WORDS AND WORLDS

Robert A. Amsler
Artificial Intelligence and Information Science Research
Bell Communications Research
Morristown, New Jersey 07960

For several years now I have been concerned with how artificial intelligence is going to build the substitute for human world knowledge needed in performing the task of text understanding. I continue to believe that the bulk of this knowledge will have to be derived from existing machine-readable texts produced as a byproduct of computer typesetting and word processing technologies which have overtaken the publishing industries. However, there are many obstacles to the acquisition of world knowledge from text.

There are some, I am sure, who would argue that world knowledge of the form needed in text understanding will have to be hand-coded and cannot be derived from existing reference books or other texts. My basic argument against those who hold this view is that they are ignoring the magnitude of the task ahead. Whether measured in terms of bytes or man-years, the sum of recorded knowledge is so massive that it is unlikely to be capable of being recoded in anything less than man-centuries.

Put another way, there currently exist sizeable publishing empires in this country which every day employ hundreds of people involved directly in the coding of information for new reference texts and revised editions of older reference works. To attempt a recoding of world knowledge solely for use in AI would eventually become an attempt to parallel this effort. It would become a major industry in itself. Thus, it is more likely that, instead of a new knowledge-base industry, we will see an evolutionary change in the methods used by the existing publishing empires to record knowledge in a manner that is of use in producing text both for human consumption and as knowledge bases for computers. Researchers in AI and computational linguistics therefore have some responsibility to determine how the existing printed knowledge can evolve into usable computational world knowledge

Now, of course, I do admit there are subclasses of world knowledge that evidence to date has not shown to exist in print at all. Jerry Hobbs is attempting to codify one such subclass in his work on TACITUS (Hobbs et al. 1986). There are others as well, such as some forms of linguistic knowledge. However, I am concerned about the very large body of knowledge that we try to communicate to people through books, newspapers and other texts. This knowledge of the outside world, of experiences in which the individual has not and in fact may never personally be involved, is nevertheless shared knowedge known to all of us through reading and listening to the words of others.

Another assumption, and one that has been guiding my work for many years now, is that natural language systems cannot *understand* text for which they do not possess the lexicon. This seems so elemental an assumption that I find it hard to see how to ignore the fact that we do not have a lexicon of any real world text as common as a newspaper.

What is in this missing lexicon? The problem has several parts. First, it now seems clear that even unabridged dictionaries miss sizeable amounts of the lexicon needed to do lexical recognition in a newspaper such as *The New York Times.* Earlier results (Walker & Amsler 1986) have shown that some of this lexicon was excluded from the dictionaries by choice, such as the proper nouns, but more recent research has revealed that even here the problem is more complex.

Proper nouns are not quite lexical in nature. They possess a grammatical structure which some researchers have noted (Carroll 1985). This is to say that a typical proper noun has a variety of forms which tend to make the use of a single lexical entry for the proper noun less computationally useful than for a common noun.

Thus we recognize,
"International Business Machine Corporation's Thomas J. Watson Research Center at Yorktown Heights, New York,"
as the same thing as
"IBM Yorktown"
or
"IBM's Watson Research Center."
What is going on here is that we have a mini-grammar for these type of utterances which allows us to contract their separate parts independently (and even to transform the order of the constituents). Thus, "International Business Machines Corporation" is contractable according to the rules for corporations, namely to forms such as "IBM Corp." or just its initials, "IBM" (but it cannot be "International B. Machines," for instance). "Thomas J. Watson" is a person's name, and already has a contracted middle initial. People occasionally can have their names contracted the same way as corporations, (e.g. "JFK" or "J.R." of Dallas fame) but more typically they contract to forms such as "T.J. Watson" and "Watson." Geographic locations, such as "Yorktown Heights, New York" can contract to forms such as "Yorktown Heights, NY" and "Yorktown." "Research Center" is a common noun, and as such is lexical, not undergoing this type of contraction and grammatical restructuring.

Finally, one should note that the order of the proper noun constituents in the original full expression was,
<Corporation-owner> <Person-Name> <common-noun> <geographic-location>.
However, "IBM Yorktown Heights Research Center" has rearranged this ordering. This capability for rearrangement is clearly grammatical in nature.

This and many other examples show that (a) most proper nouns have several forms derivable from their most complete representation, and (b) these forms obey a simple grammar of permissible contractions and transformations dependent upon the types of proper nouns involved as constituents of the entire proper noun expression.

Another important aspect of this observation is that it was made as a direct consequence of massive data collection. If one encountered a single form of a proper noun in text, one might be tempted to believe it could be treated by including it in a dictionary just as a common noun. But, examining a very large corpus of text in just the right ways (such as with a *proper-noun extraction program* and a concordance of its output) shows the proper nouns to stand out as quite distinct from common nouns. There are almost always a dozen different forms for each proper noun, scattered alphabetically according to the initial word of each form. Yet these multiple forms show a pattern of recurrence based upon standard contraction, abbreviation and transformation operations.

Consideration of a *proper noun extraction program* points out how important it is to use textual sources in the right way. We now know that counting the instances of isolated words in text is a horrendous misuse of the raw data. To encounter "New York City" and decide that "New" has made an appearance in the text is unacceptable. Lexical *events* often consist of phrases which bear little more than a historic relationship to the individual blank-delimited words of which they are composed. In what sense is "New York City" *new*? What does "soap" have to do with "soap operas"? These are historic artifacts and much the way chemical compounds may bear little relationship to the properties of their elemental constituents, observing the spectrum of elements in phrases in isolation doesn't reliably reveal the whole story about the phrases themselves. The problem of detecting such phrases in text and deciding whether they legitimately need their own lexical entries makes clear a distinction between three different degrees of specificity of information and their intended uses.

The first degree of minimal specificity is that contained in published dictionaries which present information for human readers who can be assumed to have a rather complete grasp of world knowledge. Dictionaries offer definitions that are the minimal specification of the meaning of a word capable of evoking its conceptual meaning in the mind of a reader. Dictionaries are so myopic in this regard that they are often inappropriately used to try to teach children the meanings of words, ignoring whether the children possess the accompanying world knowledge needed to understand the dictionary definitions. George Miller (1985) at Princeton has revealed just how little of what a conventional dictionary says can be understood by a child. Dictionaries also tend to split compounds into their constituent isolated words without concern for how the reader will manage to put the right senses of the words back together again. However, this should not be taken as a complete repudiation of dictionaries. They are excellect indexes into the world knowledge needed; they just make no committment themselves to supply that world knowledge.

The second degree of intermediate specificity is that needed by computational linguistics to build lexicons to be used by programs for parsing, generation, and translation. Computational linguists are required to provide in their lexicons everything necessary to substantiate their program's linguistic competence. If compounds are described by separate entries for each of their components, then rules for the combination of these components must also be included. More likely, the compounds themselves will be given their own entries, since being completely sure a rule is correct requires a great deal more knowledge of the lexicon than is available today. However, parsing, generation, and translation do not necessarily require their programs to construct conceptual structures for the lexical objects they process. One can build parsers, such as the Fiddich Parser, which blithly guess at syntactic categories for words they do not have in their lexicons--and do so so successfully that they complete most parses with acceptable grammatical structures. However, it is clear such a level of *understanding* is not adequate for more advanced artificial intelligence applications.

The third degree of highest specificity is needed to support artificial intelligence where one must be able not only to parse text, but to understand the meaning of the concepts to which the text refers. Understanding text may require other aspects of knowledge such as visual imagery or knowledge of physical laws, but above all it requires the ability to match incoming lexical entities with stored knowledge about the concepts of which the lexical entities are descriptions. This means that one needs to go significantly beyond linguistic competence.

These three levels of representation directly affect what needs to be stored in a lexicon, and nowhere more than in the nature of what needs to be stored about phrases. For example, whereas the lexicographer can dismiss 'elephant house' as the ordinary sense of 'elephant' and a sense of 'house' which means "a habitation for animals," the computational linguist needs to distinguish how we know that 'cat house' and 'dog house' are not instances of this rule, and the knowledge-base researcher needs to distinguish 'elephant house' as representing a real world building which appears in zoos, whereas 'kiwi house' has no such referrent or significance. Thus, lexicographers might defend not having an entry for 'elephant house' in their lexicons, covering its meaning with a special sense of 'house' suitable for this purpose. However, computational linguists would be very critical of the failure to accompany that special meaning with a caveat excluding forms such as 'cat house' and 'dog house' and perhaps uneasy about the fact that 'kiwi house' would pass through the parser. The knowledge-base researcher would find the computational linguist's possible problem an absolute obstacle and have to have an explicit entry for 'elephant house' that noted its location in a zoo and other details such as that elephants do not happen to own mortgages on their houses the way people do, just to start off.

What I am implying here is that whereas printed dictionaries have served us well for a few hundred years, it is very likely that we will have to greatly expand their explicitness for computational linguistic needs and even further expand the recorded information for knowledge-base needs. To do this we will have to return to the source materials from which the dictionaries were written, to the text that carries much of our world knowledge. We will have to extract the compound lexical items from these texts and make new decisions about the need to include them in new dictionaries which will serve the needs of AI programs. It is time both to increase the rigor of the lexicographic decisions about including multi-word entries in printed dictionaries, so they will be more usable by computational linguists, and to describe new tests of the adequacy of entries for more advanced knowledge representation disciplines.

REFERENCES

John M. Carroll (1985) *What's in a Name?: An Essay in the Psychology of Reference.* W. H. Freeman and Company, New York.

Jerry R. Hobbs, William Croft, Todd Davies, Douglas Edwards, and Kenneth Laws (1986) Commonsense Metaphysics and Lexical Semantics. *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, 10-13 June, 1986, Columbia University. ACL, Morristown, New Jersey, pages 231-240.

Donald E. Walker and Robert A. Amsler (1986) The Use of Machine-Readable Dictionaries in Sublanguage Analysis. *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*, edited by Ralph Grishman and Richard Kittridge. Lawrence Erlbaum Associates, Hillsdale, New Jersey, pages 69-83.