Interlingual BRICO

by K. Haase

BRICO is a broad-coverage ontology built by combining a variety of on-line resources. The initial English ontology has recently been extended to include Spanish, Italian, French, German, and Dutch, and additional extensions are planned. This paper discusses the creation and extension of the interlingual ontology, together with some prototype applications of the database.

In this paper we describe BRICO, a broad-coverage ontology built by combining a variety of on-line resources. These resources include the WordNet on-line lexical thesaurus,[1–3] a public-domain version of Roget's 1911 thesaurus, and the publicly available "top level" of the CYC common-sense knowledge base.[4] BRICO was built to support work in natural language understanding and analogical representation. It has also been a test case for the FramerD[5] knowledge base infrastructure.

The word BRICO is a shortened version of the French word *bricolage*, which denotes a functional but haphazard assemblage of components that solve a problem. The spirit of BRICO is the assemblage and interconnection of knowledge resources, rather than the characterization of any over-arching primitives or principles. We believe that this sort of architecture is more like the human mind than the elegant logical lattices of traditional knowledge bases. And we believe that we will learn more about human knowledge by constructing such a composite knowledge base.

BRICO's core consists of knowledge about the meaning of words. It has been recently extended to include languages other than English—initially Spanish, Italian, French, German, and Dutch, with more to come—by an algorithm that uses translation dictionaries to provisionally assign foreign words to BRICO concepts. This provisional assignment is then refined and corrected by human informants. This paper describes the extension process, the resulting knowledge base, and possible applications.

The organizing paradigm for BRICO's representation of word meanings is the notion of "synsets," which it inherits from WordNet. A synset represents a meaning, or word sense, that may be independently named by several different words. For instance, one synset in WordNet is comprised of the nouns *example*, *instance*, *illustration*, and *representative*; another is comprised of the verbs *read*, *take*, *learn*, and *study*. Synsets in WordNet are organized by part of speech and related to other synsets for the same part of speech. For the most part, there are no relations between synsets for different parts of speech.

Synsets are organized based on substitution relations in natural language sentences. The fact that two words are assigned to some synset can be read as saying "there exist a set of sentences in which these words can be substituted for each other without significant loss of meaning." Thus, two sentences such as:

WordNet is a typical example of an ontology.
WordNet is a typical instance of an ontology.

have significantly similar meanings (given certain contextual assumptions) and so the substitutable words (example and instance) define a synset. Synsets are not definitions, in a conventional sense, but are empirical "identifications" of meanings.

Synsets in WordNet are related to one another in a number of ways. One such relation, *hypernymy*, is a generalization relation, so that the synset:

@(NOUN.COGNITION "example" "instance"
"illustration" "representative")

has the hypernym:

@(NOUN.COGNITION "information")

which (for instance) distinguishes it from another synset containing the word "example":

@(NOUN.EVENT "case" "example" "instance")

with its hypernym:

@(NOUN.EVENT "occurrence" "happening"
("natural" "event"))

The inverse of the hypernymy relation is called *hyponymy*, so that whenever X has Y as a hypernym, Y has X as its hyponym.

Other relations inherited from WordNet include part-whole relationships. For instance, the synset:

@(NOUN.ARTIFACT "telephone" "phone"
("telephone" "set"))

has parts relationship to the synsets:

@(NOUN.ARTIFACT ("telephone" "receiver")
"receiver")
@(NOUN.ARTIFACT "mouthpiece")

but is itself in a part-of relationship to:

@(NOUN.ARTIFACT ("phone" "system") ("telephone"
"system"))

which itself has the parts:

@(NOUN.ARTIFACT "line" ("electrical" "cable")
("transmission" "line") "cable")
@(NOUN.ARTIFACT "telephone" "phone"
("telephone" "set"))
@(NOUN.ARTIFACT "plugboard" "switchboard"
"patchboard")
@(NOUN.ARTIFACT "central" "exchange"
("telephone" "exchange"))

Two other kinds of part-whole relationships are carried over into BRICO: stuff-of/ingredients (e.g., sand is the stuff-of beaches) and member-of/members (e.g., a professor is a member of a faculty).

In WordNet, only the hypernymy/hyponymy relations are thickly populated. One ancillary goal of BRICO is to flesh out the other relations in an effort to convert a lexical database into a world knowledge base.

Except for a small number of imported words, the words in WordNet are limited to American or British English. In order to explore concept- and understanding-based applications across languages, we undertook to extend BRICO to other languages.

Among our goals for the extension process were to make it as automatic as possible and to rely on the English language WordNet for our initial structure. These goals distinguished this effort (in part) from the "EuroWordNet" project funded by the European Commission.[6] This project, completed in 1999, coordinated the parallel development of independent ontologies for several different languages with the goal of identifying core terminologies that the independent efforts could share. It largely succeeded in those goals.

The EuroWordNet project employed substantial human effort in developing independent ontologies for different languages. Our goals were to:

- Develop an approach that could be easily applied to many different languages, taking advantage of lexical resources (translation dictionaries) that had probably already been assembled.
- Learn from the anomalies revealed by the naive alignment of different languages

Method

In April of 1999, we conceived a simple way to combine a translation dictionary with a version of the WordNet ontology to yield a "rough cut" of an ontology for other languages. The intuition was to take

the conceptual structures encoded in WordNet synsets and identify corresponding structures in other languages. We took advantage of the fact that most translation dictionaries translate each word into several words in the target language. The initial version of our algorithm simply assigned a foreign word to each synset containing more than one of its translations into English. For instance, the Spanish word *modelo* translates into both the English word "model" and the English word "example." This prompts its provisional addition to the synsets:

```
@(NOUN.COGNITION "model" "example")
@(NOUN.COGNITION "exemplar" "model"
   "example" ("good" "example"))
```

We extended this core algorithm to handle the simple case where a single translation had a single corresponding synset. For example, the Spanish word *canalla* has the single translation "cad" in English, which has the single meaning:

```
@(NOUN.PERSON "hound" "dog" "blackguard"
   "heel" "bounder" "cad")
```

allowing for a single interpretation.

Our initial extension of WordNet was based on automated reference to four translation dictionaries: Spanish, Italian, German, and French. Subsequently we downloaded the freeware Ergane dictionaries from http://www.travlang.com/Ergane/ and extended this (with varying degrees of coverage) to Dutch, Danish, Swedish, Finnish, Portuguese, and Swahili.

The "rough cut" generated in this way had a wide range of errors. Some of these errors derived from actual errors in the translation dictionaries we used, others derived from the occasional places where these dictionaries translated words across parts of speech (e.g., nouns into verbs). In addition to these "noise" problems, our simple heuristic further failed in a number of recurring cases:

- Some words—particularly very simple words—had single translations, providing no support for the heuristic, leading many common words to not be assigned.
- WordNet synsets derived by metaphor or simile tended to impose the same metaphor or simile on the second language, resulting in misassignments of words to synsets.
- Some concepts that might be common between languages are lexicalized as nouns in one language and verbs in another.

We were anticipating these problems and planned to manually clean up the rough cut generated by the heuristic. This effort is still in progress, but we will discuss the interface and strategy behind it.

We are working to precisely evaluate the effectiveness of our rough-cut algorithm by looking at the number of corrections that need to be made by human informants in proofing the database. Informal discussion suggests that perhaps 90 percent of the algorithm's attributions are correct and a large number of errors are the direct consequence of errors in the original translation dictionaries.

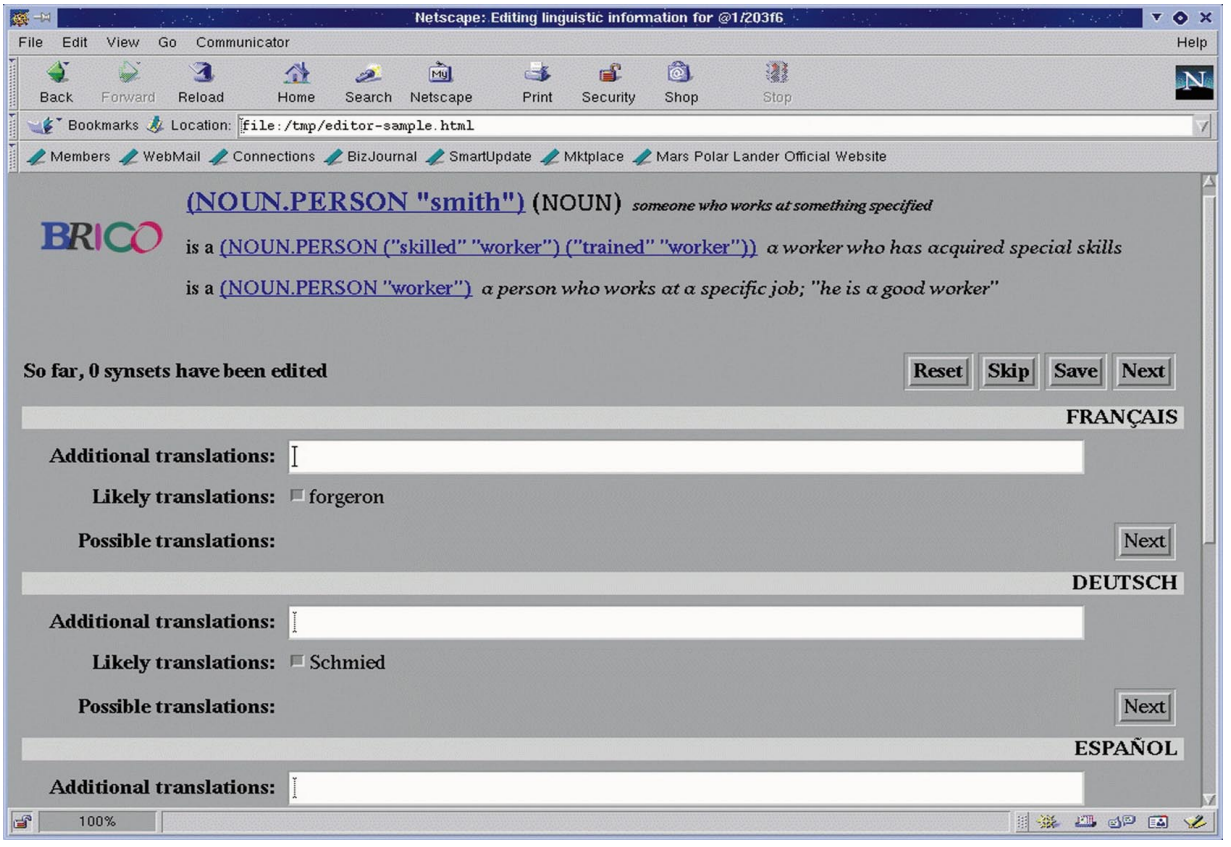
Polishing the rough cut. In order to address these deficits, we began manual evaluation and proofing of the extended BRICO. This was done with a live Web interface, shown in Figure 1. The interface displays a synset with possible translations into the target language(s). The possible translations include dictionary translations for all of the words in the synset. Of these possible translations, the likely translations (those picked by the algorithm described) are selected by default.

This interface is used, by a native speaker of the target language, to correct and extend the cross-linguistic annotations. The native informant can add translations (checking boxes), remove translations (unchecking boxes), or add entirely new words (through a text entry field). So far, we have begun this proofing and extension for Spanish and Italian, with more languages planned for the future.

Initially, we used an interface that picked individual words in the foreign language and proposed matching synsets. This turned out to be more difficult for the native informant, because it required interpreting several English language synsets. The current interface requires interpreting a single English language synset and then adding or removing words from the native language. This has been much more productive in terms of "relations per person-hour."

Given that the applications based on the extended BRICO were being developed in parallel, we also gave some thought to the order in which we improved the quality, e.g., which synsets we tried to improve first. Our initial approach was a bias toward synsets that had short words assigned to them, attempting to get

Figure 1 Web interface for editing Interlingual BRICO



at familiar words. A later approach used "number of meanings" as an index to familiarity (following Jastrezembski[7] and Zipf[8]), disambiguating synsets that might be readily confused with others. Still a third alternative was to march breadth-first down the hypernymy hierarchy, starting with the more general concepts. We are currently pursuing a combination of the second and third strategies to order our clarifications of WordNet.

We are considering exposing this interface to the Internet community in order to assist in this interpretation. The advantage of such an exposure is to leverage the contributions of the worldwide community. The disadvantage is the need to catch errors and do quality assessment.

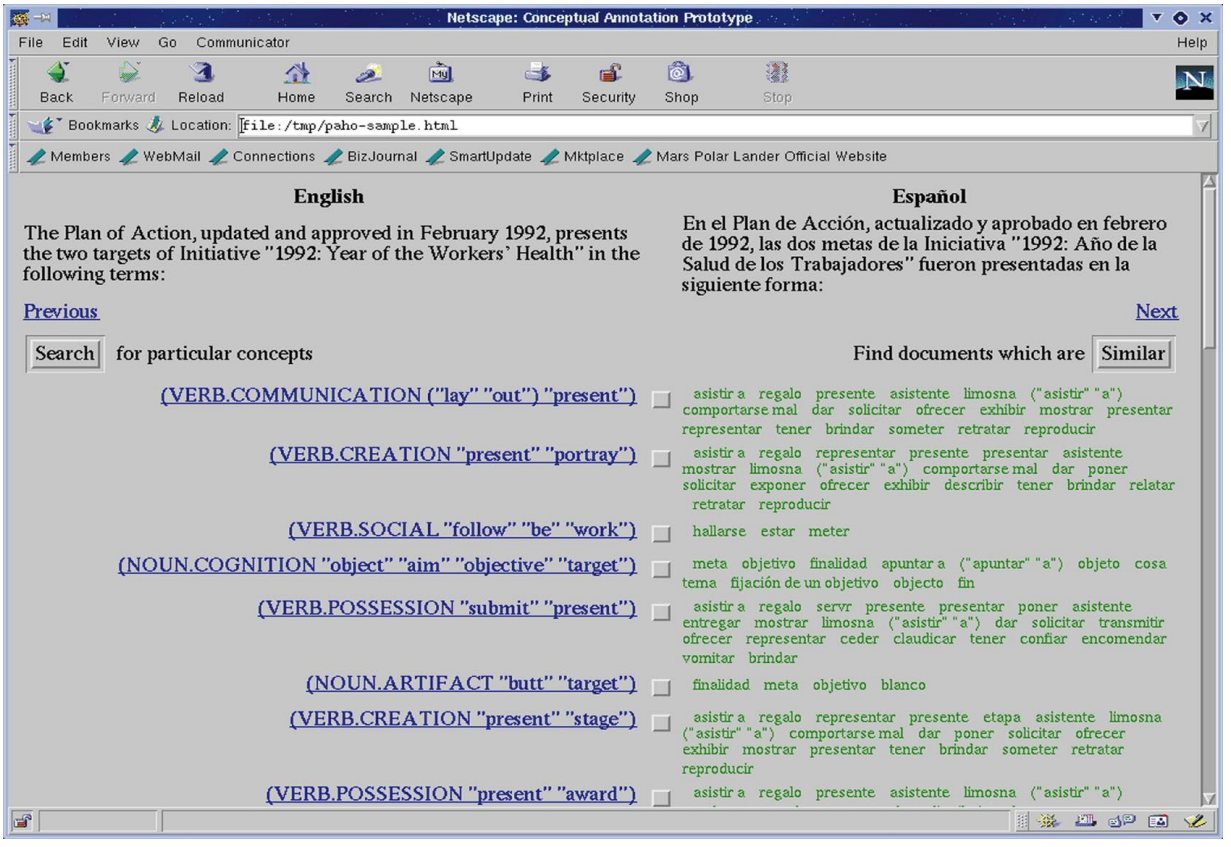
One way to evaluate contributions—both by our staff and by Internet-based contributors—would be to have multiple informants for each language and cross-correlate their word assignments. Another approach would be to use application performance (for instance, the semantic tagging of aligned corpora, described later) as feedback to the quality of information and informants.

Applications

There are many possible applications of an interlingual WordNet. We describe two that we have experimentally implemented: sense-tagging of translated documents and interlingual annotation of images.

Sense tagging. One major problem in text interpretation and information retrieval is the determination of word senses from word forms. Experiments with WordNet[9] have shown that retrieval using concepts can be much more effective than retrieval using keywords alone. However, the problem is that documents typically consist of words and not disambig-

Figure 2 Screen image of the interlingual search engine



uated synsets. Disambiguation is a very hard problem and poor disambiguation is usually worse than no disambiguation at all.

An interlingual ontology, however, affords an interesting possibility for disambiguating translated documents. When a document has been translated from one language into another, a human translator has made judgments about which foreign words capture the meaning of the words in the original document. An interlingual ontology allows us to reconstruct this reasoning and determine the meaning as originally disambiguated by the human translator. Briefly, if a human translator translated the English word "example" to the Spanish word *modelo*, we can determine that the word "model" actually had one of the three meanings:

```
@(NOUN.COGNITION "example" "instance"
   "illustration" "representative")
@(NOUN.COGNITION "model" "example")
@(NOUN.COGNITION "exemplar" "model"
   "example" ("good" "example"))
```

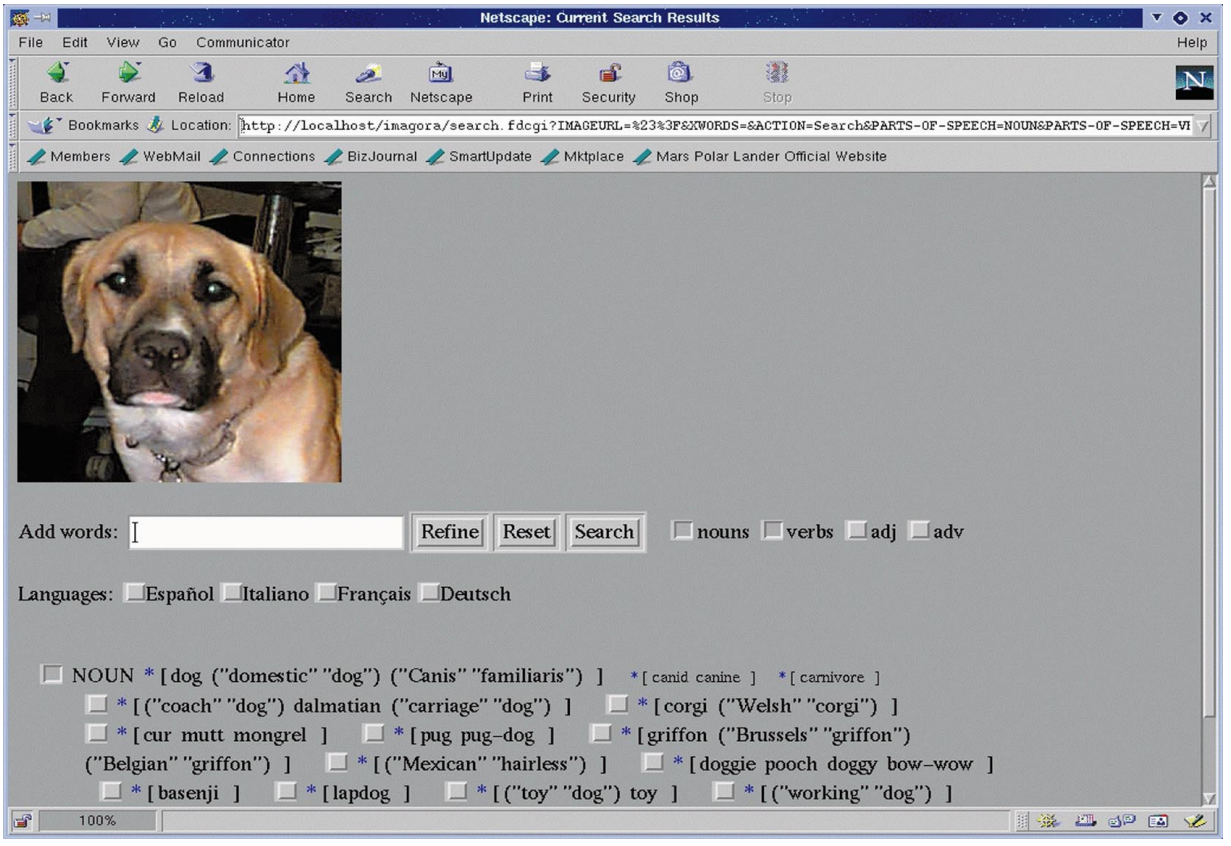
rather than one of the six meanings:

```
@(NOUN.COGNITION "example" "instance"
   "illustration" "representative")
@(NOUN.EVENT "case" "example" "instance")
@(NOUN.ACT "example" "exercise")
@(NOUN.COGNITION "model" "example")
@(NOUN.COGNITION "exemplar" "model"
   "example" ("good" "example"))
@(NOUN.COMMUNICATION ("object" "lesson")
   ("deterrent" "example") "example"  "lesson")
```

which can help in searching, indexing, or browsing.

We have experimentally used this approach to assign word senses to documents of the Pan American Health Organization translated between Spanish and

Figure 3 Interlingual image browser (set to English)



English. Figure 2 shows our interlingual browser. The concepts listed beneath the paired paragraphs have been provisionally assigned. Selecting particular concepts (by checking the boxes) can be used to search for other paragraphs mentioning the same concepts (by clicking the "Search" button on the left-hand side of the screen). The concepts can be used together to find similar documents (those that combine mentions of the same concepts) by clicking the "Similar" button on the right-hand side of the screen.

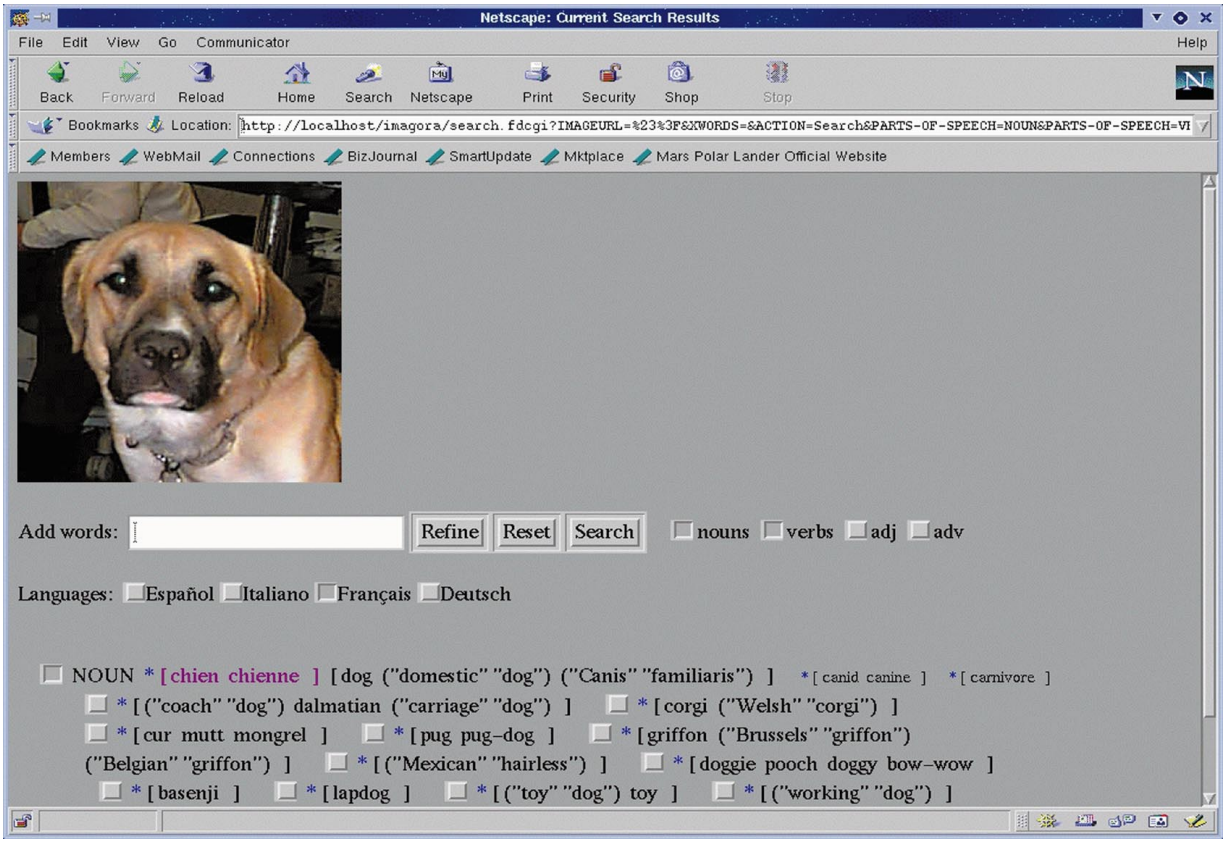
We are further evaluating this with some of the online documents of the European Commission, where the availability of extra languages may further help the disambiguation process by reducing the overlap of senses between languages.

Interlingual annotation. We have developed an experimental application for annotating images with synsets, allowing the annotation and browsing process to occur across languages. Images can be annotated in English (for instance) and then retrieved in Spanish or French (or *vice versa*). The annotation process involves some degree of disambiguation, but this has the added advantage of greatly improving retrieval recall and precision.

The prototype interface is shown in Figure 3. The image being annotated is on the top and the assigned categories are listed below. Figure 4 shows exactly the same data, but with the French language selected.

This prototype leaves many development issues and research questions unanswered. Two particularly pressing questions are how to best streamline the annotation process (images per person-hour) and the relative importance of conceptual annotations (e.g., "dog, chien, perro") and visual annotations (e.g.,

Figure 4 Interlingual image browser (set to French)



"balanced," "dark background"). We look forward to exploring these issues in future work.

One interesting way to evaluate this interface is to have different users annotate the same images from different languages. By correlating the annotations, we can evaluate the degree to which different users found the same concepts. This is similar to the evaluation criteria used by Davis[10] to evaluate iconic video annotation systems.

Future work

One clear area of future work is the further proofing and refinement of the multilingual BRICO. We will be continuing this work using native informants and are exploring the possibility of inviting the international Internet community to join the effort.

A second area is the evaluation of matches and mismatches across languages. Interesting results in language differences and language evolution might emerge from the systematic study of how conceptual structures carry (or fail to carry) across languages.

A third area would be to provide standard inferences that use the knowledge of word meanings encoded in BRICO. The search and retrieval mechanisms above can make some use of BRICO's generalization relationships (hypernymy) and other work[11] has used these links as the basis for analogical reasoning. But more work could be done in this area and the integration of common sense reasoning into BRICO's broad ontology is an exciting prospect.

Finally, the application areas we have discussed could certainly be expanded. We are considering starting an international "Children's Image Database" using interlingual annotation to allow children from different nations and languages to share the images of their homes and lives. And there is certainly the

possibility of using the same technology in a commercial setting to make media resources available internationally.

Cited references

1. *WordNet, An Electronic Lexical Database*, C. Fellbaum, Editor, MIT Press, Cambridge, MA (1998).
2. G. A. Miller, "WordNet: A Lexical Database for English," *Communications of the ACM* **38**, No. 11, 39–41 (1995).
3. G. A. Miller, "WordNet: An On-Line Lexical Database," *International Journal of Lexicography* **3**, No. 4, 235–312 (1990).
4. D. B. Lenat and R. V. Guha, *Building Large Knowledge-Based Systems: Representation and Inference in the CYC Project*," Addison-Wesley Publishing Company, Reading, MA (1990).
5. K. Haase, "FramerD: Representing Knowledge in the Large," *IBM Systems Journal* **35**, Nos. 3&4, 381–397 (1996).
6. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*, P. Vossen, Editor, Kluwer Academic Publishers, Dordrecht, Netherlands (1998).
7. J. E. Jastrezembski, "Multiple Meanings, Number of Relationed Meanings, Frequency of Occurrence, and the Lexicon," *Cognitive Psychology* **13**, 278–305 (1981).
8. G. K. Zipf, "The Meaning-Frequency Relationship of Words," *Journal of General Psychology* **33**, 251–256 (1945).
9. E. Vorhees, "Using WordNet for Text Retrieval" (Chapter 12), *WordNet, An Electronic Lexical Database*, C. Fellbaum, Editor, MIT Press, Cambridge, MA (1998).
10. M. Davis, *Media Streams: Representing Video for Retrieval and Repurposing*, Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA (1995).
11. K. Haase, "A Model of Poetic Comprehension," *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, Portland, OR (August 4–8, 1996), pp. 156–161.

Accepted for publication May 12, 2000.

Kenneth Haase *MIT Media Laboratory, 20 Ames Street, Cambridge, Massachusetts 02139-4307 (electronic mail: haase@media.mit.edu).* Dr. Haase is the Chief Scientist of the Media Laboratory's News in the Future consortium and a visiting associate professor at the Media Laboratory. He received his Ph.D. degree in 1990 at the MIT Artificial Intelligence Laboratory, working with Marvin Minsky and philosopher Thomas Kuhn on models of computer creativity. He is also a part-time professor of journalism and mass communications at the University of Tampere in Finland. His research interests include knowledge representation, natural language processing, information retrieval, and creativity.