Proceedings of the 6th AGILE
April 24th-26th, 2003 – Lyon, France

ONTOLOGY, NATURAL LANGUAGE, AND INFORMATION SYSTEMS: IMPLICATIONS OF CROSS-LINGUISTIC STUDIES OF GEOGRAPHIC TERMS

Extended Abstract

David M. Mark[1], Werner Kuhn[2], Barry Smith[3], Andrew Turk[4]

[1] Department of Geography, NCGIA and Center for Cognitive Science, University at Buffalo, Buffalo, NY 14261, USA.

[2] Institute for Geoinformatics, Universität Münster, Germany

[3] Department of Philosophy, NCGIA and Center for Cognitive Science, University at Buffalo, Buffalo, NY 14260, USA and Institute for Formal Ontology and Medical Information Science, University of Leipzig, Germany

[4] School of Information Technology, Murdoch University, Perth, Western Australia 6150, Australia

1. ABSTRACT

Ontology has been proposed as a solution to the 'Tower of Babel' problem that threatens the semantic interoperability of information systems constructed independently for the same domain. In information systems research and applications, ontologies are often implemented by formalizing the meanings of words from natural languages. However, words in different natural languages sometimes subdivide the same domain of reality in terms of different conceptual categories. If the words and their associated concepts in two natural languages, or even in two terminological traditions within the same language, do not have common referents in the real world, an ontology based on word meanings will inherit the 'Tower of Babel' problem from the languages involved, rather than solve it. In this paper we present evidence from a preliminary comparison of landscape terms in English with those in the Yindjibarndi language of northwestern Australia demonstrating that this problem is not just hypothetical. Some possible solutions are suggested.

2. INTRODUCTION

Ontology as a branch of philosophy deals with "the nature and the organisation of reality" (Guarino and Giaretta, 1995). , The term 'ontology' has been used more recently in information systems (IS) and knowledge representation (KR) to refer to "a logical theory which gives an explicit, partial account of a conceptualization" (Guarino and Giaretta, 1995). If the conceptualization in question is consistent with the reality to which it is applied, then these two approaches to ontology are complementary. But the two kinds of ontology are used for very different purposes. Ontology in the philosophical sense attempts to discover conceptualizations that match the true nature of the subdomain of reality under examination. It can be characterized as seeking a *theory of reality*, specializing primarily in the preparation of taxonomies of the types of entities existing in a given domain (including the types of relations which unify these entities together into complex wholes of different sorts). On the other hand, ontology in the IS sense normally begins with conceptualizations

developed by human beings for particular scientific and non-scientific purposes, and seeks to formalize such conceptualizations in ways that make them implementable in computer applications. Another way to characterize the difference is as follows: ontological research in the philosophical sense is pure or basic research carried out with the aim of increasing human knowledge; ontological research in the IS sense is directed toward the design of software applications for use in information systems and database management systems. Ontology is seen by both camps as part of the solution to a "Tower of Babel" problem, which threatens the interoperability of databases that are constructed independently, since such independent compilations are likely to have differing definitions of entities and relationships. Since geographic information science is in essence a branch of information science, it is not surprising that this is a domain that has seen much recent interest in ontology (Winter, 2001; Mark et al., in press).

One approach to constructing ontologies is to derive them from natural language texts that describe human activities. Kuhn (2001) used an analysis of the German traffic code as a case study, and illustrated informally the processes involved in deriving ontologies from such sources. Such a research program aims to produce formalized conceptual structures that people can commit to when communicating about reality using a GIS or other means of communication. The method appears to work well for administrative domains, where many concepts are established by fiat, but has not been tested on natural geographic domains.

A particularly important situation in information systems arises when such systems must be used by speakers of more than one natural language. If terms in one language do not subdivide a subdomain of reality in the same way as do the terms in another language, this poses a significant threat to interoperability of such systems across linguistic boundaries.

3. YINDJIBARNDI LANDSCAPE TERMS

Australia was selected for a case study of multilingual landscape terminology. The Yindjibarndi language in particular was selected because of Andrew Turk's contacts in the Yindjibarndi-speaking community (Turk and Trees, 1999, 2000). Yindjibarndi is classified by linguists as one of eight Coastal Ngayarda languages, and is in the southwestern subgroup of the Pama-Nyungan Australian languages (SIL, 2001). At the time of European contact, the language was spoken mainly in the middle part of the basin of what is now known as the Fortescue River (Tindale, 1974). Today, most Yindjibarndi speakers live closer to the coast, in and near Roebourne, Western Australia.

In November 2002, Andrew Turk and David Mark conducted library research and fieldwork regarding landscape terms in the Yindjibarndi language. A list of geographic or landscape terms and definitions was compiled from several published and semi-published dictionaries (Wordick, 1982; Anderson, 1986; von Brandenstein, 1992; Anderson and Thieberger, n.d.). Then, in Roebourne itself, Turk and Mark discussed many of these terms and definitions with local language experts, showing them landscape photographs and asking them what kinds of geographic features were depicted. Turk visited Roebourne again in January 2003 and showed many other landscape photographs to native speakers of Yindjibarndi.

At this time, results are at a preliminary stage (Mark and Turk, submitted). However, it appears that at the basic level – that is, at the level of the most commonly used vocabulary elements – Yindjibarndi landscape terms do not match landscape terms in English, German, or other European languages.

First, consider water in the landscape. In English and other European languages, the two most important distinctions appear to be between flowing water courses and static water bodies, and between fresh and salt water. However, in Yindjibarndi, the dichotomy between permanent and temporary (intermittent, ephemeral) appears to be most salient. This makes sense considering the semi-arid environment Yindjibarndi speakers lived in before the Europeans arrived. Permanent pools (*yinda*) are distinguished from temporary or seasonal

pools (*thula*); similarly, *jinbi*, which are permanent springs and the permanent trickles of water that emanate from them, are categorically distinguished from temporary springs and trickles called *yjirdi*. Note also that both *jinbi* and *yijirdi* combine elements of both static water bodies and water courses. The terms *pool* and *spring* cannot be correctly translated from English into Yindjibarndi without knowing whether or not the feature referred to has water all year round.

As a second example, consider convex topographic features. In English, most such features are categorized as either *mountains* or *hills*, depending primarily on size. Similarly, Yindjibarndi has two such terms, *marnda* and *bargu*, again distinguished mainly by size. *Marnda* is the common Yindjibarndi term for most hills and for any mountains, ridges, and mountain ranges, thus covering the meanings of several terms in English. A *bargu* is a small hill. Seen from the other side, hill in English covers some meanings of *marnda* and perhaps all meanings of *bargu*. A key point is that we cannot reliably translate *hill* into the correct term in Yindjibarndi, nor *marnda* into the correct term in English, without knowing something more about the characteristics of the landform in reality. This presents a serious challenge for compilers of bilingual dictionaries, and for those building information systems that would be used by speakers of more than one language.

The research findings have practical implications. For instance, if the current Ngarluma-Yindjibarndi native title land claim is at least partially successful, it may well lead to joint management arrangements between the Yindjibarndi people and the Western Australian State Government for large national parks in their country (Turk, 1996; Walsh and Mitchell, 2002). If a GIS were to be used to support this management, it would probably be based on the digital version of the relevant 1:100,000 topographic maps, which incorporate (western) landscape categories rather than the Yindjibarndi ones discussed above.

An approach to overcome such practical difficulties could be to establish a generic 'mapping' between the landscape terms from the two linguistic traditions, to facilitate automated geographic data conversion. This could be done through reference to observable physical characteristics that differentiate categories of landforms within each of the two conceptual systems. For instance, with respect to convex geographic features (mountains, ridges, hills vs marnda, bargu) such physical characteristics might be size, shape, and perhaps surface materials.

The next phase of the fieldwork will extend the range of examples of real world features identified by Yindjibarndi speakers as fitting into the various categories. Photographs of the same features will be shown to English-speaking subjects to confirm classification into English-language terms. These parallel efforts may lead to a geometric definition of the ranges of sizes and shapes of features fitting into each category. Such physical definitions could then be combined with digital elevation model (DEM) data to produce automated classifications of features within each of the conceptual (language) systems.

Alternatively, one could ask the Indigenous people themselves to distinguish between categories, producing terrain feature maps of their lands. If such a procedure were undertaken for the Ngaluma /Yindjibarndi native title claim area (covering thousands of square kilometres), it would require many years of effort. However, such an effort could also help address another significant issue - the availability of proper names for landscape features from each cultural tradition. During discussions with our Yindjibarndi collaborators, they frequently mentioned that significant geographic features are usually referred to by their individual (proper) names, rather than by generic terms. Knowing the names for pools, mountains, etc. is an important part of Indigenous Australian culture (Ieramugadu Group Inc., 1995; Rijavec et al, 1995). Some of this information is being collected during native title land claims and would be available for use in GIS developed for joint management of national parks.

4. CONCLUSIONS

The brief sketches of Yindjibarndi landscape terms for two geographic subdomains, water and convex landforms, clearly show that the basic level words in Yindjibarndi and English in these domains do not correspond to the same categories. There is no single term in English that corresponds to all *marnda* and only to *marnda*, nor any single word in Yindjibarndi that corresponds to all *hills* and only to *hills*. If we want to store the category for each feature in a spatial database, or depict it on a map, as is commonly done with feature code or entity type systems, we would have to store both the Yindjibarndi and English categories.

The conceptual system underlying the Yindjibarndi and English language landscape terms, however, appear to be similar at the superordinate level (water or land, concave or convex, etc.). Thus it may be possible to store a single database of topography, land cover, and other characteristics, and determine the categories for individual features based on rules. This would be useful for the development of a common spatial database, from which maps and other GIS products could be produced in relevant languages, using culturally-appropriate conceptual systems.

One approach to this problem is to identify primitives that are universal to all languages, and express other concepts in terms of those universals. Wierzbicka has found a relatively small set of conceptual primitives and successfully tested universality across natural languages for many of them (Wierzbicka, 1996). Her work shows that primitive concepts, at a higher than the basic level, are a feasible goal of semantic analysis. They constitute candidates for concepts at superordinate levels, concepts that can be used to explain the meaning of many natural language terms.

For artificial languages, such as those used in and between information systems, the task of identifying semantic primitives poses a formidable challenge, but may be in some respects easier than for natural languages. The terms used in information systems (e.g., attribute values in databases or operator names of processing languages) are subject to explicit and documented agreements among their designers and users. The notion of information communities (OGC, 1998) is intended to capture this fact and to make the search for primitives tractable by establishing a hierarchy of subsequently more specific levels (Bishr et al., 1999). These agreements form one basis of information system semantics and are, though often implicit and imprecise, available for inspection, revision, and codification. Feature-attribute catalogues, conceptual data models, descriptions of work procedures, and other sources reflect the agreements and can be mined in the process of defining information system ontologies (Kuhn, 2001). Our examination of some landscape terms in the Yindjibarndi language suggests that the prospects for resolving incompatibilities between terms used in both natural and artificial languages might be improved in addition by finding ways to compare languages in relation to an independent description of the reality – in this case geographical reality – towards which the terms in question are directed.

5. ACKNOWLEDGMENTS

This material is part of a project “Geographic Categories: An Ontological Investigation” supported by the U. S. National Science Foundation under Grant No. BCS-9975557. Support of the National Science Foundation and also of the Wolfgang Paul Program of the Alexander Humboldt Foundation is gratefully acknowledged. Members of the Roebourne community, especially Allery Sandy, Trevor Soloman, Marion Cheedy, and Nita Fishook provided invaluable assistance regarding the Yindjibarndi language.

6. REFERENCES

[1] Anderson, B., 1986. *Yindjibarndi dictionary*. Photocopy.

[2] Anderson, B., and Thieberger, N.. *Yindjibarndi dictionary*. Document 0297 of the Aboriginal Studies Electronic Data Archive (ASEDA) Australian Institute of Aboriginal and Torres Strait Islander Studies, GPO Box 553, Canberra, ACT 2601, Australia.

[3] Bishr, Y., Pundt, H., Kuhn, W. and Radwan, M., 1999. Probing the Concept of Information Communities - A First Step Toward Semantic Interoperability. M.F. Goodchild, M.J. Egenhofer, R. Fegeas and C.A. Kottman. *Interoperating Geographic Information Systems* (Proceedings of Interop'97). Kluwer: 55-71.

[4] von Brandenstein, C. G., 1970. *Narratives from the north-west of Western Australia in the Ngarluma and Jindjiparndi languages*. Canberra, AIAS 1970.

[5] von Brandenstein, C. G., 1992. *Wordlist from Narratives from the north-west of Western Australia in the Ngarluma and Jindjiparndi languages*, Canberra, ASEDA Document #0428..

[6] Guarino N., and Giaretta P., 1995. Ontologies and Knowledge Bases: Towards a Terminological Clarification. In N. J. I. Mars (ed.), *Towards Very Large Knowledge Bases*, IOS Press.

[7] Ieramugadu Group Inc., 1995. *Know the Song, Know the Country: The Ngarda-Ngali story of culture and history in the Roebourne District*. Roebourne, Western Australia: Ieramugadu Group Inc

[8] Kuhn, W., 2001. Ontologies in support of activities in geographical space. *International Journal of Geographical Information Science*, 15 (7), 591-612.

[9] Mark, D. M., 1993. Toward a Theoretical Framework for Geographic Entity Types. In Frank, A. U., and Campari, I, editors, *Spatial Information Theory: A Theoretical Basis for GIS*, Berlin: Springer-Verlag, Lecture Notes in Computer Sciences No. 716, pp. 270-283.

[10] Mark, D. M., Smith, B., Egenhofer, M., and Hirtle, S. C., in press. Ontological Foundations for Geographic Information Science. UCGIS Research Challenges. In McMaster, R. B., and Usery, L., editors, *Research Challenges in Geographic Information Science*, New York: John Wiley & Sons, accepted, in press.

[11] Mark, D. M., and Turk, A., *Landscape Categories in Yindjibarndi: Ontology, Environment, and Language*. Manuscript in submitted.

[12] OGC, 1998. The OpenGIS Guide. The Open GIS Consortium.

[13] Rijavec, F., Harrison, N., and Soloman, R., 1995. *Exile and the Kingdom* [documentary film]. Roebourne, Western Australia: Ieramugadu Group Inc. and Film Australia.

[14] SIL, 2001. *Ethnologue: Languages of the World*, 14th Edition. http://www.ethnologue.com/show_lang_family.asp?code=YIJ

[15] Smith, B., and Mark, D. M., 2001. Geographic categories: An ontological investigation. *International Journal of Geographical Information Science*, 15 (7), 613-631.

[16] Tindale, N. B., 1974. *Aboriginal Tribes of Australia. Their Terrain, Environmental Controls, Distribution, Limits, and Proper Names*. Canberra: Australian National University Press.

[17] Turk, A. G., 1996. Presenting Aboriginal knowledge: Using technology to progress native title claims. *Alternative Law Journal*, Vol. 21, No. 1, 6-9.

[18] Turk, A. G. and Trees, K. A, 1999. Culturally Appropriate Computer Mediated Communication: An Australian Indigenous Information System Case Study. *AI and Society*, Vol. 13, pp. 377-388.

[19] Turk, A. G. and Trees, K. A., 2000. Facilitating Community Processes Through Culturally Appropriate Informatics: An Australian Indigenous Community Information System Case Study. In: Gurstein, M. (ed.) *Community Informatics: Enabling Communities with Information and Communication Technologies*, Idea Group Publishing. pp. 339-358.

[20] Walsh, F. and Mitchell, P. (eds.), 2002. *Planning for Country: Cross-cultural Approaches to Decision-making on Aboriginal Lands*. Central Land Council, Alice Springs, Australia: Jukurrpa Books.

[21] Wierzbicka, A., 1996. Semantics - Primes and Universals, Oxford University.

[22] Winter, S., 2001. Ontology: buzzword or paradigm shift in GI science? *International Journal of Geographical Information Science*, 15 (7), 587 – 590.

[23] Wordick, F. J. F., 1982. *The Yindjibarndi language*. Pacific linguistics. Series C. ; no. 71. Canberra : Dept. of Linguistics, Research School of Pacific Studies, Australian University.