International Cartographic Association
Association Cartographique Internationale

Grammar to Graph, an Approach for Semantic Transformation of Annotations to Triples

Dalia E. Varanka [a,*], Emily Abbott [b]

[a] *U.S. Geological Survey, dvaranka@usgs.gov*
[b] *Colorado School of Mines, under contract to the U.S. Geological Survey, eabbott@contractor.usgs.gov*

* Corresponding author

Keywords: annotation, dependency grammar, graph triple, topographic semantics

Abstract:

Linguistic representation of geographic knowledge is semantically complex and particularly challenging when using geographic information technology to automate interpreted analysis dealing with unstructured knowledge. This study describes an approach called GrammarToGraph (G2G) that applies dependency grammar rules through natural language processing to transform annotation data into structured geospatial semantic graph triples. This approach offers data handling advantages that include reducing string annotation storage needs, improving the logical specification of relations between objects, and providing reusable classes and properties that support graph queries and logic inference.

G2G was tested using topographic feature-type definitions that were normalized for string-processing consistency targeting the objectives. The basic unit of input text is a word token that is tagged with its part-of-speech (POS) and identified by their dependency relations (dep) to other tokens, as depicted in Figure 1 (Universal Dependencies 2024). The dependency relations among POS create a semantic framework based on phrase analysis libraries (spaCy 2024) in which verbs and their core arguments take the positions of objects and relations for a basic Resource Description Framework (RDF)/RDFSchema (RDFS) standard triple (Cyganiac and others 2014). And noun phrase modifiers form subgraphs in the position of object nodes.

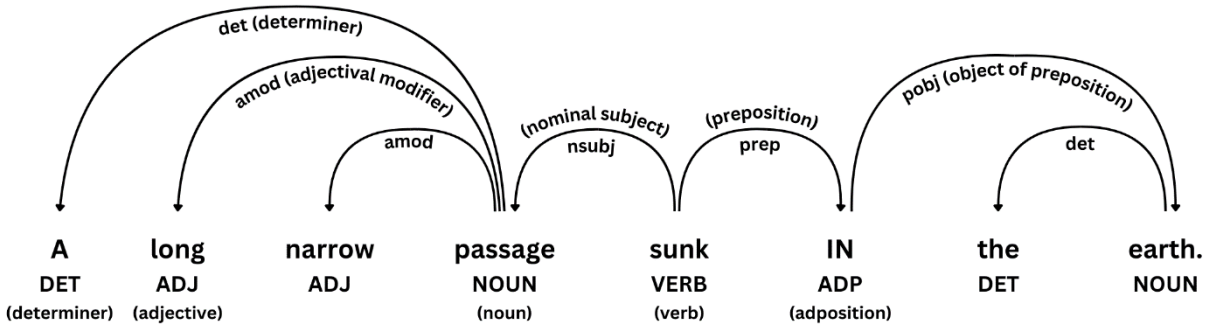

Figure 1. An example of dependency relation arc visualization between tokens of an input phrase defining the term ‘Shaft’ as ‘A long narrow passage sunk in the earth.’ Tokens are tagged with their part-of-speech. Dependency relations between two tokens appear as arcs.

Structuring the semantic schema described above was compared to Bertrand Russell's description theory, which emphasizes logical analysis to clarify linguistic expressions (Mambrol 2019). Matching resulted in a framework that utilizes predicate logic, classes and set theory, and types and variables in ontology design. While semantic triples typically involve one-place predicates (e.g., 'x is y'), more complex phrases can incorporate predicates of two or more places (e.g., "x is y and z"). Using 'is' alone can convey description, but a single proposition may not fully capture understanding.

https://doi.org/10.5194/ica-abs-7-196-2024

Comprehensive descriptions of related entities and class definitions often require multiple propositions, represented through anonymous or defined variable subgraphs. This theoretical approach allows the formulation of truth statements based on appropriate facts, including sensory data. On this basis, grammar patterns were generally aligned with RDF.

An applied ontology was designed to control grammar and logic and to respond to SPARQL Protocol and RDF Query Language (SPARQL) queries (Harris and Seaborne 2013). Multiple triple structures form a semantic graph that can be instantiated with a lexicon (vocabulary) to support composing natural language questions and queries using a SPARQL endpoint. Ontology and instance data tables were queried using virtual triplestore software that converts SPARQL queries to Structured Query Language (SQL) (ISO/IEC 9075-2:2023). Table 1 illustrates part of the data file for the sentence "An aircraft facility is an area where aircraft can take off and land, usually equipped with associated buildings and facilities." The sentence is tokenized from top to bottom.

Token	POS	Tag	Head text	dep	Lemma
An aircraft facility	NOUN	NN	is	nsubj	an aircraft facility
is	AUX	VBZ	is	ROOT	be
an area	NOUN	NN	is	attr	an area
where	SCONJ	WRB	land	advmod	where
aircraft	NOUN	NN	take-off	nsubj	aircraft
can	AUX	MD	take-off	aux	can
take-off	VERB	VB	an area	relcl	take-off
and	CCONJ	CC	take-off	cc	and
land	VERB	VB	take-off	conj	land
usually	ADV	RB	equipped	advmod	usually
equipped	VERB	VBN	an area	acl	equip
with	ADP	IN	equipped	prep	with
associated buildings	NOUN	NNS	with	pobj	associate building
and	CCONJ	CC	associated buildings	cc	and
facilities	NOUN	NNS	associated buildings	conj	facility

Table 1. Token, Part-of-Speech (POS), Tag, Head text, dependency relation (dep), and Lemma values for an input sentence.

A competency question for data validation involves asserting ontology subclasses from the feature type definitions. A linguistic concept tends to combine a label and definition, but an ontology requires a class representing the concept to be represented with label and definition annotations. Subclasses of feature type (FT) are determined from phrases by retrieving nouns of the noun subject relation combined with auxillary verbs 'is' or 'are' indicating the subsumption relation rdf:type of FT. Classes are associated with input data labels and definitions to be linked to the triples for reference. The approximate graph pattern of the SPARQL query, using prefixes for namespaces and triple syntax, for the process is printed below.

```
SELECT ?FT
WHERE { ?FT ud:hasPOS ud:NOUN .
?FT ud:dep ud:nsubj .
?FT rdfs:subClassOf g2g:FeatureType
?FT ud:hasToken ud:Token .
?FT rdfs:label ?Label .
rdfs:Label g2g:hasDescription g2g:Description . }
```

Results from SPARQL queries are iteratively evaluated for the geographic 'sense' outcomes and to refine semantic patterns. The data for the analysis are publicly available (Abbott 2024).

Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

References

Abbott, E., Grammar transformations of topographic feature type annotations of the U.S. to structured graph data.:2024. U.S. Geological Survey Data Release. https://doi.org/10.5066/P1BDPXKZ

ISO/IEC 9075-2:2023. Information technology Database languages SQL Part 2: Foundation (SQL/Foundation). ISO/IEC 9075-2:2023 - Information technology — Database languages SQL — Part 2: Foundation (SQL/Foundation)

Cyganiac, R., Wood, D., and Lanthaler, M., Eds. 2014. RDF 1.1 Concepts and Abstract Syntax. W3C Recommendation 25 February 2014. RDF 1.1 Concepts and Abstract Syntax (w3.org)

Harris, S. and Seaborne, A., Eds. 2013. SPARQL 1.1 Query Language. W3C Recommendation 21 March 2013. SPARQL 1.1 Query Language (w3.org)

Mambrol, N., 2019. The philosophy of Bertrand Russell. In: *Literary Theory and Criticism. The Philosophy of Bertrand Russell – Literary Theory and Criticism* (literariness.org)

spaCy. 2024. Industrial-Strength Natural Language Processing. spaCy · Industrial-strength Natural Language Processing in Python

Universal Dependencies, v2. 2024. Universal Dependencies.

https://doi.org/10.5194/ica-abs-7-196-2024