	Named entity	Precision
Slovene	2,946	2,228	158	194	341	0.235
English	1,823	882	351	434	156	0.516

Table 3: Term extraction results. Precision is calculated as the sum of all three positive categories (1, 2, 3) divided by the number of evaluated terms.

In addition, we evaluate our filtering methods. All nested terms (306 for English, 105 for Slovene) removed by the nested term filtering are correctly eliminated, the stop word filter did not detect any terms which should not be removed, and all near duplicates (11 for English, 22 for Slovene) detected with the fuzzy match filter are also correct (e.g., "ground-water" was detected as a duplicate of "ground water").

Lang	Not a term	Karst term	Broader domain term	Named entity
Slovene	dinarska smer	slepa dolina	naplavna ravnica	Planinsko polje
	ilovnat material	udornica	ravnovesna meja	Podgorski kras
	kataster jam	kalcijev karbonat	mehansko preperevanje	Gorski kotar
English	deepest cave	karst aquifer	sea level	Southeast Asia
	world heritage	subterranean water	carbonic acid	Castleguard Cave
	largest spring	phreatic cave	cave habitat	Central America

Table 4: Examples of term extraction evaluation categories.

Rank	Frequency	Term	Categorization
1	19269	cave	1
2	451	karst aquifer	1
3	522	karst area	1
4	459	cave system	1
5	314	dinaric karst	3
6	414	carbonate rock	1
7	348	cave passage	1
8	218	crna reka	3
9	271	karst system	1
10	209	karst feature	1
11	192	karst terrain	1
12	201	karst landscape	1
13	203	karst region	0
14	192	karst spring	1
15	564	united state	3
16	146	troglobitic specie	2
17	187	cave entrance	1
18	227	lava tube	2
19	169	cave sediment	1
20	164	karst rock	1

Table 5: Top 20 English karst term candidates with frequencies and categorization to karst terminology (1), broader domain terminology (2), named entity (3) or non-term (0).

5.1.2 Extending term lists with word embeddings

The method was tested on 47 English and 50 Slovene source terms (i.e. the terms from the gold standard list), for which out of the 20 most related words (according to the cosine distance between the source term and the related word), four per each source term were selected for evaluation (first, second, tenth and twentieth ranked words), resulting in 200 term-word pairs for English and 188 for Slovene.[11] Examples of ranked related words for five English and five Slovene terms are presented in Table 7.

[11] In this section, we intentionally name related words as words and not as terms, to contrast them to the gold standard list of terms to which they are compared. As shown in the evaluation, they can be evaluated as terms or not in the next step.

Rank	Frequency	Term	Categorization
1	1,966	nadmorska višina	0
2	9,543	jama	1
3	4,472	kras	1
4	6,359	voda	0
5	713	slepa dolina	1
6	4,481	dolina	0
7	405	brezstropa jama	1
8	2,948	apnenec	1
9	623	Pivška kotlina	3
10	2,573	sediment	0
11	3,418	dno	0
12	425	erozijski jarek	2
13	3,608	polje	1
14	2,770	rov	1
15	728	kraško polje	1
16	2,049	udornica	1
17	4,619	del	0
18	2,564	kamnina	2
19	507	suha dolina	1
20	3,882	oblika	0

Table 6: Top 20 Slovene karst term candidates with frequencies and categorization to karst terminology (1), broader domain terminology (2), named entity (3) or non-term (0).

Term	R1	R2	R10	R20
sinkhole	shakehole	suburban	sinkpoint	dump
aggressive water	aggressively	aggressiveness	qc	coldwater
epikarst zone	epikarstic	subcutaneous	cutaneous	epiphreatic
caprock sinkhole	sinkpoint	overbank	suburb	evacuation
seacave	seacoast	sealevel	vrulja	caveand
udornica	udornina	zapornica	koliševka	kamojstrnik
agresivna voda	sposoben	mehurček	skoznjo	preniči
epikras	epikraški	prenikujoč	epr	vadozen
vrtača	vrtačast	mikrovrtača	globel	neizravnan
rečna jama	reža	narečen	mohoričev	vodokazen

Table 7: Examples of ranked related words for five English (upper five examples) and five Slovene (lower five examples) terms.

The two human evaluators evaluated the related words according to two criteria:

- Is the word a term?
- Semantic similarity to the term

The first criterion is measured on a scale with four nominal classes (see Section 5.1.1), while the second criterion uses a numerical scale from zero to ten, following the evaluation procedure of Finkelstein et al. (2002), where zero suggests no semantic similarity and ten suggests very close semantic relation (fractional scores were also allowed). The inter-annotator agreement between the two evaluators (according to the Cohen's kappa coefficient) is 0.689 for the first criterion and 0.513 for the second criterion for English, and 0.594 for the first criterion and 0.389 for the second criterion for the Slovene evaluation.

Table 8 presents the results for the evaluation of embeddings-based term extension. Out of 200 English term-word pairs, 112 were manually labelled as term-term pairs by at least one evaluator, which suggests that, at least for English, embeddings can be used for extending the term list. Out of these 112 related terms, 52 were labelled as karst specific terms by at least one evaluator. For Slovenian, the results are worse, since out of 188 term-word pairs only 69 were labelled as term-term pairs, and out of these only 36 are karst specific.

Out of 112 English term-term pairs, 62 were ranked first and second and 50 were ranked tenth and twentieth according to the cosine distance similarity. Out of 69 Slovenian term-term pairs, 39 were ranked first or second and 30 were ranked as tenth or twentieth. This suggests that words that have most similar embeddings to terms according to the cosine distance (rank 1 and rank 2) are also more likely to be terms themselves than words that have less similar embeddings (rank 10 and rank 20). Similar reasoning applies to karst specific term-term pairs, where for English 30 were ranked first or second and 22 were ranked tenth or twentieth. For Slovenian, 24 out of 36 were ranked first or second and 12 were ranked tenth or twentieth.

When it comes to semantic similarity, unsurprisingly better ranked related words were manually evaluated as semantically more similar. For example, the first ranked (most similar to terms according to the cosine distance) English related words got an average semantic similarity score[12] of 4.040 out of ten, and the first ranked Slovenian related words got an average semantic similarity score of 4.468. These are larger than the semantic similarity score averages of 2.610 and 3.064 for English and Slovenian related words ranked as twentieth, respectively. Another interesting observation is the fact that the average semantic similarity score is the highest for English karst specific term-terms pairs (5.702) and much lower if all the term-word pairs are considered (3.325). If we

[12] The semantic similarity score for each related word is calculated as an average between the two semantic similarity scores given by two evaluators.

consider all term-term pairs, the average semantic similarity score is 4.710. The same applies for Slovenian term-word pairs, with semantic similarity score average rising from 3.859 when all term-words pairs are considered, to 5.536 when only term-term pairs are considered, and up to 6.722 when only karst specific term-term pairs are considered.

We also measure the correlation between cosine distances and the semantic similarity scores for term-word pairs using Pearson and Spearman correlation coefficients. The correlation is generally low, the highest being measured for Slovenian Karst specific term-term pairs where the Pearson correlation reached the value of 0.341 and Spearman the value of 0.208. There was no correlation measured on Slovene term-term pairs and surprisingly, a small negative Pearson correlation was measured on Slovenian karst specific term-term pairs and a small negative Spearman correlation was measured on English pairs which were labelled as terms.

5.2 Cognate detection and term alignment

We evaluate the approach first on the QUIKK gold standard, where 100% precision and recall above 40% were obtained. Next, we also add to the QUIKK gold standard the terms extracted using the statistical method and term embeddings that were positively evaluated. The total list of 908 English terms and 391 Slovene terms were input to the term alignment algorithm. The resulting list of 93 aligned term pairs was manually evaluated. In this experiment, the precision was 77.42% (72 term alignments out of 93 were correct), while the recall could not be calculated, as the gold standard alignment was not available.

	English				Slovene			
All words	200				188			
Avg. sem. score	3.325				3.859			
Avg. cos. dist.	0.747				0.760			
Pearson corr.	0.181				0.231			
Spearman corr.	0.136				0.194			
	R1	R2	R10	R20	R1	R2	R10	R20
Distribution	50	50	50	50	47	47	47	47
Avg. sem. score	4.040	3.540	3.110	2.610	4.872	4.468	3.032	3.064
Terms	112				69			

Avg. sem. score	4.710				5.536			
Avg. cos. dist.	0.757				0.771			
Pearson corr.	0.176				-0.018			
Spearman corr.	0.160				-0.016			
	R1	R2	R10	R20	R1	R2	R10	R20
Distribution	32	30	29	21	17	22	15	15
Karst terms	52				36			
Avg. sem. score	5.702				6.722			
Avg. cos. dist.	0.761				0.780			
Pearson corr.	0.151				-0.152			
Spearman corr.	0.070				-0.067			
	R1	R2	R10	R20	R1	R2	R10	R20
Distribution	16	14	15	7	12	12	5	7
Not Terms	88				119			
Avg. sem. score	1.563				2.887			
Avg. cos. dist.	0.734				0.753			
Pearson corr.	-0.010				0.341			
Spearman corr.	-0.110				0.208			
	R1	R2	R10	R20	R1	R2	R10	R20
Distribution	18	20	21	29	30	25	32	32

Table 8: English and Slovenian embeddings evaluation according to two criteria described in Section 4.1.2. *Avg. sem. score* stands for the average of manually prescribed semantic similarity scores for each term-word pair, *Avg. cos. dist* stands for the average cosine distance, *Pearson corr.* is a Pearson correlation coefficient between the semantic similarity score and cosine distance values and *Spearman corr.* is a Spearman correlation coefficient between the semantic similarity score and cosine distance values.

As described in Section 4.2, karst terminology contains a considerable amount of cognates. See Table 9 for cognate values for Longest Common Substring Ratio, Longest Common Subsequence Ratio, Dice Similarity, and Normalized Levensthein Distance).

5.3 Definition candidate extraction

In total, 1,320 definition candidates were extracted for English, and 1,218 for Slovene. Definition candidates were manually validated by domain experts following two criteria: whether the sentence defines the concept, and whether the concept belongs to the domain of karstology. To distinguish between definitions and non-definitions the experts checked whether the sentence explains what the concept is, either by specifying its hypernym and a set of distinguishing features (analytical), or by listing its hyponyms (extensional), or by using another explanatory strategy (e.g., functional definitions). The definition candidates were then assigned one of the following three categories:

- Definitions of karst terms (Example: *Aggressiveness is an attribute of groundwater that corresponds to a chemical potential for mobilization of a dissolved matter from the rock.*)
- Definitions of broader domain terms (biology, geology etc.). (Example: *Exploration geophysics is the science of seeing into the earth without digging or drilling.*)
- Non-definitions (Example: *The oldest rocks are the sandstones of Permian age, which are only locally present.*)

English term	**Slovene term**	**LCSTR**	**LCSSR**	**Dice**	**NormLD**
mineralization	mineralizacija	0.71	0.79	0.71	0.79
salinization	salinizacija	0.67	0.75	0.67	0.75
nitrification	nitrifikacija	0.54	0.69	0.54	0.69
aggressive water	agresivna voda	0.25	0.63	0.27	0.50
karst plateau	kraška planota	0.27	0.60	0.29	0.40
karst	kras	0.20	0.60	0.22	0.40
marble	marmor	0.50	0.50	0.50	0.50
karst drainage	kraška drenaža	0.19	0.50	0.20	0.38
karst phenomena	kraški pojav	0.13	0.47	0.14	0.20
linear stream cave	linearna epifreatična jama	0.22	0.44	0.27	0.44

Table 9: Cognate scores for a sample of Slovene and English term pairs

As presented in Table 10, for English, out of 1,320 definition candidates 218 were evaluated as karst definitions, and an additional 187 as broader domain definitions. The precision of the definition extraction on karst domain is thus 0.16 for strictly karst domain definitions, and 0.31 for broader domain definitions (incl. karst definitions). For Slovene, there are 1218 definition candidates, out of which 260 are karst definitions and 166 are from broader domain. The precision for definition extraction for Slovene is thus 0.21 for strictly karst domain, and 0.35 for karst and broader domain.

	English	**Slovene**
Karst definitions	218	260
Broader domain definitions	187	166
Non definitions	915	792
All definition candidates	1320	1218

Table 10: Number of extracted definition candidates, evaluated as karst definitions, broader domain definitions and non-definitions.

The karst definitions were then used by domain experts and linguists in the scope of the TermFrame project for a fine-grained, annotation process, following frame-based terminology principles (Faber, 2015). The annotation principles and results are presented in Vintar et al. (2019), where several annotation layers are proposed: definition element layers (definiendum, definitor and genus); semantic categories (top level concepts are landforms, processes, geomes, entities, instruments/methods) and relations (16 relations, such as has_form, has_cause).

5.4 Triplet extraction

The English subcorpus yielded 80,564 triplets. Below we list selected examples of relevant triplets that are closely related to the karst domain:

- <Karst areas, commonly lack, surface water>
- <Karst areas, have, numerous stream beds that are dry except during periods of high runoff>
- <Sinkholes located miles away from rivers, can flood, homes and businesses>
- <Karst areas, offer, important resources>
- <Some collapse sinkholes, develop, where collapse of the cave roof reaches the surface of the Earth>

The extracted triplets are analysed according to the most common relation patterns, to estimate their potential for extending predefined definition patterns. From the relation phrase part of the triplet, the verb is identified, showing the most frequent verb structures. We remove all stopwords from the relation phrase using a general list of 174 English stopwords. Table 11 lists 20 most frequent verb structures found in the processed 24 documents. The results show that many karst-specific relations can be detected (e.g., verbs related to different geological processes, such as *occur*, *develop* and *form*) but still many general verbs are also frequent. The frequent relations from triplets will be discussed in relation to the predefined set of relations used in definition frames annotation (cf. Vintar et al., 2019).

	verb	count		verb	count
1	found	1451	11	appear	336
2	occur	1347	12	consist	323
3	use	878	13	represent	321
4	form	811	14	locate	313
5	develop	787	15	include	312
6	know	646	16	contain	310
7	provide	528	17	made	306
8	show	428	18	result	295
9	take	397	19	depend	273
10	describe	337	20	extend	272

Table 11: 20 most frequent verb structures compiled from 80,564 triplets. Note that stopwords were removed from verb structures.

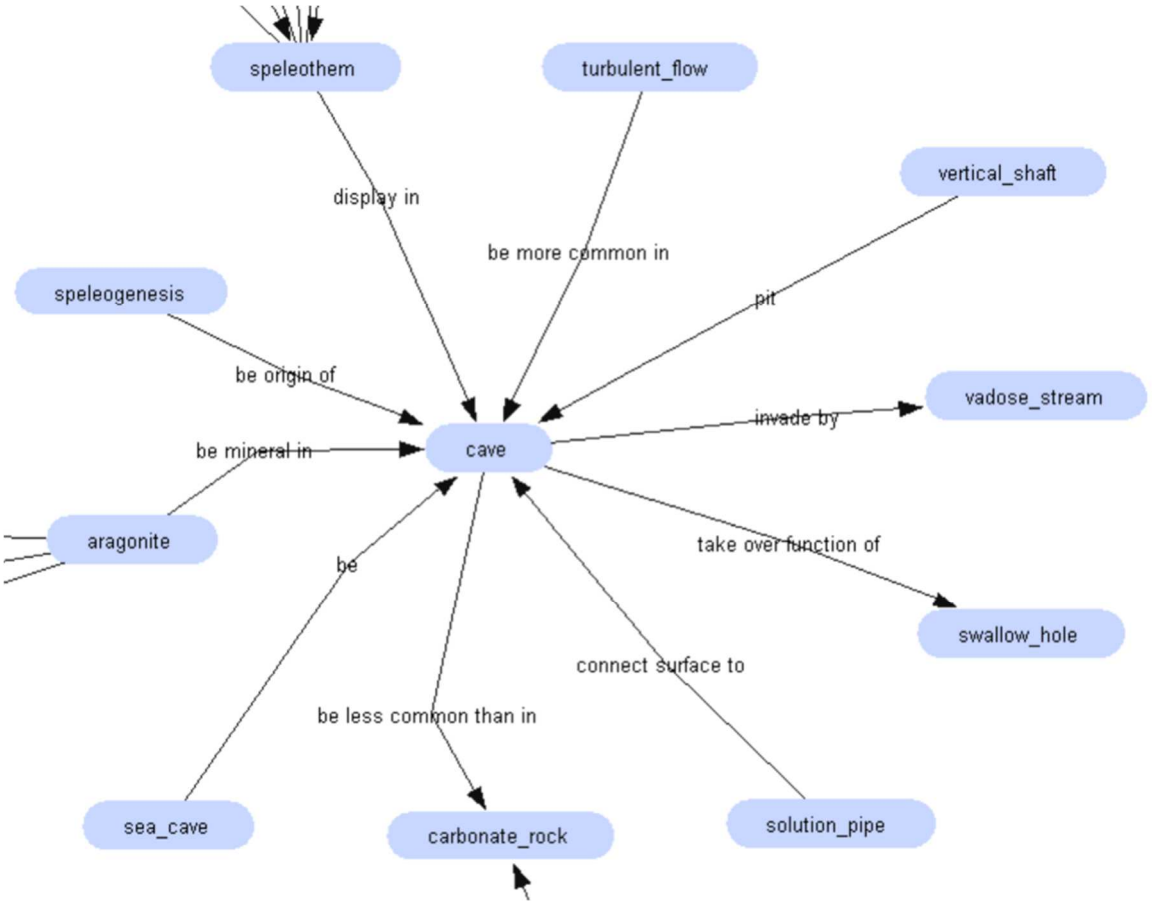

Figure 1: Visualization of a part of the triplet network. Prior to the visualization, relation phrases were lemmatized and the triplets were filtered according to the short gold standard list of Karst domain extended with an additional evaluated list of terms.

For visualization, after filtering the triplets by keeping only the ones where in a triplet <argument1, relation phrase, argument2> the two arguments are karst terms[13], we construct a network where arguments are used as nodes and relation phrases as arcs. A visualization of a part of the triplet network obtained using Biomine network visualization tool (Eronen & Toivonen, 2012) is shown in Figure 1.

6. Conclusion and further work

We model domain knowledge utilizing a range of natural language processing techniques, including term extraction (using statistical methods, filtering and word embeddings), term alignment and cognates detection, definition extraction and triplet extraction. The proposed techniques form a pipeline for contemporary terminological work, relying on semiautomated processes for knowledge extraction from specialized domain corpora. Several modules in the pipeline rely on existing techniques, which were refined for the purposes of this work (e.g., term extraction), while we believe that the use of embeddings and triplets has not yet been sufficiently explored in the context of lexicography and terminography. The hypothesis was that embeddings offer not only a possibility of extending a list of terms, but also of grouping them to semantically related concepts, which can be of great value in the organization of domain knowledge (in term bases and similar resources), and also in contemporary lexicography resources.

We apply the proposed pipeline to a corpus of karst specialized texts. The main value of the evaluation steps of term and definition extraction is to obtain new gold standard karst knowledge resources that will be used in the scope of the TermFrame project for fine grained analysis and novel visual representation corresponding to the cognitive shifts in recent terminology science approaches. On the other hand, we believe that the evaluation of word embeddings opens new perspectives to e-lexicography and terminography, as it shows that popular techniques from natural language processing are relatively successful for automatically extending the gold standard term lists (cca. half of English and one third of Slovene terms being valid terms). The evaluation also shows that the semantic similarity score is higher for the closest matching words (considering cosine similarity between embeddings) than for the lower ranked words, which suggests that embeddings do in fact manage to capture some semantic relations despite a relatively small training corpus. On the other hand, the correlation between cosine similarity and manual similarity score is weak, which might indicate high variance in cosine similarity for related words for different terms. We believe that semantic information has a huge potential for contributing to the organization of term bases and visually interesting knowledge maps. In the same line, we illustrate how triplet extraction in combination with term matching can serve as a knowledge representation module used for visualization.

[13] QUIKK terms and manually evaluated terms from Section 5.1.1.

In future work, we will consider extending the corpus by using web-crawling techniques. Next, our aim is to merge the pipeline to a set of services to support users in a knowledge extraction process, for populating term bases, as well as in knowledge visualization. We believe that such tools will contribute to better understanding of similarities and differences in terminological expression between languages, and support representations reflecting dynamic culture and language specific knowledge.

7. Acknowledgements

The work was supported by the Slovenian Research Agency through the core research programme (P2-0103) and research project Terminology and knowledge frames across languages (J6-9372). This work was supported also by the EU Horizon 2020 research and innovation programme, Grant No. 825153, EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). The results of this publication reflect only the authors' views and the EC is not responsible for any use that may be made of the information it contains. We would also like to thank Š. Vintar, U. Stepišnik, D. Miljković and other members of the TermFrame project for their collaboration.

8. References

Aker, A., Paramita, M. & Gaizauskas, R. (2013). Extracting bilingual terminologies from comparable corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pp. 402–411.

Amjadian, E., Inkpen, D., Paribakht, T. & Faez, F. (2016). Local-Global Vectors to Improve Unigram Terminology Extraction. In *Proceedings of the 5th International Workshop on Computational Terminology (Computerm2016)*, pp. 2–11.

Baisa, V., Ulipová, B. & Cukr, M. (2015). Bilingual terminology extraction in Sketch Engine. In *9th Workshop on Recent Advances in Slavonic Natural Language Processing*, pp. 61–67.

Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5, pp. 135–146.

Cabré, M.T. (1999). *Terminology: Theory, Methods, and Application*. Amsterdam, The Netherlands and Philadelphia, USA: John Benjamins Publishing.

Cabré Castellví, M. T. (2003). Theories of Terminology: Their Description, Prescription and Explanation. *Terminology 9 (2)*, p. 163–199.

Ciaramita, M., Gangemi, A., Ratsch, E., Šaric, J. & Rojas, I. (2005). Unsupervised Learning of Semantic Relations Between Concepts of a Molecular Biology Ontology. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI'05)*, pp. 659–664.

Davidov, D. & Rappoport, A. (2006). Efficient Unsupervised Discovery of Word

Categories Using Symmetric Patterns and High Frequency Words. In *Proceedings of the 21th International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*. Sydney, Australia, pp. 297–304.

Diaz, F., Mitra, B. & Craswell, N. (2016). Query expansion with locally-trained word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin, Germany, p. 367–377.

Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1), pp. 99–117.

Eronen, L. & Toivonen, H. (2012). Biomine: predicting links between biological entities using network models of heterogeneous databases. *BMC Bioinformatics*, 13(1), pp. 1–21.

Etzioni, O., Fader, A., Christensen, J., Soderland, S. & Mausam (2011). Open Information Extraction: The Second Generation. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume One (IJCAI'11)*. Barcelona, Catalonia, Spain, pp. 3–10.

Faber, P. (2015). Frames as a framework for terminology. In H. Kockaert & F. Steurs (eds.) *Handbook of Terminology*. John Benjamins, p. 14–33.

Faber, P., León-Araúz, P. & Reimerink, A. (2016). EcoLexicon: new features and challenges. In *Proceedings of GLOBALEX 2016: Lexicographic Resources for Human Language Technology in conjunction with the 10th edition of the Language Resources and Evaluation Conference*, pp. 73–80.

Fader, A., Soderland, S. & Etzioni, O. (2011). Identifying Relations for Open Information Extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP'11)*. Edinburgh, United Kingdom1: Association for Computational Linguistics, pp. 1535–1545.

Faralli, S. & Navigli, R. (2013). A Java Framework for Multilingual Definition and Hypernym Extraction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Sofia, Bulgaria: Association for Computational Linguistics, pp. 103–108. https://www.aclweb.org/anthology/P13-4018.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G. & Ruppin, E. (2002). Placing search in context: The concept revisited. *ACM Transactions on information systems*, 20(1), pp. 116–131.

Fišer, D., Pollak, S. & Vintar, Š. (2010). Learning to Mine Definitions from Slovene Structured and Unstructured Knowledge-Rich Resources. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta, pp. 2932–2936.

Frantzi, K., Ananiadou, S. & Mima, H. (2000). Automatic recognition of multi-word terms:. the C-value/NC-value method. *International Journal on Digital Libraries*, 3(2), pp. 115–130.

Frantzi, K. T. & Ananiadou, S. (1999). The C-Value/NC-Value Domain Independent Method for Multi-Word Term Extraction. *Journal of Natural Language*

Processing, 6(3), pp. 145–179.

Gaizauskas, R., Aker, A. & Yang Feng, R. (2012). Automatic bilingual phrase extraction from comparable corpora. In *24th International Conference on Computational Linguistics*, p. 23.

Gaussier, E. (1998). Flow Network Models for Word Alignment and Terminology Extraction From Bilingual Corpora. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (Coling-ACL)*, pp. 444–450.

Gil-Berrozpe, J., León-Araúz, P. & Faber, P. (2017). Specifying Hyponymy Subtypes and Knowledge Patterns: A Corpus-based Study. In I. Kosem et al. (eds.) *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference.* Brno: Lexical Computing, pp. 63–92.

Granger, S. (2012). Electronic Lexicography-from Challenge to Opportunity. In S. Granger & M. Pacqot (eds.) *Electronic Lexicography*, chapter Introduction. Oxford University Press, p. 1–15.

Hearst, M. A. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 2 (COLING'92)*, pp. 539–545.

Henry, S., Cuffy, C. & McInnes, B. T. (2018). Vector representations of multi-word terms for semantic relatedness. *Journal of biomedical informatics*, 77, pp. 111–119.

ISO 1087-1:2000 (2000). International Standard: Terminology Work — Vocabulary — Part 1: Theory and Application. Standard cited from the Glossary of Terminology Management of DG TRAD – Terminology Coordination Unit of European Parliament (Last accessed June 17, 2019). Standard. http://termcoord.wordpress.com/glossaries/glossary-of-terminology-management/.

ISO 12620:2009 (2009). International Standard. Terminology and Other Language and Content Resources — Specification of Data Categories and Management of a Data Category Registry for Language Resources. Standard cited from ISOcat Web Interface (Last accessed December 1, 2013). Standard. https://catalog.clarin.eu/isocat/interface/index.html.

Jackson, H. (2002). *Lexicography: An Introduction*. Routledge.

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

Kupiec, J. (1993). An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL)*. Columbus, OH.

Lefever, E., Macken, L. & Hoste, V. (2009). Language-Independent Bilingual Terminology Extraction from a Multilingual Parallel Corpus. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, pp. 496–504.

Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10, p. 707.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In *Proceedings to The International Conference on Learning Representations 2013*.

Miljković, D., Kralj, J., Stepišnik, U. & Pollak, S. (2019). Communities of related terms in Karst terminology co-occurrence network. In I. Kosem et al. (eds.) *Proceedings of eLex 2019*, pp. 357-373.

Miljković, D., Stare, T., Mozetič, I., Podpečan, V., Petek, M., Witek, K., Dermastia, M., Lavrač, N. & Gruden, K. (2012). Signalling Network Construction for Modelling Plant Defence Response. *PLOS ONE*, 7(12), pp. 1–18. https://doi.org/10.1371/journal.pone.0051822.

Myking, J. (2007). No Fixed Boundaries. In A. Bassey (ed.) *Indeterminacy in Terminology and LSP: Studies in Honour of Heribert Picht*, chapter 6. Amsterdam, The Netherlands and Philadelphia, USA: John Benjamins Publishing, pp. 73–91.

Nastase, V., Nakov, P., Séaghdha, D. Ó. & Szpakowicz, S. (2013). Semantic Relations Between Nominals. In G. Hirst (ed.) *Synthesis Lectures on Human Language Technologies*. London: Morgan & Claypool Publishers, pp. 1–119.

Navigli, R. & Velardi, P. (2010). Learning Word-Class Lattices for Definition and Hypernym Extraction. In *Proceedings of the Forty-Eighth Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden, pp. 1318–1327.

Pearson, J. (1998). Terms in Context. In E. Tognini-Bonelli & W. Teubert (eds.) *SCL Series, Vol. 1*. Amsterdam, The Netherlands and Philadelphia, USA: John Benjamins Publishing.

Pollak, S., Vavpetič, A., Kranjc, J., Lavrač, N. & Špela Vintar (2012). NLP workflow for on-line definition extraction from English and Slovene text corpora. In J. Jancsary (ed.) *Proceedings of KONVENS 2012*. ÖGAI, pp. 53–60. Main track: oral presentations.

Repar, A., Martinc, M. & Pollak, S. (2018). Machine Learning Approach to Bilingual Terminology Alignment: Reimplementation and Adaptation. In A. Branco, N. Calzolari & K. Choukri (eds.) *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Paris, France: European Language Resources Association (ELRA).

Repar, A., Podpečan, V., Vavpetič, A., Lavrač, N. & Pollak, S. (2019). TermEnsembler: An ensemble learning approach to bilingual term extraction and alignment. *Terminology*, 25(1).

Roller, S., Kiela, D. & Nickel, M. (2018). Hearst Patterns Revisited: Automatic Hypernym Detection from Large Text Corpora. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 358–363. URL https://www.aclweb.org/anthology/P18-2057.

Sclano, F. & Velardi, P. (2007). TermExtractor: a Web Application to Learn the Common Terminology of Interest Groups and Research Communities. In *Proceedings of the 9th Conf on Terminology and Artificial Intelligence TIA 2007*,

pp. 8–9.

Svensen, B. (1993). *Practical Lexicography: Principles and Methods Of Dictionary Making*. Oxford University Press.

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: 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 I. Kosem et al. (eds.) *Proceedings of eLex 2019*, pp. 305-318.

Zhang, Z., Gao, J. & Ciravegna, F. (2017). SemRe-Rank: Incorporating Semantic Relatedness to Improve Automatic Term Extraction Using Personalized PageRank. *arXiv preprint arXiv:1711.03373*.