DENORMALIZATION AND CROSS REFERENCING IN THEORETICAL LEXICOGRAPHY

Joseph E. Grimes
DMLL, Morrill Hall, Cornell University
Ithaca NY 14853 USA
Summer Institute of Linguistics
7500 West Camp Wisdom Road
Dallas TX 75236 USA

ABSTRACT

A computational vehicle for lexicography was designed to keep to the constraints of meaning-text theory: sets of lexical correlates, limits on the form of definitions, and argument relations similar to lexical-functional grammar.

Relational data bases look like a natural framework for this. But linguists operate with a nonnormalized view. Mappings between semantic actants and grammatical relations do not fit actant fields uniquely. Lexical correlates and examples are polyvalent, hence denormalized.

Cross referencing routines help the lexicographer work toward a closure state in which every term of a definition traces back to zero level terms defined extralinguistically or circularly. Dummy entries produced from defining terms ensure no trace is overlooked. Values of lexical correlates lead to other word senses. Cross references for glosses produce an indexed unilingual dictionary, the start of a fully bilingual one.

To assist field work a small structured editor for a systematically denormalized data base was implemented in PTP under RT-11; Mumps would now be easier to implement on small machines. It allowed fields to be repeated and nonatomic strings included, and produced cross reference entries. It served for a monograph on a language of Mexico,[1] and for student projects from Africa and Asia.

I LEXICOGRAPHY

Natural language dictionaries seem like obvious candidates for information management in data base form, at least until you try to do one. Then it appears as if the better the dictionary in terms of lexicographic theory, the more awkward it is to fit relational constraints. Vest pocket tourist dictionaries are a snap; Webster's Collegiate and parser dictionaries require careful thought; the Mel'chuk style of explanatory-combinatory dictionary forces us out of the strategies that work on ordinary data bases.

In designing a tool to manage lexicographic field work under the constraints of Mel'chuk's meaning-text model, the most fully specified one available for detailed lexicography, I laid down specifications in four areas. First, it must handle all lexical correlates of the head word. Lexical correlates relate to the head in ways that have numerous parallels within the language. In English, for example, we have nouns that denote the doer of an action. Some, such as *driver*, *writer*, *builder*, are morphologically transparent. Others like *pilot* (from *fly*) and *cook* (from *cook*) are not; yet they relate to the corresponding verbs in the same way as the transparent ones do. Mel'chuk and associates have identified about fifty such types, or lexical functions, of which S_1, the habitual first substantive just illustrated,[1] is one.

These types appear to have analogous meanings in different languages, though not all types are necessarily used in every language, and the relative popularity of each differs from one language to another, as does the extent to which each is grammaticalized. For example, English has a rich vocabulary of values for a relation called *Magn* (from Latin *magnus*) that denotes the superlative degree of its argument: Magn (*sit*) = *tight*, Magn (*black*) = *jet*, *pitch*, *coal*, Magn (*left*) = *hard*, Magn (*play*) = *for all you're worth*, and on and on. On the other hand Huichol, a Uto-Aztecan language of Mexico I have been working on since 1952, has no such vocabulary; it uses the simple intensives *yéme* and *vaücáa* for all this, and picks up its lexical richness in other areas.[2]

Second, a theoretically sound definition uses words that are themselves defined through as long a chain as possible back to zero level words that can be defined only in one of two ways: by accepting that some definitions -- as few as possible -- may be circular, or by defining the zero level via extralinguistic experiences. Some dictionaries define *sweet* circularly in terms of *sugar* and vice versa; but one could also begin by passing the sugar bowl and thus break the circularity. The tool must help trace the use of defining words.

Third, the arguments in the semantic representation of a word have to relate explicitly to grammatical elements like subjects and objects and possessors: *his projection of the budget* and

----

[1] NSF grant BNS-7906041 funded some of this work.

[2] Huichol transcription follows Spanish except *ü* high back unrounded, *'* glottal stop, *´* high tone, VV long syllable, *.* rhythm break, *x* voiced retroflex alveopalatal fricative, *r* retroflex flap, *cuV* labiovelar stop.

please turn out the light each involve two arguments to the main operative word (him and budget, you and light), but the relationship is handled in different grammatical frames.

Finally, the tool must run on the smallest, most portable machine available, if necessary trading processing time for memory and external space.

II RELATIONS

Relations were proposed by Codd and elaborated on by Fagin, Ullman, and many others. They are unordered sets of tuples, each of which contains an ordered set of fields. Each field has a value taken from a domain -- semantically, from a particular kind of information. In lexicography the tuples correspond, not to entries in a dictionary, but to subentries, each with a particular sense. Each tuple contains fields for various aspects of the form, meaning, meaning-to-form mapping, and use of that sense.

For the update and retrieval operations defined on relations to work right, the information stored in a relation is normalized. Each field is restricted to an atomic value; it says only one thing, not a series of different things. No field appears more than once in a tuple. Beyond these formal constraints are conceptual constraints based on the fact that the information in some fields determines what can be in other fields; Ullman spells out the main kinds of such dependency.

It is possible, as Shu and associates show, to normalize nearly any information structure by partitioning it into a set of normal form relations. It can be presented to the user, however, in a view that draws on all these relations but is not itself in normal form.

Reconstituting a subentry from normal form tuples was beyond the capacity of the equipment that could be used in the field; it would have been cripplingly slow. Before sealed Winchester disks came out, floppies were unreliable in tropical humidity where the work was to be done, and only small digital tape cartridges were thoroughly reliable. So the organization had to be managed by sequential merges across a series of small (.25M) tapes without random access.

The requirements of normal form came to be an issue in three areas. First, the prosaic matter of examples violates normal form. Nearly any field in a dictionary can take any number of illustrative examples.

Second, the actants or arguments at the level of semantic representation that corresponds to the definition are in a theoretical status that is not yet clear. Mel'chuk (1981) simply numbers the actants in a way that allows them to map to grammatical relations in as general a way as possible. Others, myself included, find recurring components of definitions on the order of Fillmore's cases (1968) that are at least as consistently motivated as are the lexical functions, and that map as sets of actants to sets of grammatical relations. Rather than load the dice at this uncertain stage by designating either numbered or labeled actants as distinct field types, it furthers discussion to be able to have Actant as a single field type that is repeatable, and whose value in each instance is a link between an actant number, a prcposed case, and even possibly a conceptual dependency category for comparison (Schank and Abelson, 1977.11-17).

Third, lexical correlates are inherently many-to-one. For example, Huichol quíi 'house' in its sense labeled 1.1 'where a person lives' has several antonyms: Ant (quíi 1.1) = taa.cuáa 'space in front of a house', quii.ru'áa 'space behind a the house', tei.cuáríe 'space outside the fence', and an adverbial use of taa.cuáa 'outdoors' (Grimes, 1981.88).

One could normalize the cases of all three types. But both lexicographers and users expect the information to be in nonnormal form. Furthermore, we can make a realistic assumption that relational operations on a field are satisfied when there is one instance of that field that satisfies them. This is probably fatal for joins like "get me the Huichol word for 'travel', then merge its definition with the definitions of all other words whose agent and patient are inherently coreferential and involve motion'. But that kind of capability is beyond a small implementation anyway; the lexicographer who makes that kind of pass needs a large scale, fully normalized system. The kinds of selections one usually does can be aimed at any instance of a field, and projections can produce all instances of a field, quite happily for most work, and at an order of magnitude lower cost.

The important thing is to denormalize systematically so that normal form can be recovered when it is needed. Actants denormalize to fields repeated in a specified order. Examples denormalize to strings of examples appended to whatever field they illustrate. Lexical correlates denormalize to strings of values of particular functions, as in the antonym example just given. The functions themselves are ordered by a conventional list that groups similar functions together (Grimes 1981.288-291).

III CROSS REFERENCING

To build a dictionary consistently along the lines chosen, a computational tool needs to incorporate cross referencing. This means that for each field that is built, dummy entries are created for all or most of the words in the field.

For example, the definition for 'opossum', yéuxu, includes clauses like cayúu.yúurime pücuá'aa 'eats things that are not green' and pücuáaxi.méese 'its tail is bare'. From these notes are generated that guarantee that each word used in the definition will ultimately either get defined itself or will be tagged yuunáitü mepíimáate 'everybody knows it' to identify it as a zero level form that is undefinable. Each note tells what subentry its own head word is taken out of, and what field; this information is merged into a repeatable Notes field in the new entry. Under the stem yuuri B 'be alive, grow' appears the note d (yéuxu) cayúu.yúurime pücuá'aa 'eats things that are not green'. This is a reminder to the lexicographer, first that there needs to be an entry for yuuri in sense B, and second that it needs to account at the very least for the way that stem is used in the definition (d) field of the entry for yéuxu.

Cross referencing to guarantee full coverage of all words that are used in definitions backs up a theoretical claim about definitional closure: the state where no matter how many words are added to the dictionary, all the words used to define them are themselves already defined, back to a finite set of zero level defining vocabulary. There is no claim that such a set is the only one possible; only that at least one such set is possible. To reach closure even on a single set is such an immense task -- I spent eight months full time on Huichol lexicography and didn't get even a twentieth of the everyday vocabulary defined -- that it can be approached only by some such systematic means.

There are sets of conformable definitions that share most parts of their definitions, yet are not synonyms. Related species and groups of animals and plants have conformable definitions that are largely identical, but have differentiating parts as well (Grimes 1980). The same is true of sets of verbs like ca/tei 'be sitting somewhere', ve/'u 'be standing somewhere', ma/mane 'be spread out somewhere', and cáa/hée 'be laid out straight somewhere' (the slash separatesunitary and multiple reference stems), which all share as part of their definitions 'ee.püréu.téevi X-síe cayupatátü xaü.-síe 'spend an extended time at X without changing to another location', but differ regarding the spatial orientation of what is at X. Cross referencing of words in definitions helps identify these cases.

Values of lexical functions are not always completely specified by the lexical function and the head word, so they are always cross referenced to create the opportunity for saying more about them. Quíi 1.1 'house' in the sense of 'habitation of humans' (versus 'stable' or 'lair' or 'hangar' 1.2 and 'ranch' 1.3) is pretty well defined by the function S_2, substantive of the second actant, plus the head verb ca/tei 1.2 'live in a house' (versus 'be sitting somewhere' 1.1 and 'live in a locality' 1.3). Nevertheless it has fifteen lexical functions of its own, including the antonym set given earlier, and only one of those functions matches one of the nine that are associated with ca/tei 1.2: S_1 (ca/tei 1.2) = S_2 (quíi 1.1) = quíe.cáme 'inhabitant, householder'.

Stepping outside the theoretical constraints of lexicography proper, the same cross referencing mechanism helps set up bilingual dictionaries. Definitions are always in the language of the entries, but it is useful in many situations to gloss the definitions in some language of scientific discourse or trade, then cross reference on the glosses by adding a tag that puts the notes from them into a separate section. I have done this both for Spanish, the national language of the country where Huichol is spoken, and for Latin, the language of the Linnean names of life forms. What results is not really a bilingual dictionary, because it explains nothing at all about the second or third language -- no definitions, no mapping between grammatical relations and actants, no lexical functions for that language. It simply gives examples of counterparts of glosses. As such, however, it is no less useful than some bilingual dictionaries. To be consistent, the entries on the second language side would have to be as full as the first language entries, and some mechanism would have to be introduced for distinguishing translation equivalents rather than just senses in each language. As it is, cross referencing the glosses gives what is properly called an indexed unilingual dictionary as a handy intermediate stage.

IV IMPLEMENTATION

Because of the field situation for which the computational tool was required, it was implemented first in 1979 on an 8080 microcomputer with 32K of memory and two 130K sequentially accessible tape cartridges as an experimental package, later moved to an LSI-11/2 under RT-11 with .25M tapes. The language used was Simons's PTP (1984), designed for perspicuous handling of linguistic data. Data management was done record by record to maintain integrity, but the normal form constraints on atomicity and singularity of fields were dropped. Functions were implemented as subtypes of a single field type, ordered with reference to a special list.

Because dictionary users expect ordered records, that constraint was added, with provision for mapping non-ASCII sort sequences to an ASCII sort key that controlled merging.

Data entry and merging both put new instances of fields after existing instances of the same field, but this order of inclusion could be modified by the editor. Furthermore, multiple instances of a field could be collapsed into a single nonatomic value with separator symbols in it, or such a string value could be returned to multiple instances, both by the editor. Transformations between repeated fields, strings of atomic values, and various normal forms were worked out with Gary Simons but not implemented.

Cross referencing was done in two ways: automatically for values of lexical functions, and by means of tags written in while editing for any field. Tags directed the processor to build a cross reference note for a full word, prefix, stem, or suffix, and to file it in the first, second, or third language part. In every case the lexicographer had opportunity to edit in order to remove irrelevant material and to associate the correct name form.

Besides the major project in Huichol, the system was used by students for original lexicographic work in Dinka of the Sudan, Korean, and Isnag of the Philippines. If I were to rebuild the system now, I would probably use the University of California at Davis's CP/M version of Mumps on a portable Winchester machine in order to have total random access in portable form. The strategy of data management, however, would remain the same, as it fits the application area well. I suspect, but have not proved, that full normalization capability provided by random access would still turn out unacceptably slow on a small machine.

V DISCUSSION

Investigation of a language centers around four collections of information that computationally are like data bases: field notes, text collection with glosses and translations, grammar, and dictionary. The first two fit the relational paradigm easily, and are especially useful when supplemented with functions that display glosses interlinearly.

The grammar and dictionary, however, require denormalization in order to handle multiple examples, and dictionaries require the other kinds of denormalization that are presented here. Ideally those examples come out of the field notes and texts, where they are discovered by an automatic parsing component of the grammar that is used by the selection algorithm, and they are attached to the appropriate spots in the grammar and dictionary by relational join operations.

VI REFERENCES

Codd, E. F. 1970. A relational model for large shared data banks. Communications of the ACM 13:6.377-387.

Fagin, R. 1979. A normal form for relational databases that is based on domains and keys. IBM Research Report RJ 2520.

Fillmore, Charles J. 1968. The case for case. In Emmon Bach and Robert T. Harms, eds., Universals in linguistic theory, New York: Holt, Rinehart and Winston, 1-88.

Grimes, Joseph E. 1980. Huichol life form classification I: Animals. Anthropological Linguistics 22:5.187-200. II: Plants. Anthropological Linguistics 22:6.264-274.

------. 1981. El huichol: apuntes sobre el léxico [Huichol: notes on the lexicon], with P. de la Cruz, J. Carrillo, F. Díaz, R. Díaz, and A. de la Rosa. ERIC document ED 210 901, microfiche.

Kaplan, Ronald M. and Joan Bresnan. 1982. Lexical-functional grammar: a formal system for grammatical representation. In Joan Bresnan, ed. The mental representation of grammatical relations, Cambridge: The MIT Press, 173-281.

Mel'chuk, Igor A. 1981. Meaning-text models: a recent trend in Soviet linguistics. Annual Review of Anthropology 10:27-62.

------, A. K. Zholkovsky, and Ju. D. Apresyan. in press. Tolkovo-kombinatornyj slovar' russkogo jazyka (with English introduction). Vienna: Wiener Slawistischer Almanach.

Schank, Roger C. and Robert P. Abelson. 1977. Scripts, plans, goals and understanding: an inquiry into human knowledge structures. Hillsdale NJ: Lawrence Erlbaum Associates.

Simons, Gary F. 1984. Powerful ideas for text processing. Dallas: Summer Institute of Linguistics.

Ullman, Jeffrey D. 1980. Principles of database systems. Rockville MD: Computer Science Press.

Wong, H. K. T. and N. C. Shu. 1980. An approach to relational data base scheme design. IBM Computer Science Research Report RJ 2688.