Definition frames as language-dependent models of knowledge transfer

Špela Vintar & Larisa Grčić Simeunović

Abstract Definitions are an important means of structuring knowledge in a domain. We explore definitions in the domain of karstology from a cross-language perspective with the aim of comparing the cognitive frames underlying defining strategies in Croatian and English. The experiment involved the semi-automatic extraction of definition candidates from our corpora, manual selection of valid examples, identification of functional units and semantic annotation with conceptual categories and relations. Our results comply with related frame-based approaches in that they clearly demonstrate the multidimensionality of concepts and the key factors affecting the choice of defining strategy, e.g. concept category, its place in the conceptual system of the domain and the communicative setting. Our approach extends related work by applying the frame-based view on a new language pair and a new domain, and by performing a more detailed semantic analysis. The most interesting finding, however, regards the cross-language comparison; it seems that definition frames are language- and/or culture-specific in that certain conceptual structures may exist in one language but not the other. These results imply that a cross-linguistic analysis of conceptual structures is an essential step in the construction of knowledge bases, ontologies and other domain representations.

Keywords definitions, frame-based terminology, karstology, cross-language comparison, definition types, defining strategies, English-Croatian corpus

1 Introduction

„Any scientific research in any field of study strives to establish a maximum of certainty and control in the field of categorization.“ (Andersen 2007: 3) Definitions represent the core of conceptual structuring in a domain. According to the Aristotelian scholastic principle concepts must be categorised into classes and definitions represent the formal link between the concept (*definiendum*), its parent concept (*genus*) and the differentiating set of properties, allowing its assignment into a particular class (*differentia*). In the general theory of terminology (GTT; Wüster 1974 or Felber 1984) the definition plays an equally central role of concept delineation, in other words, the pinning down of meaning before the assignment of the (linguistic) designation.

While this neat and logical approach is still useful in practical tasks such as terminography, recent decades have brought dramatic shifts in the understanding of concepts, definitions, and the linguistic reality of (intercultural) communication. Central points include the following:

1. The transient and dynamic nature of concepts (Temmerman 1997, 2000, Kageura 2002).
2. The inherent mismatch between the term definitions written in natural language and concept definitions written in a formal language (Roche et al. 2009).
3. The inadequacy of definition typologies (Seppälä 2007).
4. The inadequacy of the onomasiological approach to tackle variation, register, style (Pecman 2014, De Santiago 2014).
5. Insights from cognitive science providing evidence that concepts are in fact layers of associative networks which are "reloaded" each time they are evoked (Faber 2009, 2012).
6. The fuzziness and indeterminacy of definitions (Leitchik/Shelov 2007).
7. Multilingual and cross-cultural aspects (Faber 2012).

In the present study, we do not dwell on all of these aspects, although in our opinion, they open up exciting new spaces to explore. Instead we focus on the theoretical vs. applied nature of definitions across registers and languages, and explore the multiple formal and semantic dimensions of definitions through a corpus-based analysis. The empirical part of our study is based on the cognitive model of terminology as proposed by Faber (2009), because we find that definition frames provide a helpful tool in exploring the multidimensionality of concepts, especially if we seek to demonstrate that the cognitive-semantic components chosen to define a concept in a particular context vary. This view is in line with Sager's (1990: 16) postulate that the value of a concept is defined as a range, which means that it occupies a set of points on a given axis.

In a previous monolingual study (Grčić Simeunović/Vintar 2015) we explored the multidimensionality of concepts in karstology and their dependence on register, subdomain and author. Here we examine definitions in English and Croatian didactic and scientific texts in the domain of karstology, and by comparing different definitions of the same concept, we show that the choice of the defining strategy is influenced by a number of factors, including the perspective from which the concept is described in an

interdisciplinary domain, register (didactic vs. scientific) and language (English vs. Croatian).

The described analysis is novel in that it applies the principles of frame-based terminology to a new domain and a less-researched language pair. Our specific aim is to demonstrate the multidimensionality of concepts and the dependence of definition frames on context and register. In particular we aim to compare selected dimensions of definition frames across two languages. As part of this analysis we also devise a methodology for semi-automatic definition extraction and a detailed framework for semantic annotation.

2 Defining definitions

Formulating a definition is an essential step of terminographical description and traditionally relies on nominalist philosophy and logical positivism, both clearly discernible in Wüsterian principles of terminography. With the advent of corpus-based methods as well as language technologies targeting the automatic identification of terms and definitions in texts, the highly abstract *in vitro* view of definitions was no longer viable. Pearson (1998: 33) writes: "We think that a definition which does not take text function and target readership into account will run into difficulty because authors write for a purpose and for a readership and they tailor their language accordingly."

Apart from Pearson (1998), who first proposed a descriptive and corpus-driven approach to terminological description, numerous authors (Cabré 1999, Temmerman 2000, Gaudin 2002, Kageura 2002, L'Homme 2004, Faber 2012) have challenged the rigidity of the traditional approach. The purpose of defining is still to delineate the meaning of the term and to embed the concept in the conceptual network of a domain, but this embedding inevitably takes place under the influence of a number of pragmatic factors, such as discourse type, expected level of prior knowledge in the target audience, culture, and language (cf. Blanchon 1997, Diki-Kidiri 2000, Madsen/Thomsen 2008).

There is little consensus among linguists about what constitutes a definition and how to classify different definition types. The majority of classification attempts stem from lexicography, although various other categories are defined in philosophy depending on their function (Parry/Hacker 1991, Copi/Cohen 2009: 88). For lexicographical purposes, the most common type is the analytical definition, usually

expressed in a single phrase. The variability of defining strategies is illustrated by Svensen (1993: 117), who distinguishes between true (intensional) definitions, paraphrases (also including synonyms and near synonyms), combined definitions, and definitions by describing the use of the defined term. Subtypes of intensional definitions include the following: (i) relational definitions, in which terms are defined by their relation (other than synonymy) to other terms; (ii) operational definitions, which state that a term is applied correctly to a given case if the performance of specified operations yields a specific result; (iii) functional definitions, which define a term by explaining its use; (iv) typifying definitions, which define a term by means of its typical properties.

In related literature (Shelov 1990, Béjoint 2000, Westerhout 2010: 37) we find other categories, such as quantitative definitions, which describe the dimensions (size, weight, length, age, etc.) of the definiendum (e.g. "A mountain is a peak that rises over 2,000 feet"), qualitative definitions which state the qualities, characteristics, or properties of the definiendum, enumerative or extensional definitions which list all subordinate concepts of the definiendum, and contextual definitions. Seppälä (2007) describes several criteria that characterise definitions in order to show that definition typologies, as they exist in the literature, are insufficient to understand the real nature of terminological definitions.

Empirical analyses of authentic texts confirm that defining strategies can be multifarious and highly dependent on register, domain, and style of writing (Pollak 2014). Pollak (2014) explored definition types in an English and Slovene corpus of language technologies as a step preceding the design of a definition extraction algorithm. Not only did she identify over 20 definition types, but she also arrived at the conclusion that almost 40 % of the definition candidates were borderline cases which could be regarded as definitions or not. A validation experiment with 20 students who were required to mark sentences as either definitions or non-definitions resulted in inter-annotator agreement of 0.36 (kappa), which is very low.

Automatic extraction of definitions from text is a well-researched topic within Natural Language Processing. Many early approaches to definition extraction relied on morphosyntactic patterns presupposing the analytical definition type (Klavans/Muresan 2001), later extended with more sophisticated grammars or lattices (Navigli/Velardi 2010). Several approaches use machine learning techniques to distinguish between

definitions and non-definitions (Fišer/Pollak/Vintar 2010), and the combination of a base grammar and a classifier proved most successful than either of these techniques used alone (Degórski/Marcinczuk/Przepiórkowski 2008, Westerhout 2010). A common problem to all these attempts is low recall and/or low accuracy when extracting definitions from highly unstructured noisy corpora.

In our own approach we were concerned first with extracting definition candidates from text, then with distinguishing between definitions and non-definitions and finally with identifying the semantic constituents of each definition. The awareness of the high variability of definitions and the knowledge of different typologies helped us both in the semi-automatic extraction and the validation phase. While the use of lexico-syntactic patterns inherently assumes certain formal characteristics of defining contexts, we deliberately also included some lexical triggers which did not presuppose a certain syntactic structure (e.g. *term*; see section 4.1). Still, we were frequently faced with the dilemma of whether a certain candidate sentence was to be considered a definition in the given context; our selection criteria are discussed in section 4.1.

3 Frame-based Terminology and its application to a cross-language study of definitions

Frame-based Terminology is a relatively recent attempt to reconcile the conceptual/cognitive layers of specialised knowledge and the textual reality. It responds to several of the pressing issues mentioned in the introduction, including the inadequacy of the traditional approach to handle variation, multidimensionality, and cross-language-related phenomena.

Frame-based Terminology uses a modified and adapted version of Fillmore's Frames (Fillmore 1976) coupled with premises from Cognitive Linguistics to configure specialised domains on the basis of definitional templates and to create situated representations of specialised knowledge concepts (Faber 2002, 2012, Faber et al. 2006). The definition templates are based on corpus evidence from which typical concept features and relations are extracted and subsequently mapped to a framework of categories.

The definition patterns of individual conceptual categories are represented by combining dynamic semantic roles such as AGENT, PATIENT, INSTRUMENT,

LOCATION etc. on the one hand with concept classes such as ENTITY, EVENT, PROPERTY or PHYSICAL OBJECT on the other. The conceptual structure of the domain is described via events or situations governed by non-hierarchical semantic relations between the concept classes, e.g. *causes, measures, has_function, has_form*. Such semantic frames represent possible cognitive structures used to define the meaning of a terminological unit. However, instead of being abstract, they are based both on past experience and expert knowledge and on the frequency of contextual patterns.

We adapted this model to the domain of karstology, which seems particularly well suited to such categorisation. Firstly, the domain is interdisciplinary in that it may be studied from a geographical, geomorphological or hydrological perspective with the possibility of further extensions into ethnology, agriculture, history and many other fields (Laurini 2013). One of our intentions was to demonstrate the multidimensionality of definitions with regard to the perspective of description chosen in a particular context. For instance, special attention is given to examples where the same concept is defined via different genus concepts. This is in consonance with Faber's Frame Semantics approach which proposes elaborating hierarchies of meaning within lexical fields.

Secondly, the process-oriented view seems a natural and intuitive way of modelling karst phenomena, where multiple environmental factors (agents) affect limestone rocks and result in various typical landforms. A prototypical event in karstology could be modelled with the following frame:

Natural AGENT: *erosion, tectonics* → causes process: *dissolution, sedimentation* → affects PATIENT: *rock, limestone* → results in: *uvalas, dolines, caves*.

For our cross-language analysis of definitions in the karstology domain, we adapted the model proposed by Faber (2012) for EcoLexicon (ecolexicon.ugr.es) and introduced several additional concept categories and semantic relations. However, as Faber (2012: 120) points out, any categorization of concepts into classes is in all likelihood fuzzy and dynamic, which is why we should expect concepts to appear in several categories, and specific dimensions of concepts may be activated in specific contexts. We aim to demonstrate this aspect through corpus-based evidence, and even more importantly, we

wish to compare these dimensions across languages. In the following sections we show that cognitive patterns, insofar as they can be discerned from definition frames, are also language-dependent.

4 Empirical analysis of definition frames across registers and languages

Our corpus-based analysis of definitions was performed on a comparable English-Croatian corpus of karstology, where for each language the corpus consisted of two subcorpora, one containing scientific texts (doctoral dissertations, scientific papers, conference proceedings) and the other, didactic texts (textbooks and lecture notes). Both corpora are comparable in size: the Croatian corpus contains 881,174 tokens, whereas the English corpus has 913,416 tokens (see table 1). The corpus was compiled within the framework of the doctoral research carried out by one of the authors of this paper (Grčić Simeunović 2014) and contains authentic, relevant and contemporary works on karstology, which were selected with the help of a domain expert. The English and the Croatian corpora can be considered comparable in terms of domain and text types included, but the number of tokens in the subcorpus of scientific texts is larger in Croatian.

Table 1: Basic corpus data

		English	**Croatian**
Scientific	Number of texts	23	9
	Tokens	499,422	628,138
Didactic	Number of texts	17	9
	Tokens	413,974	253,036
Total	Number of texts	40	18
	Tokens	**913,416**	**881,174**

Both corpora received standard pre-processing including tokenisation, PoS-tagging and lemmatisation. For Croatian, pre-processing was performed by Nikola Ljubešić with a recently developed tagger (Agić/Ljubešić/Merkler 2013). For the pre-processing of English and for corpus querying we used the SketchEngine facilities (Kilgarriff et al. 2014).

Our analysis involved the following steps:

- extraction of definition candidates using lexico-syntactic patterns,
- validation of definition candidates, and
- annotation of definitions with semantic categories and relations.

In the following subsections these steps are described in more detail.

4.1 Extraction and validation of definition candidates

Definition candidates were extracted using a set of lexico-syntactic patterns, designed specifically for each language on the basis of previous research into definition extraction (Fišer/Pollak/Vintar 2010, Pollak 2014). Some of these patterns assume the traditional analytical definition (*[NP]-is-a-[NP]*), while others may contain only a trigger word or phrase (*term, be-defined-as*), and will therefore frequently capture definitions of an entirely different format. Croatian and English patterns are similar but not completely parallel. For example, the trigger word *term* has two near-synonyms in Croatian, and we used all three (*termin/naziv/izraz*).

Clearly these pattern lists are not exhaustive and other potentially fruitful expressions could also be used, but since we were not aiming for total recall, their yield was deemed satisfactory. Table 2 lists the patterns for Croatian and English, the number of candidates yielded by each pattern, and the number of definitions retained after manual validation.

Table 2: Definition extraction patterns and their productivity[1]

Croatian	**# candidates extracted**	**# defini-tions**	**% defini-tions**	**English**	**# candidates extracted**	**# definitions**	**% defini-tions**
naziv	455	84	18.46	*term*	444	64	14.41
izraz	169	5	2.96				
termin	25	14	56				
N-*biti*-N	345	28	8.12	N-*is-a*-	98	21	21.43

[1] The data in all the tables are sorted by frequency in English corpus.

				N			
				N-*be-used*	92	4	4.35
N-*predstavljati*	219	31	14.15	N-*represent*	81	3	3.70
nazivati-se	24	17	70.83	*be-called*	74	20	27.03
N-*biti*-A-N	134	12	8.95	N-*is-a*-A-N	71	9	12.68
definirati-se-kao	1	1	100	*be-defined-as*	45	32	71.11
sadržati	61	10	16.39	N-*contain*	40	4	10.00
N-*značiti*	133	3	2.25	N-*mean*	27	2	7.41
zvati-se	12	2	16.67	N-*refer-to*	15	3	20.00
N-*sastojati-se*	18	2	11.11				
možemo-podijeliti-na	6	0	0				
proces-Ng	106	11	10.38				
Total	1708	220			987	162	

The manual validation performed by the authors of this paper was not an easy task, especially considering the variability of definitions discussed in section 2. We retained sentences which contained an explanation of the definiendum in any form by giving at least one distinguishing feature. In this way, several sentences were retained although they contained no genus. As can be observed in table 2, the majority of candidate sentences were still discarded, and several cases, not listed above, were either marked as borderline or as KRC (knowledge-rich context). In the end we limited our analysis only to true definitions and ignored semi-definitions and KRCs, even though they also

contained important conceptual relations. Some definitions were extracted via several patterns. After removing duplicates, the final data set consisted of 191 examples for Croatian and 142 for English.

4.2 Annotating definitions with conceptual categories and relations

For this step we first needed to define the domain-specific categories and relations to be used in annotation. A preliminary classification of karstology concepts into semantic classes had been previously performed by Grčić Simeunović (2014), which was a useful starting point. For pragmatic reasons semantic classes were added during annotation in case the need arose. As a result, the final inventory consisted of 30 classes, including: *limestone area, landform, water cycle, opening, process, measure, method, layer, minerals, rock characteristics, substance, territory, physical phenomenon, information system, situation, geographical boundary* etc.

In each definition we first identified the *definiendum* (the concept being defined) and the genus (superordinate concept), when present. Those two concepts were then assigned to a semantic class in accordance with the information contained in the definition. For the remaining part of the definition, which in most cases represents the *differentia*, no further semantic classes were assigned. Instead, we identified the semantic relations activated by the context. The following example illustrates this procedure:

Definition sentence:

Less permeable rock below an aquifer that keeps groundwater from draining away is called a confining bed (also known as aquitard or aquiclude).

Table 3: Example of semantic categories and relations found in a definition

Definiendum:	*confining bed / aquitard / aquiclude*
Definiendum class:	**hydrological form**
Genus:	Rock
Genus class:	**Mineral**
Differentia:	less permeable
Relation:	**has_attribute**
Differentia:	below an aquifer

Relation:	**has_location**
Differentia:	keeps groundwater from draining away
Relation:	**has_function**

Thus, a *hydrological form* is defined by specifying the *attribute*, *location* and *function* of its genus.

We described our dataset with a total of 23 relations; they are listed with the frequencies for each language in table 4. The total number of relations in the dataset was 509.

Table 4: Semantic relations and their frequencies

Semantic relation	**CRO**	**ENG**	**LL**
has_location	51	47	1.12
has_form	52	42	0.16
has_attribute	20	29	**5.40**
defined_as	5	23	**18.49**
has_function	21	22	1.26
caused_by	29	19	0.18
result_of	6	18	**10.36**
contains	1	15	**19.20**
made_of	16	10	0.19
has_result	3	8	**4.08**
has_time_pattern	3	8	**4.08**
has_origin	0	7	**11.93**
causes	6	6	0.26
performed_as	6	4	0.03
similar_to	7	3	0.68
computed_as	6	0	**6.67**
transforms_into	0	2	3.41
has_part	6	2	1.08
time_of	0	1	1.70

used_for	1	1	0.04
controlled_by	0	1	1.70
affected_by	0	1	1.70
depends_on	0	1	1.70

Determining the semantic relation governing the relationship between the concept and its specific properties is not always straightforward, and in many cases, the distinctions between categories are difficult to draw. Our annotation preserved the AGENT – PATIENT and CAUSE – EFFECT dimensions of the karstological event, which is why we differentiate between the *causes* and *has_result* relations. This is illustrated by the examples below. In (1) *rainfall excess* is the natural agent causing *flooding*, while in (2) the *exposure of the river to the surface* happens as an effect of the *underground cavern collapsing*. In the first definition, we thus identified the relations of *causes* and *defined_as*, while the second definition contains the relations *caused_by, has_form* and *has_result.*

(1) *Threshold runoff has been defined as the amount of rainfall excess of a given duration necessary to cause flooding on small streams.*

(2) *Short steep-sided valleys caused by collapse of an underground cavern and exposing the river to the surface are called karst windows.*

Looking at the frequencies and the statistical significance as measured by log-likelihood ($p < 0.05$) of individual relations occurring in the Croatian versus the English corpus, there are many differences, especially as some relations occur in one language but not the other (see table 4). However, there are also a number of similarities. Both languages share the two most frequently observed relations *has_form* and *has_location*. For Croatian, the list continues with *caused_by, has_function* and *has_attribute*, and for English, with *has_attribute, defined_as* and *has_function*. These figures indicate that LOCATION, CAUSE, FUNCTION and ATTRIBUTE (physical or other) represent the key semantic properties of concepts in the domain of karstology regardless of language or register. In the Croatian dataset, 154 of 191 definitions contain at least one of the above relations, and in the English set, the number is 117 of 142. The main exception is

the group of definitions labeled with the relations *defined_as, computed_as, performed_as*. These sentences usually contain instrumental definitions, which explain the meaning of a formula, measure or experimental method (see examples 3 and 4).

(3) *Biodegradation is a complex process, but may be approximated by:* $ct = c_0 e^{-kt}$*, a decline analogous to radioactive decay, where ct = concentration of the degradable tracer at time t,* c_0 *= concentration of the conservative tracer, k = constant of decay.*

(4) *Groundwater tracing is a method of investigating underground water and contaminant transport by labelling water with identifiable tracer substances or physical properties.*

Once the dataset was annotated, we performed a quantitative and qualitative analysis of the results comparing definition frames across the two languages.

5 Analysing cross-language aspects of definition frames

5.1 Quantitative observations

The frequencies of semantic categories for the concepts defined in our karstology corpus reveal some thematic differences between our subcorpora (table 5). Apparently the Croatian texts contain a larger proportion of definitions for landforms (e.g. *hum, klanac, škrip, čučevac*), while the English texts seem to place a slightly greater emphasis on hydrological phenomena and forms as well as on different types of limestone areas, mainly karst itself or karst types (e.g. *karst, epikarst, bradikarst, fluviokarst*). These differences point to the irregularities of the term formation process where some realities are given a stable name or term in one language but not in the other. Interestingly, the English dataset does not contain a single definition of an underground form, which appears eight times in the Croatian dataset. In the English corpus, such concepts are only described in form of partial definitions or in KRC.

Table 5: Most frequent concept categories

Concept category	**CRO**	**EN**	**LL**
limestone area	10	26	**12.90**
Landform	50	24	3.25

water cycle	6	8	1.19
hydrological phenomenon	6	13	**5.13**
opening	13	7	0.49
process	5	7	1.19
measure	4	6	1.21
method	3	5	1.27
layer	2	4	1.40
minerals	8	4	0.44
rock characteristics	0	4	**6.82**
underground form	8	0	**8.89**

We were also interested in the distribution of semantic relations across the concept categories. The assumption that a certain conceptual category will be more likely defined via a specific set of relations, thus constituting typical definition frames for each category led us to formulate a cognitive model of the selected domain. While we might expect such frames to be universal (e.g. a landform may be described by its form regardless of language or register), we were particularly interested in verifying this assumption with our bilingual dataset.

Table 6: Cross-language comparison of relations occurring with selected concept categories

landform	CRO	EN	***process***	CRO	EN	***limestone area***	CRO	EN
has_form	31	14	*caused_by*	3	1	*has_location*	4	10
has_location	21	8	*has_location*	2	0	*result of*	0	10
caused_by	13	6	*has_attribute*	1	1	*caused_by*	3	7
made_of	9	1	*defined_as*	1	3	*has_attribute*	3	6
has_attribute	7	2	*computed_as*	1	0	*has_form*	3	5
similar_to	4	0	*has_result*	1	4	*has_result*	0	4

contains	0	5	*has_time_pattern*	0	2	*contains*	1	4
result_of	0	4	*causes*	0	1	*made of*	2	4
has_function	3	2				*has_function*	2	2
has_part	1	2						
causes	1	0						

Table 6 shows the relations occurring in a particular concept category typical of each language. For *landform* it seems that definition frames are universal at least in the top three relations. As might be expected, a landform is typically defined by specifying its form, location, and the natural process that contributed to its formation. The lower part of the list seems less aligned though, and there seems to be little correspondence between languages. A similar impression is conveyed by the list for *process*. Processes are usually not described in terms of their form or their composition, which explains the absence of relations such as *has_form, similar_to, made_of* and *has_part*. On the other hand, a process may be defined or even computationally modelled, and may exhibit a time pattern. The category *limestone area* was more frequent in the English subcorpus, but apart from this difference, we were surprised to find the *has_result* relation in English but not in Croatian. This relation is usually expected to occur with processes and not territories or areas.

Our study also found that for concept categories occurring fewer than 15 times, such quantitative cross-language comparisons bear little significance. We nevertheless detected the general patterns from which definition frames can be discerned, and certain observed differences provide clues for further exploration.

Before describing the qualitative analysis of a selected concept in both languages, we discuss the size of a typical definition frame, in other words the number of specific properties expressed through semantic relations. In our annotated corpus, the average definition contains two relations, and four of our definitions contain as many as four. This refers to examples from the English subcorpus, and, interestingly, all of them come from scientific (as opposed to didactic) texts. We suspected that the definition frame might be larger if the *definiendum* was a more sophisticated concept, however,

this does not seem to be the case. In fact, our most complex definitions were for *tidal creek, karst* (see example [5])*, hazard,* and *gravine*.

(5) *Karst is defined as a terrain, generally underlain by limestone or dolomite, in which the topography is chiefly formed by the dissolving of rock, and which may be characterised by sinkholes, sinking streams, closed depressions, subterranean drainage and caves.*
Relations: *made_of/has_form/contains/result_of*

5.2 Qualitative analysis of karst *and related terms*

An interesting observation mentioned above was the appearance of some agent-like relations in the context of defining concepts that we consider static, such as landform or terrain. Nevertheless, this only occurred in the English subcorpus. We thus decided to take a closer look at definitions of *karst* and related terms in both languages in order to see whether the resulting cognitive models of the domain overlapped.

The Croatian corpus contains 13 sentences defining either *karst* (*krš*, 3) or types of karst (*klastokrš* 2, *tektokrš*, *škrapavi krš, linearni krš, fluviokrš, boginjavi krš, hidrotermokarst, obalni krš*). While all of these *definienda* belong to the same category of *limestone area*, their *genus* concepts fall into two groups. More specifically, eight of the examples define *karst* or karst type as a 'kind of terrain, area or relief form', while four definitions choose the genus *pojava* ('phenomenon'). The *differentia* of the definitions contain the following relations: *has_location* (9), *made_of* (5), *caused_by* (3), *result_of* (2), *has_part* (2), *develops_from* (1).

The three Croatian definitions of *karst* (examples [6–8]) illustrate the context-dependence and multidimensionality of the concept *karst*. In example (6), *karst* is defined as a 'relief form developing on soluble rock' (*limestone, dolomite* etc.). This is not surprising since this definition belongs to the didactic part of our corpus, which is more specifically composed of textbooks. Example (7) is a less typical definition in that it focuses on the processes and agents contributing to the formation of karst. On the other hand, example (8) defines *karst* as a 'group of morphological and hydrological phenomena found on soluble rock'.

(6) *Krš je specifičan oblik reljefa koji se razvija na topivim stijenama (vapnenac, dolomit, sol, gips).*
[Karst is a specific relief form which develops on soluble rock (limestone, dolomite, salt, gypsum).]

(7) *Krš kao reljef na topivim stijenama predstavlja rezultat raznolikih i međusobno uvjetovanih čimbenika kao npr. litološkog sastava, kemijskih procesa, pukotinske cirkulacije vode, tektonskih pokreta, klimatsko-bioloških čimbenika, a u novije vrijeme sve više dolazi do izražaja i utjecaj čovjeka.*
[Karst as relief on soluble rocks represents the result of various and mutually interactive factors, such as the lithological composition, chemical processes, water circulation in crevasses, tectonic movements, weather- and biology-related factors, and in recent times increasingly human interventions.]

(8) *Krš je specifičan skup morfoloških i hidroloških pojava u topivim stijenama, prije svih vapnenačkim i dolomitskim […].*
[Karst is a specific set of morphological and hydrological phenomena occurring on soluble rocks, mostly limestone and dolomite […].]

The English subcorpus has as many as 25 definitions for *karst* (10) or its subtypes: *hydrothermal karst, hypogene karst* (2), *endokarst, epikarst* (3), *contact karst, bradikarst, ore-bearing karst, anomalous hydrothermal karst, heterogeneous karst, fluviokarst, doline karst, thermal karst*. The majority of the genus concepts used to define these terms belong to the categories ‘limestone area’, ‘territory’ or ‘relief form’, just as in Croatian. A surprising observation, however, was the fact that four definitions describe *karst* (or its subtype) as a ‘process’ (examples [9–12]), and as a ‘consequence’, *has_result* is one of its relations.

(9) *In the broadest sense, hydrothermal karst is defined as the process of dissolution and possible subsequent infilling of cavities in the rock by the action of thermal water.*

(10) *Here we introduce the working term "anomalous hydrothermal karst" to describe the hydrothermal process developing in zones where the steady-state thermal field of the hydrosphere is disturbed.*

(11) *In the most general terms, karst may be defined as a process of interaction between soluble rocks and different waters, as a result of which characteristic features develop on the Earth's surface and underground.*

(12) *Hypogene karst is defined as the formation of caves by water that recharges the soluble formation from below, driven by hydrostatic pressure or other sources of energy, independent of the recharge from the overlying or immediately adjacent surface.*

This observation supports the view that the cognitive structures governing knowledge presentation in a specialised text are not universal and depend not only on context, register, or the author's beliefs, but also on the language in which the definition is formulated. Our corpus-based evidence shows that the definition frame [limestone area] *is_a* [process] *has_result* [result] is possible in English, but not in Croatian. Quite possibly, the concept of karst activates slightly different layers of meaning for a speaker of Croatian (or Slovene), because the term originates from the geographical area Kras and thus bears a strong associative link to a (static and physically identifiable) landscape.

This finding was unexpected in the context of our study, however the relationships between language, thought and natural landscapes have been addressed by several authors. Smith and Mark (2003) for example explore the concept of MOUNTAIN in the context of building a universal ontology of geographical forms, and discuss the difficulties of unifying geographical concepts because the meanings associated with them are essentially linked with the human experience of landscapes. According to this, a MOUNTAIN may be perceived as an obstacle or a place of shelter, making it clear that our understanding of landscape forms is inevitably intertwined with our cultural perceptions. Burenhult and Levinson (2008: 138) go even further by arguing that landscape features "do not come presegmented by nature", and they demonstrate how the concept of MOUNTAIN evokes such diverse features of meaning in different languages that any attempt at a universal ontology of landscape forms must fail.

Returning to Smith and Mark (2003), who fruitfully combine philosophy and geography, we find a nice explanation of our karst-as-process finding:

> Since contemporary geomorphology is almost entirely concerned with understanding the processes that shape the Earth's surface, and with the question of how local elevations and slopes control the spatial distribution of those processes and their impacts, landforms-as-objects are in practice irrelevant to most subfields of geomorphology. (Smith/Mark 2003: 18)

6 Conclusions

In a previous monolingual study (Grčić Simeunović/Vintar 2015), we explored the multidimensionality of karstology concepts and the effects of register, context, and style on the range of concept properties chosen for the definition. This study extends those findings into the space of cross-language comparison. The results obtained seem to indicate that cognitive structures underlying knowledge transfer, of which specialised texts are a surface representation, are influenced by language and culture. While the concept of *karst* can only be defined as a type of terrain in Croatian, in English within certain contexts, it is described as a process.

This observation is interesting for a number of reasons. Firstly, it challenges the efforts to build language-independent domain representations, such as ontologies or semantic networks of the WordNet type. Secondly, it could have important implications for multilingual terminography, which for the most part remains rooted in the traditional concept-oriented approach and has so far included language or translation-specific information mostly in the form of collocations and phraseology. Finally, it would be worthwhile to fully understand the reasons why such profound differences in cognitive frames come to exist, even in the realm of specialised discourse. In the case of our experiment, we suspect that the relation between the "donor" and "receiver" language regarding the origin of terms may play a certain role, in the sense that karstology concepts might have initially evolved in a close relationship with the geographical (and cultural and linguistic) reality represented by Karst as a region. Given the dynamic nature of concepts, the layers constituting the cognitive boundaries of a concept may be restructured or modified through the transfer and expansion of knowledge to other languages and cultures, as well as through interdisciplinarity, the layers constituting the cognitive boundaries of a concept may be restructured or modified.

References

Agić, Željko/Ljubešić, Nikola/Merkler, Danijela (2013): „Lemmatization and Morphosyntactic Tagging of Croatian and Serbian." *Proceedings of BSNLP (The 4th Biennial International Workshop on Balto-Slavic Natural Language Processing)*. Sofia: Association for Computational Linguistics. 48–57.

Andersen, Øivin (2007): „Indeterminacy, Context, Economy and Well-Formedness in Specialist Communication." *Indeterminacy in Terminology and LSP: Studies in honour of Heribert Picht*. Ed. Bassey E. Antia. Amsterdam/Philadelphia: Benjamins. 3–14.

Béjoint, Henri (2000): *Modern Lexicography: An Introduction*. Oxford: Oxford University Press.

Blanchon, Elisabeth (1997): „Point de vue en terminologie." *Meta. Translators' Journal* 42.1: 168–173.

Burenhult, Niclas/Levinson Stephen C. (2008): „Language and landscape: a cross-linguistic perspective." *Language Sciences* 30.2: 135–150.

Cabré, Maria Teresa (1999): *Terminology: Theory, methods, applications*. Amsterdam/Philadelphia: Benjamins.

Copi, Irving M./Cohen, Carl (2009): *Introduction to Logic*. 13th ed. Upper Saddle River NJ: Prentice Hall.

Degórski, Lukasz/Marcinczuk, Michal/Przepiórkowski, Adam (2008): „Definition Extraction Using a Sequential Combination of Baseline Grammars and Machine Learning Classifiers." *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*. Marrakech: ELRA (European Language Resources Association).

De Santiago, Paula (2014): „De la forma al contenido, del contenido a la definición." *Normas: Revista de Estudios Lingüísticos Híspanicos* 6: 28–44.

Diki-Kidiri, Marcel (2000): „Une approche culturelle de la terminologie." *Terminologies Nouvelles – Rifal (Réseau international francophone d'aménagement linguistique)* 21: 58–64.

Faber, Pamela (2002): „Terminographic Definition and Concept Representation." *Training the Language Services Provider for the New Millennium*. Ed. Belinda Maia/Johann Haller/Margherita Ulyrich. Porto: Universidade do Porto. 343–354.

Faber, Pamela/Montero Martínez, Silvia/Castro Prieto, María Rosa/Senso Ruiz, José/Prieto Velasco, Juan Antonio/León Arauz, Pilar/Márquez Linares, Carlos/Vega Expósito, Miguel (2006): „Process Oriented Terminology Management in the Domain of Coastal Engineering." *Terminology* 12.2: 189–213.

Faber, Pamela (2009): „The Cognitive Shift in Terminology and Specialized Translation." *MonTI – Monografías de Traducción e Interpretación* 1: 107–134.

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

Felber, Helmut (1984): *Manuel de terminologie*. Paris: UNESCO, Infoterm.

Fillmore, Charles J. (1976): „Frame Semantics and the Nature of Language." *Origins and Evolution of Language and Speech. (Annals of the New York Academy of Sciences 280).* Ed. New York Academy of Sciences. 20–32.

Fišer, Darja/Pollak, Senja/Vintar, Špela (2010): „Learning to Mine Definitions from Slovene Structured and Unstructured Knowledge-Rich Resources." *Proceedings of the Seventh International Conference on Language Resources and Evaluation: LREC*. Ed. Nicoletta Calzolari et al. Malta, Valletta: ELRA/ELDA/ILC. 2932–2936.

Gaudin, François (2002): *Socioterminologie: Une approche sociolinguistique de la terminologie.* Louvain-la-Neuve: Duculot De Boek Université.

Grčić Simeunović, Larisa (2014): *Methodology of Terminological Description for the Purposes of Specialized Translation.* Unpublished PhD thesis (in Croatian). Zadar: University of Zadar.

Grčić Simeunović, Larisa/Vintar, Špela (2015): „Domain Modelling: Comparative Analysis of Definition Styles." *Od Šuleka do Schengena.* Ed. Maja Bratanić et al. (in Croatian). Zagreb: Institut za hrvatski jezik i jezikoslovlje. 251–266.

Kageura, Kyo (2002): *The Dynamics of Terminology: A Descriptive Theory of Term Formation and Terminological Growth*. Amsterdam/Philadelphia: Benjamins.

Klavans, Judith L./Muresan, Smaranda (2001): „Evaluation of DEFINDER: A System to Mine Definitions from Consumer-oriented Medical Text." *Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries*. New York: ACM (Association for Computing Machinery). 201–202.

Kilgarriff, Adam/Baisa, Vít/Bušta, Jan/Jakubíček, Miloš/Kovář, Vojtěch/Michelfeit, Jan/Rychlý, Pavel/Suchomel, Vít (2014): „The Sketch Engine: Ten Years on." *Lexicography* 1: 7–36.

Laurini, Robert (2013): „A Conceptual Framework for Geographic Knowledge Engineering." *Journal of Visual Languages and Computing* 20.1: 2–19.

Leitchik, Vladimir M./Shelov, Serguey D. (2007): „Commensurability of Scientific Theories and Indeterminacy of Terminological Concepts." *Indeterminacy in Terminology and LSP: Studies in honour of Heribert Picht*. Ed. Bassey E. Antia. Amsterdam/Philadelphia: Benjamins. 93–106.

L'Homme, Marie Claude (2004) *La terminologie: principes et techniques*. Montréal: Presses de l'Université de Montréal.

Madsen, Bodil Nistrup/Thomsen, Hanne Erdman (2008): „Terminological Principles Used for Ontologies." *Managing Ontologies and Lexical Resources: 8th International Conference on Terminology and Knowledge Engineering*. Ed. Bodil Nistrup Madsen/Hanne Erdman Thomsen. Copenhagen: Litera. 107–122.

Navigli, Roberto/Velardi, Paola (2010): „Learning word-class lattices for definition and hypernym extraction." *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Uppsala: ACL (Association for Computational Linguistics)*. 1318–1327.

Parry, William Thomas/Hacker, Edward A. (1991): *Aristotelian Logic*. Albany: University of New York Press.

Pecman, Mojca (2014): „Variation as a Cognitive Device: How Scientists Construct Knowledge through Term Formation." *Terminology* 20.1: 1–24.

Pearson, Jennifer (1998): *Terms in Context*. Amsterdam/Philadelphia: Benjamins.

Pollak, Senja (2014): *Semi-automatic Domain Modeling from Multilingual Corpora*. Unpublished PhD thesis. Ljubljana: Department of Translation Studies, Faculty of Arts.

Roche, Christophe/Calberg-Challot, Marie/Damas, Luc/Rouard, Philippe (2009): „Ontoterminology: A New Paradigm for Terminology." *International Conference on Knowledge Engineering and Ontology Development*. Madeira, Portugal. 321–326.

Sager, Juan C. (1990): *Practical Course in Terminology Processing*. Amsterdam/Philadelphia: Benjamins.

Seppälä, Selja (2007): „La définition en terminologie: typologies et critères définitoires.” *TOTh Terminologie et Ontologie: théories et applications.* Annecy: Institut Porphyre. 23–44.

Shelov, Serguey D. (1990): „Typology of term definitions (comparison of normative and non-normative terminological dictionaries).” *Nauchno-Tekhnicheskaya Terminologiya 4* (in Russian): 16–27.

Smith, Barry/Mark, David M. (2003): „Do Mountains Exist? Towards an Ontology of Landforms.” *Environment & Planning B: Planning & Design* 30.3: 411–427.

Svensen, Bo (1993): *Practical Lexicography: Principles and Methods of Dictionary Making*. London: Oxford University Press.

Temmerman, Rita (1997): „Questioning the univocity ideal. The difference between socio-cognitive Terminology and traditional Terminology.” *Hermes. Journal of Linguistics* 18: 51–91.

Temmerman, Rita (2000): *Towards New Ways of Terminological Description. The Sociocognitive Approach*. Amsterdam/Philadelphia: Benjamins.

Westerhout, Eline (2010): *Definition extraction for glossary creation: A study on extracting definitions for semi-automatic glossary creation in Dutch*. (Lot Dissertation Series 252.) Utrecht: LOT (Landelijke Onderzoekschool Taalwetenschap).

Wüster, Eugen (1974): „Die Allgemeine Terminologielehre – ein Grenzgebiet zwischen Sprachwissenschaft, Logik, Ontologie, Informatik und den Sachwissenschaften.“ *Linguistics* 119: 61–106.

Špela Vintar
University of Ljubljana
Faculty of Arts
Aškerčeva 2
1000 Ljubljana, Slowenien
Tel: +386 1 2411000
Fax: +386 1 4259337
spela.vintar@ff.uni-lj.si

Larisa Grčić Simeunović
University of Zadar
Sveučilište u Zadru
Ul. Mihovila Pavlinovića
23000, Zadar, Kroatien
lgrcic@unizd.hr