Michelle Vanni
Dept of Defense

Pre-Workshop on Interlinguas
Position Paper

In this paper, I would like to address the second and third issues presented in the announcement, namely, "What information is captured by an adequate interlingual representation system?" and "How can interlingual representation systems be built or scaled up?" My basic position is that while interlinguas are comprehensive and well-designed, there is a heavy reliance on the lexicon to carry meaning feature information to the interlingua. Too little investment is made in processing the interaction either among overlapping lexical feature values or between such values and the output of analyzers at other levels. Investigation of new approaches to such processing holds great promise for improving the efficiency and effectiveness of MT systems using an interlingual design.

It seems that we have managed to consult our semantics and pragmatics books and design into our systems nearly all those aspects of meaning which may be relevant to text understanding. Interlinguas are quite well-developed. In ULTRA'S Intermediate Representation (IR), referential, rhetorical, and intentional aspects of communicative acts are fully described with up to 52 possible fillers for some slots! Robustness of design is not a problem. But, how does one determine where the values for the features are encoded in the source language? Where is the research being done to accomplish this?

One way in which the assignment of feature values can be optimized is through investigation of overlapping values and the provision of defaults in the case of overgeneration. During preliminary stages of the Pangloss project, analyses were performed on candidate IRs output by the ULTRA parser for a set of 25 sentences. In the course of this work, it was noticed that most of the variations between IRs for a particular sentence were binary alternations in the filler (value) of a particular feature slot.

For example, since frequently in Spanish there is no difference between the written representation of a declarative sentence and that of an interrogative one (except for punctuation), almost all sets of candidate IRs for a sentence would include an alternation based on the choice of declarative v. interrogative. Add to this the general v. specific and the existential v. unique alternations for articles in a given sentence and you already have at least eight possible combinations. Analysts invoked the work of Halliday and Hasan (1976) suggesting that principles of textual cohesion which predict the values of particular slots or a high probability for a value might be integrated into the internal processing of IRs as constraints to limit their overgeneration.

The interaction between analyzer output and the lexicon is another area which could benefit from further research. Much of the information represented in one major interlingual design, the Mikrokosmos Text Meaning Representation (TMR), is initially coded in the ten zones available for each lexical entry: grammatical category, user information, orthography, phonology, morphology, syntactic features, syntactic structure, semantics, lexical relations, and pragmatics. I would like to suggest that values for some of these zones are misplaced in the lexical structure and that they can be better obtained at other levels of processing.

For example, in the Mikro lexical structure, indications of irregular morphological formation are supposedly noted in Zone 5, morphology. But, it is unclear what purpose this zone actually serves. In order to perform lexical search in a reasonable way, the lemma to be searched needs to be determined, even if the form from which it is derived is irregular, prior to lexical processing. Irregular forms, usually finite in number for a given language, are easily identified when listed in a morphological analyzer which returns a lemma with morphological features attached. It is only then that the form is efficiently searched in the lexicon.

A more effective use of the proposed morphology zone might be

the identification of a unique sense of a form, that is, an attachment to an ontological node of a lemma associated with a particular set of morphological features which cannot be assigned to another form of the same lemma. An example might be be the different tense and aspect feature values associated with forms of the verb, _conocer_, in Spanish which point to different places in the ontology, as indicated in Guillen-Castrillo (1996).

Now it may be that the designers of this particular lexical structure had a specific linguistic issue or phenomenon in mind in creating their zones and it is not my intention to take issue with the features of any single interlingual design. In fact, in this case, there is surely a "microtheory" planned to address the issue. I only want to point out that, if we want these designs to be viable, we need to take a closer look at how we plan to populate them and seek to align them with what occurs in actual textual data.

Accounting for linguistic phenomena occurring in actual text has long been a goal of the computational corpus linguistics (CCL) community. In fact, the value of rational models for parsers designed to access linguistic data in corpora, given that those models account for a small percentage of the phenomena which actually occur in text, is regularly assessed. Once the limitations of the existing models are understood, issues revolve around the extent to which the models are valuable, the quantity of additional phenomena to be represented, and the criteria for determining which phenomena are appropriate to represent.

The IL MT and the CCL communities share the imperative to provide a measure of linguistic coverage in their system designs. Once it is realized that current IL models beg the question of how to access the value information with which to fill the feature slots, or at least address it only minimally at present, the real work of prioritizing such "microtheory"-type investigations and performing linguistic corpus analysis to determine how the values can be derived, can begin. One promising approach is the development of "Construction Grammars" proposed by Levin and Nirenburg (1994) for augmenting current lexicons with information to be gleaned from identification of the construction in which the lexical item occurs. It is my belief that extensive corpus analysis will lead to the discovery of more than a few construction types which can then be inventoried in a system to serve as a feeder to the IL of semantic feature-value information.

In summary, my interest in the pre-workshop regards deriving from the input text the information provided for in proposed IL structures (ULTRA IRs, Pangloss TMRs, and UNITRAN LCSes,

etc). Integration of morphological information, corpus-based linguistic analysis, and work on developing construction grammars can be effective. While it is true that such work is provided for within the "microtheory" concept for Mikro, few substantial results or attempts at implementation of these ideas have been forthcoming possibly because no standard vehicle or way of talking about such progress has been established.

I think we need to (1) prioritize "microtheory"-type work, (2) work with corpora to find sense differences triggered by different morphological forms and grammatical constructions, and (3) progress toward development of a common language for reporting on findings. Already-robust IL designs can be enhanced with results of investigations of semantic phenomena occurring in actual textual data.