Data Collection and Definition Annotation for Semantic Relation Extraction

Jasna Cindrič, Lara Kuhelj, Sara Sever, Živa Simonišek, Miha Šemen

Department of Translation, Faculty of Arts, University of Ljubljana
Aškerčeva cesta 2, SI-1000 Ljubljana
jasna.cindric@gmail.com
larakuhelj@gmail.com
seversara@gmail.com
ziva.sim@gmail.com
miha.semen@gmail.com

Abstract

This paper presents the process of data collection, definition extraction and annotation for the purpose of semantic relation extraction based on English and Slovene texts related to geology, glaciology, and geomorphology. Automatic semantic relation extraction is an important task in NLP; its potential applications include information retrieval, information extraction, text summarization, machine translation, and question answering. This approach was based on the TermFrame project. The texts for the corpora were collected manually, while definitions were identified through targeted queries in SketchEngine and then semantically annotated using the WebAnno tool. Our research showed some significant differences between languages resulting in some difficulties during the annotation process.

1. Introduction

This paper describes the process of definition extraction, annotation and curation based on corpora created for a research project carried out by Master's students as part of the module Corpora and Localisation at the Department of Translation Studies, Faculty of Arts (University of Ljubljana). Translation students collaborated with their peers from the Faculty of Computer and Information Science (University of Ljubljana) on a project focusing on the automatic extraction of semantic relations, which required the creation of an English and a Slovene corpus and the provision of an additional data set annotated for semantic relations. We describe the process of corpus building, the identification and extraction of definitions, followed by the annotation and curation using the WebAnno annotation tool. Finally, the paper illustrates the results and obstacles as well as discusses possible further work and research.

Corpus-based automatic semantic relation extraction has become one of the main topics in corpus linguistics. Domain-specific annotated corpora are the basis for the design of many NLP systems for relation extraction (Thanopoulos et al., 2000) and are considered knowledge sources on natural language use. It is imperative to obtain corpora large enough to provide a sufficient number of instances of relation pairs for extraction (Huang et al., 2015). This is especially true for Slovene, a language with complex morphology and free word order, which currently lacks readily available large domain-specific corpora (Pollak et al., 2012).

The layout of the project relied heavily on a similar dataset, TermFrame[1] – a trilingual knowledge base that contains Karst terminology in English, Slovene and Croatian. The knowledge base was developed on the basis of the frame-based approach in terminology (Pollak et al., 2019; Vintar et al., 2021; Vintar and Stepišnik 2020; Vintar et al., 2019; Vrtovec et al., 2019), a cognitive approach to terminology that considers context, language and culture and focuses on specialised texts (Faber and Medina-Rull, 2017). Frame-based terminology is mainly used for the creation of multimodal specialised knowledge bases, where "frames" are used as a "representation that integrates various ways of combining semantic generalisations about one category or a group of categories" (Faber, 2015). Additionally, "templates" are used as a representation of parts of one category, and "templates" cover the cultural component (Faber, 2015).

Following the process of the TermFrame project, the team began with compiling an English and a Slovene domain-specific corpus, then extracting definitions and annotating them using the WebAnno tool (Castilho et al., 2016). This paper describes these steps in detail, followed by an analysis of the annotated definitions. It also highlights the obstacles the team faced during the conversion of texts and the annotation process.

The main goal of the project was to create an English and a Slovene corpus covering the fields of geomorphology, glaciology and geology, which would serve as a basis for definition extraction, annotation and curation.

2. Building the corpora

2.1. Text collection

For the purposes of our research, the linguist team compiled two corpora, one Slovene and one English. The entire project lasted for approximately one month.

The first step was to search for texts in both languages covering predefined topics, namely geology, glaciology, and geomorphology. These areas were chosen because they were semantically related to the domain of karstology, but had not yet been used in the TermFrame database. More specifically, the texts from neighbouring domains to karstology were assumed to contain the same semantic relations, so that our to-be-created data set could be fully compatible with the existing ones.

The linguist team was particularly interested in collecting scientific texts (scientific papers, articles, books,

[1] https://termframe.ff.uni-lj.si/.

doctoral and master's theses). Many of these texts can be found through the Digital Library of Slovenia[2] or through the Co-operative Online Bibliographic System & Services – COBISS[3], and through ResearchGate, a social networking site for scientists and researchers[4]. Ultimately, our team proposed 32 Slovene texts and 26 English texts as candidates. The proposed titles were validated by a domain expert and assessed as relevant.

The next step was to ensure that the texts were in a format that could be read by Sketch Engine[5], which proved difficult in some cases. Fortunately, most of the texts on dLib.si are available in TXT and PDF format. As a result, the team was able to access the texts in the appropriate format using Notepad. Texts that were suitable to the topic but could not be accessed in the correct format were omitted. Document conversion and text cleaning proved cumbersome (see Section 2.2). The team had one week to prepare the texts according to this process.

2.2. Creating the corpora

After collecting a sufficient amount of documents and successfully converting them into the appropriate formats, the team proceeded to create the corpora. As all team members had full access to Sketch Engine, we decided this would be the most efficient and straightforward tool for corpus creation and subsequent querying. Table 1 provides an overview and detailed information about both corpora.

	English	Slovene
Tokens	1,588,085	493,107
Words	1,284,564	358,731
Sentences	52,147	18,373
Documents	26	32

Table 1: Data on the English and Slovene corpus.

As can be seen from Table 1, the Slovene corpus was significantly smaller. This was due to the fact that longer Slovene texts were harder to find, which was to be expected, considering there are not as many Slovene sources as there are English ones.

As previously mentioned, arguably the most important challenge the team faced occurred after selecting the texts for the Slovene corpus. As most of them were in the form of PDF files, the team had to ensure they were searchable before converting them into text (TXT) files. Due to some language-specific characters, particularly diacritics, such as *č*, *š*, and *ž*, most of the widely available online converters failed to produce satisfactory results.

After a few unsuccessful attempts, we managed to convert them with Notepad++, but we still had to review the files and manually correct some errors before adding the documents to the corpus. Since the final text was corrected manually, man-made errors such as the inclusion of some elements, like the table of contents, English abstracts and reference lists that were unintentionally added to the final version of the corpus caused some difficulties when searching for potential definitions. Ultimately, it was impossible to rely entirely on conversion tools – this seemingly undemanding step required additional time and attention.

3. Definition extraction

In order to obtain the sentences containing definienda, definitors and genera, we had one week to extract the definitions from the corpora using targeted queries in SketchEngine. Searching for typical definition-like sentences can be done by searching for specific words or phrases and by CQL queries.

To some extent, the structure of definitions can be predicted. Typical definition structures in Slovene include "X je Y", "Y imenujemo (tudi) X". "izraz X pomeni Y", "izraz X označuje Y", "med Y štejemo (tudi) X" etc., while typical definition structures in English include "X means …", "X is a Y", "X is a kind of …", "The term X is …" or "X is defined as". (...) In this context, X is typically a hyponym and Y is a hypernym. Sketch Engine allows searching for such definitions in multiple ways. One method is to use a simple Sketch Engine query and search for words or phrases that are often included in the definitions, such as "imenujemo" or "izraz" in Slovene and "is a" or "is a term used to describe" in English. We were able to identify multiple definitions using this method, for example *"Tip kraškega površja, kjer je prevladujoča oblika vrtače, imenujemo vrtačasti kras."*

Another method is to use a CQL query in Sketch Engine and check for definitions with advanced filtering commands such as [tag="S.*"][word="je"][tag="S.*"] in Slovene or [tag="NN"][word="is"][word="a"]?[tag="N.*"] in English. This command combines a search for a specific part of speech (S.* – noun) and a specific word (je). An example of a definition identified by using the CQL query in Slovene is *"Uvala je večja kraška globel skledaste oblike z neravnim dnom in sklenjenim višjim obodom."* Another example in English is *"A coral reef is a ridge or mound built of the skeletal remains of generations of coral animals, upon which grow living coral polyps."*

Since not all definitions fit these typical structures, we used another strategy. We checked the keywords suggested by Sketch Engine and search for them with a simple query. In this way, we were able to identify various definitions which could not be found otherwise. An example of such a definition is *Slovenska kraška terminologija navaja, da je vrtača: depresijska oblika okroglaste oblike, navadno globoka več metrov in je bolj široka kot globoka*.

In addition to these strategies, the English team also utilised a glossary from the English corpus and extracted some of the definitions from there.

By combining all of these strategies, we were able to identify definition candidates suitable for annotation. The selected definitions were then verified by a terminology specialist. Some of the definitions were judged to be unsuitable, either due to their wording or for semantic reasons. After discarding the inadequate definitions, we retained 100 definitions from the Slovene corpus and 104

[2] https://www.dlib.si.

[3] https://www.cobiss.si.

[4] https://www.researchgate.net/.

[5] https://www.sketchengine.eu/.

definitions from the English corpus. All of them were then uploaded to WebAnno[6] to be manually annotated.

4. Definition annotation

The definitions were annotated using WebAnno – a web-based annotation tool, which allowed for a faster collaborative annotation process as well as a comparative evaluation of the annotations (Castilho et al., 2016). The annotation process took approximately ten days.

Altogether, the team annotated 100 Slovene and 104 English definitions, whereby four layers of information were considered. The layers were introduced to the linguist team by the course instructor and were, in term, selected because they had already been used in the TermFrame project (Vintar and Stepišnik, 2020). We believed that relying on the same categories that had already been adapted to karstology – a domain closely related to the ones chosen for this research – would ensure a straightforward annotation process with little to no ambiguities. Furthermore, the resulting data set would be fully compatible to the existing one in the TermFrame project. The layers of information include:

1. **Semantic category:** This layer covers the main semantic categories for **A. Landform** (A.1 Surface Landform, A.2 Underground Landform, A.3 Hydrological Landform or A.4 Other)**, B. Process** (B.1 Movement, B.2 Loss, B.3 Addition or B.4 Transformation)**, C. Geome, D. Element/Entity/Property** (D.1 Abiotic, D.2 Biotic, D.3 Property and D.3.1 Geolocation) and **E. Instrument/Method** (E.1 Instrument or E.2 Method). The semantic category was defined primarily for the definiendum and genus. Semantic categories are presented in Figure 1.
2. **Definition element:** Here, the term in question was marked as DEFINIENDUM, its hypernym or superordinate term as GENUS, the defining phrase (the phrase between the DEFINIENDUM and the GENUS – e.g. the phrase *is a*) as DEFINITOR and any of its hyponyms or subordinate terms as SPECIES.
3. **Semantic relation:** A set of 15 relations was used for annotating different features of the defined term: AFFECTS, HAS_ATTRIBUTE, HAS_CAUSE, CONTAINS, COMPOSITION_MEDIUM, DEFINED_AS, HAS_FORM, HAS_FUNCTION, HAS_LOCATION, MEASURES, HAS_POSITION, HAS_RESULT, HAS_SIZE, STUDIES and OCCURS_IN_TIME.
4. **Relation definitor:** This layer is associated with semantic relations and marks words or phrases that precede particular semantic relations (e.g. *in the* ocean).

WebAnno also offers an additional layer for the **canonical form**, which is used to ensure the full form of a term when it appears in an elliptic construction. The canonical form layer has been mostly used when annotating definitions in the Slovene corpus. One of the reasons for this is that ellipses are more common in Slovene. Another reason is

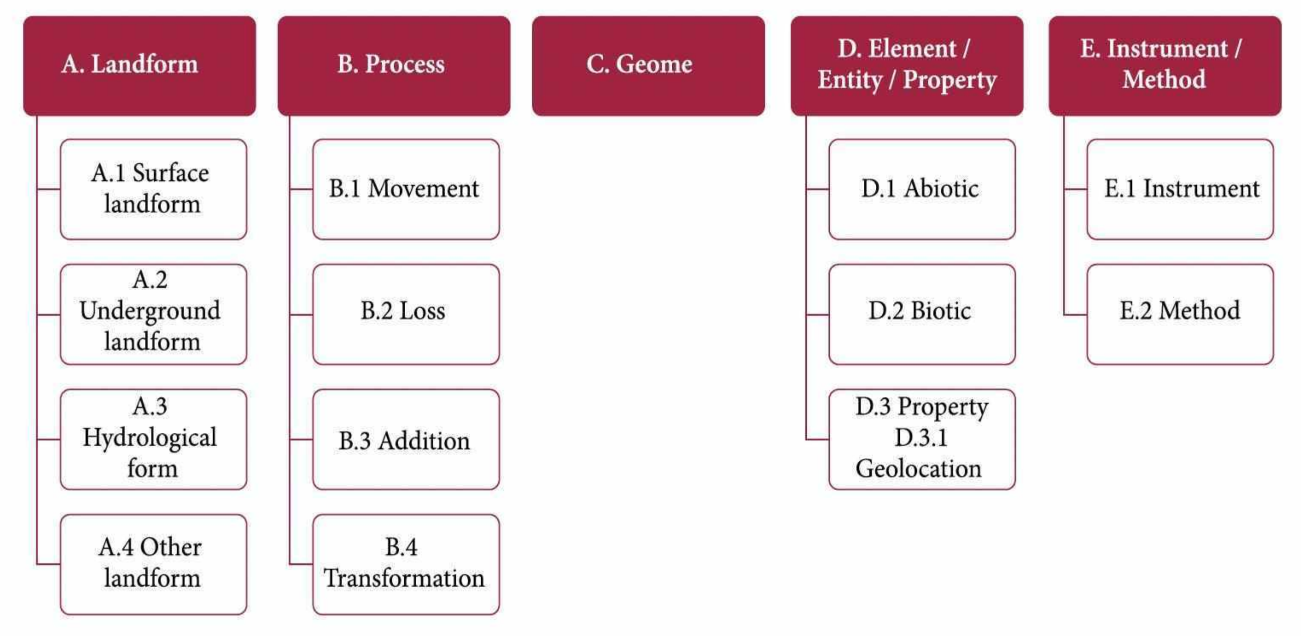
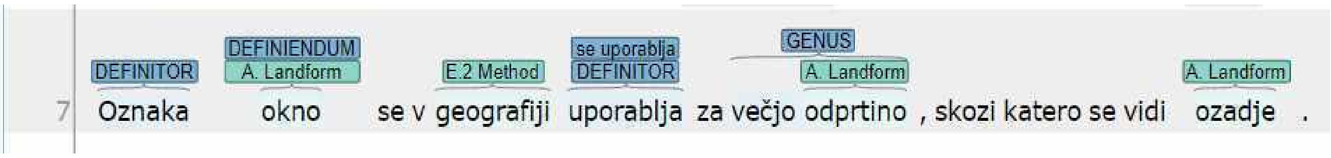


Figure 1: Semantic categories (Vintar and Stepišnik, 2021).

[6] https://www.clarin.si/webanno/login.html.



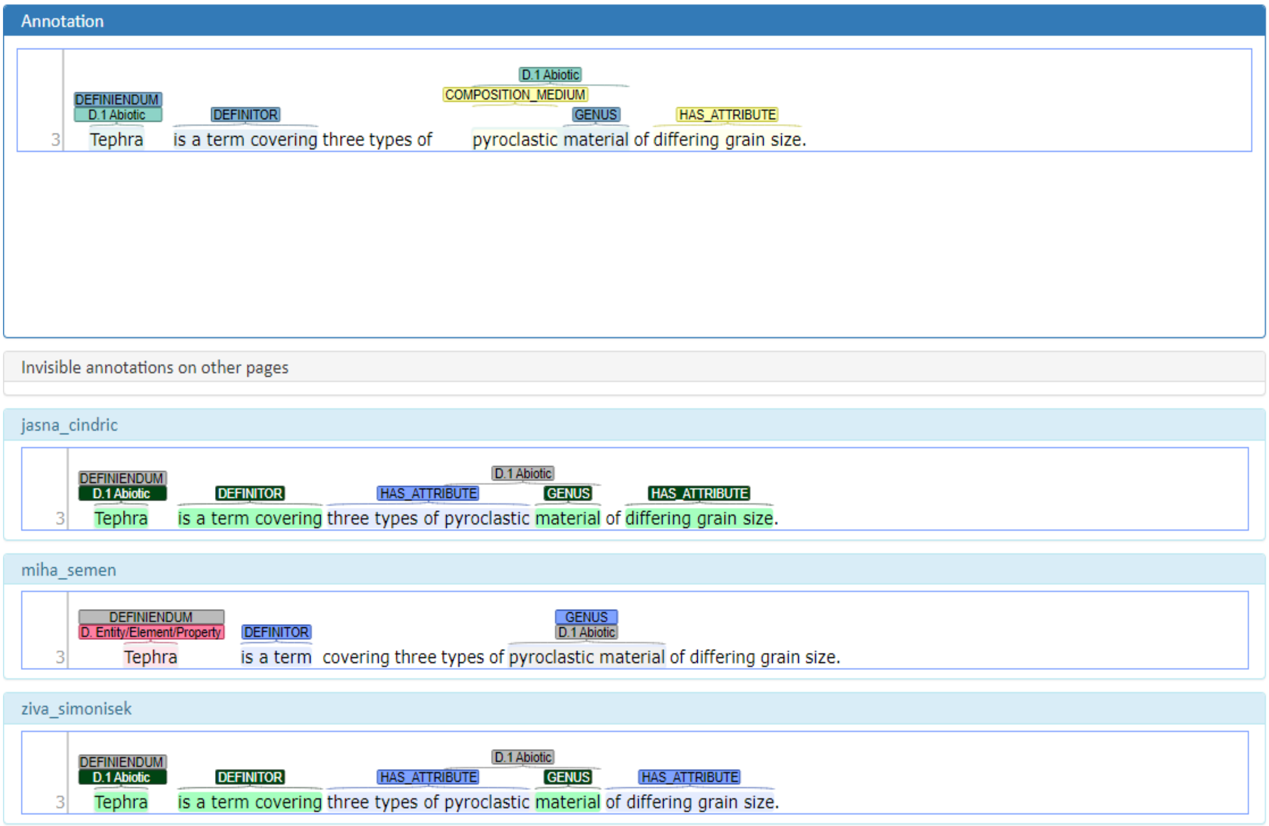
7 Oznaka okno se v geografiji uporablja za večjo odprtino , skozi katero se vidi ozadje .

Figure 2: Use of the term canonical layer for pairing the words "uporablja" and "se" to show they form a single unit.

that the predicate and the pronoun "se" are often separated by other words.

As seen from Figure 2, which shows the example of the use of the term canonical layer in the Slovene corpus, the predicate "se uporablja" consists of two words that act as a definitor. Hence, the team used the term canonical layer to pair the two words together.

For the purpose of this project, three students annotated English definitions, while two students annotated the Slovene ones. Afterward, in the process of curation, both teams jointly annotated the definitions with the course instructor's assistance. We observed that the annotation of definition elements (definiendum, genus and definitor) was the most straightforward, although the annotators' solutions still varied in some cases (See Figure 3). On the other hand, annotation of semantic categories, semantic relations and relation definitors proved to be more dubious since the annotations often differed from one another. When variations occurred, the team managed to resolve such dilemmas through discussions.



3 Tephra is a term covering three types of pyroclastic material of differing grain size.

3 Tephra is a term covering three types of pyroclastic material of differing grain size.

3 Tephra is a term covering three types of pyroclastic material of differing grain size.

3 Tephra is a term covering three types of pyroclastic material of differing grain size.

Figure 3: Curation process in WebAnno.

As Figure 3 shows, all three students who annotated English definitions chose "tephra" as the definiendum. Two students annotated the phrase "is a term covering" as the definitor and one student annotated only "is a term". The word "material" was determined to be a genus by two students, whereas one student extended the genus and annotated "pyroclastic material" – "pyroclastic" was later defined as COMPOSITION_MEDIUM.

5. Analysis

After annotating all of the extracted definitions, the linguist team wanted to take a closer look at the results. Each English definition had one definiendum, giving a total of 104 definienda, while the Slovene definitions had one or more definienda, 113 in total.

The most common definitor in English was "is a", followed by "are", and in Slovene "imenujemo" and "je". One or more genera were found in all English definitions, 112 in total, while not all Slovene definitions had a genus.

Figures 4 and 5 show the distribution of semantic categories for the annotated terms in Slovene and English. In total, 183 English and 334 Slovene terms were assigned categories. The most frequent category in English was D.1 Abiotic, followed by A.1 Surface landform. Similarly, A.1

Surface landform was the most frequent category in Slovene, followed by D.1 Abiotic.

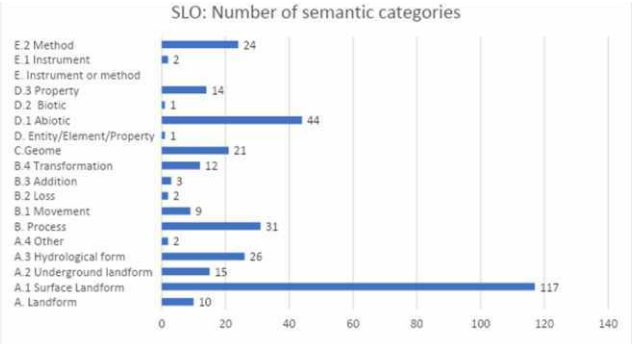


Figure 4: Semantic categories in the Slovene corpus.

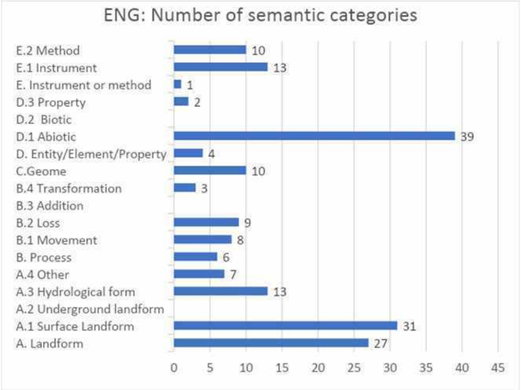


Figure 5: Semantic categories in the English corpus.

Figures 6 and 7 show the distribution of semantic relations for Slovene and English. A total of 186 relations were marked in English and 156 in Slovene. The most common relations in English were HAS_CAUSE (morphogenesis) and HAS_LOCATION (spatial distribution). On the other hand, the two most common relations in Slovene were HAS_FORM (morphography) and HAS_LOCATION (spatial distribution).

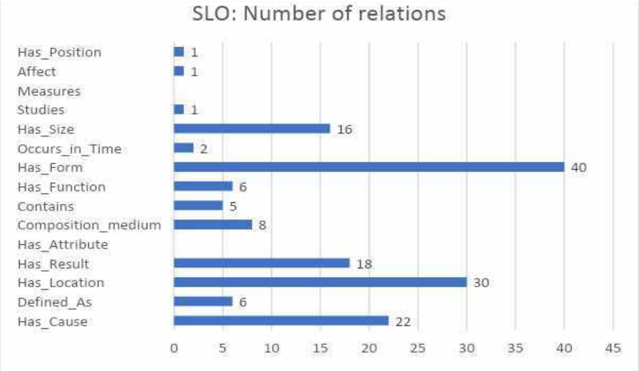


Figure 6: Number of semantic relations in the Slovene corpus.

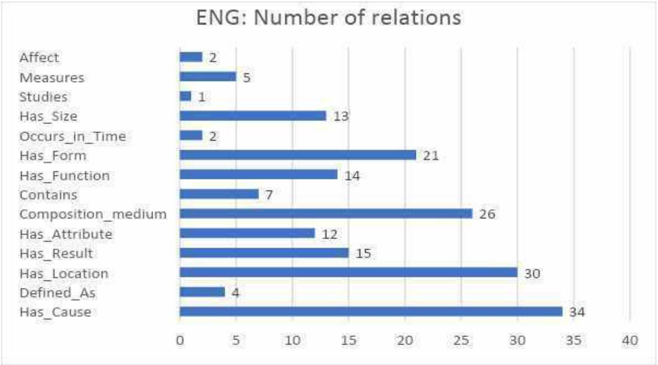


Figure 7: Number of semantic relations in the English corpus.

5.1. Annotation difficulties

During the annotation and curation process, the team encountered some complex cases, in particular when reviewing Slovene definitions, which required further discussion and careful attention. While annotating the definition element proved fairly straightforward, semantic relations posed some challenges.

The analysis showed ambiguities in 37 out of 65 sentences in the Slovene corpus. We have divided the ambiguities into the following categories.

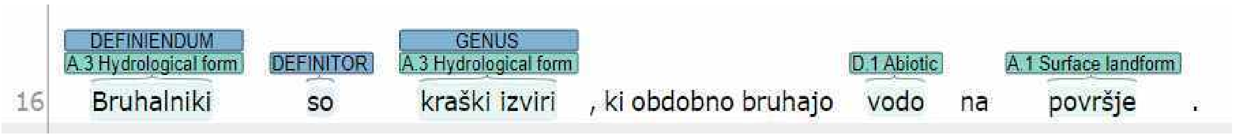
5.1.1. Phrases that could be placed in multiple categories

The most recurring ambiguity concerned phrases that could be classified into a number of categories, while others were difficult to associate with any of the possible labels. In many cases, the team had to determine how the annotators would deal with these ambiguous words and establish agreement on a consistent annotation strategy.

For example, the phrase “kraški izviri” in Figure 8 could semantically be understood as a hydrological form, a surface form, an underground form or an abiotic.

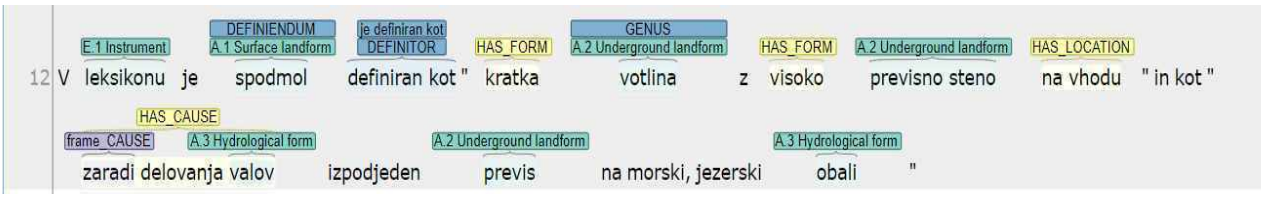
As in the previous example, the word “obala” in Figure 9 can be understood as a hydrological form, a surface form, an abiotic or a geome.

Although the word “kras” is most likely understood as geome, depending on the context, it can also be understood as karstology, the study of karst. In line with the decision to annotate “geomorphology” as a method, “kras” could therefore be annotated as a method as well as shown in Figure 10.



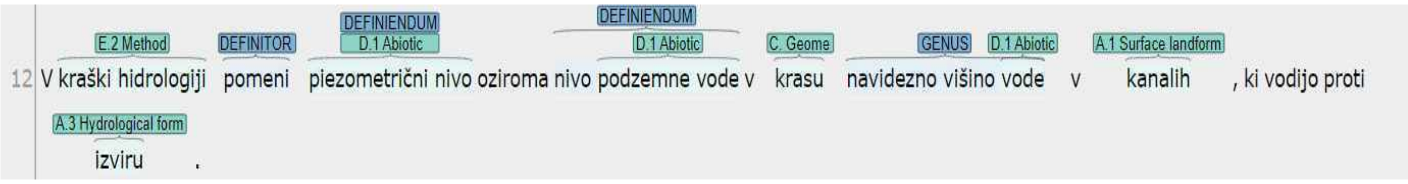
16 Bruhalniki so kraški izviri , ki obdobno bruhajo vodo na površje .

Figure 8: Example of an ambiguous annotation.



12 V leksikonu je spodmol definiran kot " kratka votlina z visoko previsno steno na vhodu " in kot " zaradi delovanja valov izpodjeden previs na morski, jezerski obali "

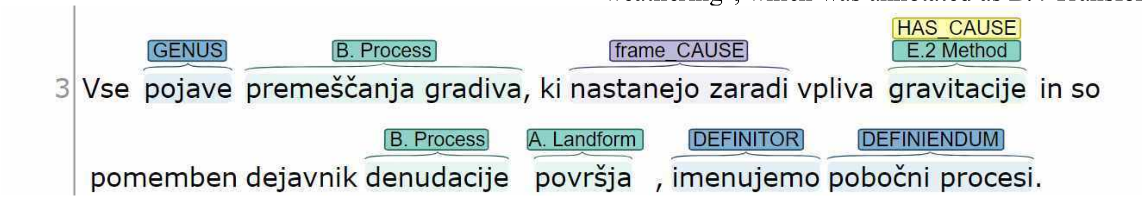
Figure 9: Example of an ambiguous annotation.



12 V kraški hidrologiji pomeni piezometrični nivo oziroma nivo podzemne vode v krasu navidezno višino vode v kanalih , ki vodijo proti izviru .

Figure 10: Example of an ambiguous annotation.

Another example was "gravitacija" (see Figure 11). It was extremely difficult to annotate a word denoting such a complex concept. In discussions with the course instructor, the team decided to annotate it as a method, as the names of the studies had to be annotated in the same way. However, it should be noted that the word could also be annotated according to other criteria.



3 Vse pojave premeščanja gradiva, ki nastanejo zaradi vpliva gravitacije in so pomemben dejavnik denudacije površja , imenujemo pobočni procesi.

Figure 11: Example of an ambiguous annotation.

5.1.2. HAS_FORM

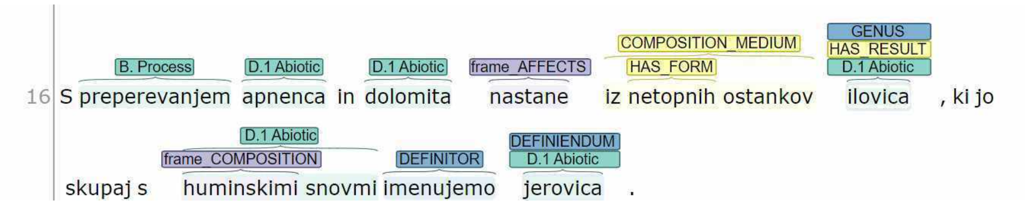
In a handful of cases annotating the Slovene definitions, it became clear that the semantic relation HAS_FORM manifests itself in different ways, as shown in Figures 12, 13 and 14.

Since HAS_FORM relations are more abstract and harder to grasp, annotation proved to be more difficult and required double-checking.

5.1.3. Annotation of genus

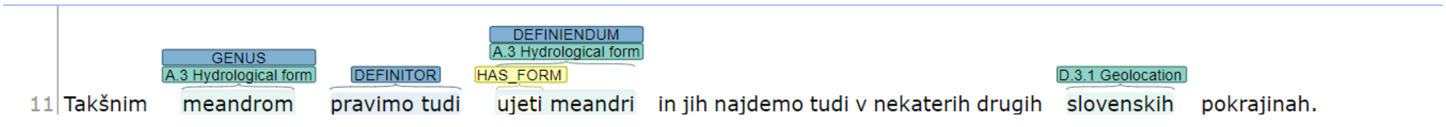
Sentences in the English corpus also posed some challenges, however their amount was significantly lower compared to their Slovene counterparts.

Before the annotation process, it was decided not to choose long phrases for the genus, but preferably just one word, e.g. "unloading of mountains" could be considered for the genus as a whole, but the team annotated only the word "unloading" as the genus. It was expected that genus and definiendum would share the same semantic category, since genus is a hypernym or superordinate term, but this was not the case for all definitions. For example, the definiendum "aquifer" was annotated as A.3 Hydrological form, but the genus "body of rock" was annotated as D.1 Abiotic in the same definition. This is because "body of rock" is not necessarily a hydrological form and can also be found on the surface. Another example is the definiendum "weathering", which was annotated as B.4 Transformation and the genus "process" was annotated as B. Process. The reason for this is that "process" is a hypernym of "transformation".



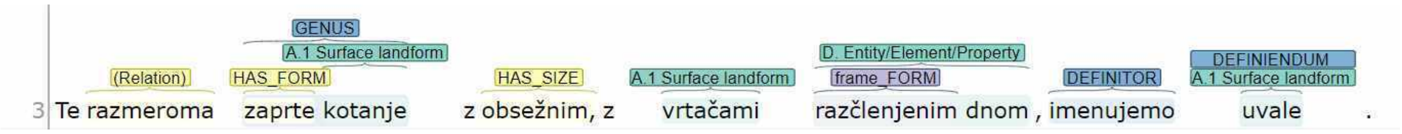
16 S preperevanjem apnenca in dolomita nastane iz netopnih ostankov ilovica , ki jo skupaj s huminskimi snovmi imenujemo jerovica .

Figure 12: HAS_FORM introduced by a preposition.



11 Takšnim meandrom pravimo tudi ujeti meandri in jih najdemo tudi v nekaterih drugih slovenskih pokrajinah.

Figure 13: HAS_FORM expressed with an adjective.



3 Te razmeroma zaprte kotanje z obsežnim, z vrtačami razčlenjenim dnom , imenujemo uvale .

Figure 14: HAS_FORM expressed with an adjective (1) and introduced by a preposition (2).

6. Conclusion

This article describes the process of corpus creation and definition annotation for semantic relation extraction. When building corpora, linguists had to pay close attention to both the format and nature of the texts. The conversion of Slovene data proved to be quite challenging and required a great deal of attention to detail. It might be useful to develop a conversion tool specifically for language-specific characters, such as diacritics, to facilitate the study of data originating from languages, namely Slovene.

Definition extraction, on the other hand, did not pose any significant challenge.

In contrast, definition annotation followed by the curation entailed a great deal of debate and additional research. Since the team consisted only of linguists/translation students lacking domain-specific terminological knowledge, it was sometimes difficult to comment on the nature of the extracted terms. For any similar research endeavours, it could be useful to seek expert's input so as to facilitate the annotation process and prompt better results. Overall, definition elements were easier to identify and annotate than relation definitors and semantic categories and relations. The result of this work is a dataset with multi-layer semantic annotations in English and Slovene which can be used for future relation extraction experiments. It complements the TermFrame dataset and will be added to the Clarin.si repository.

The paper also draws attention to the differences between the two languages. English seems to favour shorter and more concise definitions, such as "is a" or "are", while Slovene tends to introduce longer structures, namely "imenujemo" and "se uporablja", and sometimes shorter ones, such as "je".

This research provides insight into the various language-specific barriers that arise when studying smaller languages that do not enjoy the same exposure and presence as widespread world languages such as English.

Further research could examine how definitions in both languages manifest themselves in different contexts and domains.

Large data collections serve as a basis for the development of tools for automatic semantic relation extraction. Semantic relation extraction can be used to create different computer applications that can make domain-specific knowledge more accessible, not only to experts but to the general public as well. The corpora that were built during this project can be used for future creation of specialised knowledge bases on geology, geomorphology and glaciology.

7. References

Richard Eckart de Castilho, Chries Biemann, Iryna Gurevych, and Seid Muhie Yimam. 2014. WebAnno: a flexible, web-based annotation tool for CLARIN. In: *Proceedings of the CLARIN Annual Conference (CAC) 2014*, pages 4505–4512, Soesterberg, Netherlands.

Pamela Faber. 2015. Frames as a framework for Terminology. In: H. Kockaert and F. Steurs, (eds.) *Handbook of Terminology*, Vol. 1, pages 14–33. John Benjamins, Amsterdam/Philadelphia.

Pamela Faber and Laura Medina-Rull. 2017. Written in the Wind: Cultural Variation in Terminology. In: M. Gryviel (ed.) *Cognitive Approaches to Specialist Languages*, pages 419–442. Cambridge Scholars, Newcastle upon Tyne.

Chu-Ren Huang, Jia-Fei Hong, Wei-Yun Ma, and Petr Šimon. 2015. From Corpus to Grammar: Automatic Extraction of Grammatical Relations from Annotated Corpus. In T'sou & Kwong (eds.) *Journal of Chinese Linguistics Monograph Series*, Vol. 25, pages 192–221. Chinese University of Hong Kong Press, Hong Kong.

Senja Pollak, Andraž Repar, Matej Martinc, and Vid Podpečan. 2019. Karst exploration: extracting terms and definitions from karst domain corpus. In: *Proceedings of eLex 2019*, pages 934–956. Lexical Computing CZ, s.r.o., Brno

Senja Pollak, Anže Vavpetič, Janez Kranjc, Nada Lavrač, and Špela Vintar. 2012. NLP workflow for on-line definition extraction from English and Slovene text corpora. In: J. Jancsary (ed.) *Proceedings of KONVENS 2012 (Main track: oral presentations)*, Vol. 5, pages 53–60. ÖGAI, Vienna.

Aristomenis Thanopoulos, Nikos Fakotakis, and Georg Kokkinakis. 2000. Automatic Extraction of Semantic Relations from Specialized Corpora. In: *Coling 2000, 18th International Conference on Computational Linguistics*, Vol. 1, pages 836–842. Universität des Saarlandes, Saarbrücken.

Špela Vintar, Vid Podpečan, and Vid Ribič. 2021. Frame-based terminography: a multi-modal knowledge base for karstology. In: *Proceedings of eLex 2021*, pages 164–176. Lexical Computing CZ, s.r.o., Brno.

Špela Vintar, Amanda Saksida, Uroš Stepišnik, and Katarina Vrtovec. 2019 Modelling specialised knowledge with conceptual frames: the TermFrame approach to a structured visual domain representation. In: *Proceedings of eLex 2019*, pages 305–318. Lexical Computing CZ, s.r.o., Brno.

Špela Vintar and Uroš Stepišnik. 2020. TermFrame: A Systematic Approach to Karst Terminology. In: *Dela*, Vol. 54, pages 149–167. Znanstvena založba Filozofske fakultete Univerze v Ljubljani, Ljubljana. https://doi.org/10.4312/dela.54.149-167.

Katarina Vrtovec, Špela Vintar, Amanda Saksida, and Uroš Stepišnik. 2019. TermFrame: Knowledge frames in Karstology. In: *Proceedings of ToTh 2019*, pages 109–126. Presses Universitaires Savoie Mont Blanc, Chambéry