REPRESENTING KNOWLEDGE ABOUT WORDS

J. TERRY NUTTER

TR 89-22

REPRESENTING KNOWLEDGE ABOUT WORDS

J. TERRY NUTTER
Department of Computer Science
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061 U.S.A.
nutter@vtopus.cs.vt.edu

ABSTRACT

Most on-line lexicons contain only syntactic information. Semantic information is usually stored elsewhere, in a form inconsistent with the representation of the syntactic information. This paper reports on research toward developing a large on-line lexicon from machine-readable dictionaries, which contains both syntactic and semantic information in a uniform style. The fundamental theory is that of the relational lexicon; we describe relational lexicons, discuss our extensions to the usual theory of relational lexicons, rehearse very quickly some of the relations we are dealing with, and show how information for some simple entries is stored.

1.0 INTRODUCTION

Lexical relations provide a formal mechanism for expressing relations among concepts. Traditionally, lexical relations have been approached as a means for representing semantic information about words. A wide study of such lexical semantic relations was launched in the U.S.S.R. in connection with the development of the *Explanatory Combinary Dictionary* [Apresyan et al. 1969]. Lexical relations became part of each entry in the unilingual Russion dictionary, and play a key role in Mel'cuk's "meaning <=> text" theory [Mel'cuk 1973; Mel'cuk 1988].. Mel´cuk has recently led a similar dictionary effort in Canada for French, carried out through the intellectual efforts of linguists.

.Beginning from the work of Mel'cuk and others, Evens has studied lexical semantic relations extensively, beginning began by considering what knowledge would be needed in a lexicon to support question-answering [Evens and Smith 1979]. The resulting class of relations, along with those of the ECD, motivated an empirical study of the usefulness of lexical-semantic relationships in information retrieval (IR) [Fox 1980]. Evens *et al.* used the same relations as in [Fox 1980] in similar experiments with a different test collection [Evens et al. 1983]. Both sets of studies used hand-generated data for words and relations, with a relatively small data set. In recent combined work, the IIT group and an enlarged Virginia Tech group have been working on building a large relational lexicon from machine-readable dictionaries for use with information retrieval systems [Evens et al. 1985], [Fox et al. 1988], [Nutter et al. 1989] using the *Collins Dictionary of the English Language* and *Webster's Seventh Collegiate Dictionary*. We plan to extend these results to work on the *Oxford Dictionary of Current Idiomatic English* and *Oxford Advanced Dictionary of Current English*. The overall project has essentially four subtasks: determining the relations to be included, finding representations for the information to be extracted, extracting the relations from the dictionaries, and identifying effective uses of the lexicon in IR applications. While we target IR

as the first domain of application for our lexicon, we intend it as a multi-purpose tool, to support also natural language understanding and generation. This paper reports on the first two subtasks of the larger project: refining our theory of lexical relations, and using it to represent lexical information.

A fundamental concept behind relational lexicons is that as much as possible of the information in the lexicon is represented directly in terms of lexical relations among words (or word senses). Classical work in the literature concentrates on semantic relations. A relational lexicon which contains *only* semantic information, however, would essentially reproduce the current bifurcated arrangement for on-line lexicons, with syntactic and semantic information segregated from one another. We believe that this is inefficient, artificial, inadequate, and error prone. The obvious schemes for separating access store all strings twice: once in the "semantic lexicon" and once in the "syntactic lexicon". While this could be avoided by a single string entry with separate access paths for semantic and syntactic information, it will nonetheless require some duplication. Since we are talking about a lexicon which holds several dictionaries' worth of information, that overhead starts to have significant costs attached. Separation is artificial, because the line between syntactic and semantic information is notoriously ill-defined. Mel'cuk's original relations included several, such as the imperfective relation, which reflect syntactic facts with semantic repercussions. Locating the line is hard enough. Maintaining it in a principled fashion within a bifurcated system in difficult at best. The separation also strands numerous facts which require both semantic and syntactic information for their representation.

Hence we are not simply developing a relational syntactic lexicon, to use in conjunction with a separate, syntactic lexicon. Instead we are building a single, unified lexicon to contain both kinds of information, with representations which allow facts to have both semantic and syntactic components. From a representational standpoint, there is no essential difference between syntactic and semantic information. The mechanisms that work on either work on both.

This report is structured as follows. Section two discusses our contributions to the theory of lexical relations. Section three describes the knowledge representation scheme which we have adopted for implementing our lexicon. Section four describes progress to date. Section five presents conclusions and new directions.

2. Theory of the Relational Lexicon

The theory of lexical relations was first developed by lexicographers to help construct what they viewed as a completely new kind of dictionary (see [Apresyan et al. 1969]). They looked for their relations not in existing dictionaries, but in the language itself (in this case, Russian), to let create dictionary entries which they hoped would be clearer, more precise, more complete, and more perspicuous than those in other dictionaries. Critically, they intended the definitions in their dictionary to reflect all the lexical information needed to form a text, laying out the various stages of text generation as an integral sequence of steps. Their effort should thus be viewed not as an investigation of what has happened to land in various kinds of dictionaries up to a given time, but as a linguistic effort to define the semantic structures governing word use. The result is a theory of meaning, which derives its motivation and to some extent validation through work with dictionaries, but which by its nature belongs to theoretical linguistics. [Evens and Smith 1979] describes several variations on this theory.

In this regard, our work in identifying lexical relations from the texts of dictionaries which were not formulated with that theory in mind provides a measure of confirmation for the theory. (For details on the process of locating relations in dictionaries, see [Ahlswede and Evens 1988] and [Fox et al. 1988].) One form of evidence that lexical relations reflect real phenomena is to find them regularly and identifiably "in the wild" (i.e. in natural language texts). Because lexical relations constitute information about language, we will most readily find them in texts that deliberately talk about language. Dictionaries are a particularly rich source of texts about words, albeit a less than fully natural one. We feel, therefore, that the relations we find both in the theory and very strongly indicated in dictionary texts can be viewed as validated in terms of reflecting cognitively real relations among the concepts which the words in question represent. Further validation can be found in the anthropological field work of Casagrande and Hale [Casagrande and Hale 1967] studying dialect variation in Papago and Pima. Our work has identified over a hundred such relations [Nutter 1989].

One of the primary results of this work is that the lexical relations we work with are themselves related, and form a rich hierarchy, which has not previously been developed. The work of Mel'cuk's group [Apresyan et al. 1969] produced an unstructured list of relations. While relationships among the relations are often hinted at and occasionally made explicit (especially among what we call situation-verb relations), there is little to no suggestion of an overall structure within which the relations sit. In the work of Evens and Smith suggested nine categories of relations, which they viewed as internally unstructured, their members sharing only some commonality [Evens and Smith 1979]. Our results have shown a far richer structure than has hitherto been suspected, and which we feel represents a substantial contribution to the theory of lexical relations.

The hierarchy in its current form consists of over a hundred relations, classified at the top level as essentially semantic relations, morphological and syntactic relations, and factive relations, with a catch-all category for a very small number of relations that we don't know how to classify. The hierarchy goes beyond a simple partition of relations; in many areas, the tree depth is around five. By contrast with the nine categories in Evens and Smith, we have over twenty substantial categories at the next-to-leaf level. We have found enough grouping of relations into natural families to allow for advantages in representing the relations hierarchically, and enough difference at the leaves not to want to collapse it. This hierarchy allows sophisticated representation of relationships among words that may not be immediately evident in the data from which relations are extracted. A pared down outline of part of the hierarchy is presented in the appendix.

In addition to the distinctions among lexical relations reflected by the hierarchy, there is a difference between those which routinely appear among terms from any domain, such as taxonomic relations, and those which are specific to a particular domain (such as specialized relations among substances in medical terminology. Our research indicates that no set of lexical relations can be considered complete, because most domains contain specialized relations of their own. This is hinted at in [Evens and Smith 1979] especially in the relations which occur very naturally in children's stories (e.g., animal to its characteristic sound) but only rarely in other text. As a further example, a medical dictionary could be expected to reflect such relations as "symptom of", "counteragent to", and the like (for more information on lexical relations in a medical domain, see [Ahlswede and Evens 1984]. Ultimately, a realistic lexicon for information retrieval may need many such relations, and may also want to know the domain(s) in

which the relation applies. We have not added classes in our hierarchy for domain specific relations; we suspect that in a complete theory, such classes should be recognized, and that the resulting hierarchy will be to some extent tangled.

Comparing Mel´chuk's pioneering work on lexical relations arising from studies of Russian with work based on English and other languages suggests a further distinction, in that it reveals a large class of language independent lexical relations, and a much smaller class of language-dependent relations. An example of the latter is Mel´chuk's Perfective, which arises naturally in Russian because a distinction most Indoeuropean languages make by inflection is made in Russian by using a different verb. So in Russian this relation often links words with different roots, and gives significant information about them. In English, on the other hand, examples which are not essentially instances of regular inflection rules are virtually nonexistent.

What is interesting is not that there are language dependent relations, but that there are so few. Mel´chuk's work can be transferred to representations of English meanings with very little substantive change. This suggests that the use of Lexical Relations for representing meanings goes beyond language barriers, and incidentally raises the question of its potential for applications in machine translation.

We therefore make absolutely no claim to comprehensiveness. Rather we believe that we have isolated a strong central set of relations which jointly cover many, though not all, of the relations among terms in general use. But just as we have not tried for comprehensiveness, neither have we enforced exclusivity. Our full table of relations [Nutter 1989] includes most of the relations recognized over a large subset of the literature, including some relations such as the perfective relation, whose *raison d'être*, as remarked above, seems to be a fact about Russian grammar. In other words, while our relational hierarchy does not include all possible lexical relations, neither does it exclude any discussed in the sources with which we worked and for which we could find (or other authors had found) reasonable validating evidence.

3. KNOWLEDGE REPRESENTATION

The implemented lexicon actually contains two different kinds of information: information about words and word senses (in terms primarily of lexical relations which hold between them) and information about the lexical relation hierarchy. Both kinds of information are represented together in the same semantic network. The paradigm for the network is Shapiro's SNePS [Shapiro 1979], [Shapiro and Rapaport 1987], although for the full network we will not be using the SNePS software (see section 3.2 below). There are two major issues concerning our represention. First, what are the frames for representing both kinds of knowledge, and how are the representations interrelated? Second, how do we deal with problems of scale? We take these up one at a time.

3.1 Network representations

In SNePS, nodes represent all concepts about which the system has explicit knowledge. Arcs represent unconscious structuring information about the node concepts. For the lexicon's purposes, there are essentially two classes of nodes: those representing individuals, and those representing propositions about individuals. The individuals themselves can be of any of several types: they may represent strings (spellings), head words, senses, or non-linguistic conceptual constants. Because the system has knowledge of relations (and also because at least some of them are ternary

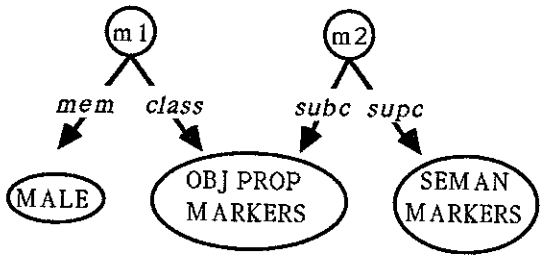

Figure 1. Representation for Hierarchical Facts. "Male-of is an object property marker, which is a kind of semantic marker"

or more), relations are represented by (individual) nodes, not arcs.

Information about the lexical relations is primarily hierarchical, and is represented most commonly by *member-class* frames (see Figure 1). In all figures, arc names are italicized, lexical relations are in all caps, word sense nodes are labeled in bold face, and proposition nodes are labeled lower case plain font. Figure 2 shows a subnetwork

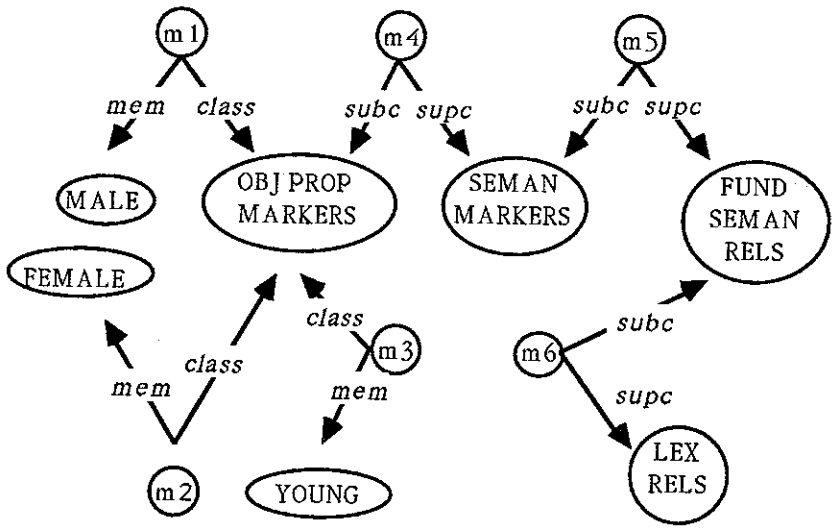

Figure 2. Subnetwork of the Lexical Relations Hierarchy. "Male-of, Female-of, and Young-of are all Object Property Markers, which are a kind of Semantic Markers, which are a subclass of Fundamentally Semantic Relations, which are a kind of Lexical Relations.

representing a small fragment of the relational hierarchy. The information about lexical relations is entered by hand, to initialize the network.

Lexical relations among words or word senses are represented by *arg1-arg2-rel* frames for non-symmetric relations such as taxonomy (see Figure 3) or by *argument-argument-rel* frames for symmetrical ones such as synonymy. Notice that inverse relations follow automatically: the same network in Figure 3 which says that the taxonomic superordinate of merino sheep, i.e., merino is a kind of sheep (*arg1* arc to *merino*, *arg2* arc to *sheep*, and *rel* arc to *taxon*), also represents the fact that a

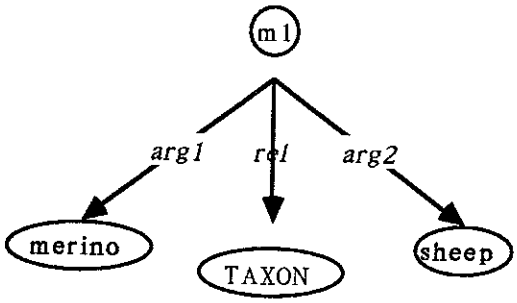

Figure 3. A Lexical Relation Instance.
"Merino" is taxonomically subordinate to "sheep".

taxonomic subordinate for sheep is merino. Since all access is by pattern matching, there is no need to distinguish in the network between relations and their inverses to be able to work equally comfortably with both.

In addition to what are usually thought of as lexical relations, dictionaries reveal a hierarchy of lexical information in the form of a sense/subsense hierarchy. This information is retained and represented using *sense-subsense* frames. Figure 4

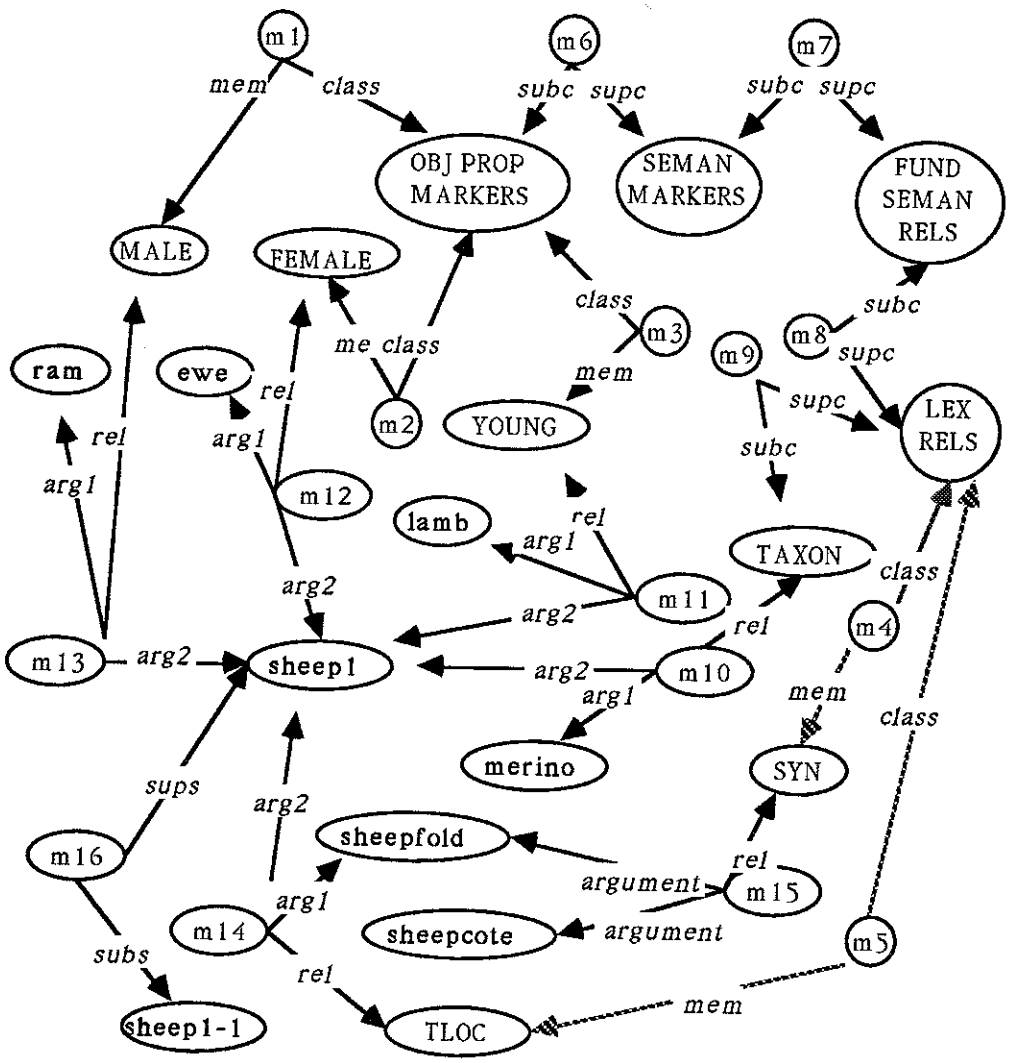

Figure 4: Subnetwork around "Sheep"

shows a subnetwork which combines hierarchical information about lexical relations, instances of lexical relations among words, and sense/subsense information on head words, using particular lexical relations derived from the definitions of "sheep" and related words in *Webster's Seventh*. In particular, proposition nodes m1 through m9 give the hierarchical information relating the various lexical relations in the example. (The dashed arcs from m4 and m5 indicate collapsed hierarchy, to simplify the figure.) Node m10 says that merinos are a kind of sheep. Nodes m11 through m13 represent that lambs are the young of sheep, rams are the male of sheep, and ewes are the female of sheep. Sheep normally live in a sheepfold (m14), which is the same thing as a sheepcote (m15). Finally, m16 reflects that there is a more specific sense of sheep (which covers only a single species). Relations are typically among senses, which may lie anywhere in the sense/subsense hierarchy.

It should be clear that the representation is uniform in the sense that all kinds of information can be accessed in the same ways. Another important fact is that the representation assumes no essential difference between syntactic and semantic information. We commented in section one that separating syntactic and semantic components of the lexicon strands some facts which require components of both kinds. An example is morphological variant relations. Morphological relations are typically treated as purely syntactic; surely this is Mel'cuk's approach [Apresyan et al. 1969]. There is a strong appeal to this view. Morphology is largely syntactic. But not entirely. For instance, "neatness" is morphologically related to "neat" (an apparently syntactic relation), *but* is only an acceptable variant of "neat" in the sense of "orderly", and not in the sense of "undiluted" (as in "a neat drink"). This kind of fact, which will matter greatly to generation systems, requires both syntactic and semantic components even to express. In our representation, there is no difficulty. Morphological relations can take three arguments: a head word representing the root, another representing the variant, and a third representing the sense(s) for which the variant is legal. This is one instance of automatic gain from uniformity.

3.2 Problems of scale: LEND

We are dealing with several dictionaries. We need base nodes for all the senses we extract, as well of course as all the relations in the lexical relation hierarchy. We need proposition nodes for all the subset-superset propositions in the lexical relation hierarchy, for all the sense-subsense relationships, and for every lexical relation we extract. We need two to three arcs per proposition node. In the long run, we intend to work with five dictionaries. A crude estimate of the number of arcs we will ultimately want to implement comes to about 2^{30}. The problem should be evident: if we work with software implemented on traditional AI languages, their hash tables won't let us have that many objects, let alone manipulate them decently. We are already integrating a prototype form of our lexicon into an IR system, CODER [Fox and France 1987; Fox 1987]. If we are to operate in an interactive environment with acceptable performance using a highly connected network of the size we already have (let alone the size we envision), we cannot rely on the Unix paging algorithm, for instance, to manage memory, especially since we have a few nodes that are connected to many, many others (consider the node for *synonym*, for instance).

These considerations have led to the design and implementation of the Large Extended Network Database (LEND) system [France et al 1989]. LEND acts as a backend, accepting a subset of the SNePS user language, with enhanced path-based inference operations, and ultimately with support for hypertext operations as well. We are not reimplementing the full SNePS functionality. In particular, SNePS supports a node-

based predicate logic style inference mechanism, which we do not anticipate reproducing, both because our present purposes do not indicate that we need it (given the enhancements to path-based inference we have specified) and because the complexity makes the problem much more difficult. The LEND system uses object-oriented programming principles to maintain node and arc managers for the various classes of nodes and arcs in the system. Within any manager, objects are accessed using perfect hashing functions [Fox et al. 1989].

4. RESULTS

While much remains in both *Webster's Seventh* and the *CED* that we have not dealt with, we have parsed definitions for about 92,000 headwords from both, leading to approximately 1,800,000 lexical relation triples (word REL word). In terms of LEND, our specification and design stages are essentially complete, and we have begun implementation. In particular, we have completed basic node and arc managers for LEND, and have loaded node bases of approximately 300,000 word senses extracted from some 150,000 lemmas representing 81,000 roots and 35,000 variants (some of these are runons in definitions, as opposed to headwords). We have loaded a relatively small number of lexical relation types, accounting among them for some 1,225,000 relation instances. In addition, one manager holds a text database of some 2,000 AIList documents, which are cross-indexed with the word base by "occurs-in-document" relations.

What we cannot yet do is get all these relations in the form we would like. That is, the relation *synonym* ought to link a word sense with a word sense. At the moment, in most instances, it links a word sense with a string. Doing better than this requires moderately sophisticated disambiguation in parsing the definition texts, certainly going beyond what we can currently do. In many cases we also must discard information in the definition that is definitely important to the meaning; for examples, see [Ahlswede and Evens 1988]. These are problems with the extraction process. So far, we have not found any essential limitations in the representation scheme.

To date, we have not completed implementation of the path-based inference methods, but we can retreive simple relations from either argument, and can follow paths by brute-force methods. LEND is implemented in the CODER distributed environment, which includes modules running on a Sequent Symmetry, two Microvax IIs, and several Mac IIs. The LEND-specific modules are primarily located on the Sequent, and run in C++. Our testbed systems, which cannot support the full base but which allow us to examine consequences of our representation, are SNePS-79 on a Microvax II in FranzLISP, and SNePS-2 in CommonLISP on a TI Explorer. (Code for the SNePS testbed systems was furnished by Stuart Shapiro from SUNY at Buffalo.)

5. CONCLUSIONS

Current on-line lexicons have three main problems. First, they separate syntactic from semantic information in ways that are artificial and inadequate to the representation of some mixed and borderline information. Second, they represent semantic information thinly if at all. Third, they are unrealistically small. We have developed a representation for information about words which suffers none of these problems. It is flexible: we are already using it for research in IR, and projects newly underway are using a scaled-down version in natural language understanding and generation [Cline and Nutter 1989]. While relational lexicons have been discussed for some time in the linguistics and lexicographic literature, this is the first on-line relational lexicon. It

thus represents advances in computational lexicography, in addition to the advances in the theory of lexical relations which resulted from its development. With this kind of change, AI systems begin to get not merely word lists with rudimentary part-of-speech information, but genuine knowledge about words, to be exploited in any context which really understanding language can help.

REFERENCES

Ahlswede and Evens 1984

Ahlswede, T. E. and M. W. Evens. A Lexicon for a Medical Expert System. *Proceedings of the Workshop on Relational Models of the Lexicon*, Martha Evens (ed.), Stanford, CA, July 1984 (to appear).

Ahlswede and Evens 1988

Ahlswede, T. E. and M. W. Evens. Parsing vs. Text Processing in the Analysis of Dictionary Definitions, *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*, Buffalo, June 1988, 217-224.

Apresyan et al. 1969

Apresyan, Y. D., I. A. Mel'cuk, and A. K. Zolkovsky. Semantics and Lexicography: Towards a New Type of Unilingual Dictionary. In *Studies in Syntax and Semantics*, ed. F. Kiefer, 1-33. Dordrecht — Holland: D. Reidel, 1969.

Casagrande and Hale 1967

Casagrande, J. B. and K. L. Hale. 1967. Semantic Relations in Papago Folk Definitions. In Hymes and Bittle, 1967, 165-196.

Cline and Nutter 1989

Cline, B. E. and J. T. Nutter. Implications of Natural Categories for Natural Language Generation. Department of Computer Science Technical Report TR 89-7, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0106, 1989.

Evens and Smith 1979

Evens, M.W. and R.N. Smith. A Lexicon for a Computer Question-Answering System. *American ournal of Computational Linguistics*, Microfiche 83: 1-93, 1979.

Evens et al. 1983

Evens, M. W., J. Vandendorpe, and Y.-C. Wang. Lexical-Semantic Relations in Information Retrieval. In S. Williams, ed., *Humans and Machines: The Interface through Language*, Ablex (Norwood, N.J.) 1983, 73-100.

Evens et al. 1985

Evens, M. W. , J. Vandendorpe, and Y.-C. Wang. Lexical Semantic Relations in Information Retrieval. In *Humans and Machines: The Interface Through Language*, S. Williams, ed., Ablex 1985, 73-100.

Fox 1980

Fox, E. A. Lexical Relations: Enhancing Effectivenes of Information Retrieval. *ACM SIGIR Forum* **15**:3 (Winter 1980) 6-35.

Fox 1987
Fox, E. A. Development of the CODER System: A Testbed for Artificial Intelligence Methods in Information Retrieval, *Information Processing and Management* **23**:4 (1987) 341-366.

Fox and France 1987
Fox, E. A. and R. K. France. Architecture of an Expert System for Composite Document Analysis, Representation, and Retrieval, *International Journal of Approximate Reasoning* **1** (1987) 151-175.

Fox et al. 1988
Fox, E. A., J. T. Nutter, T. E. Ahlswede, M. W. Evens, and J. A. Markowitz. "Building a Large Thesaurus for Information Retrieval", *Proc. Second Conf. of Applied Natural Lang. Proc.*, Austin, February 1988, 101-108.

Fox et al. 1989
Fox, E. A. , L. S. Heath and Q. F. Chen. An O(n log n) Algorithm for Finding Minimal Perfect Hash Functions. Department of Computer Science Technical Report 89-10, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0106, 1989.

France et al 1989
France, R. K., Q. F. Chen, and J. T. Nutter. LEND Design Specifications. Unpublished draft (available from the authors) 1989.

Mel'cuk 1973
Mel'cuk, I. A. Towards a Linguistic 'Meaning <=> Text' Model. In *Trends in Soviet Theoretical Linguistics*, ed. F. Kiefer, D. Reidel (Dordrecht), 1973, 33-57.

Mel'cuk 1988
Mel'cuk, I. A. *Dependency Syntaxc: Theory and Practice*. State University of New York Press (Albany) 1988.

Nutter 1989
Nutter, J. T. A Lexical Relation Hierarchy. Department of Computer Science Technical Report TR 89-6, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061.

Nutter et al. 1989
Nutter, J.T., E.A. Fox, and M.W. Evens. Building a lexicon from machine-readable dictionaries for improved information retrieval. *The Dynamic Text: Proc. ALLC/ACH 1989* (to appear).

Shapiro 1979
Shapiro, S. C. The SNePS Semantic Network Processing System. In *Associative Networks*, N. Findler, ed., Academic Press (New York) 1979, 179-203.

Shapiro and Rapaport 1987
Shapiro, S. C. and W. J. Rapaport. SNePS Considered as a Fully Intensional Propositional Semantic Network. In *The Knowledge Frontier*, N. Cercone and G. McCalla, eds., Springer-Verlag (New York) 1987, 263-315.