titulus	
	OnLiT : An ontology for linguistic terminology
huius textus situs retis mundialis	
	http://www.christianlehmann.eu/publ/lehmann_onlit.pdf
dies manuscripti postremum modificati	
	20.04.2017
occasio orationis habitae	
	Language, data and knowledge 2017. 19-20 June 2017 in Galway, Ireland
volumen publicationem continens	
	Gracia, Jorge et al. (eds.), *Language, data, and knowledge. First International Conference, LDK 2017, Galway, Ireland, June 19-20, 2017, Proceedings*. Berlin & Heidelberg: Springer.
annus publicationis	
	2017
paginae	
	42-57

OnLiT: An Ontology for Linguistic Terminology

Bettina Klimek[1](✉), John P. McCrae[2], Christian Lehmann[3],
Christian Chiarcos[4], and Sebastian Hellmann[1]

[1] InfAI, University of Leipzig, Leipzig, Germany
{klimek,hellmann}@informatik.uni-leipzig.de
[2] Insight Centre for Data Analytics,
National University of Ireland Galway, Galway, Ireland
john@mccr.ae
[3] University of Erfurt, Erfurt, Germany
christian.lehmann@uni-erfurt.de
[4] Applied Computational Linguistics,
Goethe-University Frankfurt, Frankfurt, Germany
chiarcos@informatik.uni-frankfurt.de
http://aksw.org/Groups/KILT, https://www.insight-centre.org,
http://www.christianlehmann.eu, http://acoli.informatik.uni-frankfurt.de

Abstract. Understanding the differences underlying the scope, usage and content of language data requires the provision of a clarifying terminological basis which is integrated in the metadata describing a particular language resource. While terminological resources such as the SIL Glossary of Linguistic Terms, ISOcat or the GOLD ontology provide a considerable amount of linguistic terms, their practical usage is limited to a look up of a defined term whose relation to other terms is unspecified or insufficient. Therefore, in this paper we propose an ontology for linguistic terminology, called OnLiT. It is a data model which can be used to represent linguistic terms and concepts in a semantically interrelated data structure and, thus, overcomes prevalent isolating definition-based term descriptions. OnLiT is based on the LiDo Glossary of Linguistic Terms and enables the creation of RDF datasets, that represent linguistic terms and their meanings within the whole or a subdomain of linguistics.

Keywords: Linguistic terminology · Linguistic linked data · LiDo database

1 Introduction

The research field of language data has evolved to encompass a multitude of interdisciplinary scientific areas that are all more or less closely bound to the central studies of linguistics. Understanding the differences underlying the scope, usage and content of language data provided by diciplines such as linguistics, computational linguistics, digital humanities or content analytics, requires the provision of a clarifying terminological basis which is integrated in the metadata describing a particular language resource. Moreover, the comparative use of resources

© Springer International Publishing AG 2017
J. Gracia et al. (Eds.): LDK 2017, LNAI 10318, pp. 42–57, 2017.
DOI: 10.1007/978-3-319-59888-8_4

of different languages presupposes that they use the same conceptual framework and terminology. This demand for specifying linguistic terminology has been addressed mainly by linguists in creating look-up resources such as books, e.g. the lexicon of linguistics (Bußmann et al. 1996), online registries (e.g. ISOcat[1] (Kemps-Snijders et al. 2009), the SIL Glossary of Linguistic Terms[2] (Loos et al. 2004) and the CLARIN Concept Registry (Schuurman et al. 2016) or Web pages such as the online encyclopedia of linguistics[3].

While all these resources provide a considerable amount of linguistic terms, their practical usage is limited to a look up of a term whose relation to other terms is unspecified or too general. In this respect the available data resources of linguistic terminology fail to provide a meaningful representation of a linguistic term leaving it isolated within the whole domain of linguistic terminology. Retrieving more information about linguistic concepts necessitates reading their definitions and looking up further words that are contained in it, which might be also defined terms in the database. This procedure is not only time-consuming and impractical but also results in implicit and vague specifications of linguistic terms. This is the argument from the viewpoint of usability. However, maintenance of a consistent conceptual-terminological framework likewise requires that the relations among concepts be standardized and that, for each concept, the relevant relations be specified. A set of isolated terms cannot be kept consistent.

In this paper we propose an ontology for linguistic terminology, called OnLiT, as a data model which can be used to represent linguistic terms and concepts in a semantically interrelated structure. Every terminological dataset evolving from OnLiT will result in a data graph which is easy to navigate for human users, machine-processable for semantic applications and will serve the purpose of directly and indirectly interrelating linguistic terms and concepts throughout the whole dataset. The OnLiT model is based on the *Linguistic Documentation (LiDo) database* by Christian Lehmann[4,5], who established a relational network which represents linguistic terminology that defines and delimits a term by relating it to the linguistic concept it encodes and also by including a set of specifying conceptual relations (Lehmann 1996). What is more, the proposed model is independent of the particular language of the terms and thus allows integration of terminological networks in different languages and multilingual terminological networks. By transforming the structure of the LiDo relational database to RDF, the OnLiT data model aims to provide the following contributions:

- to enable a semantic search for linguistic terms and concepts,
- to provide unique reusable and citable identifiers for each data entry,

[1] http://www.isocat.org/.

[2] http://www-01.sil.org/linguistics/GlossaryOflinguisticTerms/.

[3] http://www.glottopedia.org.

[4] A browseable version of the database is available at: http://linguistik.uni-regensburg.de:8080/lido/Lido.

[5] Christian Lehmann is the data owner of LiDo and permitted to derive the OnLiT data model from it.

- to enable the creation of conceptually consistent terminological datasets that broadly interconnect and cover linguistic terms in a required linguistic (sub)domain,
- to establish the possibility for extending the data model and enriching an OnLiT dataset with external data,
- to allow free and open reuse of the OnLiT data model.

The remainder of the paper is structured as follows. Section 2 gives an overview of relevant related work. Following an outline of the LiDo database as basis for OnLiT in Sect. 3, the OnLiT data model is presented in Sect. 4.1. Further, the purpose, domain and requirements of OnLiT are presented in Sect. 4.2 and the modelled concepts, terms and the established relations between them are discussed in Sects. 4.3 and 4.4. Finally, in Section 5 the paper concludes giving a brief summary and a prospect of future work.

2 Related Work

An investigation of available datasets (excluding the LiDo database which is presented in Sect. 3) that contain models of representing linguistic terminology, resulted in two different types of data.

(i) Linguistic term bases that offer a term look-up via a Website: Resources such as the aforementioned ISOcat registry or SIL Glossary of Linguistic Terms (GLT) are mainly aimed at human users. Their underlying semantic structure is rather flat providing definitions and very unspecific superordinate and subordinate concept relations such as *is a* or *has kinds*. In the GLT, further, terms in a term entry can be traced by the user via established links. Navigating through ISOcat is harder since it provides a wide range of different "views" and "groups" which provide linguistic terminology in general but also specify linguistic terms in a specific language data model, e.g. the "STTS group" or "CLARIN group". In this regard such linguistic term bases have no underlying data model that represents linguistic terminology in an interrelating holistic structure. What is more, the arbitrary structure of the data models, which represent the linguistic term entries in alphabetical order (as in GLT) or according to linguistic views or linguistic data groups (as in ISOcat) is neither sufficient nor suitable for gaining comprehensive knowledge about a linguistic term in the domain of linguistics. A recent project, the CLARIN Concept Registry (Schuurman et al. 2016), has taken over the work of ISOcat and promises to define terms in a stricter manner, although still providing very limited structural and relational information.

(ii) Linguistic concepts represented as Linked Data ontology: In order to enable the description of linguistic data, formalized ontological models emerged within the realm of the Semantic Web. The most significant model for the scientific description of human language is the General Ontology for Linguistic Description (GOLD)[6] (Farrar and Langendoen 2003; Farrar 2010). It provides a taxonomy of nearly 600 linguistic concepts, which have been constructed from

[6] http://linguistics-ontology.org/gold-2010.owl.

the GLT, and formalizes 83 relations (i.e. 76 object properties and 7 data properties). GOLD has been designed to support Community-of-Practice Extensions (COPEs), meaning that it is a recommended upper model for ontologies of linguistic terminology that can define their concepts as sub-concepts of GOLD concepts (Farrar and Lewis 2007). This mechanism has been adopted by several ontology providers, e.g., (Wilcock 2007; Good et al. 2005; Goecke et al. 2005). In that usage and because the terms provided by the GLT have been transformed into concepts in GOLD, linguistic terms and concepts are not distinguished any more. The concepts are only defined within the domain of linguistic description but not in the more general domain of linguistics. In addition, the variety of object properties assigned to the concepts are very specific and interrelate mostly only two concepts, which leaves the majority of the concepts unrelated. The established relations are either too specific or too general to derive the meaning of a concept within the domain of linguistics, e.g. a "grapheme" concept is defined within the taxonomy as a "FormUnit" concept, which is a "LinguisticUnit" concept, which is an "Abstract" concept. It has no further relations to other concepts, e.g. to "Character", which only implicitly states in its `rdfs:comment` that it is "similar to grapheme". Also, it is unclear why the "Character" concept ist not also modelled as a subconcept of "FormUnit". These are solvable issues, however, the development of GOLD and the community process stopped in 2010. Despite the wealth of linguistic concepts in GOLD it would be a very inconcise model for linguistic terminology, due to the lack of terms relating to the concepts and due to the complexity of relations which is aimed at a subfield of descriptive linguistics but not at representing linguistic concepts in a more encompassing scope of the domain of linguistics.

These two primary kinds of sources for a model of linguistic terminology can be summarized as being either term-focussed or concept-focussed. A coherent model of linguistic terminology, however, presupposes explicitly establishing both linguistic concepts and terms and placing them into the whole domain of linguistics. To conclude, to our knowledge there is - with the exception of the LiDo database - no data model available that appropriately describes linguistic terminology as the domain of linguistic terms that encode linguistic concepts which are interrelated in a meaningful way.

3 The LiDo Glossary of Linguistic Terms as OnLiT Pioneer

The LiDo database[7] as it is available in its current form as a browsable glossary of linguistic terms has a thirty year old history. Christian Lehmann started to collect and systematize his terminological knowledge as a general comparative linguist by introducing a documentation system for linguistics in 1976 (Lehmann 1976). Twenty years later its technical implementation in 2006 resulted in the

[7] It has to be mentioned that LiDo encompasses also bibliographical data that is referenced to the terms. This bibliographic part of the dataset is, however, not focus of this paper and, hence, not further discussed.

LiDo Web frontend which is based on a relational database[8] that has been continuously updated and extended by Christian Lehmann ever since. To date, the LiDo term and concept data encompasses more than 4500 unique linguistic concepts and over 15000 terms, most of them in English, German, Spanish and Portuguese. Moreover, each concept is interrelated to at least one other concept which yields a coherent terminological data graph. Editing and curating this considerable data size is enabled by a manageable set of relations which fulfill the self-imposed requirement to explicitly express a direct relation between two linguistic concepts (Lehmann 1996). This is achieved by the two formal relations of coordination and subordination which generate an overall taxonomic and meronomic structure and 14 subrelations of those that permit a semantically specified interrelation of concepts. As a consequence, the data structure underlying the LiDo term and concept data inheres the following criteria which we see as essential for describing terminological data:

- explicit representation of concepts and terms as separate resources,
- meaningful interrelation of concept and term data,
- an easy to use and editable data structure.

Therefore, the underlying LiDo data structure does not only permit an appropriate representation of the domain of linguistic terminology but also implicitly contains an ontological modelling of the domain. These two aspects finally motivate the reuse of the LiDo model as a data basis for creating OnLiT.

4 The Ontology for Linguistic Terminology

4.1 Components of the OnLiT Model

The OnLiT vocabulary is freely available under the URL http://lido.linguistic-lod.org/onlit.rdf[9] and open for any kind of reuse under the CC BY 4.0 license[10]. As a Linked Data model which is based on the Web Ontology Language (OWL[11]), OnLiT consists of a hierarchy of conceptual classes which represent commonality among a variety of entities, i.e. the so-called instances, individuals or resources of a dataset. The semantics of entities within the ontology is formally defined by class usage restrictions that can hold between classes and are encoded within relations. Relations are formally expressed as object properties or as data properties.

An overview of the class modelling in OnLiT is given in Fig. 1 and a detailed view of the object property structure is provided in Fig. 2. For modelling the domain of linguistic terminology, the OnLiT vocabulary contains only

[8] This database is used to render the LiDo Website but not publicly available. The database was used in order to conduct the presented research.

[9] In case of unavailability: https://github.com/AKSW/lido2rdf/blob/master/OnLiT.owl.

[10] https://creativecommons.org/licenses/by/4.0.

[11] https://www.w3.org/OWL.

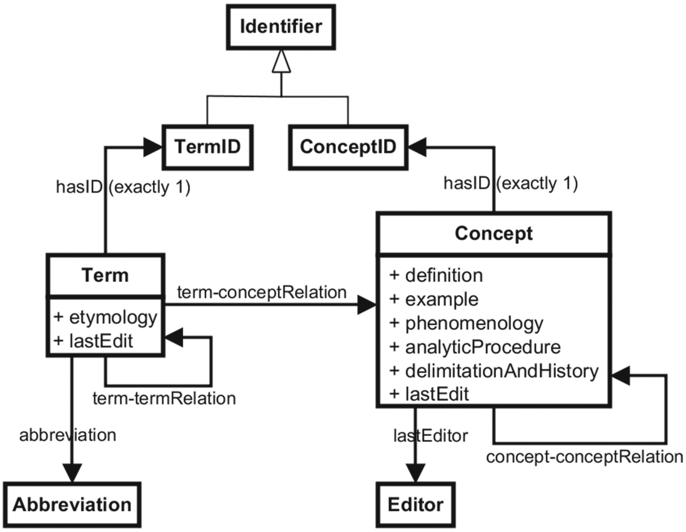

Fig. 1. Class diagram of the OnLiT model.

7 classes: `Concept`, `Term`, `Identifier` (with `ConceptID` and `TermID` as subclasses), `Abbreviation` and `Editor`. Only the first two are essential and should in any case contain instances (a more detailed presentation of their usage is given in Sect. 4.3). An `Concept` instance describes a language-independent mental entity which is encoded in different language-specific terms. As such concepts are cognitively defined as substantial meanings which are realized by a linguistic sign, which is then the term associated with the concept. In order to be able to identify and refer to such a mental (as opposed to the formal understanding of concepts as classes in OWL!) conceptual instance, it needs to be somehow denominated with a humanly readable name. This can be done by an arbitrary string identifier or by using the term expression that standardly encodes the concept in some language as, i.e., there could be a 'noun' `Concept` instance and a *noun* `Term` instance. The former, however, serves only as a conventionalized naming method[12] for a cognitive and language-independent meaning while the latter is a linguistic expression of the English language. This distinction is similar to the division of sense IDs which are associated to lexical entries in datasets such as WordNet[13]. The `Abbreviation` class is established, because linguistic terms can have various conventional abbreviations assigned. This is common practice in language description and might be, therefore, useful for some dataset creators. Meta-information provided by the `Identifier` and `Editor` classes are added for convenience, because they tend to be included in other dataset formats, such as tables and relational databases. These can be directly used in case already

[12] In the LiDo database Latin expressions are used to a large extent to denominate the concept entries.

[13] http://wordnet-rdf.princeton.edu.

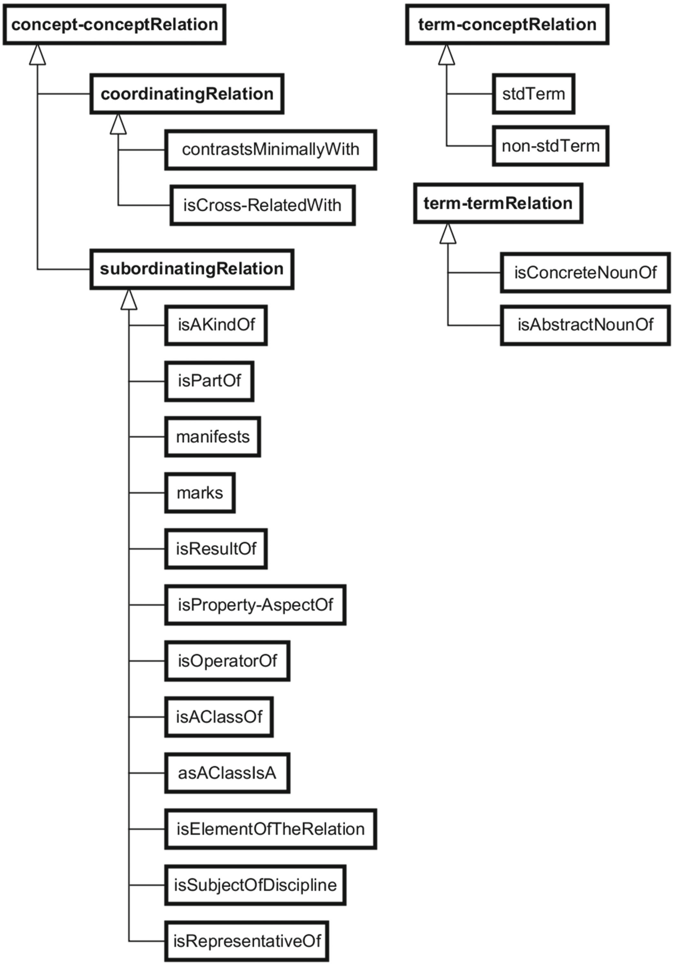

Fig. 2. Inheritance diagram of OnLiT object properties and subproperties.

existing datasets of linguistic terms in such formats shall be transferred into RDF with the OnLiT model. However, more fine-grained Linked Data vocabularies are available for representing the metadata of a dataset, e.g. DCMI terms[14] or PROV-O[15], which are easily integrable due to the interoperability of Linked Data vocabularies.

With regard to the relations, there are three main object properties established in OnLiT that interrelate instances of (1) terms with terms, (2) terms with concepts and (3) concepts with concepts. The `term-termRelation` property can be used to specify the relation between noun `Term` instances, on the one hand,

[14] http://dublincore.org/documents/dcmi-terms.
[15] https://www.w3.org/TR/prov-o.

and adjective and verb Term instances, on the other, in a dataset. That way adjective and verb terms can be included in a dataset if they are desired to be described as linguistic terms (cf. Sect. 4.3 below) and related to their corresponding noun Term resources[16] which are then interrelated to the respective Concept instance they encode. A term-conceptRelation is established in order to enable the assignment of the Term instance to its associated Concept instance. The most structuring are the concept-conceptRelation object properties. Because these are divided in the subproperties of coordinatingRelation and subordinatingRelation they add to the taxonomic and meronomic structure of the Concept data and therewith also of the Term data within an OnLiT dataset. The twelve subordinatingRelation subproperties are intended to establish a semantically more specific interrelation between concepts (a more detailed presentation of their usage is given in Sect. 4.4).

Overall the OnLiT model is of manageable size but yet provides sufficient explicitly modelled semantic interrelations to create a consistent dataset of linguistic terminology.

4.2 Purpose, Domain and Requirements of OnLiT

There are two main purposes pursued by the OnLiT model. First, it serves as the conceptual foundation for an RDF dataset of the LiDo Glossary of Linguistic Terms including the whole relational database of its Term and Concept data. Second, it provides users and creators of language data in general as well as the community of Linguistic Linked Open Data in particular with a means to easily set up and/or semantically interconnect various linguistic terminological datasets. Moreover, its basic properties are transferable to the definition of terminological datasets for other scientific disciplines.

The domain of linguistic terminology as represented by OnLiT is not restricted to a certain definition of *term*. Thus, any expression that needs to be described with OnLiT for theoretical or practical reasons can constitute a Term or Concept resource in an OnLiT dataset. As a consequence, even proper nouns denoting persons, e.g. *Noam Chomsky* or linguistically significant words of a language, e.g. the grammatical verb *be* can be entries in an OnLiT term base. In that respect Term entries in an OnLiT dataset are not limited to a narrow definition of *scientific term* as being a common noun (Kamlah and Lorenzen 1967). Rather, this definition is broadened to allow individuals' names, plain lexemes or even adjectives and verbs to be included as terminological entries. Given that OnLiT is based on the LiDo Glossary of Linguistic Terms, it meets the same criteria as outlined in Sect. 3. In addition to that and in contrast to the Lido data model, OnLiT is based on Semantic Web modelling principles. Due to that, OnLiT based datasets fulfill the requirements of semantic and structural interoperability which enable an easy reuse of data and further enrichment via interlinking to external data sources.

[16] This allows, for instance, to integrate the Term entries *homonymous* and *govern* and relate them to *homonymy* and *government*.

We assume that datasets evolving from the OnLiT model will add to the creation of a comprehensive terminological knowledge graph of the field of linguistics ranging from general and traditional linguistic terminology to the representation of newly evolving or very specifically used terms and concepts.

4.3 Linguistic Concepts and Terms

The `Term` and `Concept` classes constitute the essential classes of an OnLiT dataset since these contain the concept and term resources respectively. Two relations can be specified between them, which express that a `Term` resource is a standard or a non-standard term for a given concept. Their interrelations are illustrated in Fig. 3, which exemplifies the triples for the `Term` instance *noun* and the `Concept` instance ‘nomen substantivum’. `Concept` resources are unique, since they are mental objects which are designated by a linguistic expression, i.e. the `Term` resource. As a consequence, there can be multiple `Term` resources related to a single `Concept` resource. Thus, there is also a *Substantiv* resource stated to be the standard German term and also a *Nennwort* resource to be a non-standard term for the `Concept` resource ‘nomen substantivum’. This can be achieved by forming triples between `Term` and `Concept` resources via the two object properties `stdTerm` and `non-stdTerm`. This is the way of dealing with synonymous terms. For a homonymous term, the relation to one `Concept` resource is selected as `stdTerm`, and all the others are `non-stdTerm`. Each `Term` resource can use the property `stdTerm` for only one `Concept` resource, while it can be `non-stdTerm` for more `Concept` resources. For instance, the German term *Nomen* is standard for the concept ‘nomen’ and non-standard for the concept ‘nomen substantivum’. Further, every `Term` resource should be explicitly assigned to a language. For that purpose the language identifiers of the lexvo vocabulary[17] are reused, because they provide a precise language assignment as well as machine-readability.

The `Concept` resources can be further specified for additional information by describing the definition, delimitation and history, analytic procedure, phenomenology and example(s) via the respective datatype properties (cf. Fig. 3). This information is provided by plain text and constitutes information a linguist might have documented about a certain linguistic concept and which should be included in the database. In fact, definition and examples are frequently found in terminological datasets (e.g. in GLT or ISOcat) and can be simply transferred to an OnLiT dataset by using these datatype properties. Even though information stated in such plain text literals is not directly machine-readable and, therefore, also not semantically explicit enough for automatic data processing, it is from a human data consumer perspective very insightful. Eventually, the definitions constitute indeed a useful information source that reveals information about a concept, that can be formally modelled. The definition of the ‘nomen substantivum’ `Concept` resource states that it is “a [. . .] part of speech”, which can be formalized via a subordinating relation between the given `Concept` resource and another ‘part of speech’ `Concept` resource (this will be demonstrated in Sect. 4.4).

[17] http://www.lexvo.org.

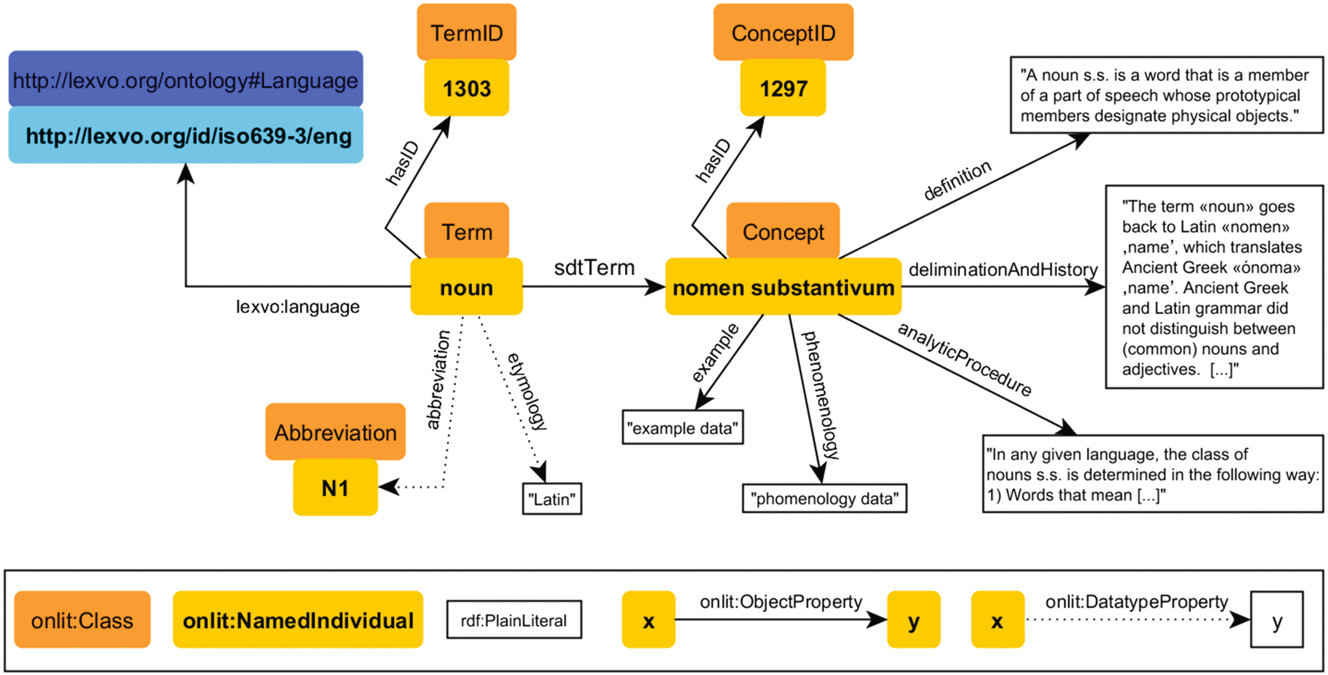
TermID
1303
ConceptID
1297
http://lexvo.org/ontology#Language
http://lexvo.org/id/iso639-3/eng
hasID
hasID
Term
Concept
"A noun s.s. is a word that is a member of a part of speech whose prototypical members designate physical objects."
definition
noun
sdtTerm
nomen substantivum
deliminationAndHistory
"The term «noun» goes back to Latin «nomen» ‚name', which translates Ancient Greek «ónoma» ‚name'. Ancient Greek and Latin grammar did not distinguish between (common) nouns and adjectives. [...]"
lexvo:language
abbreviation
etymology
example
phenomenology
analyticProcedure
Abbreviation
N1
"Latin"
"example data"
"phomenology data"
"In any given language, the class of nouns s.s. is determined in the following way: 1) Words that mean [...]"
onlit:Class
onlit:NamedIndividual
rdf:PlainLiteral
x
onlit:ObjectProperty
y
x
onlit:DatatypeProperty
y

Fig. 3. Example modelling of the English `Term` instance *noun* and the `Concept` instance 'nomen substantivum' with OnLiT. (The exemplary data for both resources used here and in other parts of this paper can be consulted under the English "noun" entry in the LiDo database Web frontend: http://linguistik.uni-regensburg.de:8080/lido/Lido.)

Hence, textual information about linguistic concepts are not only most prevalent in already existing terminological datasets, but also assist the OnLiT dataset creators in formally expressing their explicit defining relations to other concepts. Conversely, a good definition incorporates the conceptual relations specified for the concept.

To summarize, the representation of linguistic concepts and terms adheres to the requirement of providing separate resources for both. What is more, the relation that holds between a term and concept is modelled in OnLiT as a one to one correspondence between a `Term` instance (having a single unambiguous meaning) and the corresponding `Concept` instance (being the mental object of that single meaning) it designates. This ensures a disambiguated traceability and clarification of linguistic terms within the domain of linguistics. Also, the OnLiT model provides a manageable but significant set of object and datatype properties which specify `Concept` and `Term` resources in more detail and which can be easily extended with further properties if need be.

4.4 Interrelating Linguistic Concepts

As presented in the previous section, there are only two object properties that relate `Term` resources to `Concept` resources. The majority of relations is specified in object properties which are established between two `Concept` resources. While these relations could theoretically also hold between `Term` resources, this is not done for a practical reason. Because multiple `Term` instances can refer to the same `Concept` instance it is more economic to assign specific interrelations once to the `Concept` instance, instead of repeating them on every `Term` instance that is associated with the same `Concept` instance. This holds a fortiori for translations of the terminological dataset into other languages. As a result, the semantic specification is directly attached to the `Concept` resources and, therefore, also indirectly to the `Term` resources via the `term-conceptRelation` subproperties (as described in the previous Section). Figure 4 exemplifies how multiple `Term` resources can encode a single `Concept` resource, which provides further semantic specification through the `concept-conceptRelation` subproperties.

As is shown in Fig. 2, the 14 object properties which are at the lowest level of the object property hierarchy are the most specific ones. In order to create a more general taxonomic structure, these are systematized according to the superproperties `coordinatingRelation` and `subordinatingRelation`. As a result, more statements can be inferred that relate `Concept` instances on a broader semantic level. Such inferred triples are expressed in Fig. 4 via the dashed arrows connecting the `Concept` instances. There are two subproperties which yield a coordinating relation and which are described as follows:

`x isCross-RelatedWith y`: States that a concept is somehow cross-related with another concept, although the two are not sisters subordinate to a third concept.

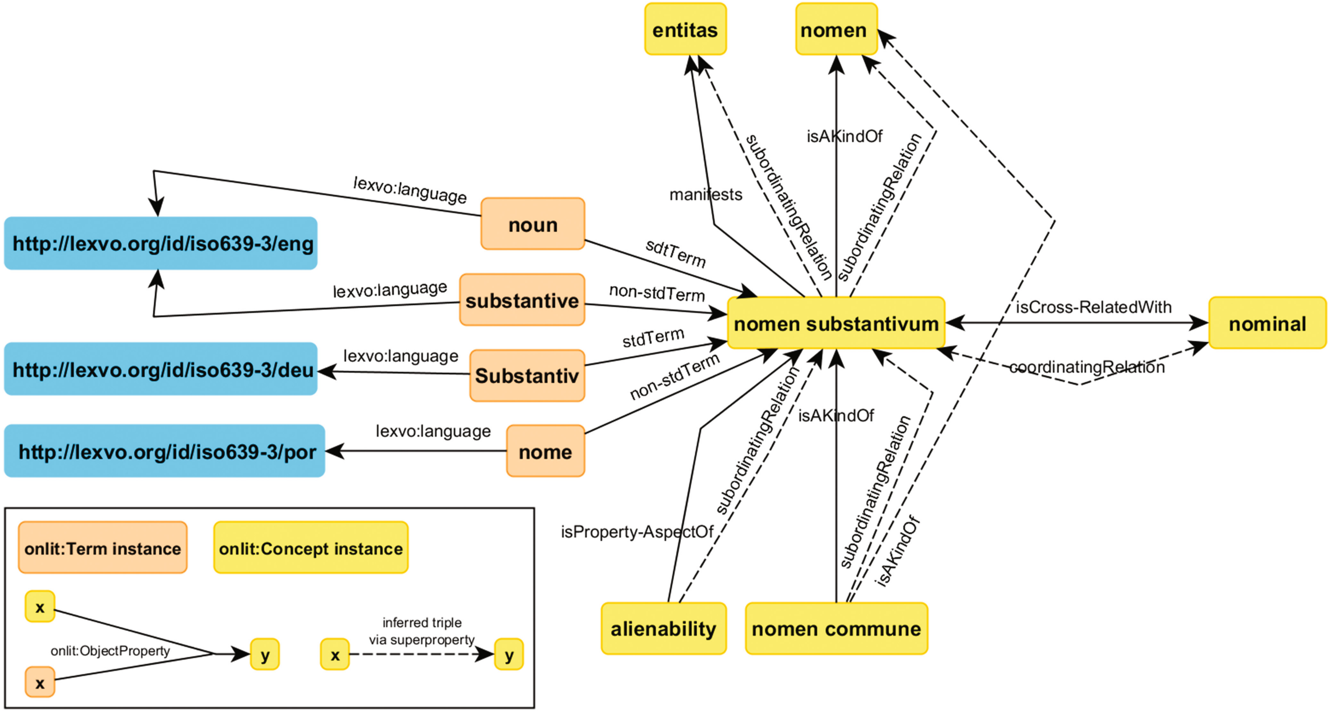

Fig. 4. Example modelling of the `Concept` instance ‘nomen substantivum’ with its interrelations to other concepts.

Example 1. nomen adjectivum (*adjective*)[18] `isCross-RelatedWith` attributum (*attribute*).

`x contrastsMinimallyWith y`: States that a concept contrasts minimally with another concept.

Example 2. aspectus perfectivus (*perfective*) `contrastsMinimallyWith` aspectus imperfectivus (*imperfective*).

The `coordinatingRelation` subproperties are symmetric properties, that group semantically similar `Concept` instances by cross-referencing.

For creating subordinating relations between `Concept` instances twelve subproperties can be used:

`x isAKindOf y`: Is the most general subordinating relation, that states that a concept is a kind of another superordinating concept. The interrelation of concepts with this property creates a taxonomy.

Example 3. linguistica (*linguistics*) `isAKindOf` scientia rerum humanarum (*human science*) `isAKindOf` scientia (science) `isAKindOf` activitas (activity).

`x asAClassIsA y`: States that if a concept x is taken to represent a class, this is a subclass of another class concept.

Example 4. nomen adjectivum (*adjective*) `asAClassIsA` pars orationis (*word class*).

`x isAClassOf y`: States that a concept represents a class.

Example 5. pars orationis (*word class*) `isAClassOf` dictio (*word*).

`x isElementOfTheRelation y`: States that a concept is an element of a relation represented by another concept.

Example 6. allomorphum (*allomorph*) `isElementOfTheRelation` allomorphia (*allomorphy*).

`x isOperatorOf y`: States that a concept is an operator of an operation represented by another concept.

Example 7. affixum (*affix*) `isOperatorOf` affixio (*affixation*).

`x isPartOf y`: States that an entity falling under concept x is a part of an entity falling under another concept. Concepts that are interrelated with this "part-whole" property will create a meronymy.

Example 8. casus (*case*) `isPartOf` declinatio (*declension*).
declinatio (*declension*) `isPartOf` flexio (*inflection*).
flexio (*inflection*) `isPartOf` systema morphologicum (*morphology*).
systema morphologicum (*morphology*) `isPartOf` systema grammaticum (*grammar*).
systema grammaticum (*grammar*) `isPartOf` systema linguae historicae (*language system*).

[18] For better comprehensibility the standard English `Term` instances corresponding to the given `Concept` instances are given in brackets.

`x isProperty-AspectOf y`: States that a concept represents a characteristic or possible aspect or property of its superordinate concept.

Example 9. arbitrarietas signi (*arbitrariness*) `isProperty-AspectOf` signum linguae (*linguistic sign*).

`x isRepresentativeOf y`: States that a person is a representative of a scientific discipline, movement or model.

Example 10. de Saussure (*de Saussure*) `isRepresentativeOf` schola Genavensis (*Geneva School*).

`x isResultOf y`: States that an entity falling under a concept is the result of an entity falling under another concept.

Example 11. vocabulum externum (*loan word*) `isResultOf` mutuatio (*borrowing*).

`x isSubjectOfDiscipline y`: States that a concept that represents some object (area) is the subject of a concept denoting the scientific discipline or a theory or model thereof.

Example 12. systema vocabulorum (*lexicon)* `isSubjectOfDiscipline` lexicologia (*lexicology*).

`x manifests y`: States that a concept denotes a grammatical or derivative category which manifests a concept that denotes a semantic, cognitive, communicative or functionally determined concept.

Example 13. tempus grammaticum (*tense*) `manifests` tempus (*time*).

`x marks y`: States that a concept represents a grammatical category which marks a grammatical relation or function represented by another concept.

Example 14. casus accusativus (*accusative*) `marks` objectum directum (*direct object*).

Figure 4 shows how the modelling of the `subordinatingRelation` property results in a taxonomic systematization of `Concept` instances. This allows for automatic reasoning over a dataset to yield insights such as ‘nomen’ is superordinate to ‘nomen substantivum’ which is superordinate to ‘nomen commune’ and, thus, ‘nomen’ is also superordinate to ‘nomen commune’. This holds also for some of the subproperties, e.g. `isAKindOf` which is a transitive property (’nomen commune’ `isAKindOf` ‘nomen substantivum’ and of ‘nomen’). What is more, the 14 established object properties are all semantically more specific than a generic “see also” relation but general enough to be broadly applied to interrelate various (and ideally all) concepts. Especially relations such as `isOperatorOf` or `marks` play a central role in the domain of linguistic terminology. In that respect, a dataset modelled with OnLiT sets every linguistic term or concept in a meaningful interrelation to relevant other terms by placing it in a navigable and coherent context within the linguistic domain a dataset describes. Finally, relations such as `isAKindOf` and `isPartOf` are general across ontologies of any science and thus serve to integrate linguistic ontologies into an all-encompassing ontology.

5 Conclusion and Future Work

The OnLiT data model for representing terminological data of linguistic domains has been created as the ontological schema basis to transfer the currently relational database of the LiDo Glossary of Linguistic Terms into an RDF dataset in the future. Moreover, the OnLiT model constitutes a valuable contribution for users and creators of linguistic data. Due to the outlined benefits of the underlying Linked Data format, evolving terminological data will be interoparable, semantically and formally explicit as well as easy to reuse and extend. Moreover, OnLiT models linguistic terminology in a meaningful and structured way that goes beyond a single term definition. I.e. the additional subordinating and coordinating relations allow to derive coherent and specific insights and knowledge about the conceptualization of linguistic terms in a given language dataset. Therefore, it can benefit producers of language data in creating their own terminological dataset or in interrelating their data to an existing OnLiT dataset (e.g. the prospective LiDo RDF dataset). Furthermore, future work includes an interconnection of the OnLiT model with OntoLex[19], which will offer more possibilities of representing and integrating OnLiT `Term` and `Concept` resources within the domain of lexical language data.

Acknowledgements. This paper's research activities were partly supported and funded by grants from the EU's H2020 Programme ALIGNED (GA 644055) and the Federal Ministry for Economic Affairs and Energy of Germany (BMWi) for the Smart-DataWeb Project (GA-01MD15010B).

References

Bußmann, H., Trauth, G., Kazzazi, K.: Lexikon der Sprachwissenschaft. Taylor & Francis, London (1996)

Farrar, S.: General Ontology for Linguistic Description (GOLD). The LINGUIST List, Department of Linguistics, Indiana University (2010)

Farrar, S., Langendoen, D.T.: A linguistic ontology for the semantic web. GLOT Int. **7**, 97–100 (2003)

Farrar, S., Lewis, W.D.: The gold community of practice: an infrastructure for linguistic data on the web. Lang. Resour. Eval. **41**, 45–60 (2007)

Goecke, D., Lüngen, H., Sasaki, F., Witt, A., Farrar, S.: GOLD and discourse: domain- and community-specific extensions. In: Proceedings of the 2005 E-MELD-Workshop (2005)

Good, J., Cysouw, M., Albu, M., Bibiko, H.J.: Can GOLD "cope" with WALS? Retrofitting an ontology onto the world atlas of language structures. Max Planck Institute for Evolutionary Anthropology (2005)

Kamlah, W., Lorenzen, P.: Logische Propädeutik oder Vorschule des vernünftigen Redens. Bibliographisches Institut (B.I.-Hochschultaschenbücher 227/227a), Mannheim (1967)

[19] http://www.w3.org/ns/lemon/ontolex#.

Kemps-Snijders, M., Windhouwer, M., Wittenburg, P., Wright, S.E.: Isocat: remodelling metadata for language resources. Int. J. Metadata Semant. Ontol. **4**, 261–276 (2009)

Lehmann, C.: Um sistema de documentação para a lingüística. Instituto de Letras e Artes, Pontifícia Universidade Católica do RGS (1976)

Lehmann, C.: Linguistische Terminologie als relationales Netz. In: Knobloch, C., Schaeder, B. (eds.) Nomination-fachsprachlich und gemeinsprachlich, pp. 215–267. Springer, Wiesbaden (1996)

Loos, E.E., Anderson, S., Day, D.H., Jordan, P.C., Wingate, J.D.: Glossary of Linguistic Terms, vol. 29. SIL International, Dallas (2004)

Schuurman, I., Windhouwer, M., Ohren, O., Zeman, D.: CLARIN concept registry: the new semantic registry. In: Selected Papers from the CLARIN Annual Conference 2015, pp. 62–79. Linköping University Electronic Press (2016)

Wilcock, G.: An OWL ontology for HPSG. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, pp. 169–172. Association for Computational Linguistics (2007)