ISSUES OF COMPATIBILITY IN THE DECENTRALIZED PRODUCTION OF ENTRIES FOR A COMPUTERIZED CONCEPT GLOSSARY

by

Kenneth Janda
Northwestern University

A Working Paper for the ISSC/COCTA Round Table on the "INTERCOCTA" Project, Sponsored by UNESCO, held in Caracas, Venezuela, June 26-30, 1983.

Abstract

This working paper reviews some issues involved in the localized preparation of concept glossaries in compatible computer-readable formats for cooperative exchange and analysis. It begins by reviewing some general issues of compatibility affecting information systems and services. It then explains the purpose and structure of a draft glossary for concepts in the field of "ethnicity." It concludes by discussing technical factors in producing and processing entries for the computerized glossary on a decentralized basis. Although the paper notes some technical difficulties, it observes that the technology has advanced sufficiently to encourage attempts at a decentralized, microcomputer-based project to produce and exchange entries for a concept glossary. The experience gained in the project should be invaluable in developing the effective use of computer technology for conceptual and terminological analysis.

Table of Contents

ISSUES OF COMPATIBILITY IN THE DECENTRALIZED PRODUCTION OF ENTRIES FOR A COMPUTERIZED CONCEPT GLOSSARY

by

Kenneth Janda
Northwestern University

A Working Paper for the ISSC/COCTA Round Table on the "INTERCOCTA" Project, Sponsored by UNESCO, held in Caracas, Venezuela, June 26-30, 1983.

1 Introduction

Computers are best known for their ability to "compute" -- to add, subtract, multiply, and divide -- at incredible speeds with unerring accuracy. For decades, social scientists have used the computer's arithmetic capabilities for statistical analysis of social, economic, and political data. More recently, social scientists have begun to use the computer's logical capabilities in their "qualitative" work as well. At many universities in the United States and abroad, for example, the computer is quickly becoming an important aid in scholarly writing -- witness my use of a computer program for "word processing" in writing this paper.

It is but a short step from such word processing applications to more sophisticated "textual analysis," which goes beyond using the computer as a superior typing device to exploiting its capacity to analyze logical relationships among words. This power makes the computer useful for preparing structured lists of names or terms and associated descriptions -- such as dictionaries, glossaries, or thesauri. Riggs has recently developed a method for using a microcomputer to produce a social science concept "glossaurus" -- a combination glossary and thesaurus for social science concepts and terms.[1] Riggs envisions

> a new kind of conceptual glossary that can provide, in a systematic (classified) way, definitions of the important concepts that are needed in a given subject field. Each concept record in such a glossary is accompanied by as many terms as may be used to specialists to designate it. This method of designing glossaries is precisely the opposite of the familiar one usually found in dictionaries where entry terms are arranged

---

1. Fred W. Riggs, "COCTA-Glossaries: The *ana-semantic* perspective," in Fred W. Riggs (ed.), The CONTA Conference: Proceedings of the Conference on Conceptual and Terminological Analysis in the Social Sciences. Frankfurt: Indeks Verlag, 1982. Pp. 234-276.

> alphabetically and followed by as many definitions as there are concepts for which that term-form may be used.[2]

1.1 A Computerized Glossaurus

Riggs has begun preparing a draft glossaurus for concepts pertaining to ethnic studies or "ethnicity." He is using a computer for entering and organizing the conceptual definitions in his prototype glossaurus. He contends:

> When compiled in an automated data base from which frequently revised print-outs can be made, this new kind of COCTA-glossary will, it is anticipated, facilitate the introduction and general acceptance, among specialists in given subject fields, of new concepts and terms that can help them communicate more intelligibly and conveniently with each other.

[3] Given the transportability of machine-readable data, it is conceivable that the ethnicity glossaurus could be distributed in electronic form for automated search and retrieval by other scholars. Indeed, other scholars might contribute their own machine-readable concepts on ethnicity to the data, improving its coverage and timeliness.

The lowered cost and increased capabilities of microcomputers provide a new opportunity for localized preparation of specialized glossaries on focused topics which could, in principle, be shared among scholars and their computers and could, in principle, be merged to produce a comprehensive "master" glossaurus for a given field or even a discipline. Although computers offer this capability in principle, there are definite problems in realizing their capabilities in practice.

2 General Problems of Compatibility in Information Systems

The problem of compatibility is not unique to the exchange of machine-readable files. Because compatibility is an issue in information transfer in any form, it is helpful to introduce some general considerations about exchange between information systems before dealing with the technical problems introduced by computerization. Fortunately, the general topic has been

---

2. Riggs, p.7

3. ibid.

treated comprehensively by Lancaster and Smith in a recent UNESCO report.[4] I will draw heavily on their work in reviewing some major distinctions and issues that pertain to our glossauri project. All citations to Lancaster and Smith will be enclosed in brackets and refer to pages in their draft manuscript, which of course is subject to change before publication.

2.1 Definition of "Compatible" and "Convertible"

Lancaster and Smith say that different information systems are compatible "if they can operate together in harmony (e.g., can communicate effectively or exchange records with a minimum of effort)." [p.1]

> If two machine-readable data bases are digitally encoded according to the same conventions and if they use the same record format, the data bases can be considered compatible. This should mean that both can be manipulated by the same computer software without any further processing. [p.2]

Lancaster and Smith identify three "levels" of compatibility: physical (e.g., using the same size computer diskettes), format (e.g., number of tracks and density of recording), and intellectual (e.g., content of the records). [p.15] If two information systems differ significantly on any of these levels, they are not strictly compatible.

Ironically, small differences between computer-based systems at the intellectual level are less likely to impair compatibility than small differences at either the physical or format level, which often frustrate attempts at information exchange. When two computer tapes are identical in intellectual content but cannot be read for technical reasons by the target institutions, it can be particularly vexing.

The problem of "technical" incompatibility at the physical and format levels can often be solved through the concept of convertibility -- the process of translating information from one form into another -- which Lancaster and Smith describe as a more general property than compatibility. [p.2] Fortunately, convertibility is often aided by the very computer technology that caused the incompatibility in the first place. There usually exists computer hardware that transfers information in one physical form (e.g., tape) to another (e.g., diskettes). Similarly, computer software sometimes exists or can be written to translate information from one format into another. In fact, software can sometimes even convert files that are intellectually incompatible by programming them for term analysis and substitution.

---

4. F. Wilfrid Lancaster and Linda C. Smith, Compatibility Issues Affecting Information Systems and Services: A Report Prepared for the Division of the General Information Programme of UNESCO. Urbana: Graduate School of Library and Information Science, University of Illinois, draft dated October, 1982.

2.2 Compatibility, Convertibility, and Standardization

Both compatibility and convertibility are enhanced by standardization. Lancaster and Smith illustrate:

> if two information centres construct thesauri that adhere closely to the standards of the International Organization for Standardization (ISO), the two thesauri will be more compatible (and, therefore, more easily merged or converted one to another) than would be true if two different standards were followed or if no standards at all had been adopted." [p.3]

On the other hand, standardization has its costs: "Since each information service has its own special community of users, with diverse needs, adherence to certain types of standards may actually reduce some aspect of effectiveness. [p.21]

Much of their report deals with standards for document representation and index languages. A considerable portion focuses on standards for "bibliographic descriptions" [p.28], treating such matters as record formats, content designators, exchange formats, script conversion, and the like. While most of this material is peripheral to our concerns with construction glossauri for social science concepts, some points are quite instructive. For example, they report experiences among libraries in the construction of an "authority file" -- "a list of the access points previously used in cataloguing" [pp.71-72] which helps insure consistency both within and between library systems. Lancaster and Smith cite Hill's conclusion that

> high-quality, consistent authority work can be performed in a decentralized manner . . . [and] successful creation of a consistent national bibliographical database depends heavily on the successful and efficient sharing of authority data.

[5] The preparation of international authority files, on the other hand, are more problematic, and Lancaster and Smith prefer instead "the compilation of compatible nationally-maintained authority files with cross-references to allow conversions to be made automatically from one form of name to another." [p.75]

2.3 Problems in Standardization

Lancaster and Smith see definite limits to standardization among information systems: "Adoption of a completely standardized vocabulary for all

---

5. J.S. Hill, "The Northwestern Africana Project: an experiment in decentralized bibliographic and authority control," College and Research Libraries, 42 (1981), p.331

specialist fields, where detailed indexing of (say) science journals is involved, seems completely impractical." [p.82] Moreover, they find little commonality in vocabulary among information centers in the same field, which helps explain information scientists' preoccupation with compatibility. [p. 83] The irony is that controlled vocabularies tend to reduce between system compatibility while promoting internal consistency within information systems:

> Consider, for example, two data bases in the biological sciences, each one consisting of bibliographic citations plus abstracts in English. Given technical compatibility (in encoding conventions and record formats), these data bases are easily merged. More important, however, a single search strategy can be used to interrogate the merged file since the terminology used in one set of abstracts, being essentially the language of scientific discourse, should not differ substantially from that used in the other.
>
> Suppose, on the other hand, that each data base is indexed by means of a different thesaurus. Given technical compatibility, the two files can again be merged. But they cannot be searched by means of an identical strategy, for a single concept might be quite differently represented in the two vocabularies. pp.83-84]

Lancaster and Smith discuss complications in converting thesaurus vocabularies between information systems in some detail. Following standards for thesaurus construction helps conversion, but structural compatibility is not the only consideration. The problem of mapping between information systems becomes increasingly complex with increases in the number of systems. For example, two systems, A and B, only need two mapping procedures: A--->B and B--->A, but full exchange among four systems requires 12 separate mapping operations. [p. 86]

Lancaster and Smith credit Neville[6] for inventorying difficulties in reconciling different vocabularies. Neville identifies six "levels of correspondence":

1. Exact correspondence. This includes singular/plural variations. Thus, AIRFIELD and AIRFIELDS are considered identical. Also included are exactly synonymous terms in different languages. FLUGPLATZE, for example, is considered to exactly correspond to AIRFIELD.

2. Synonymy. UNDERGROUND STRUCTURES, BURIED STRUCTURES and SUBSURFACE STRUCTURES can all be considered synonymous. Sometimes such synonyms are identified explicitly through cross references appearing in one of the thesauri. The reference UNDERGROUND STRUCTURES use SUBSURFACE STRUCTURES, for example, indicates that these terms are considered synonymous, at least by the compilers of

---

6. H. H. Neville, "Feasibility Study of Abstracts and Indexes in Inter-Disciplinary Areas," Journal of Documentation(London), 26 (1970), 313-336.

one particular thesaurus.

3. Specific to broader term. The term SNOWDRIFTS, for example, appearing in vocabulary B, may need to be mapped to the more generic term SNOW in vocabulary A.

4. Term mapping at different levels of pre-coordination. For example, the term FROST PENETRATION in one vocabulary is considered equivalent to two terms, FROST and PENETRATION, in a second. In a more complex and less obvious example, the term STIFFNESS METHODS in A may be taken as equivalent to the terms STRUCTURAL ANALYSIS and DISPLACEMENT in B.

5. Antonyms. The term CONTRACTION may be considered equivalent to EXPANSION in one vocabulary or to EXPANSION/CONTRACTION in another.

6. Semantic factoring. This is the most complex situation. The term THERMOMETER in thesaurus A can only be translated into three uniterms, TERMPERATURE, MEASUREMENT and INSTRUMENT, in thesaurus B. [pp.88-89]

Special problems of thesaurus reconciliation in the social sciences are treated by Sager et al.[7]

2.4 Standardized Vocabularies and a Social Science Glossaurus

These difficulties of "mapping" between vocabularies, which pose a major problem in constructing a term-oriented thesaurus should not prove as troublesome to building a concept-oriented glossaurus (as Riggs has proposed) except for the alphabetized index to terms used for concepts defined in the glossaurus. But the mapping problem is likely to confront the glossaurus project directly if and when the glossaurus becomes multilingual. Presumably, the English definitions could be readily translated into foreign languages with little difficulty, but the terms to which the concepts are linked will need to be mapped between languages.

Lancaster and Smith propose an "intermediate lexicon" as a neutral "switching" language that can be used to convert between standardized vocabularies [p. 86] They offer this example: Suppose the vocabulary of system A uses TUMORS while that of system B uses NEOPLASMS for the same phenomenon. One could assign a number, say 17904, as the "neutral" intermediate lexicon code to be used for the translation. [p.94] They suggest that the intermediate lexicon concept is especially useful in a multilingual environment. [p.95] Moreover,

---

7. J. C. Sager, H.L. Somers, J. McNaught, "Thesaurus Integration in the Social Sciences. Part 1. Comparison of Thesauri," International Classification (Munich), 8 (1981), 16-22.

they cite several studies claiming success in using "switched indexing" with intermediate lexicons in retrieval tests. [p.96] In the same vein, and of special relevance to constructing a concept glossaurus, is their reference to the same work by Neville [cited above], who classified the levels of correspondence among terms to aid his proposal for developing "'a supra-thesaurus' consisting of code numbers to represent all of the concepts represented in the various vocabularies." [p.96]

Indeed, Neville's idea of a "supra-thesaurus" for engineering seems very similar to Riggs' vision of a concept glossaurus for the social sciences. Neville's general method for directly converting any keywords from one thesaurus into the appropriate keywords of another involves a form of switched indexing using number codes to the underlying concepts as the intermediate lexicon:

> The basis for the possibility of devising such a general methods lies in the assumption that it is concepts that are indexed, the keywords merely being convenient though sometimes arbitrary labels for concepts, and that broadly speaking, thesauri covering the same subject must cater for the same concepts, although they may use quite different keywords to label them. If these concepts can be identified in each thesaurus and given unique code numbers, then the series of code numbers will enable keywords of one system to be converted into the appropriate keywords of any other participating system.[8]

Neville notes, however, that his method was only at the testing stage in 1970, and Lancaster and Smith make no reference to subsequent progress.

Lancaster and Smith conclude their discussion of vocabulary reconciliation on a hopeful note, saying that it "seems possible, whether through an intermediate lexicon or otherwise." [p.128] However, one is hard-pressed to discern the basis for their optimism, given their statement that "while research on vocabularly convertibility has proceeded for at least 20 years, actual implementations of conversion or switching projects, in a real information service environment, are practically nonexistent." [p.129] (Witness Neville's experience.) Instead, the common practice is to develop a new vocabulary when a new information service is created, thus profliferating thesauri. Even standards agencies tend to ignore the thesauri of other standards agencies. [p. 130] It would seem that the advantages of standards are not so compelling to dictate their adherence.

2.5 Terminology Data Banks

As mentioned, the Lancaster and Smith report was primarily directed to compatibility of information systems involving bibliographic materials. They touch on, but do not treat in any depth, the compatibility of terminological

---

8. Neville, p. 314.

data banks, which they see growing in importance. They quote De Besse's definition of a terminology data bank as

> a kind of living multingual electronic dictionary containing hundreds of thousands of technical and scientific terms together with the appropriate terminological information.

[9] Obviously, the COCTA glossaurus project can be viewed as a type of terminology data bank.[10] Lancaster and Smith say little about standards for terminological data banks, except noting that terminological standards are the province of the International Information Centre for Terminology (INFOTERM), with headquarters in Vienna. [p.133] INFOTERM is oriented toward the physical and natural sciences. Concerning terminology banks in the social sciences, Lancaster and Smith say only that the UNESCO project, INTERCONCEPT, was launched in 1977 with the goal of establishing a bank of terms and definitions in the social sciences. [p.134]

Of course, Fred Riggs, who served as rapporteur for the 1977 meeting that founded INTERCONCEPT, has continued to be centrally involved in INTERCONCEPT activities.[11] His connection with INTERCONCEPT should provide for compatibility between the two projects when the issue arises.

2.6 Summary and Conclusions on Compatibility

Summing up their report on compatibility in information systems, Lancaster and Smith say that the message seems clear:

> information centres cannot afford to be completely self-sufficient; adherence to well-established standards greatly improves the probability that one organization can make effective and economical use of the products and services of others. [p.225]

Throughout the report, they mentioned two alternative approaches to achieving compatibility: standardization and conversion. A high degree of standardization (physical, format, and intellectual) is sufficient for compatibility in information systems. But meticulous standardization in practice is equivalent to ridigity and can be stultifying to the creation of new knowledge rather than

---

9. B. DeBesse, "Multilingual Terminology," in Overcoming the Language Barrier: Third European Congress on Information Systems and Networks, Luxembourg, 3-6 May 1977, Volume 2. (Munich: Verlag Dokumentation, 1977) P.133

10. Terminology Data Banks. (Munich: K.G. Saur, 1980.)

11. Fred w. Riggs, "Interconcept Report: A New Paradigm for Solving the Terminology Problems of the Social Sciences." Reports and Papers in the Social Sciences, No. 47. (Paris: UNESCO, 1981.)

the accumulation of old.

> Where no standards exist or where standards are felt to be too restrictive to satisfy local needs, mechanisms for conversion may still allow a given system to achieve compatibility with another system for purposes of sharing resources. Although in principle convertibility has been recognized as an alternative to compatibility for some time, in practice it is only with new developments in technology that it has emerged as a feasible alternative to standardization for reducing many sources of incompatibility between systems. [p.238]

Although it would seem that convertibility through technology offers an answer to compatibility problems between information systems (and to our particular desire to produce entries for a social science glossaurus locally for central processing), the answer lies more in the future than in the present. Intervening between now and then are technical problems of compatibility between computer systems. As Lancaster and Smith note a few pages earlier:

> The fact is that computer and telecommunications technologies are moving too fast for the standards bodies. Progress cannot be held up while appropriate standards are developed. Standards derive from operating experience. There is a pragmatic implementation of standards within information systems, particularly those that are most innovative, long before the standardizing agencies become directly involved. [p.228]

As we will see in the next section, the technology that may ultimately make it possible to engage in a decentralized, international, process of preparing entries for social science glossauri on multiple subfields is not yet within comfortable grasp, but it has moved within our reach.

3 The Format of the Ethnicity Glossaurus

As Lancaster and Smith have noted, "standards derive from operating experience." Riggs has been gaining such experience in the process of preparing his draft COCTA-Glossary for "Ethnicity."[12] Riggs refers to his effort as a "glossaurus," for it combines features of conventional glossories and thesauri but differs fundamentally from both: "It arranges concept records systematically, as in a thesaurus, but provides definitions, as in a glossary."

Ultimately a set of such COCTA-glossaries can be combined, in series, to

---

12. Fred W. Riggs, "The Draft COCTA-Glossary for 'Ethnicity' (Ethnic Studies); Printout of 10 Feb. 1983." Honolulu, Hawaii: Political Science Department, 51 pages (mimeograph).

> constitute a comprehensive conceptual encyclopedia for the social sciences, with a cumulative index to all glossaries in in the set. The maintenance of records in an automated data base that can be continuously revised and reproduced (as illustrated in [the computer] printout) provides a mechanism for the up-dating of concepts and terms used in any subject field.

[13] Riggs' draft glossaurus on ethnicity contains 164 concepts, several hundred terms by which these concepts are known, and citations to sources that define the concepts. The definitions, terms, and sources were all entered into a computer programmed for word-processing and the output was sorted by concepts and cross-referenced (i.e., "indexed") by supporting citations and terms that have been applied to the concepts. Sample pages from the output are given in Figures 1, 2, and 3.

Reference Figure 1

Figure 1 reproduces the first page of the main portion of the draft glossaurus: a listing of the concepts and their definitions. For heuristic purposes, the concepts are grouped into broad categories:

A. Ethnic Studies (as a subject field)
B. Ethnic Markers
C. Ethnic Processes and Activities
D. Ethnic Membership
E. (Omitted, for some reason)
F. Ethnic Collectivities
G. Ethnostates
H. Transnational Ethnic Communities
I. (Omitted, for some reason)
J. Ethnographic and Ethnic Concepts

The concepts within each category are assigned ascending code numbers within brackets. Thus, concept <A1> in Figure 1 is "a subject field dealing with all phenomena and problems involving ethnicity <A2>." The conceptual definition is followed by the terms -- ETHNIC STUDIES and ETHNICOLOGY -- commonly used to label this concept in the literature. Citations to the literature follow the definition. Note that <A1> becomes the "intermediate lexicon" or "switching index" that permits unambiguous reference to the concept without using either term, "ethnic studies" or "ethnicology." Note also that concept <A1> entails in its definition another concept <A2> that is termed "ethnicity." Concept <A2> is, in turn, defined next -- just below the line of dashes, which Riggs uses to separate concept definitions.

---

13. Ibid.

A. ETHNIC STUDIES (AS A SUBJECT FIELD)

<A1> a subject field dealing with all phenomena and problems involving ethnicity <A2>

TERMS: ETHNIC STUDIES; * ETHNICOLOGY *

JA032: <ethnicity> is a principal tool utilized in the study of ethnic objects...a specialized interdisciplinary field encompassing Anthropology, Sociology, History, and Political Science

JA032: "Ethnicity"...also signifies the Social Science sub-discipline that studies <ethnic objects>.

AY001: ETHNICOLOGY: synonymous with Ethnic Studies as a subject field but more appropriate following the logic of the nomenclature of Sociology, Phycholcgy, Pharmacology, etc.

ES001: ETHNIC STUDIES: needs to include other disciplines such as Literature and Psychology if it is meant to be inclusive rather than illustrative.

<CONTRAST: compare with definition of "ethnography" at <J1>>

<A2> a generic concept (including collectivities, processes, activities, actors) of contemporary societies, distinguished by ascriptive markers <B1> and plurality

TERMS: ETHNICITY

JA032: a generic term signifying both a Social Science concept and a class of social objects--"ethnic entities"

HE000: ETHNICITY: a composite of culturally defined markers (land, language, customs) that enable individuals to perceive their ethnic group membership

---

<A3> any collectivity, process, activity, or actor that may be characterized by ethnicity <A2>

TERMS: ETHNIC ENTITY; "ETHNIC OBJECT"; "ETHNIC PHENOMENON"

JA032: ethnic entity signifies social objects

FIGURE 1: Definitions of Concepts

The most unusual feature of Riggs' glossaurus is its emphasis on the concept rather than the "term" as the main entry. Referring to concepts by alphanumeric codes takes a bit of getting used to, but it certainly shifts the focus from the "label" to the "definition," which is Riggs' intention. One can refer to concepts by terms through the use of the alphabetized term index in Figure 2. Under "ethnic studies," for example, one would be directed to concept <A1>.

Reference Figure 2

A final feature of the ethnicity glossaurus is its index to the sources of the conceptual definitions. A portion of this citation index is shown in Figure 3. One can use this index to locate the conceptual definitions culled from specific sources.

Reference Figure 3

3.1 Advantages in Computerizing the Glossarus

There are several advantages that stem from producing the glossarus on a computer. The first is the simple but important value in using the computer's editing, formatting, and printing capabilities -- what has become known as word processing. Few people who have learned to use a good word processing program ever want to go back to ordinary typing. Of special benefit to the preparation of a glossaurus is the provision for easy insertion of new concepts in updating the definitions. Extending beyond the usual notion of word processing is the computer-generated alphabetized indexes to terms and sources for the concepts. These indexes constitute a rudimentary form of computer usage for information retrieval.

Other types of textual analysis could be done with the conceptual definitions, terms, and sources once they are in machine-readable form. With proper programming, one could create a diagram of the conceptual "networks" that link the definitions, or one might evaluate the ambiguity of terms used for a multiplicity of concepts. Development of these innovative approaches to conceptual and terminological analysis will no doubt emerge from the experience gained from dealing with this new type of terminology data bank.

TERMS

The terms listed below are followed by the notation symbols for the records in which they will be found, and the page numbers containing these records. The symbols that are underlined refer to records in which the indexed term is defined, whereas those that are not underlined refer to records in which the indexed term is "entailed," i.e. used as an element in the defining text. When a given term is multivocal, it will be followed by two or more underlined notation symbols, indicating the different concepts in the field of "ethnicity" which it may designate. If all of the notation symbols following a term are not underlined, this means that the indexed term is "marginal," i.e. it has a technical meaning outside the field of ethnicity, but it is used in the definition of ethnic concepts. Since any technical term used in the definition of other terms within the field of ethnicity should be "univocal", i.e. have only one meaning within this field, it follows that all of the terms that are defined in this glossary and also used in defining other terms should be followed by one, and only one, underlined notation symbol, plus symbols that are not underlined to refer users to the definitions in which they are used.

FIGURE 2: Index to Terms Used for Concepts

BIBLIOGRAPHY

The works listed below were used as sources of data for this draft of the COCTA glossary on "ethnicity". Note that each source is preceded by a two letter code, which is used in the citations to identify the location of each quotation. Some of the sources may consist of letters, memoranda, and other unpublished documents. They support the inclusion of heuristic (scaffold) terms, marked with double asterisks, in a preliminary draft, but cannot be used to support the inclusion of terms in a published edition of this glossary.

BIBLIOGRAPHIC CODE

AY: J. A. A. Ayoade, "COCTA-glossary on Ethnicity" (unpublished memo, July 1982)

BR: Yulian V. Bromley, "Ethnos and Nation". A paper presented at the COCTA panel on Ethnicity, IPSA Congress, Moscow, U.S.S.R., August 1979.

CM: Eric Casino, "Ethnicity terms and concepts for COCTA Glossary, "(unpublished memorandum, July 1982)

ES: Ethnic Studies Program, University of Hawaii, "Ethnic studies glossary" (unpublished memorandum, July 1982)

HE: Hawaii Ethnic Studies Program, "The Basic Terminology of Ethnic Studies". A glossary.

CITATIONS

AY001 <A1>, 1
AY001 <B1>, 2
AY001 <B10>, 7
AY001 <B12>, 7
AY001 <B13, 8
AY001 <B14>, 8
AY001 <B15>, 8
AY001 <B9>, 7
AY002 <C12>, 17
AY002 <C13>, 17
AY002 <C14>, 17
AY002 <C15>, 17
AY002 <C16>, 17
AY002 <C18>, 18
AY002 <C19>, 18
AY002 <C20>, 18
AY002 <C21>, 18
AY002 <C22>, 18
AY002 <C23>, 18
AY003 <C24>, 18
AY003 <C25>, 19

BR88a <G2.2>, 30

CM001 <B5.1>, 6
CM001 <B5.2>, 6
CM002 <D3.1>, 21
CM002 <D3.2>, 21
CM002 <D3.3>, 22
CM002 <D3.5>, 22
CM003 <C5>, 11

ES001 <A1>, 1
ES001 <B6.1>, 6
ES001 <B6>, 6
ES001 <C3.2>, 9
ES001 <C4.1>, 11

GD000 <J1>, 34

HE000 <A2>, 1
HE000 <B4.1a>, 3
HE000 <B4.1d>, 4

FIGURE 3: Index to Citations for Concepts

3.2 Computerizing the Concept Glossaurus

Riggs' ethnicity glossaurus was prepared at the University of Hawaii using the SCRIPT word processing program on an IBM mainframe computer. SCRIPT was developed at the University of Waterloo in Canada and is marketed by IBM. It is a powerful text formatter that accepts input from some other editing program, typically a line-oriented editor. In contrast to the more powerful screen-oriented editors used in microcomputers, a line editor does not permit the typist to move the cursor freely to edit text anywhere on the screen. The formatting commands for SCRIPT appear to be similar in concept (but not in exact form) to those used for other mainframe and mincomputer word processing programs, such as RUNOFF and (to a lesser extent) FMT, but SCRIPT has more powerful indexing capabilities.

According to reviews of word processing software in Personal Computing (April, 1983) and Infoworld (March 28, 1983), SCRIPT is not available for microcomputers. SCRIPT's basic restriction to mainframe IBM computers seriously impairs its ability to serve as a program for processing glossaurus entries prepared in a different computing environment. Of course, concept entries in the SCRIPT format could be prepared elsewhere, even on microcomputers, without actually processing them using the SCRIPT program, but the files would likely be full of errors that would surface when processing the input. A secondary problem would occur by the need to convert microcomputer data files on 5.25" diskettes into the 9-track EBCDIC format used for SCRIPT processing in Hawaii's computing center, but this might be easily handled at Hawaii by reading the diskettes on a compatible microcomputer and communicating the file over a telephone line to the mainframe. Such a conversion is relatively standard and should be readily solved with technology on hand. The more vexing problem is the unavailability of SCRIPT for microcomputers.

3.3 Alternative Hardware for Decentralized Glossaurus Production

While the SCRIPT program is demonstrably equal to the task of preparing the draft glossaurus, it appears to be unsuited to the decentralized preparation of glossarus entries -- a stated objective of the INTERCOCTA Project.[14] If the decentralized preparation (and processing) of glossaurus entries is taken seriously, one must accommodate the pervasive trend in computing activities -- the growth in purchase and usage of microcomputers as "personal" information processing machines. This is the indisputable "wave of the future" in computing. Unfortunately, the tide is not yet in and not all the vessels are afloat.

---

14. Fred W. Riggs, "COCTA History," memorandum to Members of the COCTA Board, April 13, 1983, p. 8

The problem is that the microcomputer industry is a long way from being "shaken-down" to a few standards. The modal microcomputer for academics in the United States is still the Apple, but abroad in Europe it is likely the Commodore. Because of its corporate parentage, the IBM Personal Computer is claiming an increasing share of the academic market and may soon surpass Apple. But there are other computers in significant use: the Osborne, Commodore, Radio Shack, and the Kaypro (on which this paper is being typed) -- to mention a few.

While most of these machines have adopted the 5.25" inch floppy diskette as a means of data/program storage and transfer, nearly all of them use different formats for recording data. Of even more significance, many use incompatible microprocessors, meaning that they cannot run the same programs even if the formats could be converted. At present, the best claim for compatibility across different microcomputer manufacturers is the Z80 family of microprocessor running under the CP/M operating system for 8-bit computers. But this standard may be short-lived, as IBM takes hold and as other 16-bit computers become popular.

3.4 Alternative Software for Decentralized Glossaurus Production

Let us suppose that one devises a glossaurus distribution network for the present crop of microcomputers, deciding to use software that runs on CP/M operating systems for the Z80 family of 8-bit computers. Let us also suppose a decision to compile the glossaurus using existing commercial software, rather than writing specialized programs for the purpose. What software options are available for the task, in the absence of SCRIPT for microcomputers? It turns out that the most appropriate software is not of the word processing variety but can be classified as a data base management system.

The task of generating a glossaurus is not really a word processing problem but one of terminology (data) bank management and information retrieval. This type of problem is suited to database systems software. As defined in a "Special Report" in Interface Age (February, 1983), a database program "allows users to create and utilize files to maintain information in an organized fashion. ... Further, it should allow the user to update and inquire of those files, with the ability to create reports and lists that are organized based upon selection criteria of the user."[15] Some of the best-known and most highly regarded database management systems are DB Master, dBASE II, and Quick File III. The first two operate on CP/M computers; Quick File III is for the Apple.

A database management system operating on a microcomputer typically allows the user to design a special "form" for display on the screen to guide data-entry. This form (or "template") prompts the user to enter the right

---

15. Carl Heintz, "Guide to Database System Software," Interface Age, (February, 1983), 52-53.

information in the right order, flashing reminder messages if something is amiss in the data entry stage. Once the data are entered, a good database system offers powerful search and sort capabilities and the ability to produce both detailed and summary reports from the data base with flexibility in formatting and labeling.

It should be clear that database management software, which offers dynamic retrieval and report capabilities, is better suited to glossaurus preparation than word processing software, which mainly prints text attractively. Unfortunately, most database management systems have been devised for handling numeric data, not long strings of alphanumeric information. As a consequence, most database programs are severely limited in the number of characters that can be entered in a "field" of information. Quite commonly, the limit is 255 characters, often less. All three of the popular systems mentioned above can handle no more than 255 characters in a single field -- hardly enough to accommodate lengthy conceptual definitions. In fact, the survey of 42 database systems in Interface Age lists only 5 programs that can handle more than 255 characters in a single field. Two (PFS-File and IFO-Version II) are for the Apple, and the others (MDBS, Knowledge Manager, and ANDI) are for CP/M computers.

Few social scientists have had much experience using existing database programs for storing and retrieving long strings of natural language text as required in glossaurus construction. The advertised capabilities of the programs seem suited for the task, but their actual utility can be judged only through trial and experience. There would be considerable value in a project that would investigate the application of these commercial programs to the COCTA-glossary. They list at price ranges from $2,500 to $250 for CP/M versions and for under $200 for the Apple. If one of the less expensive programs is found suitable, we could, for a relatively low cost, purchase instant compatibility in the production, processing, and exchange of microcomputer files for conceptual and terminological analysis. We might find ourselves engaged in the decentralized production of comprehensive concept glossaries much faster than we had ever anticipated.

4 A Microcomputer-Based Glossaurus Project

Lancaster and Smith remind us that the benefits of compatibility in information transfer is resource sharing. [p.20] Modern technology has great potential for promoting information transfer:

> Computer and telecommunications technologies greatly facilitate resource sharing among information services. Indeed, some forms of cooperation, such as the exchange of very large data bases, would be virtually impossible in a non-automated environment. [p.21]

However, reliance on computers in information transfer makes compatibility increasingly important.

Lancaster and Smith explain that compatibility in information transfer is multidimensional, involving the physical medium for information transfer and processing, the format of the information being transmitted, and the intellectual content of the information itself. Riggs' draft glossaurus on ethnicity has provided a general intellectual model for analyzing the relationships between concepts and terms in the social sciences. The spread of microcomputers offers a common physical medium for transfer (and sharing) conceptual definitions and related terminology. The use of a commercial database management microcomputer program imposes a standard format for recording and sharing conceptual data. Thus it seems that the ingredients for compatibility in the decentralized production of concept glossaries are at hand.

Although the technology has not fully developed to the point of making the decentralized production of concept glossaries a routine matter, neither has the state of thinking about conceptual and terminological analysis. Yet both seem to be sufficiently advanced to explore the prospects and problems in the enterprise. We could learn a great deal about both substance and method if a few dedicated scholars were to embark on an innovative project using microcomputers and commercial database systems to produce and exchange glossarus entries in one or more subfields in social science.