

Contents

Using ‘Ontolinguistics’ for language description

Scott Farrar

1

Index of names

19

Subject index

21

Using ‘Ontolinguistics’ for language description

Scott Farrar

1. Introduction: the knowledge sort problem

An aim of descriptive linguistics is to provide an account of the observable facts concerning individual languages. As such, descriptive linguistics is primarily concerned with data that bring out notable characteristics of particular languages. But in descriptive accounts, as well as those of a more theoretical nature, it is often a problematic endeavor to determine the difference between language and what language is *about*. The difficulty is often reflected in the terminology used to analyze language data, as very often linguists mix linguistic with non-linguistic terms. For example, an outsider to linguistics might be surprised to find in our theoretical machinery such notions as ‘animate’ and ‘shape’, and especially to see them intermixed so freely with notions as ‘tense’ and ‘grammatical gender’. Such notions are juxtaposed routinely in descriptive accounts of language and usually without consideration of how such notions relate to one another, or if they even make sense as categories. The two sorts listed here differ such that the former is usually associated with what linguists call *world* or simply *non-linguistic* knowledge. The latter on the other hand falls squarely under the heading of *linguistic* knowledge. The situation with descriptive linguistics is indicative of a larger issue within the entire field: the precise ontological nature of linguistic knowledge, a primary aim of Ontolinguistics.

Indeed there exists a relatively large body of research debating the necessity of world knowledge in accounting for linguistic phenomena and whether, in the context of particular linguistic theories at least, there ought to be such a distinction in the first place. An excellent introduction to the main issues of the *knowledge sort problem*, as it is called here, is given in Peeters (2000). Theories of world knowledge have traditionally been left up to philosophers or, more specifically, to ‘ontologists’. Recently, however, the fields of knowledge engineering and artificial intelligence have emphasized the creation and use of ontologies as tools for organizing world and domain-specific knowledge for expert systems and other knowledge-rich applications (e.g., Niles and Pease 2001). Though this work (and the literature concerning Ontology proper) seems quite relevant to the knowledge sort problem in linguistics, it

has until very recently been largely separate from the linguistics literature. Bateman (1997) offers a discussion of some notable exceptions to this trend. This situation is changing, and the emergence of knowledge-rich linguistics – known in this volume as ‘Ontolinguistics’ – is one example of the change. Ontolinguistics concerns itself with describing the conceptual content behind linguistic code and how the conceptual system might affect the organization of language, and possibly vice versa. Ontolinguistics, then, must address the knowledge sort problem head on.

This chapter, then, proposes a descriptive ontology that is compatible with the aims of Ontolinguistics. This ontological account makes explicit the distinction between linguistic and non-linguistic knowledge and allows for flexible relationships among the elements of these knowledge sorts. It is hoped that by clearly defining the knowledge sort problem, it may be better formalized, and the issue will become easier to resolve. The ontology that is proposed is the General Ontology for Linguistic Description (GOLD), first introduced by Farrar and Langendoen (2003). GOLD attempts to give an account of the most basic categories and relations used in the scientific description of human language. That is, GOLD should – ideally at least – capture the knowledge of a well-trained documentary linguist. In an attempt to capture knowledge that is widely accepted – what can be considered the “standard knowledge” of the field – canonical academic sources were used in the construction of GOLD, especially for the wide variety of morphosyntactic features (e.g., Crystal 1997). As a descriptive ontology meant to be useful for real-world data, GOLD is also empirically grounded in the sense that actual data produced by field linguists has been consulted, especially for the more specific categories.

The specific goals of this chapter are as follows: Section 2. describes the methodology for ontology construction with reference to various knowledge components. Section 3. describes the GOLD ontology itself. It is concluded in Section 4. that GOLD offers a possible starting point for further development of a comprehensive Ontolinguistics framework.

2. Methodology for ontological engineering

This section will serve to introduce a concrete methodology that derives from the basic tenets of ontological engineering. As such, the details of the proposed ontology will not be taken up in the current section, rather the general

steps in its creation will be explained. The following discussion is inspired from Borgida and Brachman (2003: 379), Guarino and Welty (2003, 2004), and Franconi (2004: 30–34).

The first step in the realization of the ontology is to enumerate the basic categories found in the domain of discourse. Linguistics has been traditionally subdivided into coarse domains such as syntax, discourse, phonology and the like, but there are also subdomains within the major domains, e.g., feature theory in phonology. Linguistic (sub)domains are relatively well-delineated in particular theories, but, as discussed in the introduction, a much more far-reaching, theory-independent organization is difficult to achieve due to overlapping terminologies and the mixing of linguistic types. For example, ‘Warumungu absolutive’, ‘Russian perfective’, and ‘English past tense’ are *instances* of morphosyntactic categories, and they are usually taken to be the atoms in approaches to morphosyntax. It would be odd to categorize a notion such as ‘noun’ as an instance of a morphosyntactic category or, more to the point, as an instance at all. The notion of ‘noun’ is clearly categorial in nature, and is thus a *class*. Classes have instances, e.g., particular nouns in particular sentences. In general whether an entity is category-like or instance-like is determined by *meta-ontological* criteria. To some extent, this is an arbitrary modeling choice. For example, in some special contexts ‘noun’ could be modelled just as well as an instance of ‘part of speech’. However, the commonly accepted notion of ‘noun’ is usually construed as a category for most linguistic theories. The point here is that ontological principles can be applied to reveal what the best modeling choice is for a notion such as ‘noun’.

Once the meta-ontological status of the various domain entities is determined, the next step is to develop class taxonomies. The way taxonomies are structured ultimately derives from basic philosophical assumptions and from the theoretical assumptions in the domain. One such general methodology for constructing taxonomies is the OntoClean methodology (cf. Guarino and Welty 2002, 2004). OntoClean uses the notion of *meta-properties* to motivate distinctions within an ontology. Meta-properties are properties of properties, not of objects in the world and are used to constrain ontology development and to evaluate particular proposed ontological organizations. The meta-properties particularly important for OntoClean are: ‘rigidity’, ‘identity’, ‘unity’ and ‘dependence’. Rigidity refers to essential properties, i.e., properties that an entity cannot lose without ceasing to exist; identity refers to properties used for discriminating among entities; unity refers to the wholeness of an entity, i.e., whether it has parts, boundaries and so on; and depen-

dence determines whether an entity can exist independently or if it needs to be carried by another (e.g., the color of an object is dependent for its existence on that object, the hole of a doughnut is dependent for its existence on that doughnut, etc.). Ontoclean would ensure that a class such as COMPOUND could not be both a subclass of WORD and CONSTITUENT, at least not in the same ontological stratum.

But at a more general level, the various domain-specific classes must be merged with the upper ontology. In order to carry this out, the nature of each domain-specific class must be determined. For instance, to use the categories of SUMO (Niles and Pease 2001), whether a class is ABSTRACT or PHYSICAL is the most basic distinctions that can be made at the upper level. As most linguistic phenomena are indeed abstract – that is, other than actual utterances or printed words (see Section 3.) – a more problematic decision is whether an entity belongs to ATTRIBUTE or RELATION. To anticipate the discussion that follows, it will be argued that linguistic features are one example of a class that lends itself to classification as an ATTRIBUTE, actually INTERNALATTRIBUTE. Discourse relations are one example of a class of entities best modelled as instances of RELATION.

Once the various taxonomies are developed, it is then necessary to add the individuals that are always present in the domain, those class instances that must be there for further ontological modeling. Particular linguistic relations, such as ‘constituentOf’, are examples of this. The next step in the methodology is to take the basic relations and identify the domain and range restrictions according to the available classes. For example, the discourse relation ‘expansion’ must take two discourse entities as its domain and range arguments.

At this point, it is necessary to further refine the various ontological entities by providing definitional axioms that state what must hold given the classes, relations, and individuals in the ontology. For example, one axiom might postulate the following: syntactic elements belonging to the category ‘adverb’ never have a ‘tense’ feature. For example, *if* Greenbergian-type universals (Greenberg 1966) are to be encoded in the ontology, then it is at this point where they should be included. The amount of detail depends of course on the particular application of the ontology. In general axioms should be limited to asserting what *must* be the case versus what *can* be the case. In short, the following enumeration shows the methodology used for the current work, and is a suggestion for constructing ontologies in general:

1. Development of the overall structure of the ontology

- (a) Enumerate the entities found in all states of the domain of discourse.
- (b) Classify the entities according to whether they are a kind of class, relation, or instance.
- (c) Develop class and relation taxonomies.
- (d) Add the instances that always occur in the domain.
- (e) Devise partitions in class taxonomy.
- (f) Establish relations among classes according to available relations.

2. Axiomatization

- (a) Define internal structure of classes.
 - i. Add intrinsic properties of classes.
 - ii. Enumerate part-whole relationships and include them as relations in the relations hierarchy.
 - iii. Add equalities and inclusions for classes.
- (b) Define the internal structure of relations.
 - i. Encode the value and cardinality restrictions on relations.
 - ii. Add equalities and inclusions for relations.

3. The ontology

3.1. Strata of linguistic analysis

A complete ontological account of language must take into account its ‘multi-stratal’ nature, as noted for example by Halliday and Matthiessen (2004: 24). Language is multi-stratal because it can be analyzed from a variety of points of view: form, meaning, structure, expression, etc. (We limit ourselves here to these particular strata, but note that a complete ontological account requires a treatment of the social aspects of language as well.) For descriptive purposes at least, the separation of these different kinds of entities into various strata is necessary, because it is then possible to focus only on one stratum in an analysis, as is often done in descriptive linguistics. This section, then, argues for a separation of the concerns of linguistics into various strata, but also for a unifying entity that acts as the “glue” holding the various linguistic strata together.

Traditionally, the notion of the ‘linguistic sign’ is used for such a unifying entity (cf. de Saussure 1959; Hjelmslev 1953). In Saussurean terminology, the sign exists only by virtue of the existence of a *signifiant* ‘expression’ and a *signifié* ‘content’. That is, the construct of the sign captures the so-called ‘duality of language’, or the simultaneous merging of form and meaning. GOLD embraces a more detailed account of the sign that is closer to semiology of Hjelmslev. For Hjelmslev, a sign’s expression is actually composed of ‘expression-form’ and ‘expression-substance’, the former being the abstract (*signifié*) portion of the sign and while the latter is the physical realization of that form, i.e., the actual speaking, writing, or signing event in the world. Likewise, ‘content-form’ is the abstract thought/concept (*signifiant*) which is ‘meant’ by the expression-form, while ‘content-substance’ is the actual physical event or object in the world. Modern approaches to grammar of course recognize a more complex sign, one that also takes into account its structure, for example, any of the various generative approaches to grammar (cf. HPSG, Pollard and Sag 1994). We can say, then, that the GOLD sign is a merging of form, meaning, and structure. Including also a class for the sign’s realization, Figure 1 shows the GOLD class SIGN and illustrates its relationship to the major strata of interest for descriptive linguistics. The following sections give an account of each of the classes listed in the figure.

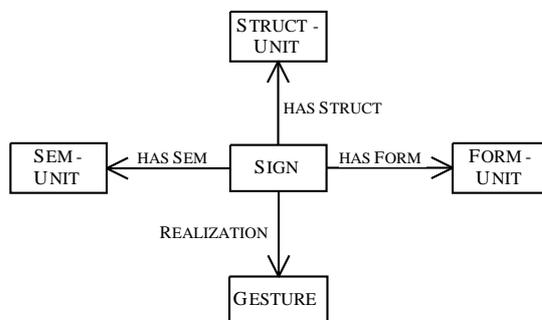


Figure 1. The SIGN as it relates to various strata

3.2. Substance and form of the sign

The first stratum to be discussed is that of linguistic form, called in the Hjelmslevian account ‘expression-form’. We propose the class SIGNFORM to subsume all kinds of expression-form units. For this preliminary investi-

gation, we focus on the type of form that corresponds to spoken expressions, PHONFORM. Using the terminology of de Saussure (1959), these types of form are the abstract sound “images” of the sign. We do note, however, that a complete account of sign form must take into account various other kinds of form units, including those based on the visual medium (e.g., various sign languages) and tactile medium (e.g., in the case of Braille). PHONFORM, and all kinds of SIGNFORM, however, have to be formulated according to expression-content, or what is physically observable in the world, e.g., the process and product of speaking (see also Bateman 2004 for a discussion). The strategy for this section is to start with various observables of the spoken sign – acoustic and articulatory – and build up various form units.

The investigation into the observables of the form-content is facilitated given GOLD’s connection to various sound- and process-related categories from the upper ontology. The general assumption is that various sorts of gestures or actions produce sounds. Following Browman and Goldstein (1992) and others, for example, we may capture various “units of action” by introducing the category SPEECHGESTURE. Such a class can be minimally defined as a subclass of SUMO:BODYMOTION. Such gestures have an inherent temporal dimension and form the articulatory basis of SIGNFORM. Subclasses of SPEECHGESTURE include gestures such as GLOTTISOPENING, TONGUEROOTADVANCING, VOCALIZING, etc. The character of various types of PHONFORM, then, can be defined according to what kinds of articulatory qualities or features are possessed by particular types of SPEECHGESTURE. But PHONFORM can also be defined in terms acoustic observables. The primary entity of concern in acoustic analysis is the sound signal, captured in GOLD as SPEECHSIGNAL, a proposed subclass of the more general SUMO SOUNDRADIATING process. We can state a preliminary dependency relation between SPEECHGESTURE and SPEECHSIGNAL: any instance of a SPEECH-SIGNAL is dependent upon an instance of SPEECHGESTURE, as there can be no sound signal without a speech gesture.

Various types of PHONFORM, then, can be defined according to combination of articulatory and acoustic observables. For example, a key type of form is the ‘phoneme’, or PHONEME in GOLD. The traditional view is that a phoneme is a ‘bundle’ of recurring articulatory or acoustic features in a language. The obvious ontological question that arises is where do the features come from? Do they carry out an independent existence, or are features dependent on some other kind of entity? We take up a more systematic

discussion of features in Section 3.5. It should be noted here, however, that an ontological treatment suggests a view of ‘distinctive features’ that goes against the traditional understanding: phonological features are “...the ultimate components [of speech], capable of distinguishing morphemes from each other” (Jakobson and Halle 2002: 3). If the class PHONEME is said to exist at all, then it is possible to create subclasses of phonemes, e.g., VOWEL and CONSONANT, based on a certain observable phenomena. But simply enumerating various subclasses of PHONFORM according to bundles of features and calling them phonemes is not sufficient. Finally, forms are either simple or complex. While the simple forms are captured by PHONEME, the complex units are composed of one or more instances of PHONEME. The simplest type of complex unit is the PHONOLOGICALWORD. Further investigation needs to be undertaken to build out these subclasses.

3.3. Content of the sign

To begin the discussion of how GOLD models the sign content, some assumptions concerning the overall approach to meaning are in order. GOLD assumes that the semantic system of a language is the bridge between its grammar and the non-linguistic knowledge which it is meant to convey. In other words the semantics ‘mediates’ (in the sense of Bateman this vol.) between the non-linguistic entities, e.g., SUMO:PROCESS and SUMO:OBJECT, and actual linguistic expressions. It is the job of semantic structure to lay out those aspects of meaning that are encoded directly in the grammar and lexicon, or simply how the speaker represents the world given the limitations of the linear linguistic signal. The link between a non-linguistic account of the world and a linguistic one expressed by the grammar is indirect. Thus, GOLD’s account of meaning adopts a form of two-level semantics in line with that of Bierwisch (1982), Lang (1991), Bateman (1997), Wunderlich (2000), and others.

Since it is the semantic system of a language that acts as an intermediary between a non-linguistic and a language-based account of the world, a language’s semantic system is necessarily unique. It must be unique because it is intimately tied to the the grammar and lexicon of a particular language. In order to accommodate this view in the current ontology, it is proposed to have language-specific extensions of GOLD for individual semantic systems. It is possible to introduce generalized ‘upper’ categories into the semantics.

Inspiration for this assumption is taken from the Generalized Upper Model (Bateman et al. 1990; Bateman, Henschel, and Rinaldi 1995) mentioned earlier. However, more research into semantic typology is necessary before a more complete account of semantics can be given.

GOLD captures all aspects of ‘sign content’ under the class SEMUNIT, as shown in Figure 1. To exemplify what kinds of categories could fit here, we explore one possible theory of meaning that is particularly complete, that which is suggested by Halliday (1985). According to Halliday the semantic system is composed of at least three levels or *metafunctions*. A metafunction can be described as a particular mode, facet, or layer of meaning. The three metafunctions in the work of Halliday are the *textual*, the *interpersonal*, and the *ideational* metafunction. The textual metafunction captures the meaning of the clause as ‘message’, or how it is used to construct a text. The textual metafunction is manifested by the theme-rheme and information structure of the grammar. The interpersonal metafunction captures the meaning of the clause as ‘interaction’, or how it is used to act in a discourse. The interpersonal metafunction is associated with the mood element of the grammar. Finally, the ideational metafunction captures the meaning of the clause as ‘experience’, or the propositional content of the sentence. The ideational metafunction can be seen, for example, in a language’s transitivity system and reflects the way the grammars classify and organize the world.

In the ontology, then, the three metafunctions are represented by three subclasses of SEMUNIT: TEXTUALUNIT, INTERPERSONALUNIT, and IDEATIONAL-UNIT. It is precisely the ideation units that are laid out in the Generalized Upper Model (Bateman, Henschel, and Rinaldi 1995). That is, classes in the GUM correspond to conceptualized units of meaning that make up the propositional content or ideation metafunction of an utterance. Ontologically, it is not possible for the meaning of an utterance to consist of only one of the above units. In many semantic theories, it is only the ideational content that is given.

3.4. Structure of the sign

Aspects of the sign’s structure are the traditional concerns of morphology, formal syntax, and to some extent discourse analysis. That is, the structural component accounts for notions such as morpheme, syntactic word, and text unit. We refer to the general category subsuming all structural units as

STRUCTUNIT, which in turn subsumes the classes MORPHUNIT, SYNUNIT, and TEXTUNIT.

The criterion used to classify STRUCTUNIT into the three major classes is the kind of relationship in which they participate. Since the kinds of relations relevant for morphological, syntactic, and discourse level units are quite different (cf. from highly structured syntactic relations to functional discourse relations), there is something fundamentally different about the structures themselves. From another point of view, a split based on types of relations makes clear the intuitive distinction (for linguists anyway) among morphology, syntax, and text. The relation of morphological constituency is called CONSTITUENT, and holds between various instances of MORPHUNIT, that is, morphological units both simple (morphemes proper) and complex (e.g., multi-morphemic signs such as STEM). The point is that structures of type MORPHUNIT do not otherwise participate in syntactic relations, and only subclasses of MORPHUNIT can participate in CONSTITUENT relations. Next, there are various syntactic relations: HEAD, ADJUNCT, SPEC just to name a few. These types of relations hold between instances of SYNUNIT, that is, syntactic constituents both simple (e.g., SYNWORD) and complex (e.g., CONSTRUCTION). Finally, there are textual constituency relations that relate instances of DISCUNIT.

As the classification of STRUCTUNIT gets more specific, the criteria of (a) structural complexity and (b) how the units relate to a system as a whole are used. Consider SYNWORD and SYNCONSTRUCT, where the former is a single structure that occupies exactly one syntactic position and the latter consists of multiple signs in standing in syntactic relations to one another. Relations also hold between the various levels of the STRUCTUNIT including, for example, those that hold between a SYNWORD and a MORPHUNIT: INFL, DERIV, and ROOT. These relations are available for use in a language description, but mainly they act to axiomatize the entities in the ontology. So, for example, SYNWORD by definition must have at least one root, or STEM must have at least one derivational component.

3.5. Features of the sign

It is a very common assumption in linguistics that signs carry features, as they are commonly called. And since the analysis of linguistic data routinely utilizes the notion of a linguistic feature, GOLD should provide for the integration of features into its overall ontological account. First of all we pro-

pose that features are attribute-like entities predicated of the sign instances themselves. That is, the various components of the sign, e.g., FORMUNIT or SEMUNIT, are not *composed* of features; rather, their make-up determines the sign’s features. There are several varieties of features which linguists find useful for language description, including formal, semantic and phonological features. The structural component provides formal, structural features, called STRUCTFEATURE; the form component provides form-related features subsumed under FORMFEATURE, e.g., phonological features; and the content component provides features that show the meaning of the sign, called SEMFEATURE.

As an illustration of the ontological treatment of features, consider the so-called ‘grammatical features’, a special type of STRUCTFEATURE. The class of structural features determining the formal behavior of signs in morphosyntax is called MORPHOSYNFEATURE. Morphosyntactic features determine the behavior of morphosyntactic units. We propose that the morphosyntactic features TENSE, ASPECT, NUMBER, etc. are instances of MORPHOSYNFEATURE. We also give a general class representing the morphosyntactic feature values, called MORPHOSYNVALUE. Subclasses of MORPHOSYNVALUE include TENSEVALUE, ASPECTVALUE, NUMBERVALUE, etc., those corresponding to the instances of MORPHOSYNFEATURE. For example, each instance of MORPHOSYNFEATURE, e.g., TENSE, has a corresponding value from MORPHOSYNVALUE, in this case, from TENSEVALUE.

3.6. Types of signs

With the major components of the sign now described, we turn to the classification and organization of the sign itself. Traditional conceptions of the sign usually focus on single words or morphemes. Following Hervey (1979) we assume that signs are not limited to words and morphemes, but encompass more complex syntactic and textual units. In short, every level of formal linguistic structure (from morphemes to whole texts) can be related to the SIGN. Various linguistic theories assume some sort of ‘ontology’ for these kinds of units, though a limited number of theories specifically refer to the sign in their account. Head-Driven Phrase Structure Grammar (Pollard and Sag 1994) is one example that does refer to the sign. In fact the GOLD account draws inspiration from and is similar in many respects to that of HPSG’s.

Signs can be classified according to a number of ontologically different, cross-classifying criteria: (1) according to the kinds of relationships in which the sign participates (e.g., syntax- vs. discourse-level relations) (Hjelmslev 1953: 41; Pollard and Sag 1994); (2) according to the complexity of the sign (cf. the complexity of a morpheme to a main clause) (Hervey 1979: 10); and (3) according to the kind of system to which the sign belongs (e.g., English signs vs. Swahili signs). Signs can conceivably be classified according to what they mean or according to their form (e.g., how they sound); however, no literature was found to support either of these possibilities. Since all these criteria play a role in the nature of signs, the problem becomes how to make the cut in a way that accords with commonly-accepted categories of modern linguistics and in a way that makes sense according to principles of formal ontology, e.g., without resorting to multiple inheritance, which leads to overly complex taxonomies and is not considered good practice in ontological engineering (cf. Guarino and Welty 2002, 2004).

The question really comes down to this: why categorize SIGN? The answer that will be defended here is that signs are categorized to capture generalizations. Since many of the formal characteristics of the sign are already captured in the structural component, a classification according to Criterion 1 is redundant. A similar argument can be made for Criterion 2, since complexity is expressed as every stratum. The generalization that has yet to be dealt with concerns the nature of the sign with respect to the overall framework of language, namely Criterion 3. Thus, some resulting subclasses would be: SWAHILISIGN, ENGLISHSIGN, RUSSIANSIGN, etc.

But having sign classes based on particular languages does not provide any insight in the ontological treatment of language itself, assuming that a class LANGUAGE can be formalized. The problem is that using the mechanisms of formal ontology, there needs to be a way to state that a sign belongs to a particular language. The problem is actually much more general, in that there needs to be a way to distinguish between the knowledge of specific languages and the knowledge of language as a general concept. Preliminary to this discussion is the ontological status of language itself. Given GOLD's sign-based approach, LANGUAGE could simply be defined, after the Saussurean tradition, as a 'system of signs'. The first step in such an endeavor is to formalize notion of a sign inventory for particular languages, that is, to formalize LANGUAGE as a set-theoretic notion. The next step would be to elaborate on the notion of 'system'. But in this chapter, because GOLD is still on-going research, we will only introduce some preliminary ideas on this

topic, leaving aside an ontological treatment of 'system'. Thus the questions to be addressed here are:

- How can the notion of a sign that is shared across a system (a speech community) be formalized?
- What criteria can distinguish individual signs from one another?

The first question concerns the difference between, for example, a word as shared among speakers of a particular language and the word used in a particular linguistic event. For example, the current argument differentiates between, for example, the sign **|DOG|** as shared by speakers of English and the instance of that sign in my saying of *Only mad dogs and Englishmen go out in the noonday sun* last summer. Fortunately, the work of Hervey (1979) already establishes a basis for such an ontological treatment and attempts to answer just this question. In Hervey's analysis, an individual sign in a given language is actually the set of all its occurrences in the language or, as he puts it, "a model for a set of speech facts" (Hervey 1979: 10), where the notion of a 'speech fact' is equated with the products of linguistic events (i.e., speaking, writing, or signing events). The model for a particular speech fact is called an 'utterance' (Hervey 1979: 11). Utterances, then, are the sign classes shared among members of a speech community. Hervey is not working in the framework of formal ontology, however, the formalization of the sign is a useful one for the purposes of GOLD. The task, then, is to alter Hervey's formalization slightly to coincide with the formal machinery of ontology: namely, the characterization of the sign in terms of classes and instances. This means that **|DOG|** is the class of signs that are similar enough in form and meaning to be recognized over and over in the speech community. If signs are classes, then every time someone utters *dog*, the sign **|DOG|** is instantiated. These instances can be represented graphically as: **|DOG|_A**, **|DOG|_B**, ..., **|DOG|_n**. Each one of these instances, by virtue of its being a sign, has the structure given in Figure 1. Thus, language specific signs can be added to the taxonomy of signs in general: the sign **|DOG|** is a subclass of **ENGLISHSIGN**; **|PERRO|** is a subclass of **SPANISHSIGN**; etc.

The second question concerns sign identity. That is, given two utterances, *mad dogs* which was uttered on January 1st, 1999, and *mad dogs* which was uttered on June 6th, 2004, how can it be determined whether or not their corresponding signs are instances of the same class? The identity criteria used by Hervey include whether or not the signs have similar forms and whether or not the signs have similar reference. (Hervey uses the term 'reference'

for what we have called SEMUNIT earlier.) A set of signs that have similar forms, but not similar references, is called a ‘form class’. A set of signs that have similar reference, but not similar forms, is a ‘reference class’ (Hervey 1979: 13). Crucially, in neither case are the signs identical. It is only when two signs have the similar form *and* similar reference that they can be classed together as instances of the same sign. Mentioning form and reference as a means to determine sign identity raises the question of why form or content cannot be used in the classification of the sign taxonomy. That is, why not group together all signs in all languages that refer to the concept DOG? Or why not group together all signs in all languages that share a similar form? In terms of form, the class of all signs that share a similar form, even across languages, is not a very interesting concept linguistically – except perhaps when considering the issue of the arbitrariness of the sign. In terms of content, such a classification already exists in the ontological model, namely in the classification of the entities that signs refer to, as in the structure of various semantic units. That is, there is no need to repeat such a structure in the taxonomy of the sign when it can be derived from the overall organization of knowledge.

4. Summary and Discussion

A descriptive ontology was discussed in the context of Ontolinguistics, namely the General Ontology for Linguistic Description (GOLD) as introduced in Farrar and Langendoen (2003). First, a step-by-step methodology for creating such an ontology was given. Of particular importance is whether or not entities in the domain of discourse are classes, relations, or instances. Finally, a detailed description of GOLD was presented which focused on the class of SIGN as the central entity that binds elements from all strata of language. Concrete suggestions were given concerning how GOLD could be linked to SUMO (Suggested Upper Merged Ontology) (Niles and Pease 2001). The resulting ontology has been presented in the context of what I have referred to as ‘the knowledge sort problem’, the problem of distinguishing linguistic from non-linguistic knowledge. It has been argued that GOLD offers a jumping off point for exploring knowledge sort problem by offering a clear formalization the kinds of knowledge needed for linguistic analysis and description. The knowledge sort problem is so fundamental to the field of linguistics, because the pursuit of a solution gets at the relationship of linguistics to other scientific disciplines. How can it be possible for the linguistic sign

to *mean* something to someone? What aspects of reality are coded in speech? How is the conceptualization of reality related to language? How does the conceptualization of reality affect the way in which these signs are arranged and interpreted? These questions are exactly what the emerging field of Ontolinguistics attempts to answer.

Acknowledgments

The initial development of GOLD was supported by the Electronic Meta-structure for Endangered Language Data (E-MELD) grant from the U.S. National Science Foundation (NSF 0094934), which is gratefully acknowledged. Subsequent support came from the Data-Driven Linguistic Ontology grant (NSF 0411348). I would also like to thank Andrea Schalley, Dietmar Zaefferer, Achim Stein, John Bateman, and Adam Pease for their comments on earlier drafts of this paper. I am indebted to Terry Langendoen, Will Lewis, Gary Simons, and Adam Pease for their help in the development of GOLD.

References

- Bateman, John A.
1992 The theoretical status of ontologies in natural language processing. In *Text Representation and Domain Modelling – Ideas from Linguistics and AI (Papers from the KIT-FAST Workshop, Technical University Berlin, October 9th–11th 1991)*, Susanne Preuß and Birte Schmitz (eds.), 50–99. (KIT-Report 97, Technische Universität Berlin, Berlin, Germany).
- 2004 The place of language within a foundational ontology. In *Formal Ontology in Information Systems: Proceedings of the Third International Conference (FOIS 2004)*, Achille C. Varzi and Laure Vieu (eds.), 222–233. Amsterdam: IOS Press.
- this vol. Linguistic interaction and ontological mediation.
- Bateman, John A., Renate Henschel, and Fabio Rinaldi
1995 Generalized upper model 2.0: documentation. Technical Report, GMD/Institut für Integrierte Publikations- und Informationssysteme, Darmstadt, Germany.
- Bateman, John A., Robert T. Kasper, Johanna D. Moore, and Richard A. Whitney
1990 A general organization of knowledge for natural language processing: The PENMAN upper model. Technical report, USC/Information Sciences Institute, Marina del Rey, California.
- Bierwisch, Manfred
1982 Formal and lexical semantics. *Linguistische Berichte* 80: 3–17.
- Borgida, Alex, and Ronald J. Brachman
2003 Conceptual modeling with Description Logics. In *The Description Logic Handbook*, Franz Baader, Diego Calvanese, Deborah McGuinness, Daniele

- Nardi, and Peter F. Patel-Schneider (eds.), 349–372. Cambridge, UK: Cambridge University Press.
- Browman, Catherine P., and Louis Goldstein
 1992 Articulatory phonology: An overview. *Phonetica* 49: 155–180.
- Crystal, David
 1997 *Cambridge Encyclopedia of Language*. 2d ed. Cambridge, UK: Cambridge University Press.
- Farrar, Scott, and D. Terence Langendoen
 2003 A linguistic ontology for the Semantic Web. *GLOT International* 7 (3): 1–4.
- Franconi, Enrico
 2004 Description logics for conceptual design, information access, and ontology integration: Research trends. [TODO:MS_OR?] Available at <http://www.inf.unibz.it/~franconi/dl/course/tutorial/>.
- Greenberg, Joseph
 1966 *Language Universals*. The Hague: Mouton.
- Guarino, Nicola, and Christopher Welty
 2002 Evaluating ontological decisions with OntoClean. *Communications of the ACM* 45 (2): 61–65.
 2004 An overview of OntoClean. In *Handbook on Ontologies*, Steffen Staab and Rudi Studer (eds.), 151–172. Berlin: Springer.
- Halliday, Michael A. K.
 1985 *An Introduction to Functional Grammar*. London: Edward Arnold.
- Hervey, Sándor
 1979 *Axiomatic Semantics: A Theory of Linguistic Semantics*. Edinburgh: Scottish Academic Press.
- Hjelmslev, Louis
 1953 *Prolegomena to a Theory of Language*. Bloomington Indiana: Indiana University Publications in Anthropology and Linguistics.
- Jakobson, Roman, and Morris Halle
 2002 *Fundamentals of Language*. Reprint of the 2d revised edition. Berlin: Mouton de Gruyter. Original edition, [TODO:PUBLICATIONLOCATION: PUBLISHER], 1971.
- Lang, Ewald
 1991 The LILOG ontology from a linguistic point of view. In *Text Understanding in LILOG: Integrating Computational Linguistics and Artificial Intelligence Final Report on the IBM Germany LILOG-Project*, Otthein Herzog and Claus-Rainer Rollinger (eds.), 464–481. Berlin: Springer.
- Niles, Ian, and Adam Pease
 2001 Toward a Standard Upper Ontology. In *Formal Ontology in Information Systems. Proceedings of the 2nd International Conference (FOIS-2001)*, Christopher Welty and Barry Smith (eds.), 2–9. New York: ACM Press.
- Peeters, Bert (ed.)
 2000 *The Lexicon-Encyclopedia Interface*. New York: Elsevier.
- Pollard, Carl, and Ivan Sag
 1994 *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press.

Saussure, Ferdinand de

1959 *Course in General Linguistics*. London: Peter Owen.

Wunderlich, Dieter

2000 Predicate composition and argument extension as general options: A study in the interface of semantic and conceptual structure. In *Lexicon in Focus*, Barbara Stiebels and Dieter Wunderlich (eds.), 247–270. Berlin: Akademie.

Index of names

Bateman, John A., 2, 7–9
Bierwisch, Manfred, 8
Borgida, Alex, 3
Brachman, Ronald J., 3
Browman, Catherine P., 7

Crystal, David, 2

Farrar, Scott, 2, 14
Franconi, Enrico, 3

Goldstein, Louis, 7
Greenberg, Joseph, 4
Guarino, Nicola, 3, 12

Halle, Morris, 8
Halliday, Michael A. K., 5, 9
Henschel, Renate, 9
Hervey, Sándor, 11–14
Hjelmslev, Louis, 6, 12

Jakobson, Roman, 8

Lang, Ewald, 8
Langendoen, D. Terence, 2, 14

Matthiessen, Christian M. I. M., 5

Niles, Ian, 1, 4, 14

Pease, Adam, 1, 4, 14
Peeters, Bert, 1
Pollard, Carl, 6, 11, 12

Rinaldi, Fabio, 9

Sag, Ivan, 6, 11, 12
Saussure, Ferdinand de, 6, 7, 12

Welty, Christopher, 3, 12
Wunderlich, Dieter, 8

Subject index

- axioms, 4
- cardinality restrictions, 5
- descriptive linguistics, 1
- features, 7, 10
 - distinctive, 8
 - grammatical, 11
 - linguistic, 4
 - morphosyntactic, 2, 11
 - phonological, 8
 - theory of, 3
 - values of, 11
 - varieties of, 11
- Generalized Upper Model, 9
- HPSG, 11
- knowledge
 - linguistic, 1
 - world, 1
- knowledge sort problem, 1
- language
 - duality of, 6
- meaning, 8
- mediation, 8
- meta-properties, 3
- metafunction, 9
 - ideational, 9
 - interpersonal, 9
 - textual, 9
- observables, 7
- OntoClean, 3
- ontology
 - descriptive, 2
 - GOLD, 2
 - methodology, 2
 - upper, 4
 - upper categories of, 8
- properties
 - intrinsic, 5
- semantics
 - system of, 8
 - two-level, 8
- semantic system, 8
- sign, 7, 11
 - classification of, 12
 - complex, 6
 - content of, 9
 - identity of, 13
 - inventory of, 12
 - linguistic, 6
 - structure of, 9
 - taxonomy of, 13, 14
- sign languages, 7
- strata
 - linguistic, 5
- terminology, 1
- typology, 9
- universals, 4
- utterance, 13

