Towards portability and interoperability for linguistic annotation and language-specific ontologies

Robert Munro and David Nathan

Endangered Languages Archive

School of Oriental and African Studies

Abstract

The central task in constructing a general linguistic ontology is mapping the phenomena of a given language to the concepts in the ontology, and extending the ontology where no viable mapping exists. Mapping well-known independent features is fairly straightforward, but mapping a new feature or language-specific ontology to a general model is a complicated problem. This is especially true when part of the meaning of a language-specific ontology derives from sociocultural phenomena – it may not be meaningful to store the data independent of its presentational context, and so the context itself must be recorded in a structured format, or with richly annotated instances of photos, diagrams, maps and/or videos.

In this study, we will survey a variety of linguistic documentation materials to report on the range of ontological categories and structures that may be required to integrate such phenomena into the GOLD framework. In particular we focus on the ability of GOLD to meet the needs of the contributors and users of endangered languages archives. We find that in order to support endangered languages GOLD needs to explicitly support uncertainty, variability and phenomena that are inherently indeterminate or complex. We propose solutions for a number of these.

Introduction

The main goal of this paper is to explore how the General Ontology for Linguistic Description (GOLD) meets the requirements of portability for language documentation and description (Bird & Simons, 2003). Our approach to road-testing here is in evaluating the ability of GOLD to meet the needs of archive users and contributors. Where we make suggestions for extensions to GOLD we will also attempt to justify the changes in terms of general ontology building.

The Endangered Languages Archive (ELAR) of the Hans Rausing Endangered Languages Project (HRELP) is uniquely placed in the language documentation community. As well as the archive, HRELP supports grants for documentation projects and an academic program focusing on language documentation. While one of the long term goals of the archive is to preserve data and provide access to it, we also train students and grantees in markup and data management strategies, multimedia development, and the choice of recording equipment.

There are many ontologies for linguistic description, but GOLD is being developed to provide an ontology specific to the issues of documenting endangered languages (Farrar & Langendoen, 2003). It is not surprising to find that it is therefore currently the most suitable ontology for supporting data portability. To date, GOLD's main goal has been in the area of providing 'datanalysis sets' (Farrar & Langendoen, 2003). We demonstrate how GOLD could further enhance its level of portability if it extended the focus on data analysis to two other areas:

- data acquisition
- data access

Data acquisition is GOLD's ability to support language documenters performing annotation, including the use of less well-known terms, modelling the uncertainty in annotating new and difficult phenomena or phenomena that are inherently indeterminate or complex. Data access is GOLD's ability to support structures that allow the navigation, rendering and interpretation of data by a variety of users, especially the speakers of endangered languages.

For all three areas interoperability is important, and a common ontology and markup system will support it. A common ontology would fill the space been the cataloguing metadata of standards such as IMDI and OLAC, and the rich metadata of linguistic annotation (Nathan & Austin, 2004).

For improving data acquisition we suggest that each concept in GOLD should be represented by a set of formal properties describing that concept (in addition to the existing definitions). For improving data access, we suggest that GOLD explicitly captures the conventions and constraints for presentation (rendering). We show that a combination of these two strategies would allow modelling of and access to concepts that are important to linguistics but currently outside of ontology building for language description, including the modelling of language-specific ontologies.

Linguistic ontologies and markups

Here, we outline the changes we propose to GOLD's representations of concepts. Crucially we explain that GOLD does not need to make assumptions about consensus

among linguists or the boundaries of linguistics for documenting endangered languages. We show that modelling variances and indeterminacies will, in fact, reduce the ambiguity in interpreting and applying the resultant ontology.

The relationship between ontologies and markup (annotation) systems is, at best, messy. At its strictest definition an ontology is about what we agree can exist. A strict linguistic ontology will need to capture phenomena that are inherently variable and continuous, but not uncertainty. For such a definition, the extent to which the ontology correctly captures domain knowledge can be evaluated simply by its utility. On the other hand, a *markup* represents knowledge, and is therefore inherently uncertain – it is indeterminate to the extent of the authors' confidence in their markup, conflicting interpretations, and the limits of conceptual frameworks. Therefore, an ontology and a markup converge only when there is consensus and complete confidence. However, when analysing languages, in particular little-studied endangered languages, there is rarely full confidence in the classification of new hard-to-classify phenomena. Academic debate and increasing understanding of the diversity of languages mean that we also lack consensus – concepts drift, divide and are born new. The last of these is well recognised as a possibility in language documentation, which is why open ontologies are used. For example, GOLD provides extensions through Community of Practice Extensions (COPEs). There are many epistemological reasons for an ontology to take account of indeterminacy, and while these can make for interesting discussion, our justifications for modelling indeterminacy are based only on utility and portability.

As an ontology linked to annotation systems, GOLD needs to represent several kinds of indeterminacies: uncertainty of annotation, and phenomena that are inherently continuous and variable. It is not enough that a language ontology conforms to an endurantist or perdurantist model of reality (Niles & Pease, 2001): for language documentation, an ontology also needs to capture uncertainty in making an annotation, variance in the definitions of categories and the limitations of categorical frameworks. In general, builders of ontologies outside of linguistics have been reluctant to accept variance and indeterminacy as integral to their models:

> In some cases, the incompatibilities [between ontologies] can be smoothed over by tweaking definitions of concepts or formalizations of axioms; in other cases, wholesale theoretical revision may be required. (Niles & Pease, 2001)

So, what about explicitly modelling the 'incompatibilities'? After all, if they can be identified they can be modelled. While the intentions for building deterministic models are sound, they can lead to poor practices. Inconsistencies should not be ignored, nor should the significant revision of an ontology be undertaken lightly. A general linguistic ontology therefore needs to allow the linking of concepts in ways that allow differences to be formally represented, and for these differences to become part of the ontology itself. If some phenomenon in a language does not easily fit into an existing ontology, it is likely to be something linguistically interesting and worthy of attention.

Bateman (1992) distinguishes between *formal-philosophical, logical-semantic* and *linguistic* ontologies (Bateman 1992), arguing that most ontologies are a combination. GOLD seems heavily influenced by formal-philosophical ontologies:

> GOLD makes a clear distinction between grammar and semantics and, furthermore, clarifies the connection between linguistic and world knowledge …

> The first requirement, then, for a general ontology for linguistics is a theory-neutral model of the world (Farrar 2003b)

Arguments for theory-neutral ontologies cannot apply to broad linguistic ontologies. An ontology for language description cannot make a clear distinction between grammar and semantics, or between linguistic and world knowledge, before theories of linguistic description do. One of the goals of language description is to better understand the relationships between syntax and semantics and between linguistic and world knowledge. An ontology needs to support this exploration, not presuppose its answers. We hold the more common view that there is no theory-neutral model of world. In fact, GOLD's current goal of a theory-neutral ontology has led it to make some very theory dependent theories:

> "Assumption 1: humans perceive, act in, and conceptualize their environment in the same way regardless of which language they happen to speak … In terms of ontology, strong linguistic relativism would imply that there could be no common upper model to mediate between various languages, rendering machine translation impossible." (Farrar 2003b)

Rejecting the linguistic contributions of Sapir and Whorf is not theory-neutral and limiting weak linguistic relativism is at odds with many current linguistic theories of inter-speaker/genre/register variance (Foley, 2003) and natural language processing. Machine translation has largely abandoned language independent models in favour of pair-wise language-to-language instantiations. While this is rarely regarded as evidence for strong linguistic relativism, it could be considered evidence against the utility of language independent modelling in natural language processing. Modelling invariable language independent concepts cannot be justified on an empirical basis.

The first requirement for a general ontology is not a theory-neutral model of linguistics, but a *pan*-theory model. We are highlighting these flaws in GOLD because they are assumptions that do not need to be made.[1] Since it is not feasible to create a theory-independent ontology, the best solution is to create an ontology that allows variation. Currently, GOLD externalises this by separating its language independent concepts from language-specific and community-specific variations by COPEs. Therefore, while GOLD currently admits variance, it doesn't seek to explicitly model it.

We propose a simple solution that can capture indeterminacies. Each concept in GOLD should be represented by a set of properties that describe that concept. A property would have three possible values to mark whether a given legacy ontology or language holds the property for a given concept: 'Yes', 'No', or 'Undefined' (default). For the ontology to accurately represent variance, it only needs to include enough properties to distinguish terms; however, for portability, it should seek to describe as many properties as possible.

'Yes' can potentially be expanded to include whether the property is mandatory or optional for the concept. It could also capture dependencies between properties for a

---

[1] We have found the problems that we have identified to be common to many linguistic ontologies. It is an unfortunate inevitability of the conference's focus that we single out one of the better ontologies for criticism.

concept. For example, property X is mutually exclusive with property Y for concept C. Variation within 'Yes' should be separated from the 'Yes', 'No', 'Undefined' distinction as, when annotating, we may know that a certain property exists but not yet know whether it is mandatory, non-mandatory, epiphenomenal or have complicated dependency relationships.

Gruber (1993) describes ontology concept definitions as:

> definitions … associate the names of entities in the universe of discourse (e.g., classes, relations, functions, or other objects) with human-readable text describing what the names mean, and formal axioms that constrain the interpretation and well-formed use of these terms. (Gruber, 1993)

Our suggestions here are Gruber's 'formal axioms that constrain the interpretation'. Currently, GOLD has few formal axioms constraining the interpretation of categories. If each concept in GOLD should be represented by a collection of formal properties. This allows inter-language variance to explicitly represent differences between concepts. It is common for ontologies to represent concepts as a set of formal properties, allowing the more explicit linking of concepts (Cysouw et al, 2005) and greater data portability, as knowledge of the full set of concepts is contained within the ontology (Bird & Simons, 2003).

While an ontology needs to capture classifications and their definitions, it does not need to capture the justifications for them. Exactly *why* a given definition is proposed is external to the ontology and to be argued elsewhere.

What is in a noun?

The current definition of 'noun' within GOLD is as follows:

> Noun Definition: A noun is a broad classification of parts of speech which include substantives and nominals (Crystal 1997:371; Mish et al. 1990:1176). (http://emeld.org/gold-ns/description.html#Noun, last checked 23/05/2003)

Unfortunately, because Crystal 1997 and Mish et al.1990 are not included in GOLD, a researcher does not know if their definition of noun corresponds to that of others also using GOLD. Even if they did know, how could they ensure that what they call a 'noun' will correspond to the same concept in future annotations by other researchers? Reference to information external to the ontology should not be required. And because Crystal's definition of a noun does not refer in turn to Mish et al, GOLD's definition of a noun refers to two different definitions. Should we 'average' between them or choose their common ground? If we choose the middle ground, then GOLD's definition of 'noun' has actually not yet been written, but only refers to an interpretive blending of the two existing definitions.

By allowing different users of GOLD to choose their own definitions of the categories, a single ontology can express multiple concurrent interpretations and the differences between them. This is analogous to defining multiple senses in lexicography, where different but related senses are described in parallel rather than being "averaged".

The argument that the core idea of a 'noun' and other such concepts will not change much is insufficient. The core of what is declared to be a noun might remain constant

for a long time, but the *boundaries* between a noun and other categories are less stable. Definitions of terms will change and the ontology must have the capacity to accommodate these changes.

With the definitions being formally contained within GOLD, there is less uncertainty as to whether GOLD does or does not capture a certain phenomenon. A researcher can easily check to see if GOLD's definition is the same as theirs. If they feel that theirs is narrower, they can still use GOLD and reduce the number of properties so that their definition remains formalised. If theirs is broader, they can add appropriate properties. In both cases GOLD becomes a richer resource.

Representing knowledge

In language documentation, there is currently an emphasis on separating form from content. This poses a paradox. If form is separated from content, then data structures, ontologies and annotation systems become more flexible; however, the more flexible a structure becomes, the less information it conveys. There is the potential for information to be lost or impoverished when resources are stored independent of display properties. Of course, for some materials, especially video, there are currently no standards for descriptive markup that come close to successfully abstracting the information from its form.

For portability, best practice currently suggests that resources are supplemented with all auxiliary software resources necessary for display (Bird & Simons, 2003). While this addresses the problems of preservation, it limits interoperability. Ideally, descriptive markup should include rendering information. For linguistic information this would include conventions such as italicising part of speech categories in dictionaries, and alignment in interlinear transcriptions. This might seem unnecessary, as these conventions will unlikely be forgotten and there is little heritage value in the form, but for language specific phenomena the presentational format may otherwise be lost.

The simplest solution is to record a thorough but informal description of the conventions for rendering. While this addresses data acquisition, it does not address data access as specific interfaces to the data will need to be created. Ideally, the conventions should be recorded in a structured format so that interface tools can customize the interaction with the data according to instructions within the data. Note that these form two extreme poles of interface design: one is a thick-interface that has hard-coded rendering, and the other takes its layout from the data itself.

The current convention for many archives is to simply provide the same interface for every language, thus ignoring any language specific conventions or constraints. It is not clear which users this supports: linguists will usually wish to search, not navigate, for cross-language phenomena. On the other hand, speakers of endangered languages are not advantaged by the fact that another language is accessible the same way. Ontologies and data structures are not navigational structures. Creating access to an archive by simulating 'transparent' access to structures is a convenient but fairly arbitrary interface design decision.

Supporting fieldworkers: data acquisition

The most valuable linguistic information

The most valuable information can be the data least likely to be recorded. All linguistic knowledge should preferably be formally captured, even if the linguist has not decided on the linguistic classification for a concept.

Most linguists will have working conventions for indicating their uncertainty about a given annotation or concept, ranging from writing a question mark to noting several alternatives (Holton, 2003). Hard-to-classify phenomena are the ones that field-workers will give the most thought to, but they may defer annotation until a decision is reached. A researcher may consider it preferable to leave a concept unannotated rather than expose themselves to the potential scrutiny of other researchers, where their data is to be deposited in an archive or made public. To guard against criticism, adding a question mark (or a level of confidence) may be sufficient, but it is not informative. On the other hand, by selecting properties they are sure about, a fieldworker can indicate that they are sure that certain properties do exist, certain ones do not, and leave the rest of the properties undecided. In this way, they are able to represent uncertainty (or more precisely, the extent of their certainty) before making choices of categoricity.

Supporting indeterminacy

There are two kinds of indeterminacy in linguistics:

- confidence in assigning a category (uncertainty)
- phenomena that are inherently variable, probabilistic, gradient or continuous.

Why we need markups of uncertain concepts:

1. To record our knowledge. For example, Bergqvist (2005) is currently investigating the possible categoricity of a word *7uhch* in Lakanon Maya. From recently collected field data, Bergqvist hypothesised that it functions as temporal-modal deictic and expresses participant frames and speaker's footings. Clearly, he has given *7uhch* more thought than other terms in the language; and so this is the term we would like him to annotate in the most detail, but until he reaches a decision it is actually the one most likely to remain unannotated.

2. To allow other people to find them. Perhaps a similar hard-to-classify concept has been encountered by another researcher. If an archive implementing an ontology with uncertain categories exists, then someone may have previously recorded the phenomena in a structured way able to be searched. If this problem is truly new, then at least this uncertain concept can be captured in a structured format that will allow future researchers to find it.

3. To reach certainty. A markup, even an indeterminate one, can allow the statistical/computational analysis of the distributional tendencies of a term. This can inform a decision about its appropriate category.

4. To highlight problems with descriptive frameworks. A feature may only appear to belong to multiple (or no) categories because the descriptive framework does not yet account for it.

5. Because the concept is inherently indeterminate. The concept may be inherently fuzzy but not previously encountered as an indeterminate feature of a language.

It is also easy to justify the need to capture phenomena that are inherently variable, fuzzy, gradient or continuous. As Himmelmann discusses, linguistic description is often criticised for ignoring such phenomena, and so ontologies cannot prohibit them (Himmelmann, 1998). The suggestion that continuous concepts can be adequately represented by discrete (symbolic) categories is false, as is well-known to anyone who has need to model both in some data modelling task.

In linguistics, continuous phenomena are described by a variety of terms: cline, gradience, squish, continuities, contiguities, vague, fuzzy, probabilistic. The distinctions between them are important but outside the scope of this paper. Here, we lump them together under 'inherently indeterminate', meaning that they are phenomena that cannot be adequately represented by discrete categories.

GOLD will need to be able to model continuous phenomena in order to capture prosodic information. In addition, if it is to keep with the current theories for morphosyntax then it will need to allow researchers to explicitly model morphosyntactic phenomena that are inherently indeterminate (Aarts, 2004; Bayen, 2003; Manning, 2003). Even for discrete concepts, researchers need to model continuities in order to model interspeaker variance.

If concepts are represented by a collection of properties, then the requirements of current formalisms for modelling linguistic gradience can be met (Aarts, 2004), but not all inherently indeterminate concepts can be represented this way. Perhaps the "ContinuousObject" concept of SUMO (Niles & Pease, 2001) can also be used to represent continuous phenomena in language. We leave the modelling of inherent indeterminism as an open problem that we have addressed only in part.

Incorporating new categories

New categories will emerge from language documentation and researchers need to be able to place them within the ontology.

A question for any ontology is: How do we know that a given category is not the same as another one identified elsewhere? In linguistics, it is especially complicated: what might be two distinct concepts in one language might be conflated in another. Formalising definitions as collections of properties provides an easy way of identifying where two concepts are really two names given to the same concept. Currently, we can identify duplicated concepts in two ways: compare the definitions given for both of them, or compare both the distributional tendencies. The first is the most thorough, but can be complicated if the two definitions are from different linguistic frameworks, which is likely in the case of conflicting terminology. The second, comparing the distributional tendencies, is a good solution, but can be problematic as the distributional tendencies may be subject to non-critical criteria that would skew calculations between languages. Genre, register, inter-speaker variation or intra-speaker variation may also skew calculations within a language. By formalising definitions we can simply compare the properties for the two – if two categories share the same properties, and therefore the same formal definition, we know they are the same category. On the other hand, if two categories share the same properties but we are sure that they are distinct, then we know that we need to extend the properties to express these differences.

For a fieldworker defining what they consider to be a new concept, software supporting GOLD could easily implement such a comparison of properties and suggest a number of existing concepts that might provide the solution.

In addition to reducing the likelihood of repeated concepts, the formal modelling of properties also reduces the need to make decisions about categoricity. If we are explicitly modelling the differences, then it does not matter whether or not we choose to call a new phenomenon a category in its own right or a variant of an existing one – we can show the dependencies between them in the shared properties and defer the discussions and decisions of the relationships to later analysis.

Importantly, the new categories found may simply not exist in any language (or are not yet discovered) and if we are able to create new categories and link them to other categories through shared properties then we can build models that work towards capturing more complex information than categorical labels.

Incorporating structures

Linguists capture much richer information than stand-alone labels for concepts. Beyond inherently discrete phenomena and inherently indeterminate ones, there is a third kind: concepts that are complex structures.

In computational linguistics there has been a tendency to ignore all linguistic phenomena other than those that can be represented as discrete categories. For example, Penton et al describe tables of discrete data as 'linguistic paradigms' (Penton et al, 2004). Such a 'linguistic paradigm' captures the *simplest* of linguistic phenomena: discrete categories. Linguistic phenomena can also be inherently indeterminate (as discussed), or inherently complex structures, including complicated language-specific ontologies.

Supporting speakers: data access

The largest (and growing) user group for endangered languages materials are the speakers of endangered languages.[2] These speakers are rarely interested in linguistic categories or navigating a corpus or archive via them.

Supporting language-specific ontologies means supporting information-rich structures for navigation, which are more appropriate to the task than databases or abstract ontologies. There is some conflict here between genericity and informativity:

> Markup may also cause problems for rendering. As we have seen, resources employ descriptive markup to maximize portability across computer systems and potential uses … However, such resources fail to cross the gap from computer to human if there is no meaningful way to display them. …Thus the best practice is one that supplements the information resource with all the auxiliary software resources that are needed to render it for display. (Bird & Simons, 2003)

---

[2] Strictly speaking, this should be extended to hearers (Grinevald, 2003) and the descendents of speakers looking to recover cultural or linguistic practices.

This trade-off between rendering and markup is likely to remain a problem for those working on data representation.

Case Study: Yolngu kinship

The Yolngu are a group of clans living in north-east Arnhem Land, Northern Territory, Australia. According to Yolngu world-view, language originates from the land. The Yolngu languages themselves have extensive kinship terminology called Gurrutu, also called by some linguists "kinship algebra" (Nathan 1996), with 27 terms that identify individuals and sets of individuals in terms of moiety, generation, gender, and patriline or matriline. The terms extend infinitely through cyclicity:

> These links work in a circular fashion: a woman's mother's mother's mother is her waku—the same as her own son or daughter. Her mother's mother's mother's mother is her yapa—her 'sister'. (Christie & Gaykamangu 2003).

A conventionalised method for representing Yolngu kinship is a graphic system showing lineage and generations together with patriline/matriline linking symbols that allow the Yolngu terms to be mapped onto configurations of individuals. Christie et al (2001) developed a full implementation for querying and rendering the kinship system via either the terms, a graphical interface, or a general wordlist.

But more interesting for the purposes of this paper is that the kinship system pervades the whole Yolngu world and acts as a classificatory and naming system. A person might regard and call a particular plant species or an area of land "mother":

> If for example, the Djambarrpuyngu men marry Gupapuyngu wives, there will be successive generations of Djambarrpuyngu children who call everything `Gupapuyngu 'mother'. For them, Gupapuyngu land, language, songs, totems, etc will all be called mother. The group into which the Gupapuyngu men marry (e.g. Garrawurra) will be the mother of the Gupapuyngu, and the märi (mother's mother) of the Djambarrpuyngu. (Christie & Gaykamangu 2003).

Finally, completing the symmetry and cyclic nature of the Yolngu world, individuals draw from the same sets of kinship relations to describe their relationship to the Yolngu lands (from whence the languages arose in the first place).

We thus see not only that simple translation of Yolngu kin terms is non-trivial, but also that the system, in actual discourse, enables the usage of kin terms to refer to anything in the Yolngu universe, and interpreting them as referring terms will depend on knowledge of the speaker's (and perhaps addressee's) place in the system, and the wider context of communication. Clearly, it is inadequate to suggest that one can annotate such linguistic concepts independently of language-specific ontologies (here, the kinship system) and so we need to work towards ways of allowing linguistic ontologies to include such complex concepts and formally relate them to more well-known phenomena.

Conclusion

Beyond supporting analysis, ontologies for endangered languages need to be evaluated in terms of access and acquisition as they support not just 'datanalysis sets' but also 'datacquisition sets' and 'dataccess sets'. In order to meet some of these requirements

we suggest that each definition of a concept in GOLD be formally represented by a set of properties. This will allow us to more accurately map the relationships between concepts and allow the variance in our definitions of concepts to be explicitly modelled. It will also meet the requirements of modelling our uncertainty in annotation, and some of the requirements of modelling inherently indeterminate phenomena. In enabling us to create new concepts for GOLD it also enables us to represent phenomena that are inherently structures, like language-specific ontologies. Further, we suggest that in order to partly address the trade-off between markup and rendering, an ontology should store some of the rendering conventions.

These extensions are needed in order to provide portability and interoperability for language documentation materials. To achieve this, GOLD only needs to remove some strong but unnecessary assumptions from its current model. There are certainly many implementation issues associated with the extensions we propose, but they are not particularly complicated to solve, and can be addressed by those already working on ontologies for endangered languages.

References

Aarts, B 2004 Modelling linguistic gradience. *Studies in Language*, 28(1):1–49.

Bateman, J 1992 The theoretical status of ontologies in natural language processing. In *Text Representation and Domain Modelling – ideas from linguistics and AI*, Technische Universität Berlin

Bayen, H 2003 *Probabilistic Approaches to Morphology* In Bod, R., Hay J. and Jannedy, S. (eds). *Probabilistic Linguistics.* MIT Press.

Bergqvist, H 2005 Semantics of temporal deictics in Lakandon Maya. Presentation given at the ELAP-ELAR seminar series, SOAS, London.

Bird, S & G Simons. 2003. Seven Dimensions of Portability for Language Documentation and Description, *Language* 79/3: 557-582.

Christie, M & W Gaykamangu 2003. “Kinship, moiety, land & language in Arnhem Land”. In *literacy link.* Australian Council for Adult Literacy, vol 23, no 5 Oct 2003.

Christie, M, W Gaykamangu & D Nathan. 2001. *Yolngu Languages and Culture: Gupapuyngu*. Faculty of Aboriginal and Torres Strait Islander Studies, NTU [Multimedia CD-ROM]

Crystal, D. *1997 A dictionary of linguistics and phonetics*. 4th edition. Cambridge, MA: Blackwell

Cysouw, M, J Good, M Albu & HJ Bibiko 2005 *C*an GOLD “cope” with WALS? Retrofitting an ontology onto the World Atlas of Language Structures. *Proceedings of the E-MELD 2005*

Farrar, S. & D. T. Langendoen. 2003. A linguistic ontology for the Semantic Web. *GLOT International* 7 (3), 97-100.

Farrar, S. 2003a Markup and the GOLD ontology. *Proceedings of the EMELD 2003*

Farrar, S. 2003b An ontological account of linguistics: extending SUMO with GOLD. *Proceedings of the 2003 IEEE International Conference on Natural Language Processing and Knowledge Engineering.* Beijing

Foley, W A 2003 Genre, register and language documentation in literate and preliterate communities. In Peter K Austin (ed.) *Language Documentation and Description vol 1*

Grinevald, C 2003 Speakers and documentation of endangered languages. In Peter K Austin (ed.) *Language Documentation and Description volume 1*

Gruber, T R. 1993 A translation approach to portable ontologies. *Knowledge Acquisition,* 5(2), 199-220

Himmelmann, N P 1998 Documentary and descriptive linguistics. *Linguistics* 36. 161-195. Berlin: de Gruyter.

Holton, G 2003 Approaches to digitization and annotation: A survey of language documentation materials in the Alaska Native Language Center Archive. *Proceedings of the EMELD 2003*

Manning, C. 2003 *Probabilistic Syntax* In Bod, R., Hay J. and Jannedy, S. (eds). *Probabilistic Linguistics.* MIT Press.

Nathan, D. (ed) 1996. *Australia's Indigenous Languages.* Adelaide: SSABSA

Nathan, D and P K Austin (2004) Reconceiving metadata: language documentation through thick and thin. In Peter K Austin (ed.) *Language Documentation and Description Volume 2.*

Niles, I & A Pease. 2001. Towards a standard upper ontology. *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems* (FOIS-2001)

Penton, D, C Bow, S Bird & B Hughes. 2004. Towards a General Model for Linguistic Paradigms. *Proceedings of EMELD 2004*