Bridging the Geological Lexicon and Corpus with Focus on MWEs Extraction

Biljana Rujević
University of Belgrade
F. of Mining and Geology
Belgrade, Serbia
biljana@jerteh.rs

Cvetana Krstev
Association for Language
Resources and Technologies
Belgrade, Serbia
cvetana@jerteh.rs

Mihailo Škorić
University of Belgrade
F. of Mining and Geology
Belgrade, Serbia
mihailo@jerteh.rs

Relevant UniDive working groups: WG2

1 Terminology extraction: from corpus to lexical entries

The geological domain corpus *GeoSrpKor* is compiled from 69 documents written in Serbian and used as interpreters of basic geological maps. It comprises 1,316,646 tokens (1,067,583 words) and is available for search on NoSketch (Kilgarriff et al., 2014) instance[1] for authorized users.

The terminology extraction in this research study relies upon the rule-based automatic multi-word term extraction and lemmatization used in several domains (Krstev et al., 2015; Stanković et al., 2016). We use *LeXimirka* lexical database, a robust system that not only manages electronic dictionaries (Serbian Morphological Dictionary – SMD), but also enables a connection with corpora, as well as systems for automatic single and multi-word terminology extraction (Stanković et al., 2018; Lazić and Škorić, 2020).

Patterns for extracted and approved MWEs from corpus *GeoSrpKor*, along with counts of approved terms (in brackets) and examples, are listed below:

AXN – an adjective followed by a noun; the adjective and the noun must agree in gender, number, case, and animateness (1162), *ugljevita glina* 'coal-clay'.

N2X - a noun followed by a word that does not inflect in the MWE (309), *mineral gline* 'clay mineral'.

N4X – a noun followed by two words that do not inflect in the MWE (190), *izrada geološke karte* 'creation of a geological map'.

2XN – a noun preceded by a word that does not inflect in the MWE (70), *alevrit glina* 'clay aleurite'.

AXN2X – a noun preceded by an adjective that agrees with it in four grammatical categories and followed by a word that does not inflect in the MWE (55), *centralni deo masiva* 'central part (of a) massif'.

2XAXN - an adjective followed by a noun that agrees in all four grammatical categories and preceded by a word that does not inflect in the MWe (35), *jezersko-barski sedimenti* 'lake-swamp sediments'.

AXAXN – a noun preceded by two adjectives that agree with it in four grammatical categories (19), *lecki andezitski masiv* 'Lece adensite massif'.

NXN – a noun followed by a noun that agrees with it in number and case, where the separator can be a hyphen (13), *facija mrtvaja* 'oxbow facies'.

AXN4X – a noun preceded by an adjective that agrees with it in four grammatical categories and followed by two words that do not inflect in the MWE (12), *ugljevita glina sa proslojcima* 'carbon-clay with layers'.

N6X - a noun followed by three words that do not inflect in the MWE (4), *krečnjak sa proslojcima rožnaca* 'limestone with layers of chert'.

The challenges that arose during the evaluation of MWE terms are inherently related to the manual evaluation of numerous candidates and the dilemma of determining what does qualifies as a valid term and what does not. Some of the specificities related to the geological domain in the Serbian language involve the use of specific morphological forms for the plural of nouns that denote materials (*naftni peskovi* 'tar sands').

It is important to note that this system has been used for terminology extraction from other domains, such as power engineering (Ivanović et al., 2022), library and information science (Trtovac and Andonovski, 2014), mining (Tomašević et al., 2018), spatial planning (Milinković, 2022), etc.

2 Lexical Database: from entries to corpus

Figure 1 presents a *LeXimirka* panel for the lexical entry *magmatska stena* 'igneous rock'. On the main panel, we can see references to entries in other dictionaries stored in the local database (Ser-

[1] https://noske.jerteh.rs/#concordance?corpname=GeoSrpKor

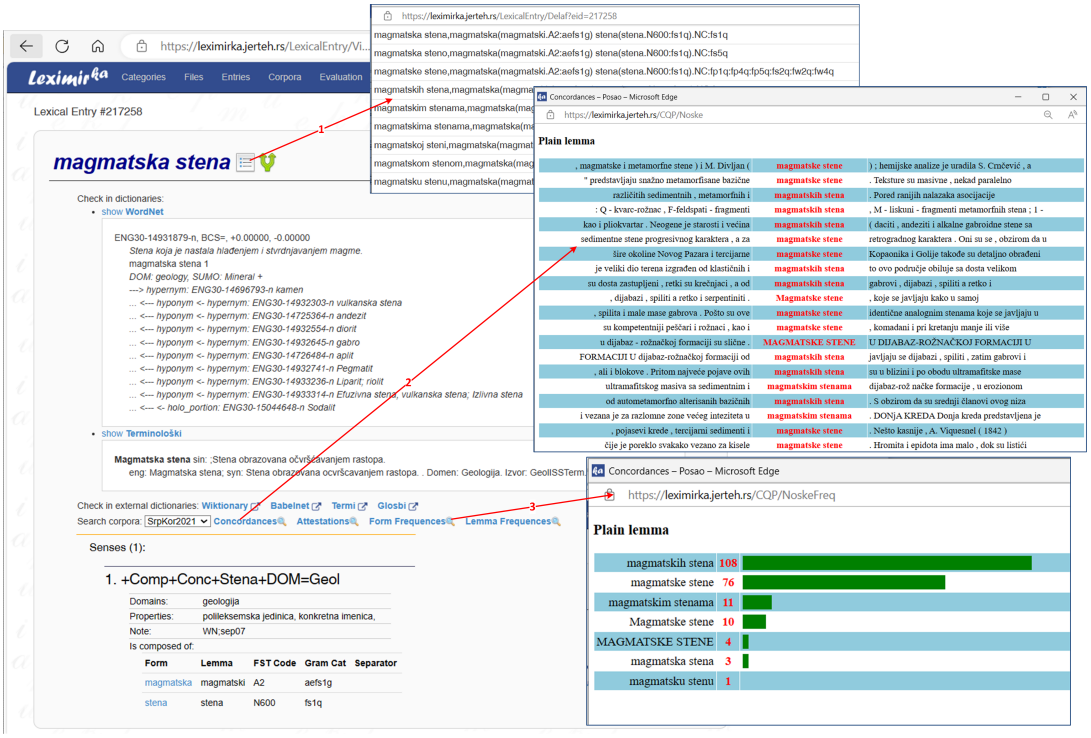

Figure 1: Panel from *LeXimirka* – a lexical resources management system

bian *WordNet* and *Terminološki*), shortcuts to entries in external dictionaries (*Wiktionary*, *BabelNet*, *Termi*, *Glosbi*), and a section with senses presented in the form of semantic, derivational, domain etc. markers.

Using the domain marker *DOM=Geol*, 2634 simple words and 1869 multi-word expressions from SMD were marked, belonging to the domain of geology. The most frequently used semantic markers in this SMD dictionary subset among simple words are *+Stena* denoting rock (131),[2] *+Mat* denoting material (178), *+Mineral* denoting mineral (109) and *+Time* denoting geological time (25). The most frequently used semantic markers related to MWEs are: *+Oro* denoting oronym (40), *+Process* denoting process (45), *+Hyd* meaning hydronym (10), *+Conc* representing something concrete (6) and *+Mat* denoting material (3). A user with adequate access rights can correct an entry, if necessary, or add additional information (e.g. markers, links to other entries).

In the Figure 1 we can see 3 additional pop-up panels. The first of them shows inflected forms with corresponding grammatical information. These are generated using the finite state transducer (FST) *NC_AXN*, where NC stands for a noun compound and AXN refers to its type depicts (adjective-noun). For the components that inflect in a MWU, FST requires information about their lemmas (in our case, *magmatski* and *stena*), their FSTs (*A2* and *N600*) and values of grammatical features of forms used in a MWU lemma (*magmatska:aefs1g* and *stena:fs1q*). The values of these features are: *a* – positive degree, *e* – same for definite and indefinite forms, *f* – feminine grammatical gender, *s* – singular number, *1* – nominative case, *g* - same for animate and non-animate nouns, *q* – inanimate.

The second pop-up panel shows concordances for the compound word in the selected corpus, in this case *GeoSrpKor*, retrieved from the NoSketch platform. These concordances contain all morphological forms of MWEs in the selected corpus. They are the result of a morphological query expansion.

The third panel shows frequencies of inflected forms in the selected corpus (again *GeoSrpKor*) for a MWE lexical entry (for single-word entries it also possible to choose a syntactic pattern, e.g. adjective-noun *A(N)*, where a noun is the current lexical entry). These frequencies are also retrieved from *NoSketch*. The interaction between lexical entry and corpus for the second and third panels is achieved via API calls from *LeXimirka* to *NoS-*

[2]Numbers in parentheses refer to the numbers of lexical entries.

ketch, triggered by a user through the interface. In the same section of the main web form, users can compare the same entry occurrences in alternative corpora that are listed.

3 Conclusion

Bridging Lexicon and Corpus is designed to facilitate dictionary use for the end users. It allows users to check examples in corpora concordances directly from the lexical entry, which saves them time and is suitable for those who may not be familiar with corpus queries. Geological domain lexicon and corpus used for MWE extraction are meant for geologists and geology students who may not have expertise in corpus linguistics.

References

Tanja Ivanović, Ranka Stanković, Branislava Šandrih Todorović, and Cvetana Krstev. 2022. Corpus-based bilingual terminology extraction in the power engineering domain. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 28(2):228–263.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: ten years on. *Lexicography*, 1(1):7–36.

Cvetana Krstev, Ranka Stanković, Ivan Obradović, and Biljana Lazić. 2015. Terminology acquisition and description using lexical resources and local grammars. In *Proceedings of the 11th Conference on Terminology and Artificial Intelligence, Granada, Spain, 2015*.

Biljana Lazić and Mihailo Škorić. 2020. From DELA based dictionary to Leximirka lexical database. *Infotheca*, 19(2):81–98.

Milena Milinković. 2022. Application of TXM tools for spatial plan corpus analysis. *Infotheca*, 22(1):32–51.

Ranka Stanković, Cvetana Krstev, Biljana Lazić, and Mihailo Škorić. 2018. Electronic dictionaries-from file system to lemon based lexical database. In *Proceedings of the 11th International Conference on Language Resources and Evaluation-W23 6th Workshop on Linked Data in Linguistics: Towards Linguistic Data Science (LDL-2018), LREC 2018, Miyazaki, Japan, May 7-12, 2018*.

Ranka Stanković, Cvetana Krstev, Ivan Obradović, Biljana Lazić, and Aleksandra Trtovac. 2016. Rule-based automatic multi-word term extraction and lemmatization. In *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016, Portorož, Slovenia, 23–28 May 2016*.

Aleksandra Tomašević, Ranka Stanković, Miloš Utvić, Ivan Obradović, and Ljiljana Kolonja. 2018. Managing mining project documentation using human language technology. *The Electronic Library*, 36(6):993–1009.

Aleksandra Trtovac and Jelena Andonovski. 2014. Enrichment of morphological dictionary of MWUs. In *Natural Language Processing for Serbian*, pages 27–40.

A Example Appendix

ARTEŠKI BUNAR (def. Bunar koji kaptira podzemne vode sa arteškim pritiskom koji se nalazi iznad površine terena.) Sinonimi: Samoizlivni arteški bunar; samoizlivni bunar – Eng. *Artesian well* (def. Well that penetrates groundwater with artesian pressure, which is above the surface.) Synonyms: Flowing artesian well; overflowing well

KVARCNI PESAK (def. Pesak preovlađujuće sastavljen od zrna kvarca.) – Eng. *Quartz sand* (def. Sand predominantly composed of quartz grains.)

MAGMATSKA STENA (def. Stena obrazovana očvršćavanjem rastopa.) – Eng. *Igneous rock* (def. A rock formed by solidification of a melt.)

NAFTNI PESKOVI (def. Naftni peskovi su značajna mineralna sirovina iz grupe nekonvencionalnih izvora za dobijanje nafte. Poznati su i pod nazivom „tar sands“ i predstavljaju nekadašnja (degradirana) ležišta nafte, osiromašena lakšim ugljovodonicima. Najčešće nastaju erozijom povlatnih sedimenata.) – Eng. *Tar sands* (def. Tar sands are significant mineral resources from the group of unconventional oil sources. They usually represent the former reservoirs of oil (degraded reservoirs), characterized by reduced amounts of lighter hydrocarbons. The genesis of tar sands is usually related to erosion of overlying sediments.)

PESKOVITI KREČNJAK (def. Krečnjak koji sadrži zrna kvarca kao klastičnu komponentu.) – Eng. *Sandy limestone* (def. Limestone containing quartz grains as its osteoclastic component.)

PODZEMNE VODE (def. Svaka voda ispod površine zemlje (u litosferi), bez obzira na agregatno stanje, vidove, poreklo, fizičke osobine, hemijski, radiološki i mikrobiološki sastav.) – Eng. *Groundwater* (def. Any water that can be found below ground level (in the lithosphere), regardless of their physical state, type, origin, physical properties, chemical, radiological and microbiological composition.) Synonyms: Subterranean water; underground water; subsurface water.)