Lexical and World Knowledge: Theoretical and Applied Viewpoints

John S. White
PRC Inc.
1500 Planning Research Drive
McLean VA 22102
(703) 556-1899
white_john@po.gis.prc.com

Three discussion points are addressed from two perspectives: that of an anthropological tradition in cognitive science, and that of application-oriented natural language processing. From the cognitive perspective, knowledge of the world is not as influenced by linguistic semantics as we are tempted to think. From the applied NLP perspective, the distinction between world and lexical knowledge can be determined practically on the basis of what is needed for the computational task.

Lexical and World Knowledge: Theoretical and Applied Viewpoints

This position paper expresses two perspectives to the issue of the relationship of lexical semantics and knowledge representation, one from a flavor of cognitive science and another from the applied orientation to natural language processing (NLP). The perspectives are very different in their effects: this variety of cognitive science is highly bound to how humans categorize an arguably continuous universe; that variety of NLP is devoted to how to make machines "do" language understanding just well enough to perform some delimitable tasks automatically. The contrasts in the positions are obvious in terms of fidelity to how it is that humans must actually organize their worlds and their lexicons. But there is a significant similarity as well, in that, once a fervently held view about determinism is expressed, it is possible to exploit the boundaries between lexical and world knowledge for both the ends of cognitive science and NLP, without exactly knowing what those boundaries are.

The original gauntlet thrown for the *Siglex* workshop was a series of questions which raise the relevant theoretical and applied issues for determining the difference between types of knowledge, the use of one for the determination of the other, and cross-cultural correlates of knowledge-representation/lexical representation phenomena. This paper uses three of these as a springboard (or better soapbox) for presenting my views on the Sapir-Whorf hypothesis (that is, the various deterministic positions on perception, cognition, and language structure). It is hoped that this such expression will be preaching to the choir, though it is possible that it will not be. Upon the expression of those views, however, I am more at liberty to show how, on the one hand, cultural knowledge helps solve lexical problems, and on the other, how useful lexical organizations are as expressions of world knowledge because they can be exploited efficiently.

The three discussion points are addressed here are the methods of determining the scope of lexical knowledge, cross linguistic evidence for lexical specificity, and implemented systems to lexical/non-lexical knowledge. Rather than a serial exposition of these, however, they are organized at the highest level by the methodological issue, treating the cross cultural perspective as a subordinate issue via the treatment of a cultural/lexical problem in anthropology, and then the implemented systems as examples of methodological practicalities.

1.0. Lexical and extra-lexical methodologies. From the human language point of view, there are intuitive tests in ordinary linguistic behavior for determining lexical and non-lexical knowledge. Among the variety of real and imagined failings of our children, for example, we do not accuse them of being unable to speak their native language, or, if we do, it is not because they do not know the capital of Korea (and thus occasionally miss the point of some utterance about where Reeboks come from). This judgment indicates that there is a boundary somewhere, and is probably a suitable first step in developing a usable linguistic methodology, one consistent with the way linguistic judgments are done. Another method involves performing cognitive tests on bilinguals and monolinguals in the same cultural setting (which begs numerous questions); the apparent similarities in organizing behavior among should in principle be the influence of world knowledge alone, and the differences possibly of language influence (to the limits expressed rather abundantly both above and below). This experiment is on the face of it too difficult to control; we seek a more bounded set of methods in the same spirit.

The next discussions beg the question of finding the boundary itself, but clearly exploit the boundary. First, the boundary between cultural belief and lexicon is used to determine the lexemic status of a morpheme in Tojolabal-Maya. Secondly, several examples in machine-readable dictionary work are discussed to show how these can be used self-consciously to represent both the lexical and conceptual organization necessary for natural language processing tasks as they are currently applied.

1.1. Lexical/cultural interactions and Sapir-Whorf. The following addresses the issues of cross-cultural perspectives on lexical/world knowledge within the context of the methodological issues, and in turn addresses linguistic determinism problems that arise whenever terminological structures are hypothesized to be related to the way people think about the world.

1.1.1. Tojolabal evil eye. In the study of the lexical semantics of Tojolabal-Maya (an indigenous language of the highlands of Chiapas, Mexico) I discovered a set of phenomena associated with the indigenous belief of "evil eye" which apparently shed light on the lexical unity of what seemed to be several homophonous words (White 1979). It was notable that many of the objects in the world associated with evil eye were expressed by (or counted by) the same Tojolabal string, which heretofore was regarded as several different, homophonous words.

The Tojolabal word *sat* has a broad range of translations. It occurs in two parts of speech (noun and numeral classifier), and can mean "eye," "face," "top surface," and (as a classifier) a piece of fruit,the sky, a word, or a coin. It is tempting, especially because of the apparent divergence in part of speech to regard the different sense of *sat* to be associated with different lexical items, and I had happily done just that until the people told me about the full extent of effects of the evil eye. Typically, you can infect a child with evil eye by pointing at its face (you are hereby advised against it), but one also causes fruit on the tree to rot by pointing at it. Meanwhile, involving your finger with the sky (pointing at the moon or a rainbow or counting the stars) will get you warts, and carrying a coin will prevent the dangerous effects of looking at a corpse. The interaction of the affectors and affecteds in these scenarios, along with the behavior of "ripe" and "unripe" objects, led to an observation that reducing that *sat's* to one also reduced the number of formulas for evil eye. This fact supports the hypothesis that there is one lexical entry *sat* in Tojolabal lexicon, from which the several referents are results of extension processes.

1.1.2. The relevance of Sapir-Whorf. Such investigation of cultural belief complexes and their relationship to the lexical semantics of the language spoken by that culture comes perilously close to raising the issues of influence of the one over the other, and in particular the Sapir-Whorf hypothesis. It seems intuitively reasonable to suppose, as the anthropologists of the early 20th century did, that the specificity by which a class of objects is lexicalized has to do with the cultural relevance of the concepts associated with that. Surely the the number of lexemes used to discriminate among classes and subclasses of thing must, at least much of the time, be affected by the importance of that category to a particular endeavor (e.g., breeders know more words for dogs than do non-breeders) But this presumes the primacy of the lexeme, that we categorize more where we have more unique words within the category. As we will discuss below, there are universal properties of taxonomic classes, internal to the lexical system itself, such as genericity, primary/secondary lexemes, expression of "truest" member, and so on (Berlin *et al., op.cit.*). However, none of these properties automatically say anything about people's ability to conceive of, or perceive, the object/idea itself.

The classic examples are found in cross-linguistic color terminology. What happens in languages where primary lexemic terms are quite restricted (e.g., only a black, white, red, and blue color word), is rather like what happens when I enter a paint store: the lack of terms does not reflect knowledge of the world nor perceived distinctions in color. Of course, speaking with people whose language does not have a lexemic distinction between blue and green reveals that there is no trouble at all circumlocuting to express the perceived distinction between the "blue/green" of grass and the "blue/green" of the sky, and other distinctions finer still. Relatively recent work by Kay and Kempton (1984) carried this observation to more empirical levels, to dispel the Whorfian claim that lack of a terminological distinction would change the perception of the boundaries between colors (the paper also is a good source for discussion of the Sapir-Whorf hypothesis).

Thus I reject the extremes of Sapir-Whorf interpretations (and the magic and perhaps self-exculpation inherent in any deterministic theory), and maintain instead the more pedestrian view that humans transcend all sorts of constraints, including the strictures of language and its lexical systems. This position allows a characterization of the specificity of lexical-semantic categories in cross-linguistic comparison, but does not immediately allow us to know what its relationship to world knowledge might be. For now, I resort to the explanation provided by Kay and Kempton, that lexical structure gives a channel for perception and cognition, barring other evidence or more convenient channels. To this extent, and no further, then, specificity in the lexicon influences world knowledge organization, and vice versa.

Returning to the Tojolabal case, it is tempting to suspect that the association of concepts under evil eye was influenced by the lexical unity of *sat* (i.e., supporting the Sapir-Whorf hypothesis as commonly put forth), there really is no grounds for that position any more than the opposite position (that the strength of the belief system maintained the integrity of the reference of the lexical item). Thus the phenomenon was neutral with respect to the influence of language on perception/cognition.

1.2. Natural Language Processing Methodologies.

As has been hinted at already, the distinction between what is and is not lexical knowledge can be visited practically. The distinction hopefully reduces to that between what is the least you have to know in order to do the work your lexicon has set you up to do in NLP, and what can you get away with excluding. To whatever extent one can present world knowledge in an algorithmic way that is perhaps the same way that you represent lexical knowledge for those purposes with precisely those same limitations (I refer to feature/value pairs, however computationally represented), one does it. To the greatest extent possible, we reduce world knowledge to lexical (and maybe syntactic) differences caused by "domain" or "sublanguage" and code those difference as if lexical, and segregate these entries in sub-dictionaries. The discussion of NLP systems necessarily involves the workshop issue of theoretical approaches and implemented systems and their methods of combining lexical and non-lexical knowledge.

Successfully implemented NLP systems do combine lexical and world knowledge together at some stage of their operations. The actual decision as to which component certain knowledge should go depends more on software engineering, human factors, and efficiency issues than on capturing the natural linguistic/knowledge boundary. This is almost regrettable, except that the realization allows us a clearer segregation among systems which are intended to model human behavior for its own sake, and those which are destined to take advantage of task and domain boundaries to do a particular job. It is frequently argued that no such boundaries exist, that any discourse will be bound up in world knowledge without which it cannot be interpreted completely. And I agree: differentiating between the two senses of "speaker" is only helped slightly by verb case role assignments, and hardly at all by domain delimitations (e.g., telephony) via specialized dictionaries. In implemented systems, the choice in correctly translating or otherwise processing "speaker" may lie between *ad hoc* feature value assignments (with concomitant loss of maintainability) and punting (letting the human disambiguate it at some point in the process). The criterion for judging success in these, however, is efficiency and acceptance by users/operators, criteria which are influenced by many factors beyond just the correctness of the knowledge representation.

1.2.1 Machine Readable Dictionary Taxonomies. The archetypal work in the computational manipulation of lexical structures is Robert Amsler's work in developing taxonomic structures from machine readable dictionaries. As is probably well known (Amsler 1980, Amsler and White 1979), Amsler used the ISA links expressed by the format of the dictionary entry between the word defined and the syntactic heads of the definition to build recursive, transitive taxonomic structures. Among a great many things that were discovered in that project was that there were principled ways to determine hidden structural information about the lexical hierarchies. Several of these principles were quite similar to those derived in the discipline of cognitive science called ethnosemantics.

Much work in this branch of anthropology emerged in the 1960's and early 1970's as attempts to design knowledge acquisition methods which could elicit culturally relevant organizing principles from informants without imposing the investigators own organizing principles. A landmark study by Berlin et al. (1974) had concluded that there were universal generalizations that could be drawn about how people categorize taxonomically, that categories of generalities in reference could be correlated with lexical expression. Thus in botanical taxonomies, it was observed that there are "levels" of organization that could be discovered by the lexemic status of the words occurring at various levels of generality n the taxonomic structure. Lexemic status in turn concerns whether the node labels are "primary" ("pool", "dog", "terrier") or secondary ("whirlpool", "wire-haired terrier") and whether a node at a particular level can actually go without a label. "Covert categories", as these unlabelled sets are called, were predicted to be able to occur only at certain points in the taxonomic structure.

In ethnosemantic methodology, covert categories are indicated when a great number of words with relatively specific referents seem to be defined by the informant as "a kind of" something very general. An analogous phenomenon occurred in the generation of taxonomic structures from machine readable dictionaries, in, of all places, botanical terminology. In the Merriam-Webster Modern Pocket Dictionary used in the Amsler work, specific plant varieties are often defined as "a plant related to the X's....",

as in "agave (Any of several spiny-leaved plants related to the amaryllis)", "hellebore (a poisonous plant related to the lilies...") (G&C Merriam, 1971). Along with the proliferation of very specific - to- very general links, clues such as the regularity of "related to" tip us off to covert categories, nodes on the taxonomy which have no overt label expressed by the direct ISA relation. Figures 1 and 2 respectively show a relatively messy tree generated by the ISA relation as manifest between an entry word and the head of its definition text, and the same tree with the imposed covert categories revealed by the expression "related to."

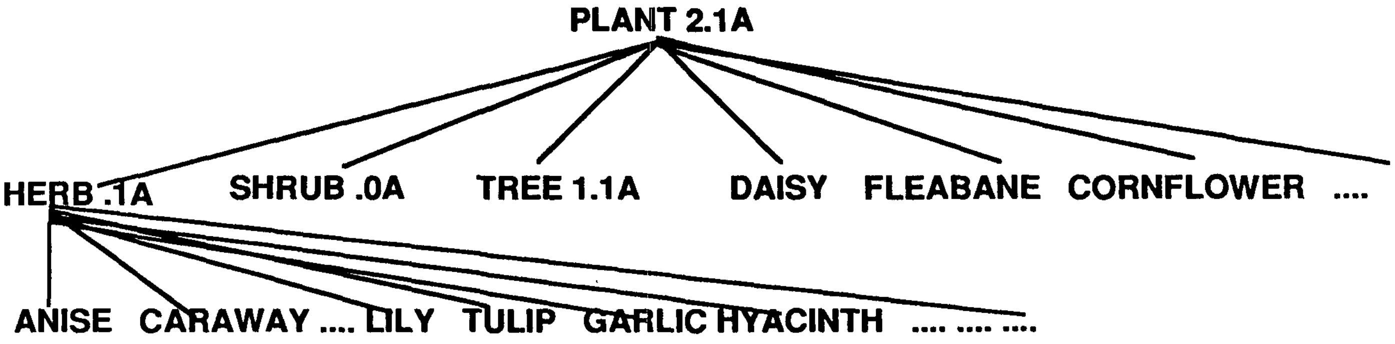

Figure 1. Partial "Plant" taxonomy derived from head terms alone

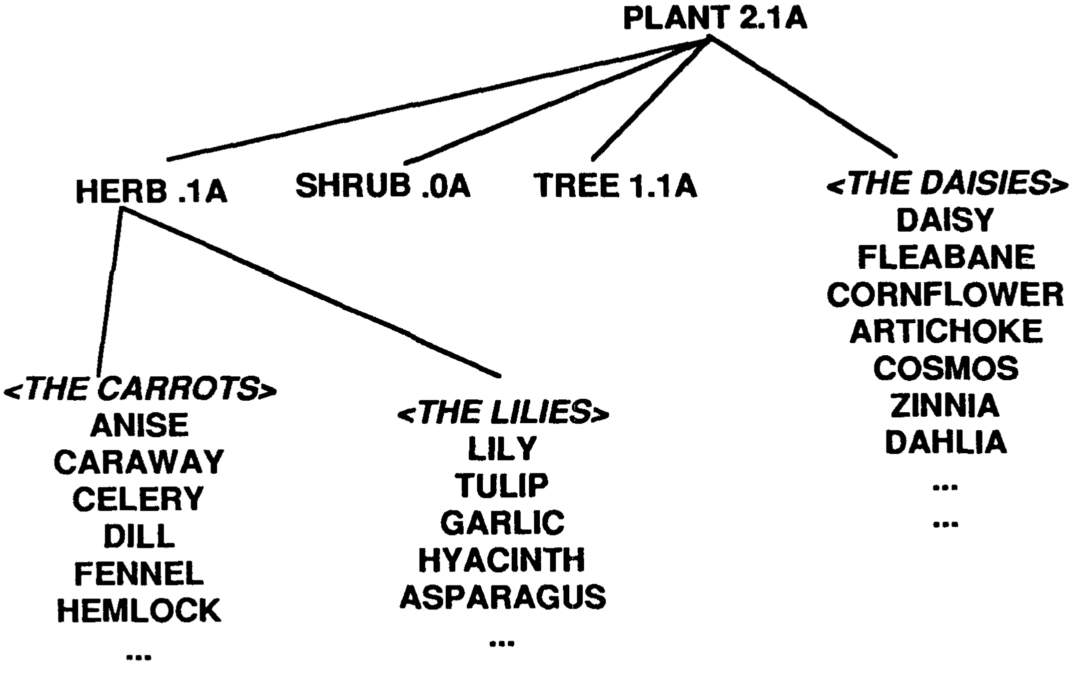

Figure 2. Taxonomy augmented with covert categories ("related to")

There was a continuing temptation to regard higher levels of the hierarchy as somehow transcendant of the language from which they were elicited, i.e., to have become knowledge representations whose nodes were labelled in an English-like meta-language. In part, this temptation arose out of the labels themselves, since such words as "cause", "thing", etc. were at the tree tops. At the time, I resisted because of the anthropological position already noted. However, I have since come to the healthier realization that the lexical structure can serve as domain knowledge for practical purposes. I address this further below.

1.2.2. Multi-lingual Lexical Structures. Later work with machine readable dictionaries used such lexical structures, not as knowledge representation models, but rather to transfer in levels of generality between two different languages, given two monolingual dictionaries and one bilingual which joined the two (White 1989). The means of transfer between one language and another (though not, of course, translation) was a matter of finding the fit for disambiguation on one side, (using relatively undifferentiated collocational cues from the definition texts) and then transferring languages in an essentially word-for-word algorithm.

The relations impose two dimensions: the ISA one between the entry term and the sense-definition head, and the relation (simply collocational) between the defining head and the other words in the definition. Obviously there are more things expressed in the definition text than just these two relations. But even this much avoids imposition of relations that are not actually in the dictionary. The collocational words have their own ISA links to other words, and their own (non-reflexive) collocational links as well. Thus it is possible to traverse lexical structures both hierarchically (ISA links) and laterally (collocation).

Heretofore, it had only been advisable to build taxonomic structures from general dictionaries, since every term in a definition is also defined, assuring some sort of closure. With the addition of the collocational dimension, however vaguely defined, it is possible to "graft" non-general dictionaries (technical, special purpose, bilingual) onto the existing general structures by matching collocations in to the non-generals to collocations in the specifics. The next step, for a Siemens translation databank experiment (both technical and multi-lingual) was to map such glossaries into dual general purpose dictionaries (one for each language). By this means generality could be traversed and collocations pursued to meet a broad range of application possibilities, including data extraction into one language from free-form text in another.

Figure 3 represents a grown link between a technical English glossary, a corresponding German one, and the general dictionaries in between from which the general links among collocates can be grown.

Figure 3. relational links across multi-lingual structures

The elaboration of elicited relations into a structure, even a two dimensional one of the explicit ISA and collocational associations, can be employed to disambiguate free text, and extract information from that text with minimal natural language processing. The MRD-

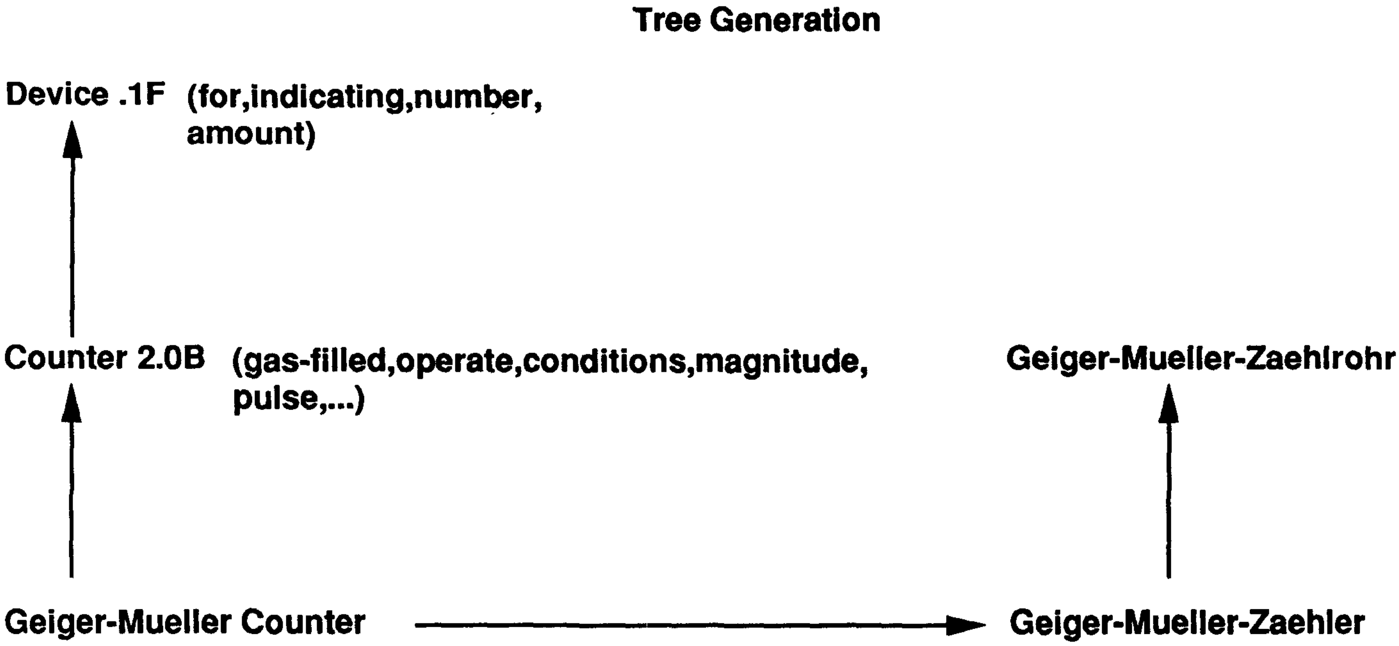
Tree Generation
Device .1F (for,indicating,number, amount)
Counter 2.0B (gas-filled,operate,conditions,magnitude, pulse,...)
Geiger-Mueller-Zaehlrohr
Geiger-Mueller Counter
Geiger-Mueller-Zaehler

Figure 3. Relational links across multi-lingual structures

derived lexical network can be used to look at free text and find a collocational fit between a datum point in the text and the other words around it, or the generic levels of the other words around it, or the collocations of the generic levels of the other words around it, or the collocation of the datum point's generic levels .

This ability to access taxonomic trees, collocates of tree members, and tree members of collocates, with no particular NLP in the ordinary senses of the term, should theoretically allow the disambiguation of a significant number, perhaps a majority, of the words the free text, and, from there, include the words which have been impossible to disambiguate into the set of known collocates.

Every term, then, may be seen as the intersection of values in three dimensions: the hierarchical (its ISA link to its immediate superordinate), the collocational (the list of words with which it occurred in definition texts), and the transfer (the corresponding entry term in the target language, itself an intersection of three dimensions). Using this structure, a collocation set can associate a term in a free text with a sense based upon text collocational matchings. An ISA path may then be exploited to conceptually generalize a term matching a lower node, be translated into the target language anywhere along that path, and proceed upward in generality in the target language.

Using this structure, a collocation set can associate a term in a free text with a sense based upon text collocational matchings. An ISA path may then be exploited to conceptually generalize a term matching a lower node, be translated into the target language along that path up to the most general ISA link in the TEAM portion, and proceed upward in generality in the target language. A general term in free text may use collocation links to direct a downward ISA path, translate along the path from the most general TEAM entry downward, and have available the ISA tree for generalization in the other language. The project design described here employs recent work in elicitation of machine- readable dictionary structures to perform multi-lingual processing. The lexical- semantic semantic structures, when optimized, form a base in which each term has a value on (at least) three dimensions. A matching term in a text, then, can be disambiguated by matching collocation, can be generalized by its taxonomic path, and can be manipulated in either of these ways in any of the languages for which it has a transfer.

This work is representative of the viewpoint of applied NLP, that is, that the consistency and the recoverability of methodological assumptions and actions is more important than the "psychological reality" that was suggested by the similarity of dictionary and folk taxonomies. The ISA relation is, or is very similar to, well attested lexical-semantic principles among speakers in many languages. The "simply collocational" relation is a convenience to avoid making wrong relational hypotheses.

The intent of this model was information extraction from messages in one language and representation in a template in another (without translating it). This effort ignored any idea that the higher levels of MRD-generated structures represented metalinguistic expressions of concepts rather than words themselves. In this case, I at best begged the question, since the language of the words at all levels of all the hierarchies was directly relevant to the functioning of the systems. So no claim was made about the organization of knowledge on dictionary taxonomies other than the words labeling the nodes had succeedingly more general references.

Again, this work is indicative of a direction in which applications can beg several questions about the purely lexical versus the purely conceptual representations. Traversing hierarchies generated by lexical processes as if they were cognitive categories evades the whole issue of the boundary that must lie between them. Further, the way that multilingual information is extracted above is completely artificial. Yet the evaluation metric will be how well the application does its job (precision, recall, efficiency), not the identification of the lexical - non-lexical boundary.

2. Natural Language Processing Systems. Application processes may be artificial, but the dictionary representation itself has valuable attributes derived from its status as a human linguistic artifact. As long as the organizing relations are derivable from the corpus itself (and not imposed from analyst interpretation), then there is an internal consistency which forms the basis of efficient knowledge representation built by recoverable principles. The most promising new NLP systems may be those which use some externally-derived lexical system to represent the lexical semantics of the system, and its conceptual structure as well. Notable among these is the work at New Mexico State in just the sort of dictionary work described above to the service of machine translation (Wilks and Slator 1987).

From the very practical perspective of doing applied NLP it has become repeatedly apparent that lexical structures turn out to be an excellent basis for knowledge representation, as long as the exact distinction between the two types of knowledge is not important for the intended application.

In the Siemens METAL machine translation system, an intentionally minimalist set of lexical semantic features coded for words could be varied for specialized sub-lexicons germane to the particular subject area of the translation task (Bennett 1988). In the Martin Marietta EQUAL database interface, general lexical semantic properties were segregated from knowledge about the database schema, but the two components were of the same data structure (frames), and linked themselves hierarchically during query generation (White 1988). In the PRC PAKTUS system (for which I can only claim affiliation without technical involvement), the description of word behavior is located in three areas in a PAKTUS system: the language-dependent lexical category networks, the language-independent, domain-independent conceptual frame structure, and an application-specific domain network. The underlying concept structure is not dependent upon any subject domain, yet, again, the parts are of the same data structure type and successful message understanding depends on the connection of those structures (Loatman and Post 1988). In each of these cases, especially in the latter two, the association of lexical-semantic information with complement-role information, and on through domain specific facts, results in a very general abstract representation rather like that of a well-articulated MT interlingua. While a combination of information from different components, the representational structure itself does not maintain the segregation of lexical and other knowledge distinctly, having thus uniform access to the facts required to do its job (query a database, populate one, or disseminate a message).

Now a great many systems to their credit do establish a component that expresses this singularity and domain information. And it can be done in a way that is dissimilar, at least conceptually, to the way lexical entries are coded for semantic properties. For example, I cite the script designs of long vintage, still exploited in numerous systems such as the CMU KBMT system (Carbonell 1988). However, this modularity does not distinctly segregate the two types of knowledge. There is nothing unconscionable about expressing knowledge of the world as values of features asserted in the lexicon, even when there is a way of expressing the same information in the script, if it happens to work better that way for the purpose of the NLP task at hand, and is more easily maintained from a software engineering position.

Having absolved the lapse in determining the true world vs. lexical nature of particular relevant bits of information, approaches which exploit the lapse generally don't grasp those problems like overlap of the worlds (domains here) in which synonymy occurs. Nor do they handle the subtleties of vagueness/homophony issues (*John plays the flute and Harry football) which turn up more than we expect in the operational use of machine translation. Similarly, the peculiar problems of database interface, where ambiguity between field name and field value is more worldly than lexical, yet not intuitively either, invite practical lexical solutions which are forgiven when effective and maintainable.

3. Conclusion. As we have seen, lexical and world knowledge must be distinguished, lest we lapse into an unenlightened determinism of human institutions. At the same time, though, interesting realizations emerge through the inability to discretely segregate the two. From the point of view of the way humans really do organize their universes, evidence from worldview can lend evidence to lexical organization, even if we do not know where the boundary is. From applied natural language processing, it is perhaps the case that NLP will perform better once the boundaries are known. But for now, lexical structures, when built from consistent , non-capricious principles, can serve as the bases both of lexical and conceptual components, serving the needs of the application as well as the needs of software engineering.

References

Amsler, R. 1980. *The Structure of the Merriam-Webster Pocket Dictionary.* Ph.D. Thesis, Austin, TX: University of Texas.

Amsler, R. and J. White, 1979. *Development of a Computational Methodology for Deriving Natural Language Semantic Structures via Analysis of Machine-Readable Dictionaries.* Technical Report TR MCS77-01315, Linguistics Research Center, University of Texas.

Bennett, W. 1988. "Methodological Considerations in the METAL Project." *Proceedings of the Second International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages.* Carnegie-Mellon University Center for Machine Translation.

Berlin, B., D. Breedlove and P. Raven. 1974. *Principles of Tzeltal Plant Classification.* New York: Academic Press.

Carbonell, J. and M. Tomita. 1985. "New Approaches to Machine Translation." *Proceedings of the First Conference on Theoretical and Methodological Issues inbMachine Translation of Natural Languages.* Colgate University.

Kay, P. and W. Kempton . 1984. "What is the Sapir-Whorf Hypotthesis?" *American Anthropologist* 86.1.

Loatman, B. and S. Post. 1988. "Natural Language Processing for Intelligence Message Analysis" *Signal Magazine.* September, 1988.

G.&C. Merriam. 1971. *The New Merriam-Webster Pocket Dictionary.* New York: Pocket Books.

White, J. 1978. "It's Impolite to Stare and Worse to Point: Linguistics and the Tojolabal Occult"*Proceedings of the 1977 Mid-America Linguistics Conference.* University of Missouri- Columbia.

White, J. 1988. "Advantages of Modularity in Natural Language Interface." *Proceedings of the Third Annual Rocky Mountain Conference on Artificial Intelligence.* Denver: June, 1988.

White, J. 1989. "Determination of Lexical-Semantic Relations for Multi-lingual Terminology Structures." In *Relational Models of the Lexicon,* ed. M. Evens. Cambridge University Press.

Wilks, Y. and B. Slator. 1987. "Towards Semantic Structure from Dictionary Entries." In *Proceedings of the Second Annual Rocky Mountain Conference on Artificial Intelligence*. Boulder, Colorado.