Ontologies and ontological methods in linguistics

Andrea C. Schalley¹,²

¹Karlstad University
²Griffith University

Correspondence
Andrea C. Schalley, Department of Language, Literature and Intercultural Studies, Karlstad University, Universitetsgatan 2, 651 88 Karlstad, Sweden.
Email: andrea.schalley@kau.se

Abstract
In the last decade, linguists have started to develop and make use of ontologies, encouraged by the progress made in areas such as Artificial Intelligence and the Semantic Web. This paper gives an overview of notions and dimensions of “ontology” and of ontologies for and in linguistics. It discusses building blocks, design aspects, and capabilities of formal ontologies and provides some implementation pointers. The focus of this paper, however, is on linguistic research and what a modelling framework based on ontologies has to offer. Accordingly, the paper does not aim at providing an overview of specific models for computational processing. To illustrate the issues at hand, an example scenario from linguistic typology is selected instead, where the aim of describing the world’s languages is approached through ontologies.

1 | INTRODUCTION AND MOTIVATION

The notion of ontology has become rather prominent in linguistics over the last decade, and, through the development of Semantic Web technologies, it has received increased attention in artificial intelligence and computer science more generally. This paper provides an introduction to the notion of “ontology” and discusses ontologies in and for the discipline of linguistics.

However, this is not a computational linguistics paper. It does not focus on ontologies for computational processing. Rather, ontologies are understood as a new exciting representation or modelling framework for linguistic research, with the ensuing computational capabilities allowing for new ways of knowledge gathering and hypothesis development by linguists. In particular, this paper does not address the automatic generation of ontologies from large data sets. It is moreover written from a typological linguistic perspective, i.e., the description of the world’s languages is chosen as an
example scenario. Yet many of the ideas presented herein are equally applicable to other linguistic sub-disciplines.

Given the selected scenario of linguistic typology, the first issue to be noted is the sheer number of languages still spoken worldwide. Ethnologue (Eberhard, Simons, & Fennig, 2019) estimates the number of currently known living languages at 7,111, and we know that these languages are very diverse in their design solutions for communicating and representing ideas. This linguistic diversity poses a huge challenge, as individual linguists can only work on (i) a limited number of languages and (ii) a limited number of linguistic phenomena, and they can only access (iii) a limited amount of data. It is impossible for any scholar to obtain a comprehensive overview even in regard to one small linguistic phenomenon. Therefore, any “linguist's knowledge can, due to the nature of the field, be compared to pieces of a puzzle” (Schalley, 2013). The question then arises as to how we can put these puzzle pieces together, in order to obtain access to the whole linguistic landscape.

The second issue with which many linguists and in particular students of the discipline struggle is the fact that there are many linguistic schools of thought, different linguistic theories and frameworks, and different ideas on how data should be annotated and analysed. Moreover, scholars describe different languages in different ways, which has led to an enormous terminological diversity as well. As Haspelmath (2010) points out, there is a tension between language-specific descriptions and more general cross-linguistic comparisons. A case in point is the notion of “evidentiality”, where confusion and disagreement is quite pronounced (cf., e.g., Boye & Harder, 2009; Brugman & Macaulay, 2010; Behrens, 2012). The discipline thus needs a vehicle that allows it to either deal with this terminological diversity or that at least provides mechanisms to bring these terminological issues out into the open in order to constructively discuss them.

A further issue motivating an ontological approach, as we will see, is the data formats linguists work with. The discipline has an impressive body of cross-linguistic data in (a) print format (books and journals), (b) many small typological corpora (collected by individual scholars with different purposes and annotation schemes), (c) a number of larger typological datasets (collected by groups of researchers who are either specialists in a specific language or in specific linguistic phenomena), and (d) corpora (large collections of primary data). Questions concerning these collections of different data formats include notions of accessibility and searchability of such data (and the flexibility of such searches), as well as the data's reusability. If these data could be linked and integrated, current computational tools would allow for targeted searches across the body of knowledge (Schalley, 2012).

This paper argues that ontologies could offer a promising approach to dealing with these issues. Initial steps have been taken and are using existing computational technologies. While adapting the technological infrastructure from computer science, the ontological approach provides a lot more than just computational tools, however, as it presents a new way of thinking about linguistic description and analysis. This paper aims at presenting the reader with this new way of thinking and thus gives an overview of this new avenue in linguistic research.

2 | NOTIONS OF “ONTOLOGY”

2.1 | Prominent notions of “ontology”

As Nickles, Pease, Schalley, and Zaefferer (2007) have shown, different disciplines have a different understanding of the notion of “ontology”. This results in different definitions across disciplines, and awareness thereof is important for interdisciplinary work (see an in-depth discussion of the notion of “communication” in Fraser & Schalley, 2009). The following description only covers the most
prominent non-linguistic disciplines, and only the most crucial differences are pointed out. Readers interested in more detailed discussions are encouraged to consult Nickles et al. (2007) and the literature cited in this section.

2.1.1 Philosophy

As a first approximation, ontology is the study of what there is. [...] we have at least two parts to the overall philosophical project of ontology: first, say what there is, what exists, what the stuff is reality is made out of, secondly, say what the most general features and relations of these things are. (Hofweber, 2018)

According to Hofweber, ontology is thus about (i) categories (and their specifying features) and (ii) relations between these categories. The discipline of philosophy, in which the concept of ontology has played a role since Ancient Greek times (although the term “ontology” itself was not coined until the early 17th century), is in particular interested in what kinds of things there are in reality, rather than what humans think or conceptualize as being there. This distinguishes it from other disciplines.

2.1.2 Artificial intelligence

The most well-known definition of “ontology” in artificial intelligence is Gruber's (1993, p. 199) definition, who defined it as “an explicit specification of a conceptualisation”, where a “conceptualisation is an abstract, simplified view of the world that we wish to represent for some purpose”. As we can see, this differs from philosophy in that the object of study is not objective reality any more, but human conceptualisation. Moreover, it is not human conceptualisation in its complexity, but only a subpart or domain, and the ontology results in a targeted model (“an abstract, simplified view”) that is there for a purpose and created by an ontology author (“world that we wish to represent for some purpose”).

2.1.3 Knowledge representation

Sowa (2003) defines “(an) ontology” for the knowledge representation context:

The subject of ontology is the study of the categories of things that exist or may exist in some domain. The product of such a study, called an ontology, is a catalog of the types of things that are assumed to exist in a domain of interest D from the perspective of a person who uses a language L for the purpose of talking about D. (Sowa, 2003)

Sowa explicitly distinguishes between the uncountable noun “ontology” as referring to the field of study and “an ontology”, using the countable noun to refer to the product of such study. I will maintain this distinction as well, as has also previously been done by Nickles et al. (2007). (Note that in a similar move, Guarino, Oberle, & Staab, 2009, pp. 1–2, distinguish between the philosophical discipline “ontology” and “an ontology as a special kind of information object or computational artefact” in computer science.) In addition, Sowa does not only include the modelling domain but also include the language that is deployed for the purpose of talking about the domain. Implicitly, this introduces
2.2 Linguistic ontologies

In the linguistic context, an “ontology” has been defined as “a cross-connected network of relevant concepts, which makes explicit, classifies and organizes the assumptions and terms of the domain in question” (Schalley, Musgrave, & Haugh, 2014). Again, this definition shows that the object of study is not reality, but human conceptualisation. This is self-evident to linguists who consider linguistic signs as representing concepts and indeed to anyone who would accept that language is represented somehow in the human mind. As in artificial intelligence and knowledge representation, we are dealing with a domain—for instance, a specific language such as Avatime, a specific linguistic phenomenon such as evidentiality, or a specific linguistic subdiscipline such as phonology. We can thus see strong links between these different discipline-specific notions of ontology, which also motivates the application of computational ontology tools—as developed in artificial intelligence and knowledge representation—to linguistic study.

In terms of field of study, there are two different levels to bear in mind in linguistics: (i) the object level—the languages of the world and their expressions and speakers, and (ii) the meta level—the linguistic description apparatus used to represent and discuss entities of the object level. To each of these, ontologies are relevant, as outlined in the following sections.

2.2.1 Object level—Speaker ontology

At this level, the object of linguistic study is in focus: linguistic expressions and language use. These provide us with the opportunity to draw conclusions on speakers’ ideas and concepts that they wish to represent and share with others through linguistic means. In the context of cross-linguistic study, these also allow to identify concepts speakers of particular languages have to entertain in order to speak their language correctly and appropriately. For instance, a language’s grammar might require the explicit marking of specific grammatical categories and hence underlying concepts. Some examples from different linguistic subsystems include the following:

- **lexicalised concepts** (e.g., the distinct meanings of English lexemes such as *sleep* and *carburettor* or of the German lexeme *Schadenfreude*);
- **concepts underlying grammar**, as reflected in
  - morphemes (e.g., the harmonic/disharmonic kin relationship expressed by the dual verbal prefixes in Dalabon—speakers have to track the kin relationship between the two people spoken about in order to use the dual prefixes correctly; Evans, Brown, & Corbett, 2001; Evans, 2008);
  - syntax (e.g., possession in Italian, which generally has the form article + possessive pronoun + noun—*la mia casa* ‘my house’, *il mio amico* ‘my friend’—except for in the case of family members, where the article is dropped—*mia sorella* ‘my sister’, *mio padre* ‘my father’; Berns, 2008); and
- **semantic/pragmatic phenomena** (e.g., anaphora resolution of sentences such as *When you try to catch a lizard, the reptile may drop the tail and escape*; Nickles et al., 2007, p. 38, where the anaphora from *the reptile* to *a lizard* succeeds because of the conceptual inheritance relation between lizard and reptile).
These examples show that, as a speaker of a language, one has to and indeed does make specific conceptual distinctions in order to appropriately use the language. Looking at language features and language use thus enables researchers to delve into the networks of concepts that speakers entertain (speaker ontologies) and how these networks differ across languages. Based on the previous definition of “linguistic ontology”, an ontology at the object level, a speaker ontology, can thus be defined as a network of cross-connected conceptualisations that are associated with linguistic signs. It represents the ontological knowledge that underpins the linguistic knowledge of speakers and groups of speakers, or, put differently, a speaker ontology is an ontology that answers the question “What kinds of things do people talk as if there are?” (Nickles et al., 2007, p. 36). Overlapping subtypes can be distinguished:

i a general, cross-linguistic speaker ontology, i.e., the network of concepts underlying linguistic coding across natural languages. An example would be the ontology of social cognition categories that are grammatically coded in the world’s languages. Example ontology concepts would include kinship categories, ethnic group categories, and interactional roles (speaker–addressee–bystander, both attending and non-attending);

ii the speaker ontology of a specific language, i.e., the network of concepts underlying a single language, for instance, the Swahili speaker ontology. Generally, concepts modelled in this ontology are conventionalised in the speaker community. Examples of ontology concepts of the Swahili speaker ontology would include the Swahili grammatical gender class system, a system which differs markedly from the English system (Corbett, 1991); and

iii an individual’s speaker ontology, i.e., the network of categories entertained by a specific individual speaker (such as the author’s speaker ontology). Concepts modelled in this ontology would naturally overlap with concepts of the above two types but would likely include a unique set of concepts, potentially comprising idiosyncratic ones (e.g., depending on the languages spoken by the individual and possible interference effects).

A prominent example resource reflecting a lexeme-based speaker ontology for English (i.e., a language-specific speaker ontology) is WordNet (Fellbaum, 1998, 2007). An ontological rearrangement and extension of this is SENSUS (Knight & Luk, 1994), which uses the Penman Upper Model (Bateman, Kasper, Moore, & Whitney, 1989) as its topmost ontology structure. The Penman Upper Model has been superseded by the Generalized Upper Model (GUM; Bateman, Henschel, & Rinaldi, 1995), a general linguistically motivated ontology that aims to provide a semantics for natural language expressions. Hence, GUM can be seen as a general, cross-linguistic top-level speaker ontology.

2.2.2 Meta level—Discipline ontology

The meta-level perspective applies to any science: In order to study and represent results, a science needs to establish a language for communicating amongst its scholars, and in an ideal world, this scientific language would be standardised with well-defined terms. The concepts that underlie the terminology in a discipline and their relationships form a discipline ontology. A linguistic discipline ontology thus is a network of cross-connected conceptualisations of the domain of language science, comprising the concepts that are used to describe linguistic phenomena, and relationships amongst these concepts. This domain-specific ontology represents an answer to the question “What kinds of
things do linguists talk as if there are?” (Nickles et al., 2007, p. 39); i.e., the focus is here on linguists’ conceptualisations, rather than on speakers’.

Again, overlapping subtypes can be distinguished:

i a general, cross-linguistic discipline ontology, i.e., network of concepts used to describe and analyse natural language. An example would be an ontology for the description and analysis of phonology phenomena, with example ontology concepts being vowel, stress, manner of articulation, tap, etc.;

ii the discipline ontology of a specific language, i.e., a network of concepts used to describe and analyse a single language, for instance, the Zulu discipline ontology—with click as an example ontology concept, which would not be part of the English discipline ontology (cf., e.g., Moran, McCloy, & Wright, 2014); and

iii an individual’s discipline ontology, i.e., the network of concepts used by the individual linguist to describe and analyse language phenomena, such as the author’s discipline ontology. Concepts modelled in this ontology will overlap with concepts of the above two types, but the ontology concepts will depend on the linguistic knowledge, experience, and school of thought of the individual.

Well-known general discipline ontologies include resources such as the General Ontology for Linguistic Description (GOLD; Farrar, 2007; GOLD, 2010) and the Ontologies of Linguistic Annotations (OLiA; Chiarcos, 2012; Chiarcos & Sukhareva, 2015; OLiA, 2012), amongst other resources. Another more specific example from within the lexical domain is Lemon, a model for modelling lexicon and machine-readable dictionaries (McCrae et al., 2012).

3  | DIMENSIONS AND ASPECTS OF ONTOLOGY

Given the above definitions and the aspects they have highlighted, we can follow Nickles et al. (2007) in positing the following variation dimensions for ontology/ontologies:

a) Field of study vs. specific account

The field of study, “ontology”, is distinguished from resulting specific accounts, “ontologies” (note the corresponding difference between uncountable noun for the field label, and countable noun as label for specific accounts). Similar distinctions are prevalent in linguistics: compare, for instance, the distinction into “syntax” as the field of study and the “syntax” of English as a specific account (in contrast to that of Warlpiri). The specific, countable notion can be further distinguished into one referring to speakers' mental ontologies, e.g., the ontologies speakers of English and Warlpiri entertain, and into another one referring to a theoretical account of these, i.e., an explicit model of the former.

b) Generality vs. specificity

While philosophy is targeting the nature of being and reality as such, knowledge representation targets a domain of interest. Thus, ontologies can be distinguished according to how wide they cast their net—whether they include “everything” or only a small domain (such as the domain of Warlpiri phonemes and their description).

c) Objectivity (realist) vs. subjectivity (idealist)

The third dimension is concerned with whether an ontology represents reality as such or a mental representation (conceptualisation) of it. In the most general case, the objective perspective is
about the entities that exist (irrespective of what we make of them), while the subjective perspective studies which entities we conceptualise as existing (and they might only exist by virtue of us conceptualising them).

Furthermore, there are other aspects of ontologies in the specific account reading, each represented by a question as follows:

d) **Origin**
   How has the ontology been developed—through an evolutionary process (e.g., the mental ontology entertained by members of a species), or by an ontology author (as in, e.g., theoretical accounts of mathematical phenomena)? If an ontology author exists, they can be known or unknown.

e) **User (and purpose)**
   Who is the user of an ontology—a human, a computer, or some other entity? Closely linked to this notion is an ontology's purpose, as an ontology capturing the inventory of Swiss German allophones is relevant to, e.g., speech synthesis applications in this language, while it is not relevant to a semanticist investigating lexemes for emotions across the world's languages.

f) **Evaluation**
   How are ontologies best evaluated? Nickles et al. (2007, p. 31) maintain that adequacy and usefulness are crucial notions, given that an ontology is a conceptualisation modelled for a purpose:

   *If there are two competing ontologies of the same domain, is it possible that one is true and the other one false? Or are there other criteria for the evaluation of ontologies? If, as we have assumed, ontologies in the simple meta-sense are conceptualisations, then they themselves cannot be true or false, they can only be more or less adequate and more or less useful. And of course they are completely inadequate if they entail false statements. […] Usefulness is a relational concept which requires a purpose. Of two competing ontologies one can be more useful than the other for one purpose and less useful for another.*

g) **Representation**
   How is the ontology formally represented? Which representation “language” or format is used? These questions are particularly relevant to ontologies from artificial intelligence and knowledge representation. Accordingly, an “ontology spectrum” was developed, covering a continuum of different representation formats, including the one focused on in this paper, formal ontologies with automated reasoning and sets of general logical constraints (McGuinness, 2003; Nickles et al., 2007; Smith & Welty, 2001).

Any linguistic ontology will be positioned in the space that is opened up by these variation dimensions. This space is not fully “accessible” to linguistic ontologies, though, i.e., not every point in the space can be reached: While the generality/specificity, user/purpose, and representation dimensions are flexible in the linguistic context, it is fair to say that linguistic ontologies are subjective and that their origin can be delimited to “evolutionary” in the case of speaker ontologies and “authored” in the case of linguistic discipline ontologies. Adequacy and usefulness of particular discipline ontologies (as represented by different schools of thought or description and annotation systems) are constantly scrutinised in linguistic research.
In this paper, the focus lies on formal ontologies, as the objective is to illustrate how linguistics might draw on available ontology tools from computer science and build explicit models. In line with this, the next section explores formal ontologies’ basic building blocks and their characteristics.

## 4 | BASIC BUILDING BLOCKS

What are the building blocks that are essential for the specification of an ontology? So far, we have defined an ontology as a cross-connected network of concepts. Accordingly, a specification requires categories for concepts as well as relationships to express the connections between these concepts. In addition, a formal representation of an ontology can comprise properties and operators. Each of these building blocks is discussed in turn, followed by a brief discussion of modelling complexity.

### 4.1 | Classes and instances

A concept (as a mental element) is represented in an ontology by way of a class (a modelling element of the ontology). Classes correspond to collections of entities (or objects). Objects that fall under a category, i.e., that are a member of a class, are the instances of the class. For example, the morpheme concept (as a mental unit) can be represented as the class Morpheme, thus becoming a category in the ontology. The English plural morpheme -s is an instance of Morpheme (see the left-hand side of Figure 1), and at the same time, it is an instance of the class Plural. Classes and instances constitute the nodes in the ontology network.

### 4.2 | Relational building blocks

Typically, there are two types of predefined relationships that enable the modelling of relations between the nodes in the ontology: instantiation and inheritance.

**Instantiation** is the relationship between an instance and a class (e.g., plural -s being an instance of Morpheme), represented graphically by a dashed arrow in Figure 1. The instantiation relationship assigns objects as instances of classes, thus specifying that these objects have all the features of class members. For example, bound morphemes could be characterised as phonologically dependent and grammatically dependent (in contrast to clitics that are only phonologically dependent). If an entity (i.e., object) is declared to be an instance of Bound_Morpheme in the model, it automatically “adopts” these features due to the instantiation relationship. If, in turn, an object shows all the features of Bound_Morpheme, it will be classified as an instance of this class by the reasoner (see Section 5.1).

**Inheritance** is the subordination of one concept under another, i.e., a relationship between two classes. For example, a Prefix is a Bound_Morpheme (as represented by the solid arrow in Figure 1). Subclasses (“children”) inherit all the features of their superclasses (“parents”). For example, Prefix is both phonologically and grammatically dependent because Bound_Morpheme has

![Diagram](attachment://figure1.png)

**Figure 1** Graphical representation of the instantiation (dashed arrow) and inheritance (solid arrow) relation
these features, due to the inheritance relation between the two classes. This allows us to model features at the level of generality where they are needed and to manage some of the network complexity. Both multiple instantiation (e.g., an object being an instance of Prefix and Plural at the same time) and multiple inheritance (e.g., a prefix being an affix and contiguous morpheme at the same time) are possible and desirable.

4.3 | Properties

The two predefined relationships discussed in Section 4.2 are the “bare bones” needed to form an ontology, i.e., to build up a class hierarchy and assign objects to classes. These bare bones can then be augmented by additional user-defined relationships. Besides the inheritance relation, part–whole relations are generally present in all domains and hence constitute additional basic ontological relations (Schalley & Zaefferer, 2007, p. 4). However, part–whole relations constitute a complex family of relations with different characteristics, such as whether the part can be removed, is mandatory, or is similar to the whole and/or other parts etc. (Schalley, 2017; Winston, Chaffin, & Herrmann, 1987).

Rather than locking ontology designers into predefined part–whole relations, thence, systems usually provide the more general notion of property, a directed binary relation that specifies class characteristics. For instance, a word can be related to its morpheme(s) through a property hasMorpheme, with the relation's domain being Word and its range being Morpheme. Such binary property relationships are more versatile and can additionally be specified as being transitive, symmetric, functional, or the inverse of another property. Through the mechanism of properties in addition to instantiation and inheritance, a fine-grained cross-connected network can be formally expressed. Other relationships specific to specialised linguistic (sub-)domains, e.g., relations relevant to morphology as in isGrammaticallyDependent or isRoot, can also be represented in the same way.

4.4 | Operators

The last building block is operators, which represent various operations on classes. Predefined are often operators such as union, intersection, complement, cardinality, and disjointness. Operators provide users with the opportunity to, for instance, define new classes based on existing classes (union, intersection, and complement), add further constraints on classes (cardinality as a restriction on the potential number of class members), or specify constraining relationships between classes (disjointness as a restriction on class membership). Disjointness, for instance, specifies whether two classes are compatible or not. One could, e.g., declare Prefix (disjoint) Circumfix, which expresses that an object that is an instance of Prefix cannot at the same time be an instance of Circumfix. Operators are particularly important for reasoning (see Section 5.1).

4.5 | Many-to-many relationships

Modelling linguistic data ontologically quickly results in a complex network. Consider examples (1) and (2) from Dalabon (Evans, 2008; Evans et al., 2001):

1 barra- h- bo -n
   3DU.HARM- ASSERT- go -PR
‘they two go’ (relatives in “harmonic” generations, e.g., two brothers, grandparent, and grandchild)

2 ke- h- bo –n
3DU.DISHARM- ASSERT- go -PR

‘they two go’ (relatives in “disharmonic” generations, e.g., father and son, aunt, and niece)

The first morpheme of (1), barra-, is a prefix, the meaning of which is tripartite—indicating 3rd person, dual, and harmonic kin. We can represent the information on barra-, extracted from the glossing that is based on the Leipzig Glossing Rules (Bickel, Comrie, & Haspelmath, 2008), in a network—an extract of which is represented in Figure 2.

Figure 2 shows on its right-hand side that barra- is a prefix and, through inference based on inheritance relations, also an affix, contiguous morpheme, bound morpheme, and morpheme. If we thus were to run a search for bound morphemes, for example, the prefix barra- would be returned as a result. Similarly, the meaning of barra- is shown to be tripartite as outlined above (harmonic kin, dual, and 3rd person), and due to the ontology extract, we know that the broader grammatical categories are those of kin, number, and person. In parallel to the above example, if we thus were to search for linguistic expressions that carry kin-related meaning or functions, barra- would be returned. The ontological approach thus allows us to extract information based on the inheritance modelling. The \{disjoint\} operator between Prefix and Circumfix ensures that if barra- were analysed as circumfix elsewhere, the resulting contradiction would be detected by a reasoner (see Section 5.1).

In a representation of both examples (1) and (2)—which only differ in their first morphemes—we already develop a network of many-to-many relationships (Figure 3). It demonstrates how linguistic subsystems can be separated and modelled in a modular way, by distinguishing realisations (R—the language-specific linguistic expressions), forms (F—the categorisation of these expressions as instances of specific linguistic forms), and concepts/meanings (C—the categorisation of these expressions as instances of specific meanings or as bearing specific linguistic functions).
5 DESIGN CONSIDERATIONS AND ONTOLOGY CAPABILITIES

There are a number of guides and resources for ontology building available (for instance, Noy & McGuinness, 2001; Uschold & Gruninger, 1996; cf. also Gandon, 2002, p. 85, for different modelling approaches). They recommend steps to be taken in the ontology development process. Design considerations and ontology capabilities are seen as important factors for successful ontology development, and they allow us to embrace and successfully deal with the complexity of linguistic modelling. For the purposes of this paper, reasoning over ontologies, ontology reuse, and—as a way of dealing with incomplete information—the open world assumption are introduced, as these are the most prominent ones in the context of the development of linguistic ontologies.

5.1 Reasoning

Reasoning refers to the process of deriving facts that are not asserted in the ontology. For this, a reasoner is used, a software that is able to carry out derivations and inferences. Being able to deploy a reasoner is a major incentive for working with a formal ontology specification. Consider this example: if (i) it is asserted that barra- is a prefix, (ii) the reasoner knows that a prefix is a bound morpheme (through an inheritance chain), and (iii) a bound morpheme is characterised as phonologically and grammatically dependent, the reasoner can then compute that barra- is also a bound morpheme and that it is phonologically and grammatically dependent.

Reasoning primarily supports three purposes: taxonomy classification (automatic subordination), class inference (automatic instantiation), and consistency checking (ensuring that no contradictions arise in the ontology). For instance, Figure 4a shows a “flat” morpheme classification—there are hardly any inheritance relations asserted. In particular, in this modelling, Prefix is not asserted as a subclass of another class. Instead, classes are defined via their properties. Figure 4b shows the same classes—after their inheritance relations have been computed based on the properties. We can see that the morpheme types, while not originally defined as subclasses of Morpheme, have been classified correctly by the reasoner. Also, Prefix has been identified as Affix (and hence BoundMorpheme) as well as as ContiguousMorpheme, showing multiple inheritance.

Similarly, Figure 5 depicts an asserted class hierarchy, without any instances (members) of the class WordClass. Figure 6 displays the result of the reasoning, listing all the different word classes that are comprised in the example data and that are newly inferred as instances of WordClass (thus
generating a list of instances and deriving new facts by way of automatic instantiation). Figure 6 also highlights an inconsistency in the definition of a test class—the InconsistentFreeAndBoundMorpheme class had been asserted as a combination of two contradictory concepts. We can hence see how reasoning can be an extremely helpful modelling tool while at the same time supporting the generation of new explicit knowledge.

5.2 | Reuse

As we have seen in Section 4.5, ontologies lend themselves to a modular design, which in turn is beneficial to and a mechanism for ontology reuse. Reusing existing ontologies results in the reuse of existing knowledge that has already been modelled. This is highly desirable, for efficiency reasons. There are different reuse mechanisms available (for more detailed information on ontology reuse, consult, e.g., Pâslaru-Bontaş, 2007):

i Ontology integration refers to the import of existing ontologies.

ii Namespaces are a means to unambiguously interpret elements. A name can only occur once in a namespace, but it can occur in multiple namespaces. For example, the class Root could occur with very different meanings in the three different domain ontologies of biology, dentistry, and linguistics.

iii Modularisation stands for a separation into different thematic ontologies, comprising, for instance, language information, society information, example data, cross-linguistic coding, and geographical region in different modules or sub-ontologies.
5.3 Open world assumption

The last characteristic under discussion is closely linked to the reasoning capabilities discussed in Section 5.1, the open world assumption. In contrast to the “closed world assumption” as used by traditional programming languages and databases, the “open world assumption” states that “if the system cannot prove a statement to be true (using the knowledge in the knowledge base and the reasoner), the conclusion cannot and must not be drawn that this statement is false (as would be done by systems using the closed world assumption)” (Schalley, 2012, p. 148). Instead, lack of knowledge results in the statement being classified as unknown. For example, if there is no information on a grammatical gender system for a language, the reasoner cannot infer that the language does not have a grammatical gender.

The open world assumption thus lends itself to dealing with incomplete information. It adequately aligns with the situation in linguistic research, where knowledge about the domain will continue to be discovered. Given the situation of the world’s languages—in particular their diversity and endangerment—complete knowledge will never be achieved. Therefore, the open world assumption is crucial for linguistic modelling.

Due to the open world assumption, the knowledge base as modelled in the ontology can be easily expanded and revised, and (even continuous) addition of new information is in principle smooth—unless new statements contradict previous ones.

FIGURE 5 Asserted class hierarchy, no objects (“Members”) are asserted as instances of WordClass on the right-hand side
6 | IMPLEMENTATION POINTERS

How can we go about actually modelling and building ontologies? How can we express and manage these complex networks and make them available to others? There is a wealth of tools available from computer science, the most prominent of which are briefly touched on in the following—the Web Ontology Language (OWL) and the ontology editor Protégé. Note that this section only provides pointers to the two most relevant tools; readers interested in more details are referred to the references given in this section. Moreover, it is important to note that these tools do not support the automatic generation of ontologies. Rather, the idea is that linguistic ontologies as discussed here are manually built in a bottom-up process, based on example linguistic data, by linguists as domain experts (see also Schalley, 2012). The tools’ role is to support this process, e.g., by managing the complexity, checking for consistency, and supporting joint collaborative work on the ontologies (Borkowski & Schalley, 2011).

6.1 | Web Ontology Language (OWL)

The Web Ontology Language (OWL) is a family of knowledge representation languages for authoring ontologies (McGuinness & van Harmelen, 2004; W3C OWL Working Group, 2012). It is the formal underpinning of the building blocks outlined in Section 4. OWL is description logic based (a description logic models concepts, roles, individuals, and their relationships), and it is a declarative

FIGURE 6  Inferred class hierarchy, with inferred instances of WordClass; inconsistency highlighted in red
language, not a programming language. For instance, classes and their members can be defined in OWL through either explicit class assertion (extensionally), e.g., by stating that barra- is an instance of Prefix, or through the definition of properties (intensionally), e.g., by specifying that every instance of the Morpheme class with the properties isGrammaticallyDependent = true, isPhonologicallyDependent = true, and hasPosition = beforeStem is an instance of the Prefix class.

Different property types are supported in OWL: data property (between a class and a data type, such as a string and integer), object property (between two classes), and annotation property (for adding meta information that is not part of the formal ontology and thus reasoning, such as a label, comment, or definition). Examples are maximumSpeakerCount: LanguageSize → Integer, isSpokenInRegion: Language → GeographicRegion, and contributor (with a string in the range), respectively.

OWL has a model theoretic formal semantics, and some of its languages allow for reasoning. It employs the open world assumption. OWL itself is built upon the Resource Description Framework (RDF), a W3C standard for representing information in the Web, and on RDF-Schema (RDF-S), a semantic extension of RDF (see Brickley & Guha, 2014; Schreiber & Raimond, 2014).

6.2 | Ontology editor: Protégé

Protégé (BMIR, 2012) is the leading open source ontology editor (Gašević, Djurić, & Devedšić, 2009, p. 194). It provides a graphic user interface that allows users to model and edit OWL ontologies, based on the above-mentioned building blocks. Protégé allows for extensions through plugins. This includes several reasoners. Figure 7 shows a screenshot of Protégé. The screenshot shows a view of one class, UniqueMorpheme, displaying the class hierarchy of which the class is part in the top left window (it is a child to Morpheme), its property in the bottom left window (hasPosition),

FIGURE 7  Screenshot of a morpheme ontology using the ontology editor Protégé
and its formal description in the bottom right window (in particular adding the attribution
\texttt{isUnique = true} to the ones inherited from other classes).

\section{CONCLUSION}

In this paper, I have given an overview of the core ideas of an ontological approach to linguistics, which opens up a new avenue in linguistic research. This has included an introduction to notions of ontology, in particular in linguistics, as well as of relevant dimensions of ontologies. Building blocks of ontologies and ontology design considerations and capabilities have been discussed, and implementation pointers provided. A typological scenario was chosen as example, as it allowed us to draw on a broad range of linguistic examples at different depths of analysis. As outlined in the introduction and throughout the text, there are a number of potential advantages to an ontological approach in linguistics: It facilitates dealing with dispersed data and more generally the complexity and incompleteness of cross-linguistic data, terminological issues, and different data formats. It may also help with generating new explicit knowledge and assist in the reuse of existing knowledge.

As we have seen, an ontological approach allows (i) the formal modelling of speakers’ cross-connected network of linguistically relevant concepts (“speaker ontology”)—thus enabling the modelling of, e.g., semantic and grammatical relationships in a fine-grained way—and (ii) the explanation of the linguistic scholars’ conceptual network—i.e., the modelling of the domain-specific ontology of the general linguistic description apparatus (“discipline ontology”). Haspelmath (2010, p. 680) referred to the latter as being characterised by “rampant many-to-many relationships”, for which “a taxonomic conceptualisation, while logically possible, only obscures matters”. However, with the new Semantic Web developments, we now have the technical infrastructure and capabilities to create and work with cross-connected networks of this complexity and to establish “a storehouse of discipline knowledge that is integrated enough to allow access from different perspectives and with different aims in mind” (Schalley, 2013). Of course, such resources will always remain simplified representations of the linguistic (sub)domains and are by no means intended to replace fine-grained and in-depth discipline discourses. They can, however, serve as common reference points, which can support the unearthing of misunderstandings and gaps in said research discourses.

\textbf{ENDNOTE}

\footnote{It should be noted, however, that WordNet as such is not a formal ontology.}

\textbf{ORCID}

Andrea C. Schalley \(\oplus\) https://orcid.org/0000-0003-1323-1548

\textbf{WORKS CITED}


*Ontologies of linguistic annotation (OLiA)*. 2012. Available online at http://www.acoli.informatik.uni-frankfurt.de/resources/olia/


**AUTHOR BIOGRAPHY**

**Andrea C. Schalley** holds the positions of Chair Professor in English Linguistics at Karlstad University in Sweden and Adjunct Professor at Griffith University in Brisbane, Australia. Her research is situated at the interface of linguistics, sociology, cognition, and computer science. Amongst her interests are ontology-based linguistics (ontolinguistics), typology, the study of natural language meaning (semantics), the relationship between language and cognition, and bi/multilingualism. She has published widely in high-quality outlets, and her publications include two monographs, six edited volumes and special issues, and numerous journal articles and book chapters. She holds a Research Masters' in Mathematics and a Secondary School Education Degree from RWTH Aachen, Germany, as well as a PhD in Theoretical Linguistics from the University of Munich, Germany.

**How to cite this article:** Schalley AC. Ontologies and ontological methods in linguistics. *Lang Linguist Compass*. 2019;e12356. https://doi.org/10.1111/lnc3.12356