For the Lexicon That Has Everything

Martha Evens, Joanne Dardaine, Yu-Fen Huang, Sun M. Li,
Judith Markowitz, Frank Rinaldo, Margi Rinaldo,
Robert Strutz

Computer Science Department
Illinois Institute of Technology
Chicago, IL 60616
csevens@harpo.iit.edu or mwe@schur.math.nwu.edu

Abstract

This paper argues that it is impossible to separate lexical and encyclopedic knowledge and describes an attempt to build a large lexical database that contains the range of information needed to make a parser or a text generation system interpret and use words and phrases correctly.

1 Introduction

Any natural language processing system needs both knowledge about words and knowledge about the world. Many natural language systems divide these two kinds of knowledge into two knowledge bases, which we call the lexicon and the encyclopedia for the purposes of this discussion. We argue that the distinction between the lexicon and the encyclopedia is difficult to maintain both in theory and in practice. We describe the design and development of a large lexical database intended to support parsing, generation, and information retrieval applications. We claim that these applications require information of many different kinds, some of which is traditionally stored in a dictionary, some in a thesaurus, and some in an encyclopedia.

We need to support not just alphabetic access to this information but access through semantic links. Bierman [1964] was one of the first to describe lexical-semantic links between words. They define the basic organization of semantic information, he claims. He paints an image of a very large single-page dictionary with language-specific nodes connected by semantic relations.

Can we distinguish between the lexicon and the encyclopedia in this context? Bierwisch and Kiefer [1970] assume that both kinds of information are contained in the same lexical entry. The distinction between linguistic or lexical and encyclopedic knowledge, they say, corresponds to the difference between the core and the periphery of a lexical entry, where:

> The *core* of a lexical reading comprises all and only those semantic specifications that determine, roughly speaking, its place within the system of dictionary entries, i.e., delimit it from other (non-synonymous) entries. The *periphery* consists of those semantic specifications which could be removed from its reading without changing its relation to other lexical readings within the same grammar. [Bierwisch and Kiefer 1970, 69-70]

The major difficulty with this criterion is its instability. As new entries are added to the system, information sufficient to distinguish one entry from another may have to be

shifted from the periphery to the core – and thus from the encyclopedia to the lexicon. For instance, suppose a new entry, "leopard – a large wild cat" is to be added. The entire lexicon must be searched for entries that mention large wild cats. If one is found, say "lion – a large wild cat," then enough information must be added to both definitions to differentiate *leopard* and *lion* from each other.

Apresyan, Mel'čuk, and Žolkovsky run into the same difficulty of distinguishing lexical and encyclopedic information in attempting to define the lexical universe of a word C0.

> The main themes dealt with under the heading 'lexical universe' are: 1) the types of C0; 2) the main part or phases of C0; 3) typical situations occurring before and after C0, etc. Thus, the section lexical universe for the word ***skis*** consists of a list of the types of skis (***racing***, ***mountain***, ***jumping***, ***hunting***), their main parts (***skis*** proper and ***bindings***), the main objects and actions necessary for the correct use (exploitation) of skis (***poles***, ***grease***, ***to wax***), the main types of activities connected with skis (***a ski-trip***, ***a ski-race*** ...) ... the sections contain only such words as are necessary for talking on the topic, and nothing else. [Apresyan *et al.* 1970]

The problem is that "what is needed for talking about the topic" depends very much on who is going to do the talking. The definition of *ski* in *Webster's New International* (2nd edition) begins:

> One of a pair of narrow strips of wood, metal, or plastic, usually in combination, bound one on each foot and used for gliding over a snow-covered surface.

Apresyan *et al.* do not provide for three of the items mentioned here: what skis are made of (wood, plastic, or metal), what shape they come in (long and narrow) and where they belong spatially (on the human foot). Yet these items could be essential to understanding implicit inferences in a story.

> It was snowing. Jim took out his skis and the can of wax. He began to wax the wood carefully. Then he looked for the poles.

It could be needed to answer questions:

> Jim skied rapidly down the mountain.
> Question: What was Jim wearing?
> slippers skis sandals

Although in English and Russian it is possible to refer to skis without knowing that they are long and narrow it is not possible in Navajo or certain African languages where physical shapes determine verb forms. While the entry in *Webster's New International* goes on at length beyond the sentence given above, it does not include all the items that Apresyan mentions. Clearly the boundaries of the lexical universe are not well defined.

The dichotomy between the lexicon and the encyclopedia is particularly hard to preserve during the updating process. Recognizing definitions phrased in ordinary English is difficult [Bierwisch and Kiefer, 1970]. This information does not come neatly packaged and marked "for the lexicon" and "for the encyclopedia". How do we tell which is which?

Addition of information to one part of the entry may necessitate updating other parts of the entry. For example, if we learn that *record* is a verb as well as a noun we need to add morphological information and describe the relation between *record* and *erase*. We should probably describe recording materials, as well. We also need to add that the verb *record* is a factive, i.e., the assertion that someone records an action implies the assertion that the action really occurreed. Which of this information is lexical and which is encyclopedic? Both theoretical and practical arguments convince us that the lexicon-encyclopedia dichtomy is not valid.

Information about semantic relationships between words - thesaurus information - is needed for many reasons. It is crucial to semantic access. Hirst and Morris [1990] have shown that it is fundamental to language understanding. Fox [1980, 1988], Nutter *et al.* [1988] and Wang *et al.* [1985] have used thesaurus information to improve the results of an information retrieval system. Eiler [1979] has shown the importance of lexical relationships in human text generation. Lee [1991] is using this kind of information to generate cohesive text by machine. Zhang [1990] is using it to generate explanations.

We need to store all this information not just for words but for phrases. Becker [1975] argues cogently that language is ordinarily generated in large swatches, not a word at a time. Commercial dictionaries include many phrasal entries. For example, approximately 16% of the main entries in *Webster's Seventh Collegiate Dictionary* are phrases and many other phrases appear as "runons" at the foot of other entries. Charniak [1972] makes it clear that "birthday party" needs an entry of its own – and shows also that its lexical universe is huge.

2 Organization of the Lexical Database

Because we want to see our lexicon used by as many people as possible, we have sought sources for our data that will permit us to distribute the database to anyone who plans to use for research purposes. Collins Publishers has generously agreed we may give a copy of our lexicon including data derived from the first edition of the *Collins English Dictionary* (*CED*) to anyone who qualifies to obtain a machine-readable copy from the Data Collection Initiative. Another valuable source of lexical data is the Brandeis Verb Lexicon constructed by Grimshaw and Jackendoff [1985]. Sven Jacobson [1964, 1978] has kindly allowed us to keyboard and distribute his *Dictionary of Adverb Placement.* We have also put into machine readable form the Adjective, Adverb, Noun, and Verb Lists developed from Householder's NSF project of twenty-five years ago [1964, 1965].

Our lexical database is organized and stored using the Oracle Relational Database Management System. Our database relations (which we will call tables to distinguish them from semantic relations) include a main table with the word, and its homograph and sense number (from the *CED*) combined into one field, the part of speech, and the source. Each word with a different homograph and sense number (assigned in *CED*) is put into a different entry for the purpose of lexical disambiguation. Words that have no homograph or sense number in *CED* are assigned a code of -1. We have also designed separate tables for each part of speech. Each table contains different information specific to that part of speech.

The noun table contains information about whether a noun is regular or irregular, abstract or concrete, count or mass, human, animate or inanimate, singular or plural, common or proper, collective or not, what gender it is, and whether it appears in an

Indiana noun list. We have a separate table for Indiana nouns (those that support *that* clauses) giving the number of the Indiana list [Bridgeman 1965]. Then there is still another table that gives the definition and an example for each Indiana noun list. We also have a separate table for nouns with irregular plurals, like *child*, *goose*, and *ox*.

There are a number of different verb tables. The main verb table tells whether a verb is regular or irregular, dynamic or stative, transitive or intransitive (or both), takes a sentential complement or not, can be put into passive voice or not. If it is in a speech act class [Wierzbicka, 1989] or a performative class [McCawley, 1979] then the class will be given. Then there is a table for strong verbs with their forms. There is a case table giving information about verb arguments. If a verb takes sentential complements it appears in a special table that tells what complementizers the verb takes, its implicative class (factive, etc.), whether it is subject to raising, and whether it appears in an Indiana verb list. The Indiana verb table gives Indiana verb classes in which the verb appears. There is yet another table that gives the defining information for the Indiana verb lists.

The adjective table indicates whether the adjective is dynamic or stative, gradable or non-gradable, inherent or non-inherent. An adjective may be intensitive. It may appear as a post-determiner. It may be a general adjective susceptible to subjective measure, a general adjective susceptible to objective measure including size or shape, or color. It may be a denominal adjective denoting material, or a denominal adjective denoting provenance or style. Information about the semantic classes an adjective belongs to is essential to determining its position in the sentence during text generation. While most adjectives can occur in both attributive and predicate positions, some are non-attributive, others are non-predicative. We also have a table for unpredictable adjective inflections and another for Indiana adjectives [Householder *et al.* 1965]. Our adverb tables have been fully discussed elsewhere [Pin-Ngern *et al.* 1990].

We also have a table listing lexical-semantic relations with definitions and examples and then several tables of lexical- semantic relationships [Ahlswede and Evens, 1988a].

Our plans include tables containing other information from *CED* such as definitions, pronunciations, and etymologies, but these have not been built since none of us is currently using that information.

3 Entries for Phrases

We have been concerned for several years with the design of entries for phrases; it seems apparent that we need to record the same kind of information for phrases as for single word entries and that they are involved in the same lexical relations as other words and more besides [Markowitz *et al.* 1988; Ahlswede *et al.* 1988]. Li and Markowitz are concentrating on questions about phrasal verbs. Problems about the kinds of constructions that these verbs take part in have often been discussed in the literature but not resolved. Markowitz has devised several series of examples that we are trying out on every passerby who happens to be a native speaker of English. The data collected so far is chaotic; it suggests that the explanations in the literature are over-simplified. *CED* contains many thousands of phrasal main entries and many more phrases appear as runons in other entries. We are trying to design programs to translate these phrasal entries into entries in the lexical database.

4 Arguments for Verbs

Information about appropriate arguments for verbs is an obvious need. We are building a table that indicates for each sense of the verb what cases it takes, how those cases are syntactically realized (as subject, object, or object of a preposition), whether it is obligaory or not, and what are the selection restrictions on the fillers of those case slots. Joanne Dardaine wrote a program to build skeleton entries for the verbs in the Brandeis Verb Lexicon [Grimshaw and Jackendoff, 1985]. Then we sit around and argue about additional examples, beginning with verbs that we are using in text generation in a tutoring system for cardiovascular physiology [Zhang *et al.* 1990] and in an explanation subsystem for an expert system for stroke [Lee and Evens 1991]. Grimshaw's new book [1990] on argument patterns has been of the greatest help. Given the theoretical disagreements between Fillmore [1970], Bresnan [1982], and Grimshaw [1990], it is not possible to come up with an ideal solution. When in doubt we try to make the finest distinctions we can, in the belief that it will be easier for others to clump our categories together than to divide them further.

Clearly much of what we are doing for verbs needs to be done for adjectives and adverbs. Much of the necessary research for adverbs has been carried out by Householder's group [1965] and by Sven Jacobson and published in very detailed and useful forms [1964, 1978]. Jacobson has generously given permission for us to include this work in our database. Sumali Pin-Ngern Conlon is using the superb computing facilities of the University of Mississippi, where she is now a faculty member, to put this material into machine readable form and to combine it with the information from the Indiana Adverb Lists [Householder *et al.* 1965]. We are trying to locate and understand more of the research on adjectives such as the work of Ljung at Goteborg, before we start to enhance our adjective tables appropriately.

5 Sentential Complements

We have split off the problem of sentential complements from other arguments for verbs because we wanted to store this information in separate database tables and because there are there are separate rich sources of information. Yu-Fen Huang has entered the verbs from Wierzbicka's list of speech act verbs. We are trying to find out if *CED* synonyms of speech act verbs are also speech act verbs and if they sometimes fit into the same speech act classes [Wierzbicka 1989] or performative classes, using McCawley's [1979] categories. Pin-Ngern wrote a program to put Indiana Verb List verbs [Alexander and Kunz 1964; Bridgeman *et al.* 1965] into tables in the database. Huang is rewriting that program to include further information and trying to correlate Wierzbicka's [1989] speech act verbs and the Indiana verbs with their *CED* homograph and sense numbers.

6 Sublanguage Information

CED contains quite a lot of information about sublanguage and register (e.g., entries begin "a legal term for" or "a slang name for"). We are trying to figure out how and where to capture this information so that we can study it more effectively and also so that we can figure out to use it to make appropriate subsets of the lexical database.

Of course, sublanguage affects the syntactic correlates of words as well as the lexical ones. It is clear that we need to relate syntactic information in the lexical database to a given sense and homographic number.

We are designing tools to help us deliver subsets of the database to potential users. Clearly we need to be able to make subsets on the basis of sublanguage information as well as from word lists given us by people who want data to match. We expect to make this kind of data available in flat files (unless the user has an Oracle Relational Database Management System). All the attributes currently recorded in the database are also defined in the database. Any user of the database will be provided with this information. We expect that most of these users will need add to information to the data that we give them. So far our lexical data acquisition tools function mainly as SQL forms [Evens *et al.* 1989]. We need to provide flat file versions of these tools.

7 Tools for Accessing and Building the Database

We a re designing two families of tools, one for building the database and one for accessing it. Database construction tools themselves fall into three categories. One group of tools is intended to collect information from human informants to make it easy to add material to the lexicon for some special purpose or to extend existing information. For example, we have a tool to examine synonyms of verbs on the Indiana Verb Lists that also take sentential complements and add them to the correct lists [Evens *et al.* 1989]. Another group of tools is intended to take explicit information from a source and put it into the right table or tables. The third group of tools, most of which were originally built for sublanguage study, is designed to tackle text, sometimes dictionary definitions, sometimes other text, and extract information from it. These tools make lists of words and phrases and count them and parse text. Frank Rinaldo has built most of these tools and is working on bigger and better ones.

Our Oracle database expert, Robert Strutz, is working on tools to access the database. These tools extract information to be used by a parser or a text generation program. Other tools in this category check the database for missing data and make reports. One tool makes a list of nouns that appear in subsidiary noun tables but not in the main noun table, for example. Still other tools make subsets of the database for different kinds of user specifications.

8 Current Applications

A small subset of the lexical database, the stroke lexicon [Ahlswede and Evens, 1988b], is being used in experiments in information retrieval and text generation. Wang *et al.* [1989] are using thesaurus information to enhance queries in an interactive information retrieval system, which operates as a separate PC program and carries out searches of the stroke literature either independently or in support of an expert system. Lee and Evens [1991] are using the stroke lexicon to generate explanations for an expert system for stroke. Information about lexical-semantic relations is used in an experiment to make that text cohesive; other lexical information is used to support the basic generation process.

9 Summary

We are trying to build a big lexical database that contains detailed information about all its entries. We argue in support of this enterprise that much information often classified as encyclopedic is needed by natural language processing programs trying to carry out tasks in parsing, generation, and information retrieval. In particular, the need for thesaurus information (information about lexical and semantic relationships between words) is becoming increasingly clear. We are convinced that phrasal entries need information that is at least as rich and detailed as that provided for individual words.

We are trying to make this lexicon usable by many different people for many different tasks, with the goal of providing it free to anyone who can use it. This means that we must build it out of pieces that are readily available to the research community, that we must be able to provide subsets of many different kinds, and that we must provide tools so that others can access these files and add to them whatever further entries and information they need for particular applications.

References

[1] T. Ahlswede and M. Evens, "Generating a Relational Lexicon from a Machine-Readable Dictionary", *International Journal of Lexicography*, 1, 3, 1988a, pp. 214-237.

[2] T. Ahlswede and M. Evens, "A Lexicon for a Medical Expert System". In M. Evens, ed., *Relational Models of the Lexicon*, Cambridge University Press, Cambridge, 1988b, pp. 97-112.

[3] T. Ahlswede, J. Anderson, M. Evens, S.M. Li, J. Neises, S. Pin-Ngern, and J. Markowitz, "Automatic Construction of a Phrasal Thesaurus for an Information Retrieval System from a Machine Readable Dictionary". *Proceedings of RIAO 88*, Cambridge, MA, March, 1988, pp. 597-608.

[4] D. Alexander and W. Kunz, *Some Classes of Verbs in English*. Indiana University Linguistics Club, Bloomington, IN, 1964.

[5] Yu. Apresyan, I.A. Mel'čuk and A. Žolkovsky, "Semantics and Lexicography: Towards a New Type of Unilingual Dictionary". In F. Kiefer, ed. *Studies in Syntax and Semantics*. Reidel, Dordrecht, Holland, 1970, pp. 1-33.

[6] J. Becker, "The Phrasal Lexicon". In R. Schank and B. Nash-Webber, eds., *Theoretical Issues in Natural Language Processing*, ACL Annual Meeting, Cambridge, MA, June, 1975, pp. 38-41.

[7] A.K. Bierman, *Logic, A Dialogue*, Holden Day, San Francisco, 1964.

[8] M. Bierwisch and F. Kiefer, "Remarks on Definitions in Natural Language". In F. Kiefer, ed. *Studies in Syntax and Semantics*, Reidel, Dordrecht, Holland, 1970, pp. 55-79.

[9] J. Bresnan, ed., *The Mental Representation of Grammatical Relations*, MIT Press, Cambridge, MA, 1982.

[10] L. Bridgeman, 1965. *Nouns before That Clauses in English*. Indiana Linguistics Club, Indiana University, Bloomington, Indiana, 1965.

[11] L. Bridgeman, D. Dillinger, C. Higgins, P. Seaman, and F. Shank, *More Classes of Verbs in English*. Indiana University Linguistics Club, Bloomington, IN, 1965.

[12] E. Charniak, "Context and the Reference Problem", In R. Rustin, ed., *Natural Language Processing*, Algorithmics Press, New York, 1972. pp. 311-331.

[13] M.A. Eiler, *Meaning and Choice in Writing about Literature: A Study of Cohesion in the Expository Texts of Ninth Graders*. Ph.D. Thesis, Dept. of Linguistics, Illinois Institute of Technology, 1979.

[14] M. Evens, S. Pin-Ngern, T. , S.M. Li, and J. Markowitz, "Acquiring Information from Informants for a Lexical Database". *Proceedings of the First International Lexical Acquisition Workshop*, Detroit, Michigan, August, 1989.

[15] C. Fillmore, "Types of Lexical Information". In F. Kiefer, ed. *Studies in Syntax and Semantics*, Reidel, Dordrecht, Holland, 1970, pp. 109-137.

[16] E. Fox, "Lexical Relations: Enhancing Effectiveness of Information Retrieval Systems". *ACM SIGIR Forum*, Vol. 15, No. 3, 1980, pp. 5-36.

[17] E. Fox, J.T. Nutter, T. Ahlswede, M. Evens, and J. Markowitz. "Building a Large Thesaurus for Information Retrieval". *Proceedings of the ACL Conference on Applied Natural Language Processing*, February, 1988, pp. 101-108.

[18] J. Grimshaw, *Arguments of Verbs*. MIT Press, Cambridge, MA, 1990.

[19] J. Grimshaw and R. Jackendoff, Report to the National Science Foundation on grant IST-81-20403. In xerograph. Brandeis Verb List, 1985.

[20] G. Hirst and J. Morris. preprint. Computer Science Department, University of Toronto, 1990.

[21] F. Householder, D. Alexander, and P.H. Matthews, *Adjectives before That-Clauses in English*. Indiana Linguistics Club, Indiana University, Bloomington, Indiana, 1964.

[22] F. Householder, W. Wolck, P.H. Matthews, J. Tone, and J. Wilson, *Preliminary Classification of Adverbs in English*. Indiana Linguistics Club, Indiana University, Bloomington, Indiana, 1965.

[23] S. Jacobson, *Adverbial Positions in English*. Dissertation, Uppsala, AB Studentbok, Stockholm, 1964.

[24] S. Jacobson, *On the Use, Meaning, and Syntax of English Preverbial Adverbs*. Almqvist & Wilksell International, Stockholm, Sweden, 1978.

[25] W. Lee and M. Evens, "Generating Coherent Text Using Lexical Semantic Relations", *Proceedings of the Midwest Artificial Intelligence and Cognitive Science Symposium*. Carbondale, IL., April, 1991, pp. 41-45.

[26] J. Markowitz, S. Pin-Ngern, M. Evens, J. Anderson, and S.M. Li, "Generating Lexical Database Entries for Phrases", *Proceedings of the New OED Conference*, Waterloo, Ontario, October, 1988, pp. 115-127.

[27] J. McCawley, *Adverbs, Vowels, and Other Objects of Wonder*, University of Chicago Press, Chicago, 1979.

[28] J.T. Nutter, E. Fox, and M. Evens, "Building a Lexicon from Machine-Readable Dictionaries for Improved Information Retrieval". *Literary and Linguistic Computing*, Vol. 5, No. 2, 1990.

[29] S. Pin-Ngern, M. Evens, and T. Ahlswede, "Generating a Lexical Database for Adverbs". *Proceedings of the Waterloo Conference on Electronic Text Research*. Waterloo, October 28-30, 1990, pp. 95-109.

[30] G.N. Wang, M. Evens, and D. Hier, 1989. "LITREF: A Microcomputer Based Information Retrieval System Supporting Stroke Diagnosis: Design and Development". *Proceedings of the 2nd Annual IEEE Symposium on Computer Based Medical Systems*, Minneapolis, June 25-27, 1989, pp. 46-51.

[31] Y.C. Wang, J. Vandendorpe, and M. Evens, "Relational Thesauri in Information Retrieval". *Journal of the American Society for Information Science*, vol. 36, no. 1, 1985, pp. 15-27.

[32] A. Wierzbicka, *English Speech Act Verbs: A Semantic Dictionary*. Academic Press, New York, 1989.

[33] Y. Zhang, M. Evens, J. Michael, and A. Rovick, 1990. "Extending a Knowledge Base to Support Explanations". *Proceedings of the Third IEEE Conference on Computer-Based Medical Systems*, Chapel Hill, North Carolina, June 4-6, 1990, pp. 259-266.