Can GOLD "cope" with WALS?
Retrofitting an ontology onto the World Atlas of Language Structures

Michael Cysouw, Jeff Good, Mihai Albu, Hans-Jörg Bibiko
{cysouw, good, albu, bibiko}@eva.mpg.de
Max Planck Institute for Evolutionary Anthropology, Leipzig

0. Introduction

The *World Atlas of Language Structures* (WALS, Haspelmath et al. 2005) is a large-scale "database of databases" consisting of 141 typological databases, covering a wide range of grammatical features, joined into one composite resource through the use of a common metadata scheme. While this metadata scheme ensures interoperability among databases across some dimensions (e.g., language names and families), it falls far short of allowing complete database interoperability. At present, a project is underway to "retrofit" an ontology onto this existing resource. Two broad questions being addressed by the project are: (i) What conceptual and design problems need to be solved in order to build an ontology "internal" to WALS which can allow for a high degree of interoperability among the WALS databases? and (ii) How can the WALS categories be related to a general ontology? Or, to put it another way, we are interested in determining (i) how we can build a worthwhile *Community of Practice Extension* (or COPE) for WALS and (ii) how the categories in this COPE can be related to categories in the *General Ontology for Linguistic Description* (henceforth, GOLD ontology; Farrar and Langendoen (2003)).

This paper is structured as follows. Section 1 gives some background information on WALS, including discussion of its history and its overall design. Section 2 discusses the basic methodology adopted by the WALS ontology project. This section also discusses some of the research results of the WALS ontology project which we believe may be relevant to the conceptual relationship between a typological COPE and the GOLD ontology. Section 3 discusses some ways in which we believe GOLD could be extended to better relate to typological concepts and some design desiderata for ontological tools which would facilitate the creation of an ontology like the WALS ontology. Finally, section 4 summarizes the current findings of this ongoing project.

1. Background: WALS

The *World Atlas of Language Structures* (WALS)[1] is a database of 141 typological databases each linked by a common metadata scheme. The databases cover a wide range of typological features across phonology, morphology, syntax, and the lexicon. A *feature*, in this context, refers to a grammatical parameter over which a given language is expected to have a *value*, an attested grammatical pattern considered to be a possible exponent of that feature. For example, one possible grammatical feature is *number of genders* and a language could have a value for that feature of *no gender*, *two genders*, *three genders*, etc. Each feature covered by WALS is associated with a particular author or group of authors, making its structure more comparable to, say, that of an edited volume than a monograph.

The original purpose of the WALS database was to facilitate creation of the WALS Atlas (Haspelmath, et al. 2005) a printed volume of maps corresponding to the data collected

[1] http://wals.info/

for the grammatical features covered by the project. Figure 1 gives one of the published maps, showing a view of the data collected for the feature *number of genders* (Corbett 2005a).

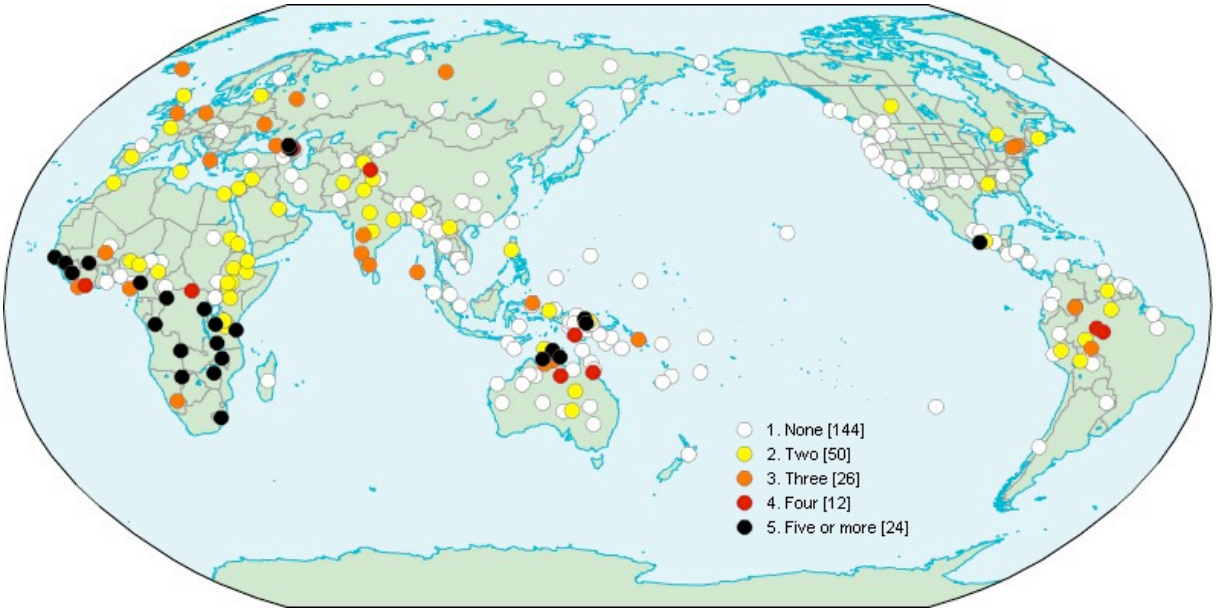

Figure 1 *Number of Genders (Corbett 2005a)*

In a typological resource designed for map production like this one, for a given grammatical feature, each language must, in some way, be reduced to a single, mappable data point. In the case of WALS, these data points are, in fact, what are stored as primary data. Each can be understood as a four-dimensional entity consisting of (i) a language code, (ii) a typological feature name, (iii) a value for the feature, and (iv) a reference for the source of the data. Each of these four dimensions can be associated with relevant metadata—for example, an English name and a latitude and longitude are associated with each language code, and each feature is associated with an author or set of authors. Of these four dimensions, the feature name is "privileged" in the sense that it was the primary dimension along which data was collected and organized. There are close to 60,000 primary data points in the database.

While it is sometimes convenient to refer to WALS as a single database, especially since it presently exists as a single FileMaker file, it is better understood as a collection of over 100 different typological databases which achieve a relatively high degree of interoperability through their use of a shared metadata scheme. This is because only the metadata was standardized during resource creation. Otherwise, the data entry for each grammatical feature was left largely up to the discretion of the authors with no provision for systematizing the meanings of the terms designating different features and values. This model for the structure of WALS (including some indication of different types of metadata used in the resource) is schematized in Figure 2.

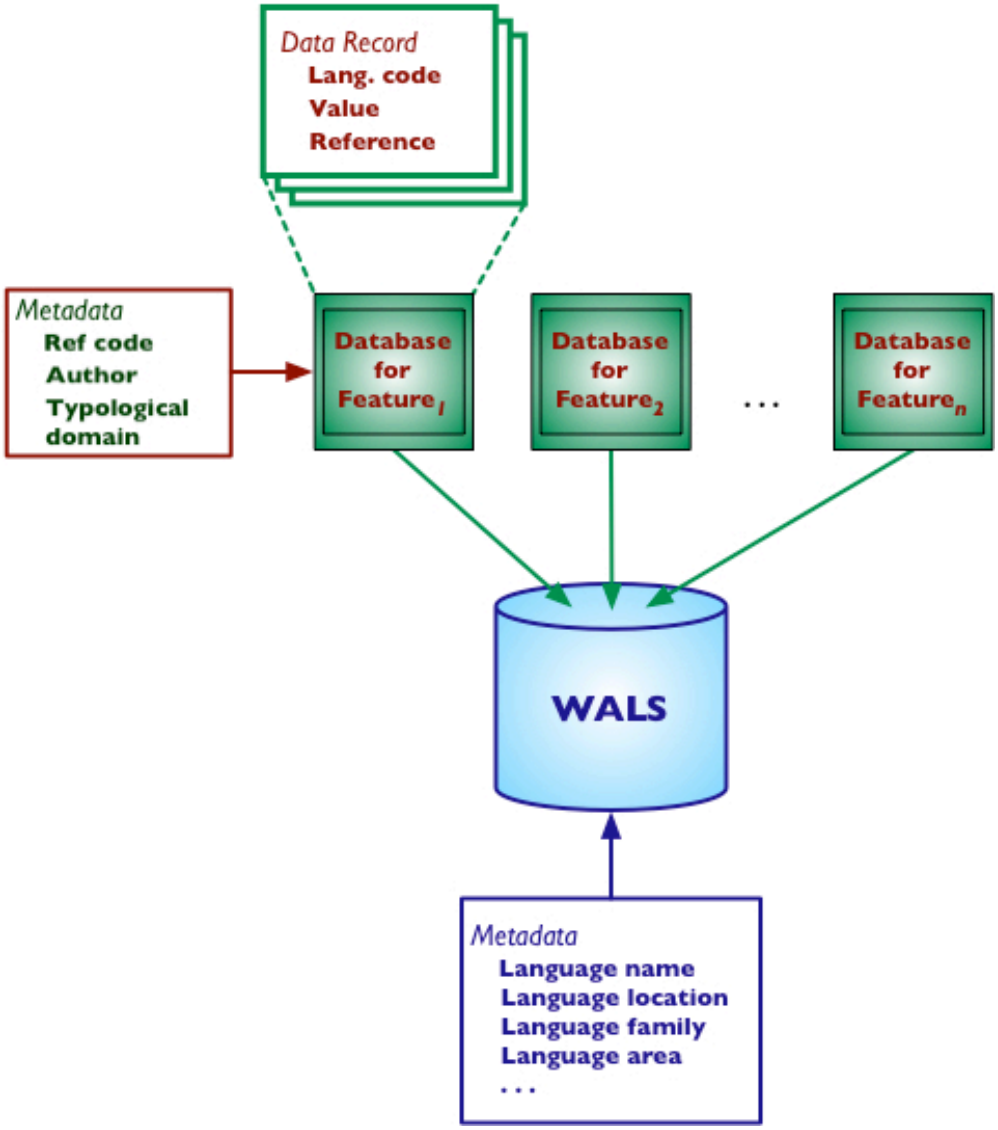

Figure 2 *The Structure of WALS, Conceptualized as a Collection of Databases*

Conceptualizing WALS a single database can lead to potentially false inferences, for example that a term designating a value for one feature has the same meaning when it is used to designate a value for another feature. This problem can be illustrated by examining the two maps produced from data in WALS seen below. The grammatical feature illustrated by the map in Figure 3 is *case syncretism* (Baerman and Brown 2005), and the grammatical feature of the map in Figure 4 is *exponence of selected grammatical formatives* (Bickel and Nichols 2005). Note that each feature is associated with a value which, naively, would seem to be interpreted to mean "this language does not have morphological case marking". In the case of the map in Figure 3 this value is *no case marking*. In the map in Figure 4, it is the value *no case*.

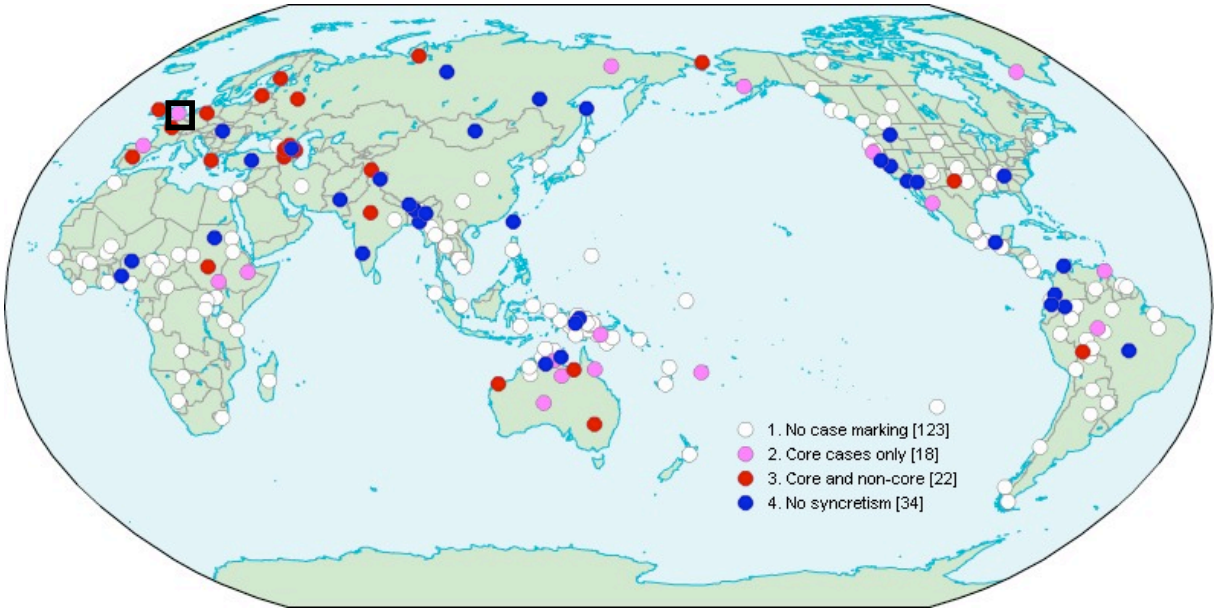

Figure 3 *Case Syncretism (Baerman and Brown 2005)*

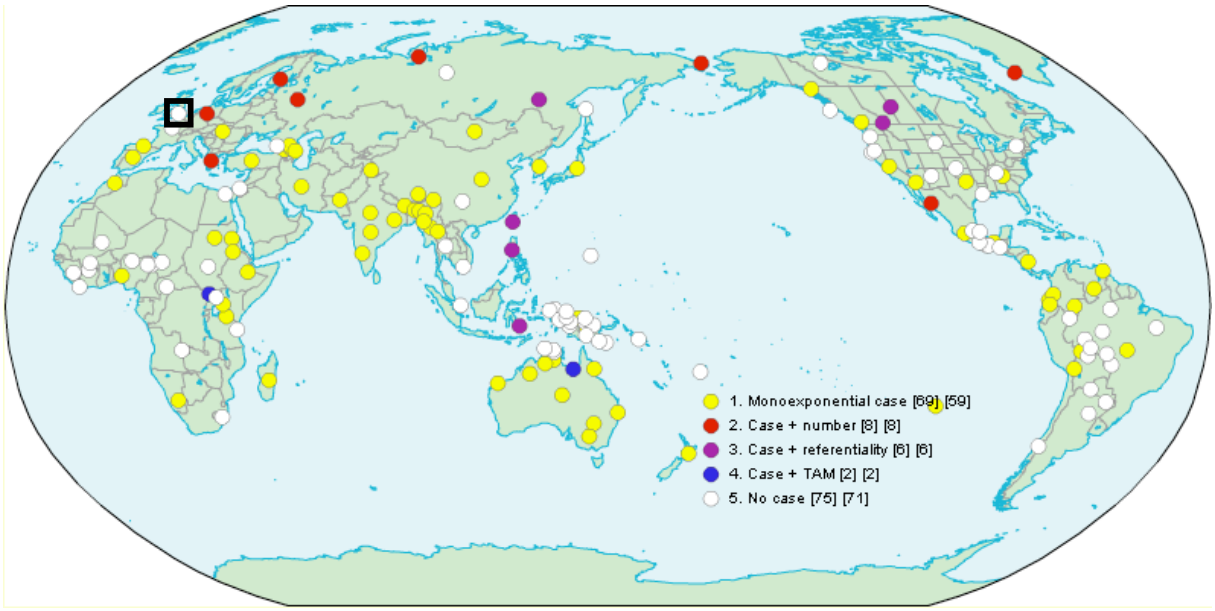

Figure 4 *Exponence of Selected Inflectional Formatives (Bickel and Nichols 2005)*

On each of the maps, the dot corresponding to English has been surrounded by a square. (For the purposes of WALS, English is localized to southern England.) Crucially, Figure 4 assigns English the value *no case*, implying that English does not have case marking. However, Figure 3 does not assign English the value *no case marking*, implying it does have case marking. This example has been specifically chosen since it should be clear to those with even a limited knowledge of English nominal and pronominal morphology why English might be categorized by one researcher as having case and by another researcher as having no case. Contradictory classifications like this would seem to have three possible causes: (i) genuine disagreement about classification, (ii) consultation of different sources/datasets, or (iii) the existence of typological "false friends". The first two of these possible causes are not

ontological in nature—that is, they cannot be fixed through the use of an ontology. They will, therefore, not be discussed further here. The third problem, however, can be, at least, partially dealt with through the use of an ontology, and we, therefore, will go into it in more detail.

In using the term *typological false friends*, we borrow from the use of the term *false friends* in foreign language instruction to refer to a pair of words in two different languages which, because of their form, are liable to being falsely considered as having the same meaning (e.g., English *actually* and French *actuellement* 'currently'). By analogy, a typological false friend refers to a case where the same term or similar terms are used in the context of two different research projects to refer to two different concepts—perhaps greatly different concepts or only subtly ones. Within the published version of WALS, typological false friends can often be straightforwardly identified by consulting a prose description of the feature and corresponding values accompanying each of the maps. This essentially follows standard research practice in linguistics of carefully defining all terms used within a work. However, while this method of term disambiguation works well for a book-based model of research, it is inadequate for database research where capabilities for automatic search and comparison are desired.

To again use the problem of English case categorization, suppose one wants to do a typological study looking for, say, correlations between morphological tense marking and case marking. In the present form of WALS, this is difficult because it is not at all clear which case classification data to use—should one treat English as a language having case or not? The actual "best" classification for English with respect to case is, of course, a matter of research, not solvable by changing the database structure. However, the database should provide information as to exactly how a term like "no case" should be interpreted so that it is completely clear, both to a human being and a machine, whether or not two terms which seem to be the same are *supposed* to refer to the same concept, and if, therefore, contrasting classifications represent disagreement or simply the use of different concepts.

Having given an overview of the structure of WALS, including justification for conceptualizing it as multiple databases, in the next sections we will discuss the major challenges that have been encountered by the WALS ontology project in attempting to construct an ontology of concepts found in the WALS data.

2. COPEing with WALS

2.0 Introduction

In this section, we discuss, in fairly concrete terms, the process through which we are developing a WALS ontology in a way which we believe will allow it to become a straightforward extension of the GOLD ontology and, thereby, serve as *Community of Practice Extension* (COPE) for typological resources. While the content of this section is rather specifically tied to the needs of the WALS ontology, we expect that projects like ours which seek to build an ontology on a legacy typological resource may be faced with comparable problems. For example, many of the problems that we are faced with are also described by Dimitriadis and Monachesi (2002), working on the Typological Database System.[3] In section 2.1, we discuss the process through which we first extracted a term set from WALS, which formed the foundation on which the WALS ontology will be built. In section 2.2, we present an example of how we envision such a term set could be used for a

[3] http://languagelink.let.uu.nl/tds/ontology/

typological COPE. The idea is to devise a typological metalanguage combining "primitive" concepts into some of the more complex concepts often used by typologists in language categorization. In section 2.3, we discuss some problems we encountered when conceptualizing how WALS concepts relate to GOLD concepts. The basic problem seems to be the notion of *linguistic system*, which is central to the practice of typology and much of linguistics in general. However, we will argue that the problems related to this do not necessarily need to have a major impact for current work on GOLD.

2.1 Extracting a term set from WALS

As a first step towards building a typological COPE, we investigated the texts that accompany each map in WALS. The authors of the maps were instructed to explicitly state the definition of their terms as precisely as possible. Of course, some authors were more particular in following these guidelines than others were. Still, the texts with these definitional statements are an interesting resource to use to investigate the usage of terminology in typology. For each map in WALS, we extracted all sentences related to the definitions of the terms used.[4] A term list was produced from these sentences, from which we removed the terms relating to segmental phonology, colour description, and orthography because we had the impression that they would not raise typology-specific problems for an ontology. This resulted in a first rough version of the term list, which in the current version consists of 400 words.[5] This list is not intended to be a definitive survey of the terms used in typology, but only as an indication of the kind of terms that should be accounted for in an ontology.

However, there are a few words of caution needed regarding this approach. A term set from typological databases is of a rather different conceptual type than a term set based on a particular language. Unlike language-specific terms, which apply to attested forms or constructions in a given language, typological terms are generalizations over a number of languages and are not intrinsically grounded in primary data. For example, in some language there might be a construction that is called the "Perfect" by a group of investigators. For the link to the ontology, it is not really important whether this construction is really a perfect, or maybe more of a simple past. There can be big debates; opinions might change; but the term can simply be linked to whatever category in the ontology seems most suitable at a particular point in time. The name "Perfect" is only a label referring to an empirically available construction in the language under investigation. For mnemonic reasons, we expect this construction to be at least something like a perfect, but this is not necessarily so: it might also have been called randomly "ZKFHS" or "construction type 253". Independent of the name, the term is empirically grounded in primary data, and it is a theoretically simple task (though obviously a laborious one) to collect all such terms as used in a linguistic sub-community and map them to an ontology.

In the case of typology, this kind of grounding is not available. It is exactly the goal of most typologies to compare the range of constructional variability found throughout the world's languages, which necessitates abstracting away from language-specific constructions. To achieve this, a typologist typically starts from a functional-semantic notion (the so-called *tertius comparationis*) and investigates the constructions used to express this notion in a sample of the world's languages. It is well-known to typologists that it is extremely difficult to find the right construction in a particular language. It is also widely acknowledged that the

[4] http://www.eva.mpg.de/~cysouw/pdf/WALSdefinitions.pdf
[5] http://www.eva.mpg.de/~cysouw/pdf/WALStermlist.pdf

terms used in the description of a particular language do not necessarily signify the same as the typologist would have expected. To overcome these problems, the object of investigation in a typological study is mostly very rigorously defined in functional-semantic terms,[6] restricted to a clearly defined syntactic domain (e.g. only within matrix sentences), and often also restricted along some morphological parameters (e.g. only inflectionally marked, but not on auxiliaries). However, even with such nicely explicated definitions, it is important to realize that all the terms used in the definition ("matrix sentence", "inflectionally marked", "auxiliaries") are not grounded. All these terms have to be understood from a general linguistic knowledge, in practice often enhanced by the presentation of various examples from widely differing languages, indicating how the definition should be interpreted.

One could say that for a typological term set, the terms do not have to be mapped onto an ontology: they *are* the terms from the ontology, though most often combined into terminological complexes. Because of the lack of grounding, the same terms used in different typologies often do not mean exactly the same thing. Although two different typologies might both claim to say something about, for example, the perfect, the definitional details will often be different, along with the constructions classified as exemplifying that concept. Both might still both be typologies of the perfect—but different perspectives on the same theme.

Such complications certainly do not argue against linking concepts in typological resources to general ontologies. However, they do suggest that research is needed into how best to model the differences between non-grounded typological concepts and grounded language-specific concepts with respect to ontological linking. Understanding such differences would seem relevant both from an abstract, knowledge engineering perspective and from the more concrete perspective of tool development—in the latter case, non-grounded categories might need to be treated differently from grounded categories with respect to smart search tools, for example.

2.2 Towards a typological meta-language

Ideally, we would like to develop a suitable meta-language to formalize the definitions used by typologists based on a restricted term set. However, this goal is far from trivial. As an example, consider the following extracts from the feature *sex-based and non-sex-based gender systems* (Corbett 2005b) given in (1). Typically for typology, the main grammatical characteristic in question (in this case "gender") is defined very precisely, which means that it is only classified as present when all definitional parameters are met. Because of this definitional detail, the usage of the term "gender" by one typologist is not necessarily the same as when it is used by another typologist—even if they both intend more or less the same concept by the term. However, the differences can often be established in great detail as long as all definitions are clear and all relevant parameters are taken into account.

[6] For example, consider the following extract of the definition of perfectives from the WALS feature by Dahl and Velupillai (2005a): 'To be interpreted as a perfective, we demand that a form should be the default way of referring to a completed event in the language in question. In many languages, there are forms or constructions that are used of completed events but only if some additional nuance of meaning is intended, for instance if emphasis is put on the result being complete or affecting the object totally. Such strong perfectives [...] exhibit relatively large variation cross-linguistically. They are often called "perfectives" in grammars but are not counted as such here.'

(1) Extracted definitions from Corbett (2005b)

--- Map title ---

31. Sex-based and Non-sex-based Gender Systems

--- Values depicted ---

1. No gender system
2. Sex-based gender system
3. Non-sex-based gender system

--- Definitions as given by the author ---

The defining characteristic of gender is agreement: a language has a gender system only if we find different agreements ultimately dependent on nouns of different types. In other words, there must be evidence for gender outside the nouns themselves. […] languages in which free pronouns present the only evidence for gender will be counted as having a gender system. […] there is no substantive difference between what are called "genders" and what are called "noun classes"; the different terms may be merely the products of different linguistic traditions. […] In many languages, nouns may be divided into groups according to the agreements they take, even when we control for other factors such as number and case. We should then ask whether these groups are arbitrary. The answer is that there is always a semantic "core" to the system. That is, there is an overlap between the nouns which take a particular set of agreements and some semantic feature. [...] Linguistic gender systems are frequently linked to biological sex. This is not the only possibility; alternatives occur, particularly in some of the larger gender systems. [...] in many languages, nouns may be divided into groups according to the agreements they take, even when we control for other factors such as number and case. We should then ask whether these groups are arbitrary. The answer is that there is always a semantic "core" to the system. That is, there is an overlap between the nouns which take a particular set of agreements and some semantic feature.

As immediately becomes clear from the complex prose of the definitions in (1), they are unsuitable for computerized processing. Ideally, the definitions should be reformulated in a semi-formal meta-language based on the term list. It is important to realize that any formalization of this description is not intended to replace the prose definitions as given by the typologist. The formalization will always be a simplification of the real complexities of typological research. However, such formalizations are needed to be able to automatically process large typological datasets like WALS. To illustrate what such a meta-language might look like, we tried to reformulate the definitions given above for gender systems. The terms are given in capitals; relators are put between angled brackets. The details of this meta-language are not important—this example is purely illustrative.

(2) Reformulated definitions for Corbett (2005b)

--- Values ---

1.<absence of> GENDER_SYSTEM

2.<presence of> GENDER_SYSTEM

<and>

<presence of> GENDER_CLASS <related to> SEX

3.<presence of> GENDER_SYSTEM

<and>

<absence of> GENDER_CLASS <related to> SEX

--- Term explications ---

GENDER_SYSTEM <defined by> <presence of>

AGREEMENT

<between> FULL_NOUN

<and> NON_NOUN

<controlled by> FULL_NOUN

<or>

AGREEMENT

<between> FREE_PRONOUN

<and> NON_NOUN

<controlled by> FREE_PRONOUN

GENDER_CLASS <defined by> <set of> NOUN <with> <coherent> AGREEMENT

SEX <defined by> <presence of> <opposition between> MASCULINE <and> FEMININE

NOUN_CLASS_SYSTEM <is the same as> GENDER_SYSTEM

NOUN_CLASS <is the same as> GENDER_CLASS

2.3 Typology as system-based linguistics

An immediately obvious problem for an ontology is that typological parameters constantly use concepts like "presence of" versus "absence of" a particular characteristic, not only in the form of relators, but also in the form of concepts like "non-noun", referring to anything that is not a noun. This aspect of typological terminology will be discussed extensively in section 3.2.1. Other difficult aspects are relators like "coherent" and "opposition". Such relators do not invoke any particular piece of data, but rather grammatical *systems* within a language. From the perspective of, for example, the case *system* of a language, one can say that for a particular language this system will be empty, so it has no cases. Or one can say that the cases form some *coherent* system for, say, argument marking, or that there is a particular *opposition* between two cases in particular.

From a typologist's perspective, it is very important for the process of building ontologies to distinguish linguistic systems from the individual forms in such a system. For example, saying that a language has an ergative system is something different from saying that it has an ergative case. If a language has an ergative case, then this language will have an ergative system somewhere, but not necessarily throughout the language (the ergative case might, for example, be restricted to pronouns). Conversely, a language might have an ergative system based on syntactic phenomena without having any nominal case forms at all.

The forms versus systems issue clearly extends well beyond WALS, and we have no concrete recommendations for its resolution here. However, we do think it is one that needs to be clearly addressed by the GOLD Community—even if it is simply addressed by an explicit statement that systems have no direct place within GOLD and that encoding them, therefore, requires the creation of a separate community.

3. Typological categories and ontological relationships

3.0 Introduction

In this section, we discuss some of the general challenges raised by WALS as we have tried to determine how specific WALS concepts should be linked to concepts in the GOLD ontology. These can be placed into three broad classes: (i) non-encoded internal structure of features, (ii) non-canonical concepts, and (iii) lateral relationships holding among concepts across different WALS component databases. We take up each class in turn.

3.1 Non-encoded internal structure

The data available in WALS is an enormously valuable source of information for linguistic research. However, in its current form it cannot be used for certain computational and statistical approaches to language typology. The problem is that the database does not encode logical dependencies between concepts referred to by terms found in the databases. Such implicit dependencies can be found both within the values for a single typological feature database in WALS and among values found in different component databases. To illustrate this problem, it is first useful to consider a case where the values for a given typological feature have no logical dependencies. For the typological feature *voicing in plosives and fricatives* (Maddieson 2005a), with accompanying map in Figure 5, four different possibilities are distinguished: (i) no voicing contrast in plosives and fricatives, (ii) voicing contrast in plosives alone, (iii) voicing contrast in fricatives alone, (iv) voicing contrast in both plosives and fricatives.

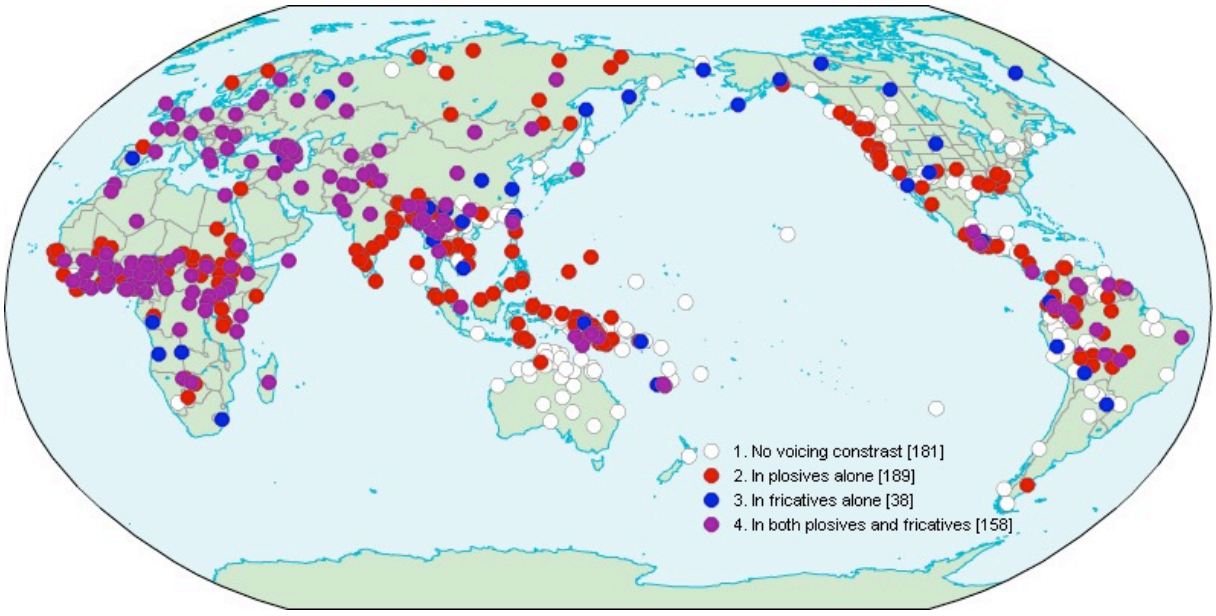

Figure 5 *Voicing in Plosive and Fricative Systems (Maddieson 2005a)*

These four possibilities clearly represent the intersection of two independent dimensions: voicing in plosives and voicing in fricatives. There is no *a priori* reason why these two characteristics should show any dependency on each other—that is, there is nothing about the definitions of *plosive, fricative*, and *voicing*, which would imply that there should be any correlation between *plosive voicing* and *fricative voicing* in the world's languages. So, the fact that there is an apparent correlation between the two parameters in the data (Fisher's Exact p = .000037 when counting genera) is an empirical observation of potential interest. However, Dryer's test (Dryer 1992) shows only marginal significance in three out of six geographic "macro-areas" (Africa, Australia/New Guinea, and South America), indicating that the overall significance is not a world-wide effect, but only regionally important. These sorts of correlations are potentially interesting—but they are only linguistically meaningful if we know that the relevant parameters are logically independent from each other.

The data represented in Figure 5 can be usefully contrasted with the data represented in Figure 6 which covers *voicing and gaps in plosive systems*. The values for this feature are: (i) missing /p/, (ii) missing /g/, (iii) missing both, (iv) none missing in /p t k b d g/, and (v) other. Unlike the data represented in Figure 5, the values for the data represented in Figure 6 show a high degree of logical interdependence. For example, a language missing both /p/ and /g/ is also a language missing /p/, but the database does not encode this. Similarly, a language missing no sounds in /p t k b d g/ cannot be a language missing /p/, missing /g/, or having both missing. From the perspective of a human user, these logical dependencies are obvious. However, a computational algorithm designed to discover correlations among values in the various databases will find spurious patterns without an explicit machine-readable encoding of such dependencies.

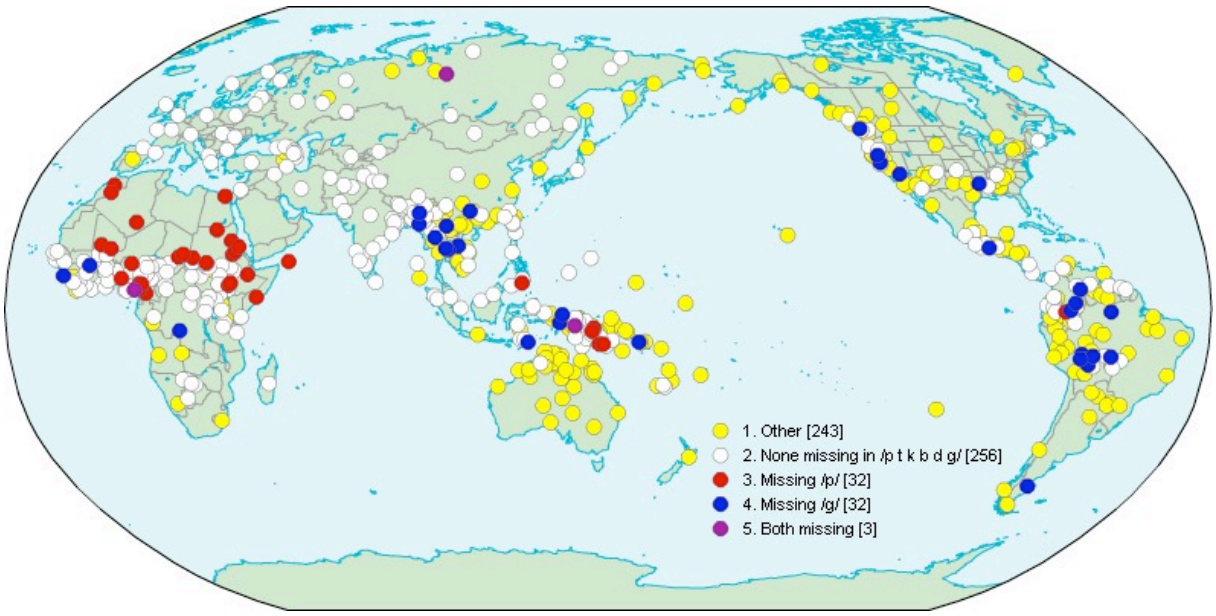

Figure 6 *Voicing and Gaps in Plosive Systems (Maddieson 2005b)*

The logical dependencies holding among the values for the feature *voicing and gaps in plosive systems* are schematized in the tree in Figure 7. Figure 7, of course, includes possible typological feature values not found in the data represented in Figure 6, and it also includes a

number of higher-level categories. The sort of information represented in Figure 7 can be easily expressed using an ontology. An important part of the WALS ontology project is to enumerate the logical dependencies holding among the concepts found in WALS and build appropriate ontological resources for encoding them.

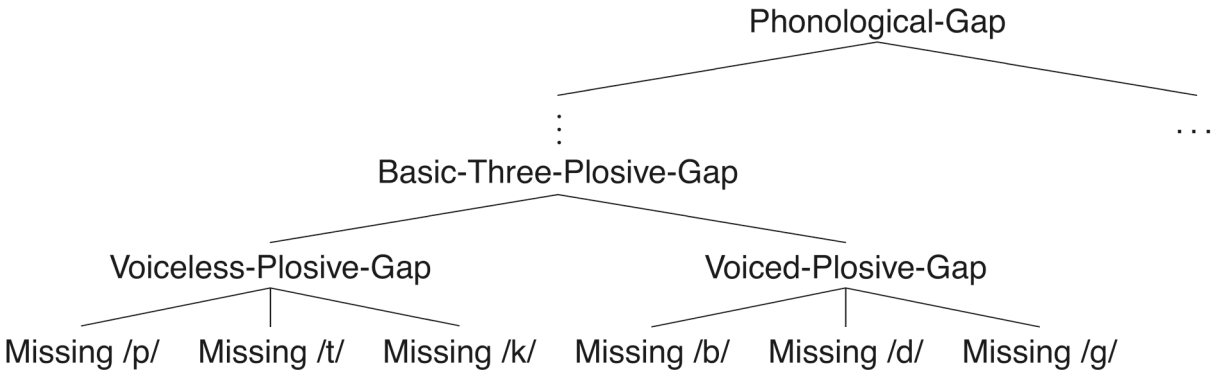

Figure 7 *Logical Structure of Voicing and Gaps in Plosive Systems*

Of the problems the WALS ontology project has encountered with respect to linking WALS concepts to a general ontology, implicit logical dependencies have required the greatest deal of human labor. However, from an ontological perspective, they are relatively easy to deal with.

3.2 Non-canonical concepts

3.2.0 Introduction

As a resource designed for use in language typology instead of use in individual language description, WALS makes use of many concepts which are quite distinct from the concepts found in a typical grammar or annotated text. Three such classes concepts seem worthy of mention here: (i) *absence* concepts, (ii) *numerical* concepts, and (iii) *fuzzy* concepts. We label these concepts as *non-canonical*. They contrast with *canonical* concepts by not being straightforwardly expressible using *instance of* relationships with respect to concepts in an ontology. We discuss each of these non-canonical concepts in turn.

3.2.1 Absence concepts

Absence concepts are found throughout WALS. They refer to a concept explicitly defined as *not* being an instance of another kind of concept. In Figure 3 and Figure 4, we already saw one such absence concept *no case marking*. In ontological terms, the concept of *no case marking* means that, in some language, there is no grammatical structure which can be claimed as instantiating *case marking*. Crucially, an absence category is quite different from inferring the absence of some grammatical phenomenon in a language simply because it is unattested or because there is no discussion of it in a grammatical description. The former is an explicit statement about the properties of a language's grammar, and can, therefore, be taken directly as linguistic data, while the latter cannot.

Some sense of the variety of possible absence concepts can be achieved through a simple enumeration of some of the ones that are found in WALS. They include: no action nominals, no adpositions, no antipassive, no bilabials, no case, no distributive numerals, no fricatives, no gender distinctions, no glottalized consonants, no grammatical evidentials, no

independent subject pronouns, no irregular negatives, no laterals, no nasals, no nominal plural, no obligatorily possessed nouns, no perfect, no person marking, no plural, no possessive affixes, no productive reduplication, no question particle, no suppletion in tense or aspect, no tense-aspect inflection, no tones, no uvulars, and no velar nasal.

Looking through the definitional statements as given by the authors, some more absence concepts can be found. They include: non-agreeing, non-benefactive, non-bound, non-declarative, non-derived, non-finite, non-head, non-human, non-iconic, non-inflecting, non-inflectional, non-number, non-obligatory, non-paradigmatic, non-periphrastic, non-possessible, non-pronominal, non-realized, non-reduction, non-referential, non-reflexive, non-relativizable, non-sex-based, non-sibilant, non-singular, non-subject, non-syntactic, and non-verbal.

Clearly, absence concepts are important for typological description.

The existence of absence concepts within WALS leads to a simple recommendation with respect to the relationship between a general ontology like GOLD and a community-specific ontology like the WALS ontology. In addition to allowing concepts in the community ontology to be relatable to the general ontology via positive relationships like *language shows instances of*, it is also necessary to allow them to be relatable via negative relationships like *this language does not show instances of*. While this would not seem to put a particular burden on the development of a general or a community-specific ontology, it would seem to put a burden on software designers building ontologically-intelligent search tools to ensure that their tools can deal with absence categories in a way which is useful to linguists.

It seems worthwhile to point out here that the general problem of encoding absence concepts may be more complex than is reflected in the WALS data. In WALS, all absence concepts are assertions about a property of a language's grammar. However, there is at least one other important kind of absence: *no information on*. Here, again, we need to contrast inference and explicit statements. If the only descriptive linguist who has worked on a particular language states that, quite simply, there is no data which would allow a language to be classified one way or another typologically, that is information of quite different value from discovering that a particular resource happens to have no information on a given topic. The former would not seem to call for looking for other resources to see if they contain the relevant information, while the latter would.

While we have encountered numerous absence concepts within WALS, the WALS ontology project has not attempted to exhaustively enumerate all possible kinds of absence concepts which might be useful as linguistic annotation. This seems like a worthwhile area for future research.

3.2.2 Numerical concepts

Another frequently occurring non-canonical concept found in WALS is the usage of features with "countable" values. For example, the feature *number of genders* as presented in Figure 1 above distinguishes languages with no gender from language with two, three, four, or five or more genders.

Such numerical concepts are found rather frequently among the WALS features. Illustrating this approach are, for example, features covering the number of distance contrasts in demonstratives (Diessel 2005), the number of cases (Iggesen 2005), the number of classes of possessive classification (Nichols and Bickel 2005), and the number of degrees of remoteness as distinguished in the past tense (Dahl and Velupillai 2005b). Such overt

examples nicely illustrate the importance of counting in typological parameters. However, there are also more covert examples of counts being used in the definitional details of typological parameters. For example, the value *no case* in the feature on *case syncretism* (as shown in Figure 3) is at first sight rather particular. Specifically, in this feature a language is treated as having the value *no case* if it its nominal paradigm has *no more than two* distinct forms. This somewhat idiosyncratic definition is justified by reference to the definition of *syncretism* in the other values.[7] To be able to talk about syncretisms of cases, there should be at least two overtly marked cases. All languages that do not fall under this criterion are irrelevant for the discussion of syncretisms, and are, thus, designated as having *no case*, as a convenient shorthand.

With respect to linking WALS concepts to the GOLD ontology, the existence of numerical concepts would seem to necessitate concept relators which can directly refer to cardinal numbers.

The usage of numbers for countable phenomena has to be distinguished from numbers used to divide more or less continuous parameters into discrete values. For example, in the feature depicting vowel/consonant ratios (Maddieson 2005c), the value *low* is defined as having a ratio of *two or less*. The fact that the cut-off point is a whole number is clearly just an arbitrary decision, as the ratio results in a quasi-continuous parameter (see Cysouw, forthcoming, for a discussion of such quasi-continuous parameters in typology). The division of such continuous parameters into discrete, numerically-defined classes is related to the notion of *fuzzy* concepts, to which we turn next.

3.2.3 Fuzzy concepts

The third class of non-canonical concepts in WALS are what we call *fuzzy* concepts. These are concepts which cannot be straightforwardly relatable to other relevant concepts via a logical relation. This is not to say they are unrelatable to other concepts—rather, some consistent policy needs to be developed for determining how to annotate such relationships which makes the "fuzziness" ontologically tractable. Some examples of fuzzy concepts found in WALS are: *small consonant inventory* (Maddieson 2005d), *complex syllable structure* (Maddieson 2005e), *borderline case marking* (Iggesen 2005), *weakly suffixing* (Dryer 2005), and *highly differentiated genitives, adjectives, and relative clauses* (Gil 2005).

The hallmark of a fuzzy concept is the use of a modifier like *small* or *-like* which is open to a subjective or relative interpretation. For example, a small consonant inventory can only be considered small in reference to all the known consonant inventories—and, even then, there is still a subjective element to determining the boundary between, say, small and moderate. As another example, consider descriptions using the modifier *-like*. Going through the definitional statements as presented in WALS, we found the following terms being used: adjective-like, agent-like, case-like, patient-like, vowel-like, we-like. These can only be understood as referring to an unidentified deviation from the more prototypical meaning of the head-term.

Fuzzy concepts, then, can be distinguished from simply idiosyncratic concepts which combine categories in unexpected ways but which, in principle, are logically definable without an explicit statement of interpretation. One such idiosyncratic concept is the value

[7] Here, it should be noted that the feature labels used in WALS necessarily had to be short enough to fit comfortably in the published maps. This, in all likelihood, largely explains why the label *no case* was used in the map seen in Figure 3 in a seemingly counterintuitive way.

pronouns avoided for politeness in the database for the feature *politeness distinctions in pronouns* (Helmbrecht 2005). This concept incorporates the notions of *pronoun*, *avoidance*, and *politeness* into a single concept in a way which would be unlikely to be specifically anticipated by developers of a general ontology. Nevertheless, assuming that an ontology contained these basic concepts, there is nothing "fuzzy" about them and it should, therefore, be possible to relate a concept combining them to a general ontology using standard logical relations.

There would seem to be two broad strategies available for linking fuzzy concepts to a general ontology. The first is to always associate the entire concept to a reasonable non-fuzzy definition and treat relative and subjective terms like *small* or *borderline* as useful abbreviatory conventions with no real ontological status. Thus, for example, a *small consonant inventory* could be defined as meaning "less than fourteen consonants" (which is, in fact, the definition given by Maddieson 2005c). A second strategy would be to relate the fuzzy modifiers themselves to a general, concrete definition. So, perhaps, *small* would be defined as "two or more standard deviations away from the average for a countable quantity".

In looking at the prose descriptions accompanying the WALS maps, what we find is that, in general, authors did, in fact, associate apparently fuzzy concepts with a non-fuzzy definition explaining, for example, how they specifically interpreted terms like *small* or *borderline* with respect to particular categories in particular languages. Thus, common practice in the creation of WALS was to take the first of the two strategies outlined above. The WALS ontology project is following this common practice and adopting it as its general strategy for dealing with fuzzy concepts.

To make the discussion more concrete, in Figure 8 we give the map representing the data collected for the feature *syllable structure*. This feature has three values in WALS: *simple*, *moderately complex*, and *complex*. These values all refer to fuzzy concepts. However, within a prose description accompanying the map in the published version of WALS, Maddieson (2005e) associates each fuzzy concept with the more concrete definitions given in (3). These definitions do not explicitly appear in Maddieson (2005e) but are adapted from his prose descriptions of the relevant categories.

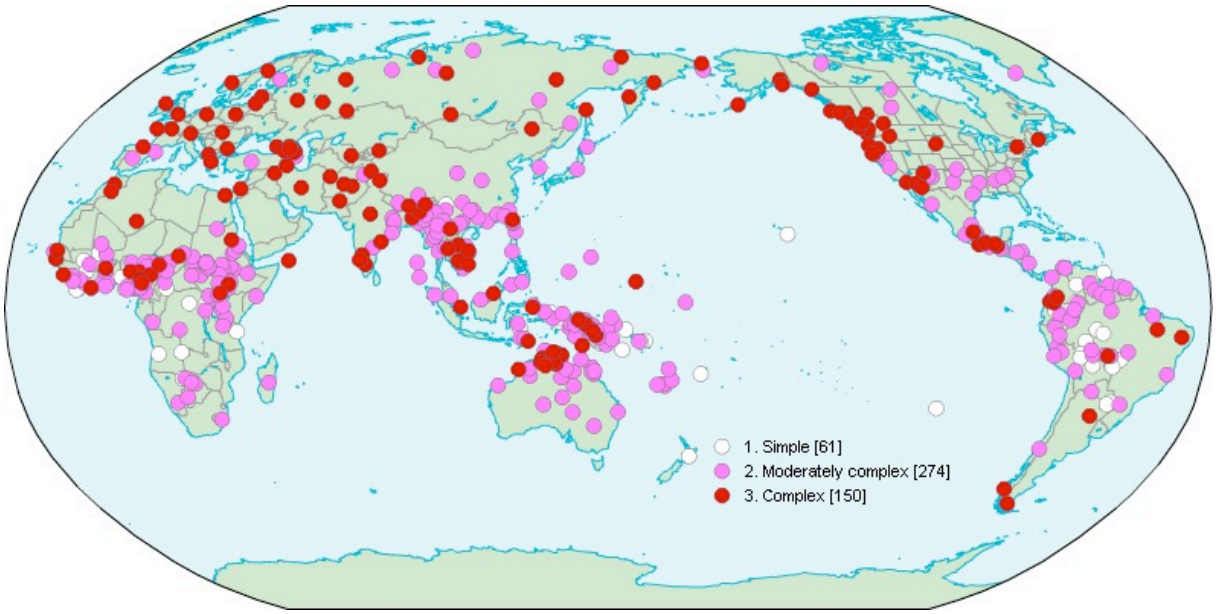

Figure 8 *Syllable Structure (Maddieson 2005e)*

(3) Concrete definitions of fuzzy categories from Maddieson (2005e)

a. Simple Syllable Structure: Describes a language which only allows syllable structures conforming to a (C)V pattern
b. Moderately Complex Syllable Structure: Describes a language which only allows syllable structures conforming to (C)V(C) or CWV(C) patterns (where W stands for a liquid or glide)
c. Complex Syllable Structure: Describes a language which allows syllable structures other than those described as permitted in simple syllable structure or moderately-complex syllable structure languages

The strategy of associating each fuzzy concept with a non-fuzzy definition, instead of devising a general definition for each attested type of fuzzy modifier, is a fairly comfortable one for the WALS ontology project since each of the typological feature databases was essentially conceived as its own internally coherent research project with terms defined for that project alone. We leave open the question as to whether or not some other project might find it worthwhile to deal with fuzzy concepts by giving concrete definitions to the fuzzy modifiers themselves. If this were considered necessary, it would seem to necessitate the inclusion of notions like *average* or *standard deviation* within a general ontology for linguistic description (perhaps linked to an upper ontology also containing such concepts).

3.3 Lateral relationships among concepts

By the term *lateral relationship*, we mean a relationship holding among two concepts in different resources. Of course, lateral relationships abound among categories in linguistic resources since it is what makes the data they contain comparable in the first place. Thus, for example, when one encounters the term *nominative case* in one resource and *ergative case* in another, one assumes a lateral relationship holding between those concepts where they both

have something in common (being instances of *case*) and something not in common (being instances of different kinds of *case*). In principle, lateral relationships can all be encoded by linking linguistic concepts to their appropriate place in an ontology—in fact, this is one of the tasks ontologies were designed for.

However, in the WALS ontology project, we have often found it useful to make note of lateral relationships among concepts. The reason for this is a simple one: Sometimes it is possible to determine a lateral relationship holding between two categories before it is possible to relate each of those categories to a higher-level ontology. Thus, encoding lateral relationships allows us to indicate some of what we know about a category at a given time even if we do not know enough about the category to link it to an ontology. For example, returning to the case exponence problem introduced in Section 1 with reference to Figure 3 and Figure 4, recall that two different typological feature databases in WALS made use of values which, superficially, appeared to both refer to the concept *no case*. However, one of the databases characterized English as making use of case marking, while the other did not. Upon making such an observation, one knows something important: the two uses of the concept of *no case* are empirically distinct. However, without more detailed study it is hard to know anything more. They could refer to exactly the same concept, and there is simply disagreement on how to classify English. Or, they could refer to different concepts and be typological false friends.

Under a typical work scenario for the WALS ontology project, it is not always advisable for an individual who has found a terminological clash like that exemplified by the case exponence problem to determine how it is best resolved for one of two reasons. First, that individual may not be qualified to devise an ontologically appropriate solution for it. Second, it may be the case that the best solution requires further research into the ontological nature of other related concepts and, thus, cannot be devised immediately.

In such cases like this one, where a researcher may discover an ontological "problem" in WALS but might not be able to devise an immediate ontological solution, encoding lateral links among concepts has proven quite useful. In Figure 9, we schematize our model for workflow involving lateral links. One linguist examines two of the WALS databases and notes a particular lateral link, which another linguist can then examine at a later time in order to determine how the two laterally-linked concepts can be related to the WALS ontology.

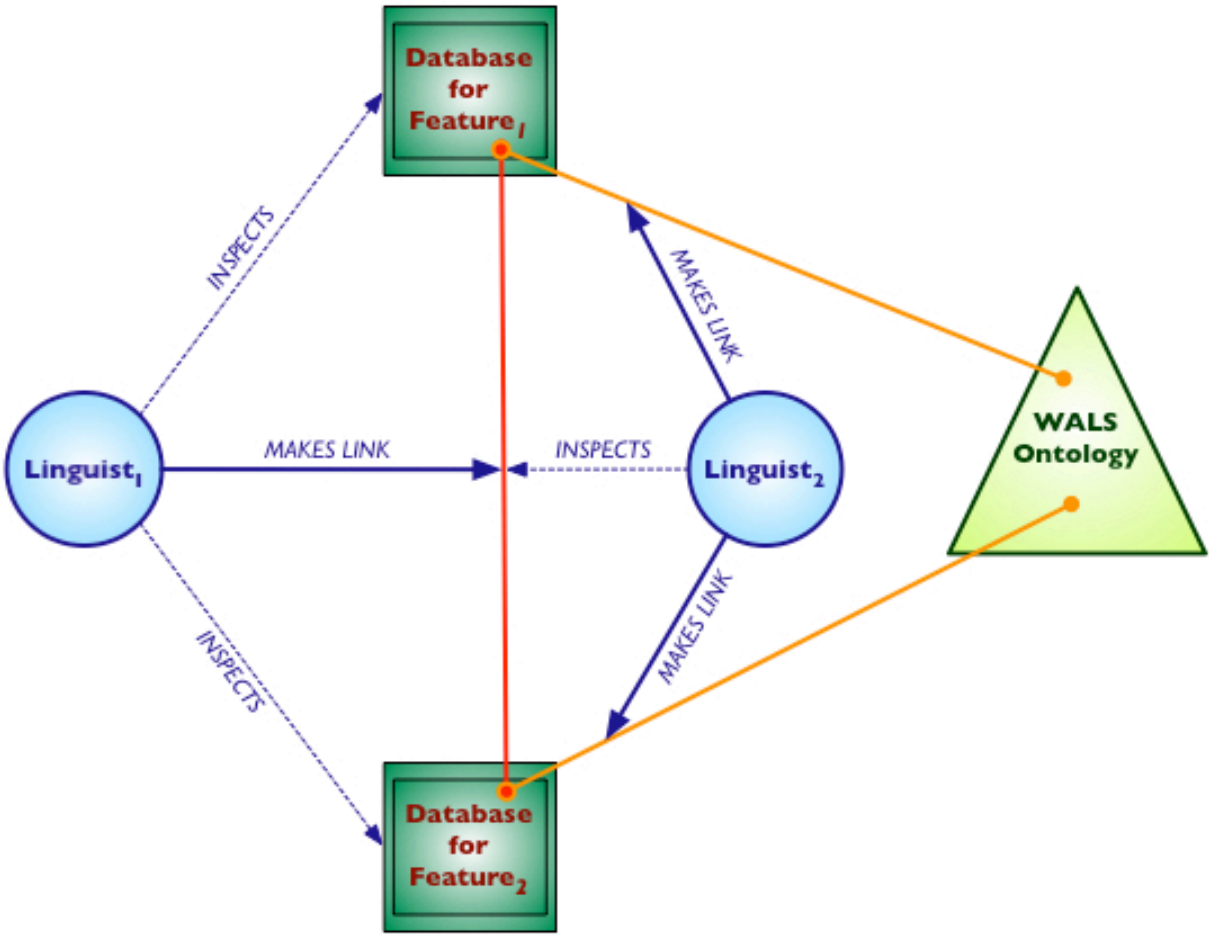

Figure 9 *Workflow making use of lateral links*

Some of the classes of lateral links we have found useful in annotation are given in (4). In principle, a controlled vocabulary for such links could be developed. However, at this time, since there is no tool designed to exploit lateral links in developing an ontology, they are always inspected by hand, and we have found no need to develop a machine-readable annotation system for them.

(4) Some classes of lateral links

a. Similar term names, but different concept
b. Theoretically (almost) same concept, but certain languages classified differently
c. Same concept, but no appropriate ontology concept(s) found

Given our own experiences on the WALS ontology project, we believe that it may be useful for any ontology tool intended to allow ontological annotations over multiple resources using different term sets to be designed to facilitate the workflow schematized in Figure 9. This would mean (i) allowing for the creation of lateral links and (ii) providing functionality to search for and examine lateral links whose associated concepts have not been linked to the ontology. Such functionality would, in all likelihood, require the creation of a controlled vocabulary for expressing lateral links, and the WALS ontology project is doing research in

this area using the informal system of lateral links that has been developed as a useful starting point.

Ultimately, as ontologies become more fully developed and as more resources are "born ontological", we expect the need for lateral links to be greatly reduced. For now, however, they have proven to be a useful strategy for migrating a legacy resource (WALS) to best practice standards. Since, at the present time, almost all linguistic materials are legacy materials with respect to ontological mark-up, lateral links potentially have the role of serving as a part of a general migration strategy.

4. Summary and recommendations

In this paper, we have discussed a number of issues that have been encountered during the development of the WALS ontology. In addition, we discussed some general strategies for dealing with those problems, trying to devise general recommendations based on our experiences with this one project. Broadly speaking our recommendations for the GOLD Community are that it should adopt: (i) recommendations for linking non-grounded concepts, of which typological categories are an important class, to linguistic ontologies, (ii) recommendations for how grammatical systems, as opposed to grammatical forms, relate to the GOLD ontology, (iii) a system of relators to linguistic ontology concepts which go beyond *instance of*, minimally included *not an instance of* and *no information on*, and (iv) recommendations for linking counting and fuzzy concepts to linguistic ontologies. In addition, we recommend that developers of tools for linking linguistics resources to ontologies consider implementing functionality for lateral links in addition to hierarchical ones.

References

Bickel, Balthasar and Johanna Nichols. 2005. Exponence of selected grammatical formatives. In Haspelmath et al., 2005.

Baerman, Matthew and Dunstan Brown. 2005. Case syncretism. In Haspelmath et al., 2005.

Corbett, Greville G. 2005a. Numbers of genders. In Haspelmath et al., 2005.

—. 2005b. Sex based and non-sex-based gender systems. In Haspelmath et al., 2005.

Cysouw, Michael. Forthcoming. Quantitative method in typology. In: Gabriel Altmann, Reinhard Köhler and R. Piotrowski (eds.). *Quantitative Linguistics: An International Handbook*. Berlin: Mouton de Gruyter.

Dahl, Östen and Viveka Velupillai. 2005a. Perfective/imperfective aspect. In Haspelmath et al. 2005.

—. 2005b. The past tense. In Haspelmath et al. 2005.

Diessel, Holger. 2005. Distance contrasts in demonstratives. In Haspelmath et al., 2005.

Dimitriadis, Alexis and Paola Monachesi. 2002. Integrating different data types in a Typological Database System. Proceedings of the International Workshop on Resources and Tools in Field Linguistics, Las Palmas, Spain.

Dryer, Matthew S. 1992. The Greenbergian word order correlations. *Language* 68: 80–138.

—. 2005. Prefixing versus suffixing in inflectional morphology. In Haspelmath et al. 2005.

Farrar, Scott and Terry Langendoen. 2003. A Linguistic Ontology for the Semantic Web. *GLOT International*. 7:97–100.

Haspelmath, Martin, Matthew Dryer, David Gil, and Bernard Comrie (eds.). 2005. *World Atlas of Language Structures*. Oxford: Oxford University Press.

Helmbrecht, Johannes. 2005. Politeness distinctions in pronouns. In Haspelmath et al. 2005.
Iggesen, Oliver. 2005. Number of cases. In Haspelmath et al. 2005.
Maddieson, Ian. 2005a. Voicing in plosives and fricatives. In Haspelmath et al. 2005.
—. 2005b. Voicing and gaps in plosive systems. In Haspelmath et al. 2005.
—. 2005c. Consonant-vowel ratio. In Haspelmath et al., 2005.
—. 2005d. Consonant inventories. In Haspelmath et al. 2005.
—. 2005e. Syllable structure. In Haspelmath et al. 2005.
Nichols, Johanna and Balthasar Bickel. 2005. Possessive classification. In Haspelmath et al. 2005.