12-1989

Data base design for research in comparative Zapotec

Anita C. Bickford

DATA BASE DESIGN
FOR RESEARCH IN COMPARATIVE ZAPOTEC

by

Anita C. Bickford

Bachelor of Arts, Carleton College, 1974

Master of Fine Arts, University of Minnesota, 1978

A Thesis

Submitted to the Graduate Faculty

of the

University of North Dakota

in partial fulfillment of the requirements

for the degree of

Master of Arts

Grand Forks, North Dakota

December

1989

This Thesis submitted by Anita C. Bickford in partial fulfillment of the requirements for the Degree of Master of Arts from the University of North Dakota has been read by the Faculty Advisory Committee under whom the work has been done, and is hereby approved.

G. Hubert Matthews, Chairperson

Robert A. Dooley

Stephen A. Marlett

This Thesis meets the standards for appearance and conforms to the style and format requirements of the Graduate School of the University of North Dakota, and is hereby approved.

Dean of the Graduate School

Permission

Title: Data Base Design for Research in Comparative Zapotec

Department: Linguistics

Degree: Master of Arts

In presenting this thesis in partial fulfillment of the requirements for a graduate degree from the University of North Dakota, I agree that the Library of this University shall make it freely available for inspection. I further agree that permission for extensive copying for scholarly purposes may be granted by the professor who supervised my thesis work or, in his absence, by the Chairman of the Department or the Dean of the Graduate School. It is understood that any copying or publication or other use of this thesis or part thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of North Dakota in any scholarly use which may be made of any material in my thesis.

Signature Anita C. Bickford

Date August 10, 1989

TABLE OF CONTENTS

Preface

I wish to thank the members of my committee, G. Hubert Matthews, Stephen A. Marlett and Robert A. Dooley, for their help in producing this thesis, and for the large number of hours that they cheerfully gave to me at the various stages of its development.

Special thanks also go to the many members of the Mexico Branch of the Summer Institute of Linguistics, for their cooperation in filling out my data questionnaire at the outset of this project, and for their interest and encouragement.

My husband, Albert, patiently provided love and encouragement during the many months that I spent on this thesis, and frequently took time from his own linguistic projects to care for our four children so that I could work. I thank him for all of that, and for the patient technical assistance that he provided to me in printing the thesis on an uncooperative computer!

Last, and most important, I thank my Lord, Jesus Christ, for creating language in all its beauty and variety, and for stimulating my interest in that particular aspect of His creation. He brought together a number of circumstances that enabled me to finish this thesis more quickly than I had at one time anticipated. To Him be all glory and honor!

ABSTRACT

This thesis explores the nature and content of a comparative data base for the Zapotec languages of Mexico that may be produced. Many questions are discussed: how to sub-divide and list the Zapotec languages and dialects; what data format will be most accessible to researchers as well as to field workers with data to add; computer software considerations (designing computational tools); how to handle non-cognates with shared meanings, non-overlap of glosses (i.e., one Zapotec gloss covers several Spanish words, or vice versa), semantic shifts and secondary meanings, and other problems such as special characters, free variation, elisions, etc.; whether to use phonetic or phonemic data forms, or some other form; what form of verbs will be useful to the study of Zapotec; and how to index and cross-index the data. While the answers to these questions, as discussed here, pertain specifically to the Zapotec languages, it is hoped that the questions themselves can also be taken as a guideline for application by other linguists who may be planning similar projects in other languages.

Introduction

The purpose of this thesis is to propose a format for a data base of cognate sets for future comparative work among the various Zapotec languages, spoken by about 500,000 people (Grimes 1988) in the states of Oaxaca and Veracruz in Mexico. The thesis includes a discussion of the many considerations involved in deciding on the format to use for such a project, as well as my rationale for the decisions proposed for the Zapotec project. This is not meant to be an actual comparative dictionary itself--hopefully that will follow in the next few years--but the final section does include seventeen sample entries, as a small pilot project preliminary to making the dictionary itself.

Beyond presenting the preliminary groundwork for the Zapotec data base, this thesis is also intended to provide a handbook useful to others who are commencing comparative studies in other languages. Although a mere "cookbook" approach to such a complex task would not be appropriate, much may be gained by seeing what someone else did in planning for a similar project. Thus, I offer this presentation of my thoughts and research.

Much of the data used for this project is based on unpublished word list questionnaires filled out for me by these members of the Summer Institute of Linguistics who work with Zapotec groups in Mexico: Joan Smith and Grace Thiessen (Western Ixtlán), David Persons (Lachixío), Wolfram Kreikebaum (Albarradas), Ted and Kris Jones (Guelavía), Neil and

Jane Nellis (Atepec, Ixtlán de Juárez), Joseph and Mary Benton (Chichicapan), David and Sylvia Riggs (Amatlán), Inez Butler (Yatzachi), Donald Olson (Sur de Zimatlán), Randy and Susan Regnier (Quiegolani), Roger Reeck[1] (Xanica), Ronald Newberg (Yalalag), Morris Stubblefield (Mitla), John and Donna Kreutz and Barbara Morse (Guevea de Humboldt), Larry and Rosemary Lyman (Choapan), Mary Hopkins and Julie Olive (Xanaguía), Charles Speck (Texmelucan), Velma Pickett (Isthmus), Michael Ward (Quioquitani), Robert and Katherine Earl (Rincón). The word list consisted of over three hundred items, and these people had to spend several hours looking up the data in their files and field notes, and/or obtaining it directly from their language informants. I am grateful for their time and cooperation. Additional data was extracted from published sources, as cited in the text.

Research for this thesis has also included examination of various other comparative projects, to learn what answers their authors arrived at for the methodological questions involved. These other projects include Yuman (Margaret Langdon, research in progress at University of California, San Diego), Uto-Aztecan (Miller 1988), Mixtec (Josserand 1983), and Indo-European (Buck 1949).

I am presuming that a computer will be used in any contemporary work of this sort. Certainly none of the processes involved is difficult to complete, but, done by hand, they are drudgerous and time-consuming, and the manipulative flexibility desired for other potential applications of the lexical entries does not accrue. Johnson (1985:285) lists the following processes for which he designed computer software for use on Margaret Langdon's Yuman dictionary: to compute the reconstructed root of each Yuman word by undoing the sound changes that have occurred in each language, replacing one segment by another; to alphabetize each modern reflex by these reconstructed roots; and to typeset the dictionary, producing photo-ready copy. He states that a VAX 11-750 minicomputer can do all of those processes in 40 minutes, for a 120 page dictionary. One can only imagine how many months or years of labor that would require by hand! In addition, once stored in the appropriate format, this data can easily be used for other projects, such as dictionaries of each of the individual modern languages involved. Most of the computer applications that I am proposing here deal with data storage in a format that provides optimum flexibility.

Chapter 1 gives an overview of the planned Zapotec project. Chapter 2 contains a discussion of how to order the data in the data base, including ordering of the languages and the data itself. Chapter 3 contains a discussion of how to gloss the data. In Chapter 4, the phonological representation of data is discussed. Chapter 5 contains discussion of data selection. In Chapter 6, the computer record design and output formats for this Zapotec project are discussed.

There are four appendices at the end of the thesis. Appendix A contains a list of the languages included in the study and their abbreviations. Appendix B contains a map of where they are located relative to each other. Appendix C contains a list of other symbols and abbreviations used in the thesis. Appendix D contains a sample dictionary with seventeen entries, with index and cross-referencing.

Notes

[1] The Xanica data was gathered by Reeck and given to Piper, who subsequently shared it with me.

Chapter 1

Overview of the Proposed Zapotec Data Base

This chapter gives an overview of the proposed comparative Zapotec data base, defining what it is and what its goals are, and introducing my proposal for the general format to be used. Rationale for decisions represented in this section will be given in later sections.

1.1 The Project and its goals, defined

The Zapotec Data Base (henceforth refered to as "the ZDB", to differentiate it from data bases in general, which will also be discussed in this thesis) needs to be defined, both in terms of what it is and what it is not.

The ZDB is conceived to be a collection of data sets from the varieties of Zapotec listed in the introduction (plus any others from which data becomes available), organized by Spanish gloss. In the early stages of the project, the data sets will not necessarily all be cognate sets, but rather lists of words that share the same, or nearly the same, glosses. As the ZDB progresses and matures, sets that are not cognates will hopefully be rearranged into cognate sets, with missing forms (due to, e.g., semantic shifts) being filled in to complete the cognate sets whenever possible. Explained in another way, it is not a historical cognate dictionary, but rather a comparative language data base from which such a dictionary may some day be developed.

The ZDB project has one goal: to make a large corpus of Zapotec data available, in a standardized form and flexible format, to anyone who wants to use it for any of a variety of linguistic or sociological purposes. Data will be contributed by anyone with reliable data to contribute. Implementation of the plans and proposals depends on availability of funding, so it is premature to predict exactly when work will commence and specifically who will be implementing the project.

1.2 Intended users of the project

The ZDB is intended to have many types of users, and is being designed for maximum flexibility to accomodate a variety of uses. The primary users of the data base will presumably be comparative and historical linguists. It could also be used for sociolinguistic studies of various kinds. It could also be very useful to field linguists, such as those members of the Summer Institute of Linguistics who are initiating language learning projects in previously unstudied varieties of Zapotec. In addition, it could form the basis of bilingual dictionaries (Zapotec/Spanish) for the individual Zapotec languages. It is hoped that both English speakers and Spanish speakers will find the ZDB useful in their studies.

Some of the primary considerations in designing the format to be used for the ZDB are flexibility for a variety of applications, ease of expansion of the data base (i.e., additions and corrections must be able to be made easily), and cross-referencing capability.

1.3 General format

It is standard practice for computer data base material to be stored in units called records. Each record consists of a number of labeled fields. (Cf. Johnson 1985 for discussion of what records and fields are.) The form in which these records are stored bears very little resemblance to the format ultimately desired for a print-out of the working data base. Example (1) below contains a blank partial record, demonstrating the form in which data is planned to be entered into the computer for the ZDB.

```
(1)  \sg
     \eg
     \gr
     \sd
     \re
     \cm
     \xm
     \xs
     \Ate-o
     \Ate-p
     \Ate-a
     \WIx-o
     \WIx-p
     \WIx-a
```

The first eight fields contain introductory material for the record. The "\sg" and "\eg" fields contain glosses of the data set, Spanish and English respectively. The "\gr" field contains an abbreviation denoting the grammatical category of the Zapotec words. The "\sd" field contains a broad semantic domain label, based on the semantic domain categories used in Buck 1949. The "\re" field is to contain Proto-Zapotec forms, postulated after further study and reconstruction has been done.[1] The "\cm" field provides space for comments about the data set. The "\xm" and "\xs" fields contain cross-references to other related data sets.

Following the introductory fields are the actual data fields, whose markers are composed of three-letter codes for the language names, a hyphen and then a single letter (then a space followed by the language data itself). There is potential for three such fields per language, an

"\-o" field containing the data item in the form in which the compiler originally received it (most likely practical orthography), a "\-p" field containing a form between broad phonetic and classical phonemic, and an "\-a" field containing a phonemic representation based on further analysis. Since there will be about twenty varieties of Zapotec included in the ZDB, there could be about sixty data fields.[2] In each of these data fields, there is the possibility of including comments and source references. The language names are to be listed in each record in a consistent, predetermined order based on genetic relatedness.

Notes

[1] I foresee the eventual need for additional introductory fields, to contain intermediate reconstructed proto-forms, e.g., proto-Northern Zapotec, proto-Western Zapotec, etc. Such fields should be added immediately following the "\re" field, and be labeled "\re-Northern", "\re-Western", etc., or some shortened codes that would not allow for confusion with the individual language names. See Section 2.1 for a discussion of the five regional divisions of Zapotec languages, as postulated by Kaufman (1983), which I propose to use for the ZDB.

[2] However, the number of data fields could be expanded to include all fifty-five dialects mentioned in Section 2.1, if that should prove desirable. Additional introductory fields may be added, too, as needed.

Chapter 2

Order in the Data Base

When dealing with a computerized data base, it may be slightly misleading to discuss questions of ordering. Within the computer, a record can be an unordered set of fields. Order is imposed on the fields only to provide convenience for data entry, presumably by making the order of fields consistent from one record to the next. However, software created for a data base like the ZDB should not depend on any particular order, except that one field needs to be designated as the record's "title", i.e., the field that marks the beginning of the record. Within the computer, the records themselves are completely unordered with respect to one another. It is only when a user wants the records sorted and/or printed out that order is imposed on them.

Thus, the word "order" as used in Chapter 2 refers to suggested orderings for printouts, not for computer internal orderings.

2.1 Order of languages

The most obvious question to answer at the outset of a comparative project is this: What languages or dialects[1] should be compared? Presumably a linguist who is beginning such a project will have some idea of which languages are related to each other sufficiently to make an interesting comparative study. The relationship could be based on geographic, political, commercial, typological or linguistic criteria. For example, one might want to compare all the languages spoken between two particular rivers or in one certain river valley, or perhaps all the languages spoken by peoples within the boundaries of a newly formed state or country, or all the dialects spoken in a particular market town, or all the OVS languages of the world, or, obviously, all the languages known to belong to a particular language stock, e.g., Oto-Manguean, or family, e.g., Zapotec.

In the case of the ZDB, the decision of which languages to include seems fairly simple: include all the varieties of Zapotec for which reliable data can be acquired. They can be sorted into about five separate language groups, generally regarded as being as different from one another as are the Romance languages.[2] Grimes (1988) lists 55 Zapotec dialects. A few of these are so similar to one another as not to need to be regarded separately. In others, no data was available. As of the time of this writing, I am working with the twenty varieties of Zapotec listed in the introduction. More may be added, or some deleted, as other linguists feed into the project. Even the languages within each of the five major divisions are quite different from each other, with low mutual intelligibility figures (Egland 1983:66-81), so they should not be regarded as being simply dialects of the five major language groups.

After the decision has been made regarding which languages to include in the study, the practical problem of how to order those languages in the data base needs to be addressed. Numerous questions arise: Is there already a consensus as to how the languages should be grouped? Are some more closely related to each other lexically than others? Has any phonological reconstruction of sound changes yet been done? Have proto-phones been postulated? Are there shared innovations?

Is there any evidence of waves of change from one influential center across a wider area; which dialects are affected?

If absolutely no previous work has been done to determine answers for any of the above questions, then some arbitrary order will need to be chosen for listing the languages in the data base. This order could be random, alphabetical, or geographically-based. For a large data base including more than two or three languages, it seems that random ordering would be impractical, making it difficult to find the data for reference. Random ordering should not be used if there are any alternatives. Geographic ordering would be useful, especially for comparative studies involving a large geographic area, e.g., Uto-Aztecan, in which geography is probably a large factor conditioning the language variation.[3] Alphabetical order is also a practical way of listing languages in the data base until another order, based on relatedness, can be established. It is vital that the computer program used allow for the data to be rearranged easily should a new order be desirable.

For the ZDB, I propose to use the following organization of the varieties of Zapotec, from Kaufman 1983, based on genetic relatedness; this classification could be changed if further study revealed a preferable order.[4] Note that the languages in parentheses are not included in the ZDB as yet.

(1) Northern
Atepec
Western Ixtlán
(Cajonos)
Yatzachi
Yalalag
(Zoogocho)
(Tabaa)
Choapan
Rincón

Western
Texmelucan
Lachixío

Southern
Quiegolani
(Mixtepec)
Quioquitani
(Coatlán)
Xanica
Xanaguía
Amatlán

Eastern
Guevea de Humboldt
Isthmus

Central
(Ayoquesco)
Sur de Zimatlán
Chichicapán
(Ocotlán)
Guelavía
Mitla
Albarradas

The ordering within each of the main five categories is Kaufman's. I arbitrarily propose to list the main categories in geographical order: northern, western, central, eastern, southern. This order yields the following list, used in the sample dictionary in Appendix D:

(2) Atepec
Western Ixtlán
Yatzachi
Yalalag
Choapan
Rincón
Texmelucan
Lachixío
Sur de Zimatlán
Chichicapán
Guelavía
Mitla
Albarradas
Guevea de Humboldt
Isthmus
Quiegolani
Quioquitani
Xanica
Xanaguía
Amatlán

The order chosen for a particular data base needs to be explained in the introduction to accompany any printed editions of that data base that are produced.

2.2 Dividing the lexical items into files

Next, it is necessary to decide how to divide the data itself into files of manageable size. Strictly speaking, all of the data in a data base such as the ZDB could fit onto a hard disk in one file. The ZDB will contain about two thousand records, each containing about one hundred fields, with an average of maybe fifteen characters per field. These are very rough estimates, which are probably quite overly generous. Even so, with these estimates, the ZDB will occupy only about three mega-bytes, i.e., three million characters. This is more than the capacity of any diskettes currently being made, but will fit comfortably onto virtually any hard disk in current use.

However, it is extremely clumsy to work with one file anywhere nearly that large; it takes a long time to load a file that big, and every single time that additions or corrections are made the entire file will need to be rewritten to the disk. Working with large files is also very risky; if any errors are made by the computer, or if a "bad block" develops on the disk, the entire file can be lost. Thus, the data needs to be divided into smaller files, probably no more than fifty thousand or one hundred thousand characters per file, depending on the speed of the computer being used. Three possible division criteria arise: chronological, by semantic domain, or alphabetical. Again, random organization can be discarded immediately in favor of any available alternative.

2.2.1 Chronological order

Data files could be set up chronologically, that is, putting the data in in the order in which it is received, e.g., one file for Tuesday p.m., another for Wednesday a.m. This might have the small advantage of making proofreading easy, allowing the compiler or a clerk to proofread, for example, all of Wednesday morning's entries at one time, rather than having to search through lists for the things that were added during that one time-block. This one small advantage, however, can be accomodated by software that allows new data to be added at the ends of previously established files and merged into the file later (after proofreading) in another order, probably alphabetical, as discussed below. Furthermore, once the word was entered into the data base, the proofreading advantage would be outweighed by the large disadvantage to users of having the data appear almost random in order. They would not be able to find an entry without searching in many files for it; even by means of the computer, such searches would be time-consuming and discouraging. The files need to be set up in such a way that new data can be entered directly into the files where they will be stored for easiest use later, not for maximum entry ease.

2.2.2 Organizing by semantic domain

Organizing by semantic domain has advantages and disadvantages. It would be very easy for fieldworkers to use such a data base in language learning, since most words needed in a particular semantic context would be conveniently available in one place. Such organization would be easy to use in gathering new data; thus it would be kind to solicit data from people in early stages of fieldwork in some sort of questionnaire organized by semantic domain. For example, lists of glosses could be headed "time terms" (hour, month, later, in two weeks, early, etc.), "farming terms" (planting, sew seeds, machete, field, ripe, harvest, etc.), "kinship terms" (maternal grandmother, son, closely-related, family, adopt, generation, etc.), and so on.

The disadvantages of organizing by semantic domain seem minor. First, such semantic lists present the data with all parts of speech (verbs, nouns, adjectives, adverbs, etc.,) mixed together. Again, however, this problem could be surmounted by including a grammatical category field in each entry, by means of which the computer could sort the data as needed for convenient analysis. In this field, each entry would be labeled as a noun, verb, or whatever, and the user could ask the computer to list out, for example, all the verbs, which could even be subcategorized by transitivity, mode, tense, or whatever might be significant to the language in question. (See discussion of the '\gr' field in Section 6.1.) Second, finding data in a printed version of a semantically organized data base necessitates using an index rather than finding the data directly in the files, as would be possible with, e.g., an alphabetical listing. Assuming that the indexes are made well, however, this is not a large problem.

Much broader semantic domains should be used as titles for files. I propose to organize the ZDB files based on the semantic domain outline used in Buck 1949, each of his chapter titles being the title of one data file. If additional files prove necessary, so as to include, for example, grammatical morphemes ("functor words"), pronominal forms, etc., they can be added. If some files need to be split into smaller files, or combined into larger ones, or omitted entirely, such modifications can be made as well. Example (3) below contains a list of Buck's semantic domain categories (Buck 1949:xix):

(3) Buck's Semantic Domain Categories

1. the physical world in its larger aspects
2. mankind: sex, age, family relationships
3. animals
4. parts of the body; bodily functions and conditions
5. food and drink; cooking and utensils
6. clothing; personal adornment and care
7. dwelling, house, furniture
8. agriculture, vegetation
9. miscellaneous physical acts and those pertaining to special arts and crafts, with some implements, materials, and products; other miscellaneous notions
10. motion; locomotion, transportation, navigation
11. possession, property and commerce
12. spatial relations: place, form, size
13. quantity and number
14. time
15. sense perception
16. emotion (with some physical expressions of emotion); temperamental, moral and aesthetic notions
17. mind, thought
18. vocal utterance, speech; reading and writing
19. territorial, social and political divisions; social relations
20. warfare
21. law
22. religion and superstition

2.2.3 Alphabetical order

As discussed above for ordering the languages, alphabetical ordering of the data seems to make it most universally accessible to all types of users. It is very practical and easy to comprehend. A decision must next be made regarding which field to use as the title for each record. Worded in another way, the question is, "alphabetical order by what?" The computer can alphabetize by any field chosen by the compiler. The obvious choices would be one of the main introductory fields in each record: Spanish or English gloss or reconstructed form. Any of those could be useful, depending on the intended use of the data base. (The computer can also sort out all the words in one grammatical category, e.g., nouns, from the data base, but one would not want to alphabetize by grammatical category.)

However, organizing files alphabetically, e.g., one file for A-C, another for D-F, etc., would not be useful, imposing awkwardness on the data entry personnel as well as on subsequent users of the data base. Alphabetical order should be used only for ordering individual records within files that are set up on other criteria.

Data for the ZDB was collected by means of a questionnaire organized mostly by semantic domain, with some random sections, and then entered into the computer chronologically, language by language as received, with each record numbered to match the number from the questionnaire. (I subsequently discarded the number field, since data from other sources would not follow my system of numbering.) One good possibility for a printout of the data is to have the computer alphabetize the records within each semantic domain file by Spanish gloss.[5] See Section 3 for discussion of glossing considerations and Section 6.1 for discussion of the keyboarding format.

Notes

[1] For simplicity, from now on I will be referring to these "languages or dialects" simply as "languages" or "varieties of Zapotec".

[2] Kaufman 1987, Butler 1980, Pickett 1985.

[3] The languages in Miller 1988 are arranged geographically, from north to south.

[4] Egland 1983 has another outline, which differs from Kaufman's somewhat.

[5] Note that distinct words can have the same spelling, with the result that distinct records can have the same title. This is not problematic, since a simple computer program can be made that lists all instances of repeated record titles. A number can then easily be added in each of those records' title fields, specifying the desired order for those particular records. For example, there could be two records whose "\eg" fields contain "rain", one for the verb and one for the noun. If the English gloss field were chosen to be the title for a given printed edition of the ZDB, these two records would have identical titles. To specify how these two records would be ordered, one's "\eg" entry could be changed to "rain1" and the other's to "rain2". Similar duplications could take place in the other two most likely title fields, "\sg" or "\re", as well, and could be handled in the same way.

Chapter 3

Glossing Considerations

3.1 Language to use in glossing of language data

What language to use in glossing the language data is another important decision to be made in preparing the data base. The choices are fairly obvious: use the national language of the country in which the target language is spoken, the main language of the primary potential users of the dictionary, or both. (Of course, in some circumstances, these "two" possibilities will be the same language, as in Margaret Langdon's Yuman dictionary, which treats languages spoken in the United States. English is the obvious choice for glossing in this case, since it is the national language of the United States, and the language used by most of the likely users of the dictionary.)

In the ZDB, I propose to gloss in both Spanish and English, making the data base accessible to educators and government officials in Mexico, to bilingual Zapotec speakers (whose second language is much more likely to be Spanish than English), and to English-speaking linguists who might desire to use the data presented therein. Depending on the elicitation situation, e.g., degree of bilingualism of the Zapotec speaker, command of Spanish and/or Zapotec of the linguist, etc., one of the two glosses might be more reliable than the other. In any case, care must be taken to gloss in both languages as exactly as possible, to avoid confusion, even if such care necessitates use of phrases rather than single-word glosses. For example, the Isthmus Zapotec word for 'flea' is 'bi?iu. The Spanish gloss would be 'pulga', which means either 'flea' or 'thumb'. Care must be taken to specify which meaning of the Spanish word is reflected in the Zapotec form 'bi?iu, e.g., 'pulga (insecto)'.

If only one gloss is available in the data, and the other gloss is merely filled in by the compiler of the data base by means of translating the gloss provided, the translated gloss should be verified as soon as possible with the source of the data (or someone else knowledgeable), and should be flagged as an uncertain gloss until it is verified. For example, if the Spanish gloss provided is 'niño', the compiler could fill in an English gloss, 'boy 2', the '2' indicating uncertainty (see Section 6.1.1 for a discussion of such reliability codes), until the gloss can be checked with someone who knows for sure whether the Zapotec word glossed 'niño' has the 'generic child' sense or if it specifically means 'boy', as in 'male child'.

It should be noted that separate editions of the ZDB can easily be printed out, one in which the English gloss is printed first, making it act as the title for each entry, and another having Spanish act as the title. If in a given situation one of the two languages is clearly preferable as a "primary" gloss (e.g., the linguist has very little command of Spanish and most of the elicitation was done monolingually so that English is the real gloss and the Spanish gloss is merely a translation of the English), then that language's edition should be regarded as being more reliable. In other situations where there is not such a clear-cut choice (e.g., the linguist supplying the data has enough knowledge of the Zapotec meanings that English and Spanish glosses can both be assigned directly from the Zapotec data), there is no basis for considering either Spanish or English as the primary glossing language; rather, both would have equal status and the choice of which edition to use would depend on the needs of the users.

Use of two or more gloss languages poses the need for multiple indexes for printouts of the data base. See Section 5.3 for a discussion of what indexes are needed and how to create them directly from the data base.

3.2 Special problems

Numerous special problems arise in comparative work, especially in regard to deciding what data sets to include in the data base. Some sets will be "cognate sets" and some will not. It is first necessary to define the term "cognate". Two linguistic forms are cognates only if they are historically derived from the same form in the parent language, and neither is the result of any borrowing (Crystal 1985, Arlotto 1972). Of course, that historic relationship often yields a similarity in form, and the differences can hopefully be compared with parallel differences in other cognate sets, allowing generalizations to be made in terms of innovations which occurred over time to produce characteristics in each modern language that make it unique from the parent language and from other sister languages.

As an example of the type of variation that can occur between modern reflexes of the same proto-form, consider the following word pairs from Atepec and Isthmus Zapotec. In each case discussed here, the Atepec data are taken from Nellis 1983 and the Isthmus data from Pickett 1980[1]. Consider first these four pairs, in which the Isthmus sound [y] corresponds to the Atepec sound [y]:

(1)

	Ate	Ist
doblado (doubled)	yech.u	ye'chu'
cinco (five)	gayu'	gaayu'
árbol (tree)	ya	yaga
olote (corn cob)	yana	yaana'

The next three pairs demonstrate a similar correspondence between Isthmus [g] and Atepec [g]:

(2)

	Ate	Ist
siete (seven)	gatsi	gadxe
camote (sweet potato)	guu	gu
chapulín (grasshopper)	guxaru'	guxharu

This correspondence occurs only when the [g] in Ist is followed by a back vowel. However, when it is followed by a front vowel, it corresponds to Ate [y] instead of [g], as shown in (3):

(3)

	Ate	Ist
cigarro (cigar)	yeeda	gueza
cuero (leather)	yeeti	guidi
fuego (fire)	yi'	gui
escarabajo (beetle)	cutu.lu ye'e	bidolagui'
olla (clay pot)	yeth.u'	guisu

Because the Isthmus sound [g] corresponds to two different Atepec sounds, [g] and [y], and because there is a consistent conditioning environment which could explain the variation in the Atepec sounds, I postulate a protophoneme, *g, from which all of the Isthmus and Atepec [g] sounds have come, and an innovation in pre-Atepec in which *g became [y] when it preceded a front vowel. I need also to postulate a protophoneme, *y, from which all of the Isthmus [y] sounds come, as well as the Atepec [y] sounds in example (1).

The *g>y innovation preceding front vowels in pre-Atepec may result in a g∼y alternation in modern Atepec, decribed by a synchronic phonological process. If so, the g∼y alternation stems from the innovation and the synchronic rule is a reflection of the innovation. Of course, a much larger data base, and data from many more Zapotec languages, would need to be studied to make definitive statements about actual protophones of Proto-Zapotec. At any rate, this short passage in the thesis is just an example of how the sound correspondence discovery process works. Anything more than speculation is beyond the scope of this thesis.

It is sets of truly cognate forms that are most interesting to the historical linguist. However, sociolinguists could be interested in any sort of data set--some with shared glosses, some with semantic shifts, etc.--whether or not the words in the set are all cognate to each other. Thus, a way must be found to include and gloss a wide variety of data sets in the data base.

In this section, I examine several types of data sets, and discuss special glossing problems that they pose. Section 3.2.1 discusses noncognates with shared meanings. Section 3.2.2 discusses synonyms. Section 3.2.3 discusses words that are similar in form but not in meaning, also including accidental homonyms within a single language. Section 3.2.4 discusses secondary meanings and semantic shifts. Section 3.2.5 discusses problems posed by non-coextensiveness of glosses.

3.2.1 Non-cognates with shared meaning

In the most obvious kind of cognate set, the modern reflexes of the word are all derived from the same word in the parent language, thereby probably retaining fairly similar form. However, sometimes lexical substitution occurs which complicates this situation, effectively robbing a cognate set of forms that might have been included. Consider, for example, several Romance languages' words for 'dog': French, *chien*; Italian, cane; Portuguese, cachorro, cão ; and Spanish, perro. These words express the same primary meaning in four clearly-related Romance languages. However, by the process of lexical substitution, apparently Spanish substituted a unique form, perro, for the proto-Romance form (presumably Latin canis), with the obvious result that its modern reflex differs greatly from the rest of the family. Portuguese appears to be in the process of substituting cachorro for cão, which will no doubt eventually remove the most common Portuguese form meaning 'dog' from the canis cognate set as well. Cachorro and perro cannot be traced back to canis by "undoing" sound changes that occurred over time. They simply come from sources other than canis. The Latin word canis, which has the same meaning as these two modern reflexes, is preserved (or re-introduced as a Latin loan word) in two Spanish words, the adjective canino (meaning 'canine') and the archaic noun can 'dog', but is not present at all in perro. (Incidentally, canis is reflected in the English adjective canine, which is clearly a loan from Latin, but obviously is not present in dog.)

Opportunities for cognate sets can likewise be missed when the parent language has two or more synonymous words, and the descendant languages take different ones, excluding and losing the others through the passage of time. Of course, cognate sets could also be incomplete if the collector has simply failed to get the cognate synonym in one or more of the languages in his study. (See Section 3.2.2 for a more complete discussion of synonyms.)

Data sets of these two types are not cognates and are thus not of interest in cognate studies, but for comparative studies and sociolinguistics they are, and thus should definitely be included in the data base and stored in some form that would allow them to be used for future analysis. How they group together may reflect generations of relatedness in the modern languages or indicate when the lexical substitution took place. For example, old forms like canis could form the basis of new cognate sets being assembled, even though the primary modern glosses and forms may not match up with those of the original language. This could most easily be done when there is considerable knowledge of the earlier language, as in the case of Latin, so that the tracing could essentially be done backwards, from the earlier language to the modern reflexes. In a Romance Data Base, the modern reflexes for the canis example above could be listed in a record titled "dog", from the English gloss field for the Latin word canis, and then other records, containing only a gloss, e.g., 'canine', and a cross-reference, e.g., "cf. dog", could be set up to direct users to the correct record to find the desired forms. The proto-form canis itself would also be placed in the "dog" record, in the "\re" field.

To summarize, such sets would be formed by sorting the data by modern gloss, searching for semantically-related words, and finding cognate forms whose glosses do not match perfectly. For this reason, exhaustive cross-referencing of data records should be done, as discussed in Chapter 6 and exemplified in Appendix D, and glosses of individual data items must be included in the data fields when they differ from the gloss of the record as a whole.

3.2.2 Synonyms

Synonyms are lexical items, within a single language, that have the same, or nearly the same, meaning. Many synonyms are not cognates, for they have nothing to do with shared linguistic parentage or similarity of form.

However, they are important to discuss here, since their existence greatly complicates the process of finding cognate sets. When one language selects one synonym from the parent language, and another language chooses another, the modern reflexes in those two languages will be completely unrelated to each other. Complete cognate sets involving those words will probably not be found. Likewise, synonyms within one modern language can confuse the process of finding cognate sets, since one speaker may use one form and another speaker another form.

The complexity increases when synonym problems are introduced from the gloss language as well. Consider, for example, the English words _pretty_, _lovely_ and _beautiful_, which are quite close in meaning and could be used interchangeably in most contexts. In seeking to learn the Spanish form used for that meaning, any of those three English words (and probably several others!) could be used as the cue, and the Spanish response could also be any one of many words, e.g. _lindo_, _bonito_, _hermoso_, _chulo_, etc. If French, Italian and Portuguese forms are also to be elicited and compared, the potential for confusion increases greatly, and if several more languages are added for comparison, astronomically! In the case of the Zapotec project, the elicitors are probably thinking and planning in English, cueing in Spanish, and then receiving data in twenty different varieties of Zapotec.

Nellis 1983 provides examples of complex synonym sets in Atepec Zapotec:

(4) bonito (pretty) - latsitte, joscu, coscu, latsitteni
grande (big) - xeni, el.la, thu, thuu, yeni
bien (well) - tse', tse'taa, joscu, coscu, latsiru, tse'ni

Note that _joscu_ and _coscu_, and the stem _latsi-_, are found in both the 'bonito' set and the 'bien' set, further complicating glossing of those two words.

Further Zapotec examples, this time from Isthmus Zapotec, are found in Pickett 1980:

(5) luego (later) - oraque, oraqueca, maca
lagartija (green lizard) - guragu', uragu', guxaaya, uxaaya, yeeta, sumbidxi

All such complex data sets will need to be sorted through carefully, by hand or perhaps utilizing a complex computer search program, so that all the cognates are correctly put into sets. The computer can print out all the records that contain certain specified words in the gloss fields or a semantic domain field. (See Section 6.1.1.) In the case of the ZDB, these gloss fields would be Spanish, "\sg", and English, "\eg", and a semantic domain field, ("\sd"). With these words all together in one place, it should be a very straightforward task to sort them into cognate sets.[2]

Such sets could then be subcategorized by Spanish gloss. Example (6) demonstrates record titles (Spanish glosses) that could be used for the Isthmus Zapotec 'lagartija' set:

(6) lagartija1 - (g)uragu'[3]
lagartija2 - (g)uxaaya
lagartija3 - yeeta
lagartija4 - sumbidxi

(Of course, each record would actually contain the cognate Zapotec forms for many or all of the twenty varieties of Zapotec, not just Isthmus as shown here.) What this example indicates, though, is that there can eventually be several data sets for a given gloss. The data forms would probably initially all be massed together under one gloss, e.g. 'lagartija', and then subsequently sorted and subcategorized into cognate sets in separate records. All such records that share the same, or very similar, gloss should be cross-referenced to each other. For example, the 'lagartija1' record above would contain the directive, "see lagartija2, lagartija3, lagartija4". Since all four 'lagartija' records would be contiguous to each other in the alphabetized data base, this may seem unnecessarily ponderous. However, the need for such cross-references becomes more obvious if the records are titled somewhat differently, e.g. 'lindo, bello, hermoso, chulo, etc.', as discussed earlier, rather than 'lindo1, lindo2, lindo3, etc.'

Buck 1949 deals with the complexity problem caused by synonyms by organizing his Indo-European data base around semantic domains rather than alphabetically. Each of the twenty-two chapters in his data base deals with one large semantic domain (Cf. Section 2.2.2 for a list of his semantic domain titles). Each individual data set is organized around an English gloss. Occasionally there are two data sets with the same gloss, but only when that English gloss word can occupy two grammatical categories (e.g., 'milk' can be either a verb or a noun) and the two forms of the words are not morphologically related. If they are morphologically related, then the two sets are combined into one, with the grammatical category specified as, eg., "vb.;sb." (i.e., verb, substantive) and the two forms given together in the data list, separated by a semicolon, (i.e., the verbal form of the word, then a semicolon, then the substantive form). This combining saves a tiny amount of cross-referencing. However, there is still considerable cross-referencing between sets, usually in the notes following a data set and pertaining to only one individual language, rather than in the data list itself. Occasionally an entire data set could fit into more than one semantic domain chapter, e.g., 'tree' could go into either chapter 1 ("the physical world in its larger aspects") or chapter 8 ("agriculture, vegetation"). The data set itself is in chapter 1, item number 1.42, so in chapter 8 there is an item number 8.60 for 'tree' which simple says "Tree=1.42".

Besides still maintaining the need for cross-referencing, Buck's organization by number poses the need for a large alphabetical index of glosses, giving the chapter and data set number of each entry. It seems also to complicate the process of expanding the data base, since new data sets would need to be fitted into the existing numbering system.[4] Minor shades of meaning might not be able to be accomodated in this way, whereas they could be if a purely alphabetic listing were used within each file rather than numbering the entries and using numbers for cross-referencing, etc. As such, organizing files by semantic domain still remains the best option for the ZDB, but pure alphabetical order within the files, with exhaustive cross-referencing by record title rather than by number, either in an index or interspersed with the actual data sets, appears preferable to Buck's system.[5] It may, in fact, turn out to be desirable to merge all the files together into a purely alphabetical listing for printing out the data, so that no indexes would need to be made to direct users to the correct chapter for a given data set. See Chapter 6 and Appendix D for further discussion and exemplification.

3.2.3 Words that are similar in form but not in meaning

Occasionally in comparative work, words may be found that appear to be cognates, since they are very similar in form. For example, the English word <u>sole</u> 'bottom of a foot or shoe' and the Spanish word <u>sol</u> 'sun' are very similar in pronunciation, except for fine phonetic detail. Obviously, however, the meanings are vastly different, and there is no evidence that the two words evolved from a common form in a parent language, nor that the meaning of one has evolved from that of the other (or from a common earlier form with yet a third meaning) by means of semantic shifting. Thus, *sole* and *sol* are not cognates, by definition. Comparison of such pairs (or larger sets) is not of interest; thus, they should not be included in the data base as pairs or sets. Of course, each of the two words will appear as part of a set, listed with others of the same meaning. (Since further study could reveal relatedness that was not at all obvious initially, the system of storing data should perhaps facilitate regrouping of such sets or pairs of words at some later time, e.g., they could be cross-referenced to each other.)

A number of pairs like *sol*/*sole* occur in comparing the wordlists in Nellis 1983 and Pickett 1980 for Atepec and Isthmus Zapotec respectively. Some examples are listed below in (7).

(7)	Ist	du	(noun)	la flor del maíz (the cornflower)
	Ate	du	(participle)	estar parado (be standing)
	Ist	guba	(noun)	vapor, hálito (vapor, breath)
	Ate	gubba	(noun)	escoba (broom)
	Ist	maca	(adverb)	en ese momento, al momento, luego (at that moment, at the time, later)
	Ate	maca	(noun)	red gruesa (thick net)

Clearly, these pairs are very similar in phonetic form but very different in meaning and, in some cases, even in grammatical category. Cross-referencing of such sets would not fit into either of the fields designated for cross-referencing, i.e., "\xs" for semantically related cross-referencing and "\xm" for morphologically related cross-referencing, so it could be put into the general miscellaneous comment field instead, or perhaps even omitted altogether.

3.2.3.1 Accidental homonyms

This subsection is relevant only to application of the data base in creating a comparative dictionary, in which the records are organized around posited reconstructed stems rather than around glosses. Homonyms as discussed here do not pose a problem to the data base format that is being proposed for the ZDB.

Frequently, two or more words within a single language have the same phonetic form but do not share a common meaning. These can be accidental homonyms, as in the English sets, *bow*, *bough*; *earn*, *urn*; *be*, *bee* (contest, e.g., spelling bee), *bee* (honey-producing insect), *bee* (second letter of the English alphabet), *Bea* (girl's nickname); etc. Such sets of words are, of course, not cognates, since they occur within the same language. There is no obvious reason why these sets of words share common phonetic forms. Yet one would need to define such words carefully when eliciting their equivalents in another language, so as to receive the desired data forms.

Initially, such words would need to be labeled in some way, to subcategorize their phonetic forms. For example, if someone were investigating English, the phonetic form listed for the entire *be*, *bee*, *Bea* set shown above, [bi], could be sub-categorized and glossed as follows:[6]

(8) [bi]1 - (noun) contest (e.g., spelling or quilting)
[bi]2 - (noun) winged insect that produces honey
[bi]3 - (noun) second letter of the English alphabet
[bi]4 - (proper noun) girl's nickname
[bi]5 - (verb) exist

Presumably the order of listing the various meanings could be random, although perhaps grouping them syntactically would be useful in a very large set. For example, in (8) above, all the nouns have been grouped together rather than inserting the one verbal meaning between two of the nominal meanings, and the one proper noun follows the common nouns rather than splitting that group. (Strict alphabetical order would also accomplish this.)

More importantly, such accidental homonyms are not peculiar to English. They would almost certainly also occur in the target language, causing potential for confusion when the roots are being compared, listed and indexed. Example sets of such accidental homonyms in Texmelucan Zapotec are shown here in figure (9), all data being drawn from Speck 1972:

(9) bey1 - nail
bey2 - cloud

baa1 - sky
baa2 - grave

In the final printing of a comparative dictionary, after the analysis is complete and the formatting for the printout is done, the numbers used to sub-categorize the homonymous stems should be removed, leaving a purer, cleaner-looking data list.

If all the cognate sets have been formed already, this whole step can be omitted. It should be noted, too, that such subcategorization may not be needed at all if the data records are organized by gloss rather than by the Zapotec forms themselves. (See the discussion earlier in Section 3.2.3.)

3.2.4 Semantic shifts, secondary meanings, etc.

Alternatively, homonymous sets can occur as a result of extended meanings, of which a speaker may or may not be conscious as he uses the words. Consider the English word <u>green</u>, which has many meanings: a color, unripeness of fruits or vegetables, recency of cutting of firewood, inexperience in a worker, etc. The 'unripe fruit' sense is probably derived from the prototypical green color of many fruits or vegetables when unripe (and prototypical 'green firewood' is green in color, too).[7] This sense is further extended to a worker who has not yet 'ripened' in experience and is therefore 'green' at his job. Yet all these meanings share the same phonetic form, [griʸn].

Similarly, in Texmelucan Zapotec (Speck 1972), there are such words, with extended meanings. For example, *kaas* means 'black, dark', the 'dark' meaning probably being an extension of the darkness of the color black. *lo* means 'face, before'. "...Zapotec extends body-part terms to...locations that stand in a relationship of figure to ground with objects, and to a few specialized locatives." (MacLaury, to appear:2) The 'face' sense is first extended to 'the front-facing side', then further extended to time, so that it means something like 'the front facing part of time, that which comes first'.

Sometimes, a semantic shift takes place in which the primary sense in such a set is lost. For example, consider the words *wife* and *Weib*, English and German, respectively. *Wife* obviously means 'female spouse'. *Weib* (phonetic form [vayp]) appears to be a cognate, yet it means 'woman', often with some pejorative overtones. They probably have the same proto-form, and have retained very similar phonetic forms, yet they have different meanings. According to Morris 1978, the English word *wife* is derived from an unattested Germanic form, *wif*, meaning 'woman'. Thus, the English word *wife* has undergone a meaning shift during the course of time, narrowing its generic 'woman' sense to the more specific 'married woman' sense.

Another English example is the word *housewife*. The Old English compound word *hūswīf* has, through regular sound changes over time, been altered to hussy, whose meaning has shifted considerably. A newly coined compound, housewife, now has replaced the word hussy to convey the original meaning of the old word, and hussy is used with the new pejorative meaning.

Again, Speck 1978 provides a Texmelucan Zapotec example. Note the semantic shift demonstrated in (10):

(10) [bidr] means 'bottle', borrowed from the Spanish word vidrio, meaning 'glass (substance)'.

An even wider meaning shift has taken place in the Isthmus Zapotec borrowing of the Spanish word seguro, meaning 'sure'. In Isthmus Zapotec, [seguru] means 'tal vez (maybe)'!

Several questions arise: How should green, kaas and lo, wife and Weib, and bidr (and other such words with extended meanings and semantic shifts) be glossed? What if the multiple meanings are paralleled in the target and gloss languages? What if they are not?

Simple awareness of these problems is half the battle in solving them, since that awareness will stimulate the gathering of careful glosses and discussion of shared components of meaning in words that are semantically related but not quite synonymous. True cognates should be grouped together in one record whose gloss best represents the shared semantic components. Specific meanings in the individual languages being compared should be entered into the data base as secondary information, if they differ from that main gloss, and cross-referencing needs to be done, as discussed in Chapter 6 and exemplified in Appendix D.

Two possible ways of dealing with cross-referencing in situations such as the wife, Weib example discussed earlier are demonstrated below. Each is appropriate at a different stage of the development of the data base.

Initially, separate partial records could be set up, organized by English gloss, and cross-referenced to each other using the "\xs" field which is intended for semantic cross-referencing, as shown in (11), (12) and (13) below.

```
(11) \eg  woman
     \xs
     \Eng woman
     \Ger Dame

(12) \eg  spouse, female
     \xs  cf. woman (pejorative)
     \Eng wife
     \Ger Frau

(13) \eg  woman (pejorative)
     \xs  cf. spouse, female
     \Ger Weib
```

This system focuses on the gloss, as is appropriate before much sorting of cognate sets can be done.

However, re-sorting with a focus on cognicity will eventually be desirable, putting the data forms which really do seem to be cognates together in sets, as shown in (14) and (15). In this system, the two cognate forms are paired in one record, and specific glosses are given in each language data field to show how the individual glosses vary from the more general title gloss of the whole record:

```
(14) \eg  woman
     \Eng wife (female spouse)
     \Ger Weib (pejorative)
```

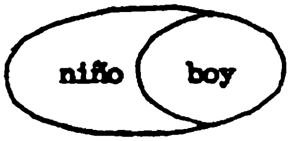
In (15), a separate record containing only cross-referencing information to faciliate finding the data, but not containing any actual data, has been included.

```
(15) \eg  spouse, female
     \xs  cf. woman
```

3.2.5 Non-coextensive glosses

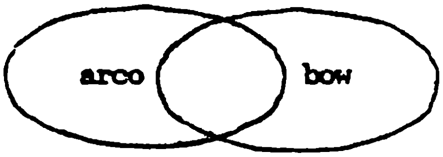
Words that share commonality of meaning between languages may not overlap completely in all components of meaning. One language's word may have more elements of meaning than another language's rough translation of that word, e.g., Spanish niño and English boy. Niño can mean 'child' in a generic sense in which gender is irrelevant, or it can mean specifically 'male child', as boy does. A diagram comparing the two words might look like this:

(16)



Alternatively, two words may partially overlap, sharing one or more components of meaning but not others, e.g., Spanish arco and English bow. Arco means 'bow (as for archery)' and 'arch', while English bow has the meaning of 'looped knot, as of ribbon' as well as the archery meaning, but not the arch meaning. The relationship between these meanings of these two words can be diagrammed as shown below:

(17)



Two words' meanings will rarely overlap completely, especially between two languages whose speakers are culturally dissimilar. I hesitate even to offer examples of such pairs, since there may be figurative meanings in either language of which I am unaware. Some possible English-Spanish examples include January, enero; and twenty, veinte.

To the extent that two words do not overlap perfectly in meaning, their glosses are non-coextensive. This section discusses various instances of non-coextensiveness, and suggests ways of handling them.

3.2.5.1 Generic to specific

In many semantic domains, words can be organized on a hierarchy from generic to specific. Of course, such hierarchies are not always identical between languages. For example, English has a generic word, *cactus*, as well as more specific terms for various individual species of cactus, e.g., *teddy bear cholla*, *sahuaro*, *organpipe*, etc., some of those terms being loans from Spanish. In contrast, Seri, an indigenous Mexican language spoken in Sonora, has no generic term for cactus, only terms for the individual species (Felger and Moser 1985:245ff):

(18) coote, sea - teddy bear cholla (p. 266)
mojépe - sahuaro (p. 247)
ool - organpipe (p. 258)

How could a set of words meaning 'cactus' be set up including both English and Seri? Presumably it could not be done; instead, terms for the individual species would be paired. One data record could be titled 'cactus', containing cross-references to the separate records containing the data for the individual species.

(19) \eg cactus
\xs Cf. sahuaro, organpipe....

Of course, the more languages that are being compared, the greater is the potential for glossing problems. When partial cognate sets can be formed, they should be preserved as sets and not ignored despite their incompleteness.

3.2.5.2 Kinship terms

Kinship terms provide good examples of non-coextensive glosses. Consider the meanings of the English words *brother* and *sister*. Ego's *brother* is ego's parents' male child, and ego's *sister* is ego's parents' female child. Ego's age relative to the brother or sister is irrelevant to the term chosen to refer to him or her, as is ego's gender. In contrast to this, however, Proto-Oto-Manguean sibling terms proposed by Merrifield (1981:21) have the following meanings:

(20)	*tu, *nu, *yu	man's elder brother
	*kwaHn, *kaHn	woman's elder sister
	*nsi-?ya	cross-sex elder sibling
	*kuHn	sibling
	*nsi-kihn, *kihn-si	younger sibling

In this situation, if the sibling being named is older than ego, ego's gender is relevant and so is the sibling's gender. However, if the sibling is younger than ego, then neither's gender is reflected in the referent term.

A similar disparity of relevance occurs between English and Proto-Oto-Manguean grandkinsman terms. Clearly, English has four terms: *grandfather*, *grandmother*, *grandson*, *granddaughter*. In each, ego's gender is irrelevant but the referent's gender is relevant, and the direction of lineal descent is reflected in the term chosen. In contrast, the proposed Proto-Oto-Manguean terms (Merrifield 1981:20) are as follows:

(21)	*seh, *hkeh	grandfather, grandson
	*Ynsan, *nan	grandmother, granddaughter

Here, gender of the referent but not of ego is reflected in the term chosen, and the direction of lineal descent is irrelevant. Thus, grandfathers and grandsons use the same reciprocal term in referring to each other, with the result that ego uses the same term to refer to both his own grandfather and his own grandson.

It is clear from the above two examples that different kinship systems group the components of meaning in different combinations. This poses a potential problem for glossing kinship terms. As was the case with semantic shifting, all these words should be grouped together in sets and given a general gloss that best represents the most common thread of meaning that they share (e.g., 'male grandkinsman' for the first line of (21) above), and individual glosses for each language should also be included if they differ from that main one. Again, cross-referencing should be done as explained in Chapter 6 and exemplified in Appendix D. In the 'grandfather' example above, the cross-referencing could be done by means of a special record, titled "grandfather" that contains nothing but the title gloss and the semantic cross-reference:

```
(22) \eg  grandfather
     \xs  cf. grandkinsman, male
```

Such a cross-reference indicates that the data forms themselves are found in the record titled "grandkinsman, male", but allows someone looking for a specific form glossed 'grandfather' to know where to look to find it.

Naturally, the non-coextensiveness of components of meaning can occur in vocabulary other than kinship terms as well. Other such problems should also be dealt with in the way described in this section.

3.2.6 What to do with incomplete sets

What should be done with incomplete data sets? In a word, keep them! They can still be useful in analysis, though incomplete, and can also serve to trigger ideas for further data gathering; other field linguists can get ideas and motivation from seeing "their" gaps and seeking to fill them in.

A special type of data set occurs when responses to a particular gloss fall into two or more cognate groups rather than all being cognate to each other. For example, the following is a list of Zapotec forms for the gloss, 'woman'.

(23)	Ate	ni'ula M-M-H
	WIx	'muilʌ
	Ytz	'no?ol(ə)
	Ylg	no?ol(e)
	Cho	nikula
	Rin	ni'kula
	Lac	una?a L-L-L
	SZm	ku'nã? L-L
	Chi	ku'n:a?:a 'LF
	Glv	ku'na?
	Mit	ku'na?a
	Alb	pi(?)nku'na? L-L-H (familiar)/penku'na? H-L-H (respect)
	GvH	nkwna? R
	Ist	ku'na?a L-L
	Qgl	u'na?a L
	Qui	kwna?a L
	Xan	u?una
	Xng	una?a
	Ama	nkwna?a

At least on superficial perusal of this list of forms, it appears that there are two cognate sets here, one involving the first six forms, and another involving the rest.[8] When this happens, synonyms (perhaps archaic forms) should be sought that are cognate to one of the other forms. In this way, one data set can be filled out into several cognate sets. For example, cognates to the forms given by the first six languages in (23) could be sought from the other languages, and vice versa. Thus, two cognate sets could be formed for the one gloss, 'woman'. Of course, there usually are minor shades of difference between the meanings of the two "synonyms", too. Cross-referencing between the various sets must be carefully done, as shown in Appendix D.

Notes

[1] Data presented here and elsewhere throughout this chapter are in the orthography used in the source(s) cited. Tone is not relevant to this discussion and has therefore been omitted, for simplicity. If a source is not specifically cited, then the data come from the questionnaires which I gathered from the various Zapotec fieldworkers, and can be assumed to be in the technical orthography described in Chapter 6 of this thesis.

[2] As mentioned in the introduction, computers can save huge amounts of time in such processes. In this case, a "fuzzy search" program could look through the data (in a specified set of records) for other words that look similar to a given data form. To do this, it must be supplied with phonological features for each segment in the language, and be told how many of those features may be varied, and by how much if the values are not simply binary. For example, if the user is looking for forms that are cognate to *bello*, he wants to look at forms beginning with [p, b, ᵽ, ƀ, m, w], etc., but not [n, k, r], etc. (such features as [-syllabic, +labial] would be used to specify this), followed by a vowel like [e, i, ɛ, ɩ], etc., but not [u, o, a], etc. (such features as [-consonantal, -back] would be used to specify this). In other words, enough features must be specified to the computer that it can intelligently search for suspicious pairs at each point in the word. These features will, to some extent, be language specific, i.e., there is need to specify which are relevant to a particular language study. Some way must also be found to allow for extra segments and syllables, as well as omissions.

The computer could then be told to list all the words that meet the specifications together in one place, for the compiler to study and compare. Of course, he will then make the ultimate decisions regarding what words are and are not cognates in a set.

"Spell" programs to locate misspelled words in a text utilize some ideas related to this; such programs could be used to begin designing software to meet this need. Exact feature specifications, and detailed software design to handle this problem, are beyond the scope of this thesis. I have included this as a suggestion for future work.

[3] See Section 4.2.1 for an explanation of the parenthesis notation used here.

[4] Some sort of Dewey decimal numbering system might work, but, because it involves a hierarchy of organization, it would necessitate a corresponding semantic theory and an analysis.

[5] Please note that the objections mentioned here to Buck's chosen format are based on the better options made possible now by computer technology. His organization of the data was excellent for 1949, given that expandability without retyping and renumbering was not feasible before computer technology became available.

[6] This numbering method was suggested by Margaret Langdon, during a personal discussion regarding her Yuman project, yet to be published.

[7] Buck (1949:vii) states that the 'unripe' sense of *green* is actually primary, being derived from the root of *grow*, and that the color sense is secondary, based on the color of growing vegetation. Whichever way it actually went, this is a good example of semantic shifting.

[8] Alternatively, it could turn out that all these terms really are cognates to each other, the *l* in the last syllable of the top six forms corresponding to the *n* in the last syllable of the rest.

Chapter 4

Data Selection

In deciding what form of data to use in a data base, various possibilities for the morphological form must be considered. Section 4.1 addresses whether to use roots, stems or inflected forms. Section 4.2 discusses special problems regarding data forms, such as free (or situational) variation, elision, culturally-determined vocabulary differences, suppletion, causatives, and possession of nouns. Section 4.3 addresses two further questions regarding kinds and amounts of data to gather.

4.1 Roots, stems or inflected forms

Before deciding what form of words to use in the data base, it is necessary to understand what the three options are--roots, stems, and inflected forms--and what each of those terms means.

Crystal (1985:267) defines the term "root" as "the base form of a word which cannot be further analysed without total loss of identity...it is that part of the word left when all the affixes are removed." Thus, _bird_, _tangle_, _ask_ and _strong_ are roots, but _birds_, _entangle_, _asked_ and _strongly_ are not, since they contain derivational and/or inflectional affixes.[1]

In contrast, a "stem" is any word form to which inflectional affixes can be attached. Crystal (1985:287) outlines three types of stems: a "simple stem" consisting solely of a single root morpheme (e.g., *man*), a "compound stem" consisting of two root morphemes (e.g., *blackbird*), and a "complex stem" consisting of a root morpheme plus a derivational affix (e.g., *manly*, *unmanly*, *manliness*).

Inflected forms usually involve addition of inflectional affixes (see note 1) to the stem, but new lexical items are not created by inflection. Sometimes there can be significant modification of the stem itself in inflection, as in the case of English *man*, *men* and the *f*-final nouns that add voicing to the final *f* in the plural form, e.g. *wife*, *wives*; *leaf*, *leaves*.

With these definitions in mind, one must decide what form of each word to cite in the data base. The choices are roots (e.g., *man*), stems (e.g., *manly*) and inflected forms (e.g., *men*, and *man's*). Of course, most roots are also stems. For the ZDB, I propose to use primarily stems. Excluding stems that are not roots (e.g., *manly*) would omit words that seem to be truly separate lexical items.

I recommend against the inclusion of inflected forms (except for verbs) that are predictable by simple rules, for two reasons: First, their inclusion would greatly augment the size of the data base, without significant improvement in the quality of the data, i.e., the extra data would simply be repeating "more-of-the-same" rather than providing anything for new insights into the languages being studied. Second, if the inflectional affixes vary from language to language, but are included in the data base on stems that are comparable, false sound correspondences can be assumed which could lead to invalid reconstruction. For example, if a set of noun stems are cognate, but possessive affixes in the various dialects have been borrowed from various surrounding languages, the segments of the non-cognate affixes could be compared along with those of the stems, yielding confusing and wrong assumptions regarding sound correspondences.

However, for the ZDB, three inflected forms of each verb stem need to be included, due to a peculiarity of many Zapotec verb stems to change depending on the aspect being used. Such variation is inflectional, but not predictable by simple rules. (See Section 4.2.4 for a more complete discussion of Zapotec verb suppletion.) Presumably, similar verbal complexity exists in other languages as well. The decision of whether or not to include inflected forms should be made anew for each comparative data base project being contemplated, focusing on the characteristics of the languages involved.

4.2 Special problems

This section addresses some of the special problems that arise in deciding what form of a word to use in the data base of a comparative project.

4.2.1 Free or situational variation

In free variation, one word can take two or more forms, both (or all) of which are completely acceptable, though perhaps their use is governed by non-linguistic factors. There is no phonological reason for the variation, and a speaker can choose to use one form one time and another the next. Such variation is especially obvious when comparing fast speech with careful speech. Often "free" variation is actually conditioned by situational factors such as relative ages of speaker and hearer, formality of the speaking context (e.g., formal public lecture versus casual conversation), etc. The variants are not multiple words that are synonyms, but rather two or more forms of the same word. Examples in English include *equatorial* and *economic*, in which pronunciation of the first vowel can vary between [i] and [ε], with no change in meaning, and *with*, in which pronunciation of the final consonant can vary between [θ], [ð] and even [t], with no change in meaning.

A few examples from various Zapotec languages follow (the "∾" symbol between data forms indicating situational or free variation):

(1)	Albarradas - Kreikebaum	
	sep'tyempr (L-LH) ∾ 'styempr (LH)	'September'
	kihk: (L) ∾ kik: (LH) ∾ kihk: (H)	'head'
	ya?s ∾ nya?s	'black'
(2)	Yatzachi - Butler	
	'kakwa? ∾ 'kaokwa?	'I will eat'
(3)	Choapan - Lyman	
	š:na ∾ šna	'red'

Which form of these various words should be used for the data base? What should be done with the rest of the forms? Three possibilities exist for answering these questions: 1) decide which form is most standard and ignore the rest; 2) choose one, perhaps because it is most standard, and list the others as secondary; 3) when there is regular free variation, include only one of the possible variations and, in an appendix to the ZDB, explain what the variations could be. These possibilities are discussed below.

Choosing one "standard" form and discarding the rest seems to be a poor choice, for the obvious reason that it involves loss of valid data. It is also very possible that the one form chosen as "standard" would turn out not to be standard at all after further study, or after the language has evolved a bit during twenty more years or so.

Choosing one "standard" form and listing the others secondarily with it is a better choice, since all of the forms are still being preserved for future use. This can be done in two ways, using parentheses or slashes. As an example of the parenthetical method, the Yatzachi words in (2) above could be listed as 'ka(o)kwa, the o in parenthesis indicating that its inclusion in the word is optional. Only the part of the word that can optionally be omitted is placed in parentheses.

However, this method will not work for the Albarradas example above, in which the variation involves substitution of one tone for another, and not merely omission of a segment. Thus, a second method is needed. I propose the use of slashes. Albarradas 'head' would be listed as kihk:[L/H]/kik:[LH].

I will probably use the parentheses form when possible, since it is slightly more space efficient and shows at a glance where the variation occurs between the forms. Thus, I would use the slash method only when the variation involves substitution rather than omission. However, one could also choose to use the slash method exclusively, for consistency and ease of explanation. (In that case, ka(o)kwa becomes kakwa/kaokwa.)

The third possibility is to explain regular free variation rules in phonology write-ups for each dialect, which will be included in an appendix to the data base. For example, Pickett says that every lenis obstruent fluctuates in voicing in Isthmus Zapotec. Rather than include two or more forms of every Isthmus word that contains lenis consonants, this fact of Isthmus phonology can simply be explained in the Isthmus phonology write-up in the appendix.

4.2.1.1 Adaptation for bilingual dictionaries

In the event that the data is adapted at some future time to create bilingual dictionaries for individual dialects (e.g., Choapan Zapotec/

Spanish), each different form will need to be listed in its respective alphabetical location, with cross-references to make all the forms equally accessible without knowledge of which form to consider "standard". For example, if someone is looking for the 'kaokwa form, and it is listed alphabetically only as kakwa/kaokwa, he would need to know that the o can be omitted, and look under 'kakwa. This problem would be greater with the Albarradas word ya?s/nya?s (see (3) above), since ya?s and nya?s are further separated in an alphabetical listing. A person could accidentally stumble on 'kakwa while looking for 'kaokwa (or vice versa), since they happen to be near each other in the alphabetical listing. Not so with ya?s and nya?s.

Thus, all the various forms should be listed in each one's alphabetical location. For example, the same Yatzachi words in (2) above could each be listed in their respective places in the alphabetical listing of words, as 'kakwa and 'kaokwa.[2] Each listing would give a reference to the other in one of two ways: ''kakwa, cf. 'kaokwa' and ''kaokwa, cf. 'kakwa', or simply ''kakwa/'kaokwa' and ''kaokwa/'kakwa'. Despite the extra work and complication involved, and the little bit more space that it requires, I consider this the best choice, since it preserves more data, with maximum accessibility. Note, however, that this redundancy of listing is only necessary in a single-dialect bilingual dictionary (e.g., Choapan Zapotec/Spanish), not in a comparative dictionary, where the modern Zapotec reflexes are not the main entry titles anyway. Since the data is listed under the gloss or the proto-form, all the variations in an individual dialect will already be grouped together on that dialect's data line for easy perusal.

4.2.2 Elision

French exemplifies the problem of variation in form caused by elision. Consider the following two French forms, written in standard French orthography:

(4) mes livres 'my books'

(5) mes amis 'my friends'

In (4), *mes* is pronounced [me], whereas in (5) the *s* is pronounced with the vowel that initiates the following noun, [mez]. Which form of the word *mes*, to be glossed 'my, plural', should be used, [me] or [mez]? This type of elision is very common in French. Nearly every time that a word written with a final consonant is followed by a vowel-initial word in the same clause, that consonant is pronounced. Otherwise, it is omitted.

Obviously, if the language has a standardly accepted orthography, as French does, the word could simply be listed as commonly spelled in that orthography, in this case as *mes*. Anyone could easily look in existing studies of French (a few of which should be referenced in the bibliography of the data base for the comparative project that includes this French data) to determine the pronunciation of the word *mes*, or, if desired, the compiler of the comparative data base could include a chapter on the orthographic conventions of each language included. This option has the drawback to be discussed in Section 5.1, that orthographic representation does not reflect the phonetic variation between dialects that makes a comparative study interesting.

Also, obviously, many languages exist for which there is no established orthography yet. If an elision problem were to arise in such a language, two methods could be used to handle this problem. As discussed above, both forms could be listed, (in their respective alphabetical locations for a single-dialect bilingual dictionary) to provide maximum accessibility to an unknowledgeable user of the data base. Some sort of brackets would be needed to indicate that there was the possibility of variation in form (e.g., me[z]). Note, however, that regular parentheses should not be used in this situation since they imply free variation rather than conditioned variation. The use of brackets versus parentheses would need to be carefully explained somewhere in the introduction to the data base. In addition, an explanation of the rules conditioning inclusion or exclusion of the bracketed consonants would need to be given.

The other possibility is to adopt an orthographic convention rather than using the modified phonetic transcription discussed in Section 5.1. For example, 'my, plural' could be listed simply as mez (choosing a z to represent the phonetic pronunciation of the consonant that occurs preceding a vowel-initial word). Again, rules governing the deletion of the z would need to be given, presumably generalized to cover deletion of other such final consonants as well. This second option is the one probably taken in developing a real orthography. (Observe that French words are spelled with those consonants present, even though they are more frequently absent than present in actual pronunciation.)

In Atepec Zapotec (Nellis 1983:352), vowel-final verb stems lose or change the vowel before a vowel-initial clitic pronoun. For example, (6) below contains two forms of each of several verbs. The first form listed is the stem form, and the second form has the third person singular morpheme -ā added, with resultant variation or deletion of the final vowel in the stem. (Tone has been omitted from this data, since it is not relevant to the discussion here.)

(6)	uhuia'	divertirse (have fun)
	uhui'ã	él se divertirá (he will have fun)
	egu'u	meter de nuevo (put in again)
	egu'ã	él lo mete de nuevo (he puts it in again)
	inne	hablar (speak)
	inniã	él hablará (he will speak)
	go	comer (eat)
	guã	él lo comerá (he will eat it)
	cueda	esperar (wait, hope for)
	cuedaã OR cuedã	él esperará (he will wait, hope for)

In the first two examples, the last vowel simply drops before -ã. In the third and fourth examples, final e and o change to i and u respectively before -ã. In the fifth example, the stem-final a can either drop or remain unchanged. Such variation is a good example of elision, but Zapotec elision is different from French elision in that the form spoken in isolation is the underlying form, whereas in French the elided form is the underlying form. Since this Zapotec variation can be predicted easily by rules, I propose to list only the underlying stem forms (the upper form in each pair in example (6) above), without the variations on the final vowel.

4.2.3 Culturally-determined vocabulary differences

It is not uncommon for a language to contain vocabulary words that only certain types of people in the culture can use. For instance, the men's term for something could be different from the women's term for it, with the result that, e.g., a woman would only use the men's term when quoting a man. Certain words could be used only by shamans and be taboo for use by ordinary laymen. Age of the speaker or the addressee could also determine what vocabulary would commonly be used. Such use limitations do not pose a major problem regarding which form to choose for inclusion in the data base. Since each of these words is a valid part of the language, albeit restricted in its use, all forms of such vocabulary items should be included if at all possible, and the glosses should include information specifying their use.

For example, in the Albarradas Zapotec word for 'child', the initial consonant and the tones vary, depending on whether the speaker is male or female.[3] Thus, these words must be listed as separate entries, with careful glosses to specify usage:

(7) šin'to? H-R 'child, men's speech'
pin'to? L-L 'child, women's speech'

Alternatively, if it turned out that only one or two dialects made this gender distinction, both forms could be merged in with the general gloss, 'child', with the usage specifications included beside the form itself:

(8) šin'to? H-R(men's speech)/pin'to? L-L(women's speech)

In this option, the gloss 'child' is only listed in the \sg field. I find this alternative less attractive because, although the field for the form does have unlimited length, the printed column might not be of unlimited width, and some of the entry could get chopped off in printing, resulting in a loss of information. (Long lines could be wrapped in printing, but this is a complex enough process for chart formats that not all software has this capability.)

4.2.4 Suppletion

The presence of suppletive allomorphy complicates the decision regarding which form to include in the data base. Obvious examples of suppletion in English are go versus went, and in Spanish, all the various forms of the verb *ser*. Which form of these verbs should go into the dictionary?

A monolingual English dictionary (Morris 1969) lists *go*, and includes *went*, *gone*, *going*, *goes* in the entry for *go*. *Went* and *gone* also are listed separately in their respective alphabetical locations.[4]

In the *Pequeño Larousse Ilustrado* (1964), a monolingual Spanish dictionary, the verb *ser* is conjugated under the alphabetical listing of *ser*, but the individual forms are not also listed alphabetically. In contrast, Williams 1978, a bilingual Spanish and English dictionary, lists every form of *ser* in its respective alphabetical location, and defines each grammatically.

Zapotec provides further good examples of this problem in its many radical stem-changing verbs. Since my data base is to include mainly stems and roots, I will not include every inflected form of these irregular verbs. However, the three forms of the stem that result from different aspectual prefixes should all be listed, along with aspectual prefixes, since there is not agreement among Zapotec analysts as to where morpheme breaks occur. They could be listed without specific gloss, e.g., simply as *u'ta:na*; *'θa:na*; *ri'ta:na*, since a convention could be prescribed in the introduction to the data base, explaining the order of the three forms for every verb, e.g., habitual, completive, potential. If one of those three forms was not available for some reason, a hyphen should be used to indicate which one was missing, e.g., *u'ta:na*; *--*; *ri'ta:na*. (Note that, for the verbs manifesting no stem variation, only one form could be listed, followed by some code letters, perhaps "nv", to indicate that there is no variation, rather than that some of the forms were just missing. Alternatively, these invariant verbs could be listed three times in the row, just for consistency and to avoid yet one more explanation of a convention.)

Examples of stem-changing verbs from Choapan Zapotec are shown in (9) below, each listed on one line as they would be on the Choapan line in the data base, in the order completive, potential, habitual:

(9) 'andar' (walk)
uta; t:a; rta

'moler' (grind)
ptu; tu; rutu

Any other suppletive forms of stems should also be included in the data base, probably in the same form as is outlined above for verbs. If morpheme breaks are controversial, then prefixes should also be included, as for verbs above.

4.2.5 Causatives

Adding some sort of affix to a verb can create a causative verb that is related to the original verb. For example, z or gu added before the first vowel of some Yatzachi Zapotec verbs (Butler 1980:129 and 127, respectively) creates a causative:[8]

(10)	čey	'se quema' (burns)
	čzey	'lo quema' (burns it)
	čao	'come' (eats)
	čguao	'da de comer, al animal' (feeds, to an animal)

(Note that data in (10) and (11) are written with Butler's orthography rather than the technical orthography recommended for use in the ZDB.)

However, not all causative verbs are related to their non-causative counterparts in such obvious ways. Often in Zapotec, there is an alteration of the initial consonant, or even a completely different stem for a causative verb (Butler 1980:129-140). For example, the following two verbs from Yatzachi Zapotec (Butler 1980:134 and 139 respectively), show consonant alteration in the first pair (*z* and *s*) and radically different stems in the second pair (*zo* and *bec*):[4]

(11)	čeza?a	'salir de nuevo' (leave again)
	čosa?a	'hacer salir de nuevo' (cause to leave again)
	čzo	'estar' (be located)
	čbec	'ponerlo' (put it, place it)

Such a complex causative situation does not pose a problem in a data base. When there is alteration in the stem, both stems should be present in the data base, glossed as two separate verbs. However, they should be cross-referenced so that the semantic relationship between the two is not lost. For example, for the *salir*.../*hacer salir*... pair shown in (11) above, a comment field for *salir de nuevo* could contain "cf. 'hacer salir de nuevo'", and likewise the *hacer salir de nuevo* comment field could contain "cf. 'salir de nuevo'". Such cross-referencing would be more important in pairs like 'sings' (*čol:*) and 'plays, as a musical instrument', (*čgol:*) where the causative relationship in the meaning of 'play--to make an instrument sing' would not be immediately obvious from the Spanish glosses, *cantar* and *tocar*.

4.2.6 Possession of nouns

Many Zapotec nouns, especially body parts and certain relatives, are obligatorily possessed. Such nouns always appear with a pronoun indicating the possessor. It appears that, at least for Atepec Zapotec (Nellis 1983:343) the third person singular form (which would always be accompanied by the pronoun *bi* when actually referring to 'his or her....'), is the stem form, as shown in example (12) below:

(12)	l.le'e bi	'el estómago de él (his stomach)'
	laya' bi	'el diente de él (his tooth)'
	yithua bi	'su nieto (his grandson)'

(Note that all data in (12), (13) and (14) utilizes Nellis' practical orthography, rather than my proposed technical orthography.) However, the first person singular form of such nouns manifests a change in the stem, thus necessitating inclusion of both forms in the data base. For example, (13) below contains the first person singular forms of the three nouns given in example (12). Compare the stems to see how much they differ. (The differences reflect the same morphophonemic changes noted in Section 4.2.2.)

(13)	l.li'a'	'mi estómago (my stomach)'
	laya'a'	'mi diente (my tooth)'
	yithua'	'mi nieto (my grandson)'

Such nouns should appear in the data base, both forms in one record in a predetermined order, e.g., third person stem form l.le'e, followed by the first person form, l.li'a'. Individual glosses, e.g., 'estomago' and 'mi estomago', need not be included.

In contrast, however, other nouns are only optionally possessed, possession being indicated in Atepec Zapotec by the preposition qui' 'of' following the head noun and preceding the possessive pronoun or noun (Nellis 1983:343). Example (14) below demonstrates this:

(14)	bia' qui' bi	'su caballo (his horse, lit., horse of him)'
	nana qui' Betu	'la mamá de Pedro (Pedro's mother, lit., mother of Pedro)

Apparently such noun stems do not vary depending on whether or not they are possessed. Thus, only one form of such nouns needs to be listed in the record, e.g., bia' 'caballo (horse)'.

4.3 Kinds and amounts of data to gather

What kinds of data, being most useful in comparative work, should be gathered for such a project? Core vocabulary that is the least likely to be borrowed from another language is the best for comparative work. Such vocabulary is culturally relevant to the indigenous speakers of the language, and probably within the command of children. It deals with concepts that are universal to the human experience. Hockett (1958:529) discusses this basic core vocabulary as follows:

> There are certain recurrent things and situations, or kinds of things and situations, for which every community of human beings, regardless of differences of culture or environment, has words. The words used by a given human group for these omnipresent things and situations constitute the basic vocabulary of the group's language. It should be noted that "basic vocabulary" is defined in semantic terms.

Anttila (1972:397) explains that core vocabulary is very useful in comparative linguistics:

> ...basic core vocabulary is very valuable in giving a quick elicitation list for those items where loans are least likely, and thus one can start comparative work conveniently from here.

Speaking specifically, then, names of the weeks and months are poor choices, as are very large numbers such as "100". However, small numbers, function words (e.g., in English, the, and, for, from, etc.), names of common foods and plants and animals that are native to the area, weather terms, verbs and adjectives describing everyday life, etc., are likely to be truly core vocabulary, not borrowed from a dominant national culture. It is important to seek data in a wide variety of semantic domains and from varied grammatical categories.

Of course, non-core vocabulary should be added, too, as long as the size of the data base does not come to exceed what is workable for its intended purposes. The ZDB's size can be very large, indeed.

Related to the question of kinds of data to gather is that of quantity. Just how much data can realistically be elicited from the field workers, and how large a data base is needed for purposes of reconstruction?

How much cooperation one can achieve in gathering data from field workers depends on how carefully one eliminates unnecessary busy-work from the task. For this reason, I propose that the compiler of the data base ask field workers only for the form in which they already have their data stored, e.g., printouts from their computer files or xerox copies of card files, etc. The compiler is then to place these forms, as received, into the appropriate fields in each record. The compiler will derive phonetic or phonemic forms (whichever one the original was not) from these original forms, and verify them with the field workers. (See Section 6.1.3 for discussion of these fields.)

This approach appropriately shifts most of the time and effort burden onto the compiler of the data base. All that the field workers have to do is share data directly out of their files, and later check the compiler's derived phonetic forms. Compared to deriving all the phonetic forms themselves and copying those phonetic forms as well as their original forms onto a compiler's questionnaire, field workers should find this a small task. As such, a much larger quantity of data can reasonably be requested by the compiler.

As for just how large a data base is needed for purposes of comparative work and recontruction, that is an open question and will depend to a large extent on the languages involved and the goals of the project. Terrence Kaufman (personal communication) suggests a thousand or more cognate sets as a minimum for significant work in reconstruction, with three or four hundred perhaps sufficing for student papers. Clearly, many more sets than that would need to be gathered to arrive at a thousand actual cognate sets. Presumably interesting sociolinguistic comparative studies could be done with much fewer than a thousand data sets, and in such studies there is less need for actual cognate sets. In any case, as much data as possible should be acquired and processed into the data base in the desired forms.

Notes

[1] Crystal (1985:89) states, "Derivational affixes change the grammatical class of morphemes to which they are attached." What is not clear from Crystal's definition is that the *in-* of *indefinite* is also a derivational affix, even though it does not change the grammatical class of the word to which it is added. Bartholomew (1983:41) also discusses this problem.

[2] See the sample Albarradas Zapotec/Spanish index in Section 6.3, example (12), showing each variation listed alphabetically.

[3] In Zapotec the difference between men's and women's speech does not involve taboos. It is simply uncommon for a man to use the women's term except in quoting a woman, and vice versa.

[4] Actually, *going* is listed separately, too, not in the "continuative aspect" sense, but rather with various nominal and adjectival senses.

[5] The *č-* on all of these verbs is a prefix which Butler glosses as 'futuro (future)'.

[6] Perhaps verbs whose stems differ so radically from their noncausative counterparts should not properly be termed causatives at all. I call them such because that is how they are treated in Butler 1980.

Chapter 5

Phonological Representation of Data

Several decisions must be made regarding the phonological representation of the data in a working data base or a comparative dictionary. Section 5.1 discusses whether to use phonetic or phonemic forms, or practical orthography. Section 5.2 discusses special notational problems, such as stress, tone, nasalization, length, retroflexing and backing, fortis and lenis distinction, various vowel phenomena and special characters on the computer.

5.1 Phonetic or phonemic forms, or practical orthography

Another set of options arises in deciding what form to use for each word in the data base and dictionary. Having chosen, for example, to use stem forms, how should the linguist spell those stems? Three obvious possibilities present themselves: phonetic representation, phonemic representation, or practical orthography.[1] Each option has advantages and disadvantages, as discussed below.

Phonetic representation has, perhaps, the best potential for uniformity, despite a variety of dialects and data collectors. The true dif-ferences between the languages' modern reflexes should be apparent with use of good phonetic representation. Unfortunately, it also carries with it the problem of redundancy, i.e., all the phonetic characteristics of the languages being compared--both relevant and irrelevant to the comparative study--are laboriously reproduced in word after word, when they are irrelevant to the comparative study. For example, if all stops are aspirated or all sibilants retroflexed, there seems no reason to notate these characteristics; a brief phonetic statement discussing these facts should suffice, coupled with broad phonetic transcription that omits these fine details. Another disadvantage of phonetic transcription is that it demands extensive use of special characters which are cumbersome to produce on a computer. Use of a broad phonetic transcription reduces this problem somewhat, but does not eliminate it altogether.

Phonemic representation of data forms is another good choice to consider. Its biggest advantage is elimination of redundancy, which produces simplicity. For example, if a vowel is predictably nasalized when it precedes a nasal, and nowhere else, there is no need to write nasalization at all. If stress is predictable by a few simple rules (for example, if stress were always penultimate), it need never be notated in the data base. What is needed is a careful phonemic statement and listing of all phonological rules in the languages being studied. Crucial orderings of the rules, as well as some details of the rules themselves, may vary from dialect to dialect--indeed such variation of rules and ordering can be the cause of the very dialect differences being studied--so statements of the rules and their order for each dialect need to be included as well.

The biggest disadvantage of phonemic representation is the variety of assumptions available for phonemic analysis. When a large number of people are involved in a comparative project, and their training as linguists has been varied in amount, quality and goal, there is little hope for uniformity of product.

The third choice, practical orthography, has the same large disadvantages as the second: lack of quality control and non-uniformity of product. Since practical orthographies are based in large part on phonemic analysis (and then further complicated by, e.g., political or sociological considerations which could vary from language to language), it seems that their use in a comparative project would be useful only in two cases: a situation involving great cooperation, probably committee-work representing all languages concerned, and strong consensus of what orthographic conventions would best serve the whole language family involved; or a case like Classical Aztec, in which the data is available only from written sources in a standard orthography. In either of these two cases, the practical orthography would need to be carefully explained in the introduction to the data base.

An even greater weakness of using practical orthography in comparative work can be explained using English as an example. The various dialects of American English vary not in orthography, which is standard throughout the United States, but in phonetic detail of pronunciation. (Even the differences between American and British spellings, e.g. -or versus -our, are minor and do not reflect the greater current phonetic dif-ferences between spoken forms of American and British English.) Listing orthographically-written data for comparative work in English would be pointless since the very differences in pronunciation that would make the study interesting would not be detectable in the practical orthography. Use of practical orthography seems even more inappropriate when thinking of comparative work in the vastly different Chinese languages, that share the same characters to notate words with shared meaning even when their pronunciations are not even remotely similar to each other.

However, it does seem that, in some circumstances, it would be appropriate to include practical orthography in a secondary data field, i.e., broad phonetic form in the primary data field, with practical orthographic form also listed in a separate field which could be ignored most of the time. If the data is from a secondary source, and is written in that source in orthography, that orthographic form should be preserved somewhere in the data base, for one obvious reason: the compiler of the data base and the user of the practical orthography might have different understandings of rationale behind orthographic decisions made, and this could result in faulty transcription of the phonetic form from the orthographic form. Given this potential for error, both forms should be preserved, the orthographic form and the derived phonetic form. See Section 6.1.3 for a discussion of the "\-o" language fields, which implement inclusion of the orthographic form in a secondary field.[2]

Specifically, my proposal is to list the data in three separate forms, one in each of three fields. The first data field is to contain the data as received. The second is to have a form between broad phonetic and classical phonemic (although refer to Note 1 in Appendix D for a suggested way to include both phonetic and phonemic forms in one field). The third field is to contain a more abstract form reflecting further analysis. See Section 6.1.3 for further discussion.

Examine (1) below, as an example of the process described above for arriving at standardized phonemic notation for use in the "\-p" field. Two words from Western Ixtlan Zapotec, are shown first in "\-o" field form, and then in "\-p" field form.

(1) 'September' 'fifteen'
sɛp'tieƀře 'ci:nu
sep'tiembre 'cinu

The first line of each word is the form given on the questionnaire filled out by Grace Thiessen and Joan Smith. (I requested and received phonetic form.) Vowel height and length are carefully notated, as well as some details concerning the consonants. In the 'September' example, the lax [ɛ] vowels turn out to be predictable allophones of the phoneme /e/, so they are rewritten with an e̲ symbol. The bilabial fricative [ƀ] is an allophone of /b/, so it can be written simply as b̲, and all ř̲'s in Zapotec are flaps, so for simplicity they can be written simply as r̲. In the 'fifteen' example, the only phonetic detail that can be eliminated is the predictable vowel length.

The second form given for each word in (1) should be written in the "\-p" field of the data base, and the first form in the "\-o" field. (See note 1 of Appendix D for another option which combines both forms in the "\-p" field.)

5.2 Notational problems

Decisions need to be made regarding notation of a few special things. This section discusses how to notate stress, tone, fortis and lenis distinction, length, nasalization, retroflexing and backing, and various vowel phenomena, and how to represent special characters (non-alphabetic phonetic symbols) on the computer. The first four of these problems can be handled using standard ASCII[3] characters on the computer keyboard.

The remaining problems can be handled in any of three ways: First, Macintosh computers are specifically designed to handle special character needs and to be "user-friendly". Macintoshes are available in a wide array of models, which vary in speed and capacity (number and size of records that can be handled). Certainly, a model with a hard disk should be selected to handle a project of this size. If a Macintosh machine is available which is adequate in terms of these other factors without being prohibitively expensive, then it would be a good choice for a comparative linguistic data base. Data base software for a Macintosh is considerably more costly than comparable software available for other types of computers. Since the machines that are fastest are also more expensive, and since the whole line of Macintoshes tends to be more expensive than other types of computers, e.g., MS-DOS, a Macintosh would probably not be a cost-effective choice if a new computer were being purchased specifically for a project like the ZDB. However, if a suitable Macintosh were available, then with the right software it would be adequate for the job.

The second option is to use upper ASCII characters[4] with the MS-DOS operating system. An upper ASCII code number must be assigned to each individual special character needed. A person keyboarding the data with such a system can push a special control key, <ALT>, similar to the typewriter key that shifts to upper case letters, and then the three digit number code that corresponds to the desired special character. Special software must be written to display the character properly on the screen (in the way that it will print rather than the standard IBM extended ASCII character assigned to that number) and to print it properly. Tools are available to facilitate the writing of this software. Software exists that translates single-key keyboard input into the four-key sequence (the <ALT> key being held down throughout the typing of three number keys) needed to communicate the upper ASCII codes to the computer, greatly facilitating keyboarding of upper ASCII characters. That is, one key can be pushed, resulting in the computer receiving the entire upper ASCII code needed for one special character. One such program is KeySwap, developed by Al Reitz of the Summer Institute of Linguistics.

A committee of linguists in the Mexico branch of the Summer Institute of Linguistics, chaired by J. Albert Bickford, is currently developing special character support on MS-DOS computers, for use in archiving texts in the indigenous languages of Mexico. If an MS-DOS computer is to be used for the ZDB, it would be most practical to use their special character support system, since it will already have been developed and is customizable for any needs peculiar to the ZDB that may not already have been anticipated.

The third option is for use with UNIX or CP/M computers, which do not support upper ASCII characters. In this option, ways must be devised to represent special characters using only the standard ASCII characters, such as are available on an ordinary typewriter keyboard. This method is cumbersome, and requires extensive explanation of the conventions chosen, e.g., digraphs, trigraphs, non-alphabetic characters, etc., which often deviate considerably from conventions in common use in the linguistic world. However, since UNIX is a more powerful tool for automated processing of data than is MS-DOS, UNIX might be the best operating system to choose for a large linguistic data base. In addition, CP/M might be used for keyboarding by people who already own a CP/M computer and want to use it to commence data entry. Thus, there is need for proposals regarding special character conventions for UNIX and

CP/M systems. Sections 5.2.5 through 5.2.8 contain proposals for such conventions.

5.2.1 Stress

Unless it is totally predictable, stress should be notated in the data base, because it can be an interesting part of comparative work. In Isthmus Zapotec, stress usually falls on the first syllable of a native stem. However, there are quite a few loan words from Spanish, and some old compounds, which do not follow this rule, so it seems that stress should be notated for the ZDB. Thus, there is need for a convention for notating stress.

Several options present themselves, including use of an acute accent over the syllable peak (e.g., tíka) of the stressed syllable, use of a superior vertical stroke preceding the whole stressed syllable (e.g., 'tika), underscoring of the stressed syllable or syllable peak (e.g., $\underline{\text{ti}}$ka or t$\underline{\text{i}}$ka), and use of upper case letters in the stressed syllable (e.g. TIka).

Use of an acute accent is common for notating tone, so I reject that option to avoid confusion. Upper case letters are sometimes used for lax vowels (e.g., [I] to distinguish the lax vowel from the tense one, [i]) and in digraphs, so that option should be avoided to indicate stress. Underscoring was used by a number of the linguists who provided my Zapotec data, to indicate nasalization on a vowel. Thus, since they are among the most likely users of the data base, I reject that option for stress notation, to avoid confusion for them.

The most viable option remaining is the vertical stroke preceding the stressed syllable. The apostrophe, being a standard ASCII character, could be substituted for the vertical stroke to facilitate keyboarding, and there is no danger of it being confused with its more normal usage for ejectives (cf. Pullum and Ladusaw 1986:216) since Zapotec has no ejectives. Thus, I propose that, in the ZDB, stress be notated with an apostrophe preceding the stressed syllable.

5.2.2 Tone

Most linguists working with Zapotec languages report the presence of tone in their languages. Thus, a consistent way of notating tones must be devised. Common notational conventions for level tones include diacritics, numbers and letters:

(2)	high	mid	low
diacritics	´	‾	`
numbers	1	2	3
letters	H	M	L

These are combined in various ways for contour tones:

(3)	falling	rising
diacritics	^	ˇ
numbers	1-3 or 13	3-1 or 31
letters	H-L or HL	L-H or LH

or whole new symbols are introduced:

(4)	falling	rising
diacritics	⌒	˘
letters	F	R

First, I propose that tone be notated beside the word (e.g., 'tika H-L or 'tika 1-3) rather than directly over the syllable peaks of the words bearing tone (e.g., 'tíkà, 't$\overset{H}{i}$k$\overset{L}{a}$, 't$\overset{1}{i}$k$\overset{3}{a}$). This serves to separate tone from the rest of the segments, which will simplify the comparative study to faciliate concentration on the segments themselves.

It is common practice in a practical orthography to choose one tone, presumably the most common one, to be unmarked. For example, if there are three tones, and the mid tone is the most common one, only high and low tones would be marked, while all unmarked syllables would be assumed to bear mid tones. There is one disadvantage of not putting the tone directly over the words: one tone cannot be chosen to be unmarked in this way. However, this disadvantage is nullified by the fact that not all the Zapotec language data was reported with tone at all, so all tones would need to be marked anyway, to differentiate between those syllables for which tone was unknown and those containing what would have been the "unmarked" tone. In any case, it is a technical orthography rather than a practical one that is being devised here, and such shortcuts are usually not taken in technical orthographies anyway.

Placing the tone off to the side rather than over the words has the effect of eliminating diacritics from consideration, since they look very strange when not oriented to a specific vowel (e.g., //\/\\/\\/)! In addition, the symbols most commonly used for tone glides (circumflex and wedge) are not fully standardized in usage (cf. Pullum and Ladusaw 1986:224-226), leaving potential for considerable confusion.

Thus, the only choice remaining is between numbers and letters. I propose the use of letters for two reasons: 1) Which number represents high tone and which represents low tone is an arbitrary decision that must be made anew and explained for each project. To further complicate matters, most members of the Mexico branch of the Summer Institute of Linguistics, who are among the most likely users of the ZDB, have traditionally numbered tones using a different convention from that used by Africanists, among others. Thus, it would be especially difficult to remember whether, e.g., "1" is high or low. In contrast, using "H" for high and "L" for low is crystal clear and easy to remember.[5] 2) Consistency and space economy are desirable in a data base. When there is no contrast of level in contour tones, (i.e., when all falling tones fall from the same higher level to the same lower level, rather than having, for example, high-to-mid glides contrasting with high-to-low or mid-to-low glides), a one-letter designation, e.g., "F" for falling and "R" for rising, can be used for economy of space in the data base, rather than always having to type at least two numbers or letters, e.g. "1-3" or "13" or "H-L". Aesthetically, I prefer a consistent all-letter or all-number convention for tone notation, rather than mixing some letters and some numbers. Use of numbers for level tones and letters for contour tones is not consistent. To accomodate my desire for both economy and consistency, I propose the use of letters, using "H, M, L" for high, mid and low, and "F" and "R" for falling and rising. Since only three varieties of Zapotec report contrastive levels in contour tones,[6] I suggest using these one-letter designations for falling and rising tones in all the other Zapotec languages, and such symbols as "HF", "LF", "LR" and "MR" for the three that do have contrastive levels in contour tones.

Finally, I propose the use of hyphens between tone designations to indicate how the tones correlate with the syllables. Thus, the four-syllable Albarradas Zapotec word for 'thunder' would be written this way:

(5) ra'ča?aku'ša? L-H-L-L

It is easy to see that the first low tone belongs with the first syllable, ra, the high tone with the second syllable, ča?a,[7] the next low tone with ku, and the last low tone with ša?.

This use of hyphens also provides for differentiation between notation of, e.g. a high-to-low glide, HL, should this be needed, and a two-syllable sequence of high and low tones, H-L.

Such tone letters and hyphens, of course, are easily available on any of the three operating systems discussed in Section 5.2.

Two last comments regarding tone notation remain to be made. First, Robert MacLaury (personal communication) reports the presence of five contrastive level tones in Ayoquesco Zapotec. He suggests notating the highest and lowest tones using "primes". For example, H' would indicate ultra-high tone, and L' ultra-low tone. This type of notation could also be used for the preliminary phonetic form of data, in which there might be five phonetic levels of tone but only three phonemic levels.

Second, Joseph Benton reports (in personal communication) that Chichicapan Zapotec only has significant tone on the stressed syllable of the root. Thus, his data was reported in that manner. In the "\-o" field, I typed the tone off to the side of the word, with a stress mark preceding it, to indicate which syllable it belonged to. For example, in (6) below, meaning 'thunder',

(6) ku'si?iyu 'L

the tone marking indicates that the low tone belongs to the stressed syllable, si?i. In the "\-p" field, the stress mark before the tone could be omitted since Benton's analysis linking tone with stress would be explained in the introduction anyway.

5.2.3 Fortis and lenis distinction

Zapotec is noted for the fortis and lenis distinction that occurs on many consonants, fortis consonants being "stronger" and longer and often voiceless, and lenis ones being "weaker" and shorter and voiced. The two most obvious choices for notating this distinction are to indicate the difference by focusing on the contrastive length component, e.g., l: as fortis and l as lenis, or on the contrastive voicing possibility, e.g., p as fortis, b as lenis. Since the voicing contrast is not always immediately representable in a basic symbol (e.g., the [l] sound does not have an obvious voiceless counterpart such as exists for the [p/b] contrast), I propose to focus consistently on the length component, notating fortis segments as long, with a colon following the symbol for the segment, and lenis as short, using just the segment symbol alone. In pairs for which the voicing contrast is relevant, voiceless allophones occur quite frequently, but voiced fortis stops, fricatives, affricates and sibilants never occur (Nellis and Hollenbach 1980). Thus, I propose to use symbols for voiceless segments in those cases, e.g. t for lenis and tt or t: for fortis, rather than d and t. It seems perhaps misleading to notate [tt] as a single fortis t:. On the other hand, use of double consonants may be preferable in cases where fortis consonants result from morphemes coming together, perhaps with phonological assimilation, e.g., where [k] and [t] come together across a morpheme boundary, they appear as tt (Kaufman 1987). Use of a colon for fortis segments allows for a notational distinction to be made between a genuine fortis segment, t:, and two identical lenis segments that happen to occur together, e.g., on either side of a morpheme boundary, tt.

Further analysis of Zapotec will presumably reveal which notation is preferable. For this thesis, I have chosen to use the colon.

5.2.4 Length

If there is any contrastive length in Zapotec other than the fortis-lenis distinction discussed above, I propose to notate it with a colon as well, e.g., *a:* for a long segment and *a* for a short segment. If length occurs in combination with other modifications on a segment, the basic segment symbol should come first, then the symbol for the other modification, and finally the colon to indicate length (e.g., *e~:* indicates a long nasalized [e].)

The colon is a standard ASCII character, and thus can be used with any of the three types of operating system mentioned earlier. The next four subsections deal with notational needs for the ZDB which cannot be handled automatically using only the standard ASCII characters. Thus, in each case a proposal is made for how to handle the problem for use with UNIX and CP/M operating systems.

5.2.5 Nasalization

Some varieties of Zapotec contain contrastive nasalization of vowels, which thus needs to be notated. Of the various options available (underlining of the vowel, e̲, tilde mark above the vowel, ẽ, Polish hook underneath the vowel, ę, etc.), I propose to use a tilde. All of the options are equally good in terms of clarity, but the tilde is probably the most standardly used in the linguistic world to notate nasalization.

With MS-DOS, a separate upper ASCII number must be assigned for each separate nasalized vowel symbol. I propose using the number assignments being developed by J. A. Bickford's committee (see Section 5.2

above).

If a UNIX or CP/M system, which cannot support upper ASCII characters, is to be used, special characters must be represented using only standard ASCII characters. Since keyboarding is much easier without utilizing the control keys necessary to produce the effect of double-striking of a single letter, I propose to adapt the *ẽ* symbol by typing the tilde after the vowel rather than over it, e.g. *e~*. To facilitate alphabetizing, it is useful to keyboard many modifications after the main segment rather than before (see Section 5.2.8). For consistency with other modifications of segments, in which the keyboarding order matters, the tilde should come after the vowel rather than before the vowel. If it is desirable to print nasalized vowels with the tilde symbol over rather than following the vowel symbol, a consistent change operation can be done after keyboarding, converting them all to that format just for printing (e.g., inserting backspaces, etc.). Note that a comma, as a close approximation of the Polish hook, could be used instead of a tilde if a given keyboard did not contain the tilde character.[8]

5.2.6 Retroflexing and backing of consonants

An under-dot under the segment is the commonest way of indicating retroflexing and backing of consonants, e.g., a dot under an *s* indicates retroflexing, and a dot under an *x* indicates backing. Again, such symbols can easily be made for an MS-DOS system by assigning them upper ASCII number codes.

As discussed above for the nasalization tildes, for use on UNIX or CP/M systems I propose to adapt these symbols by placing the dot after the retroflexed or backed segment rather than under it, e.g., *s*. and *x*..

Again, a consistent change operation can be done after keyboarding, to insert backspaces and lowering, etc., so that the dots can actually be printed directly below the symbols.

5.2.7 Vowel phenomena

There are several different contrastive qualities that Zapotec vowels can have. Vowels can be simple (i.e., no glottal modification involved), checked (i.e., the vowel is cut off by a glottal stop) or laryngealized (i.e., the vocal cords are partially closed throughout articulation of the vowel, or the vowel is broken in the middle by a glottal stop.) Since these three vowel qualities are contrastive, a way must be found to differentiate them in notation.

For use with MS-DOS systems, I propose the use of a single vowel symbol for the simple vowels, a; vowel symbol followed by glottal stop symbol for the checked vowels, a?; and vowel symbol followed by a superscript glottal stop symbol for the laryngealized vowels, a$^?$. The glottal stop symbol and its superscript form are being assigned upper ASCII codes by Bickford's committee.

For UNIX and CP/M systems, the checked vowels can simply be keyboarded using the vowel symbol followed by a question mark, a?. However, notation of laryngealized vowels involves more complexity. Superscripts should not be used for UNIX systems, which work best with a clean ASCII file, i.e., use of only standard ASCII characters and no "invisible" keys such as backspace or superscript commands. Thus, some other notation system must be used. I propose to use the vowel symbol followed by either two question marks, a??, or a question mark and a capital V, a?V. These are very non-standard notations, chosen for ease in computer sorting and alphabetizing.[9] If the more standard vowel-glottal-vowel notation, a?a, were used, the computer would interpret that as a sequence of checked vowel followed by simple vowel. If a search were to be made for all words involving laryngealized vowels, it would be much easier to ask the computer to find all instances of "??" or "?V" than all instances of "a?a", "e?e", "i?i", "o?o" and "u?u". If the more standard notation, a?a, is desirable for printing, a consistent change can be done just for printing, that changes the V or the second ? to a copy of the vowel that precedes the glottal, i.e., all instances of a?? or a?V become a?a, and all instances of e?? or e?V become e?e by the same command to replace the second ? or the V with a copy of the segment that precedes the first glottal.

5.2.8 Other special characters

The preceding seven subsections have dealt with ways of notating non-segmental phenomena in ways that utilize as few non-alphabetic characters as possible. Some segmental phonemena can be handled on MS-DOS operating systems using specially designed print matrices which have been assigned upper ASCII number codes. For the ZDB, such segment symbols include č, š, ž, ɬ, ɛ, æ, ɨ, ə, ë, đ, θ, g, tʰ, ñ, and ?.

Due to limitations of character representation possibilities on UNIX and CP/M computers, some segmental phonetic distinctions need to be represented in non-alphabetic ways for such operating systems. See Appendix C for a chart of the special characters that could be used to represent these sounds, as well as a summary of the non-segmental keyboarding conventions proposed earlier in this chapter, for use with UNIX or CP/M.

Many of the special characters suggested in Appendix C are digraphs, each digraph consisting of a lower case letter followed by an upper case letter. For example, alveopalatal sounds ([š], [č], etc.) are represented as *sY*, *cY*, etc., and fricatives ([p], [g], etc.) are represented as *pF*, *gF*, etc. This upper and lower case pattern is useful for alphabetizing. Johnson (1985) suggests the use of such a system. The computer can be told to insert a space preceding every segment in a word. The segments are defined as single lower case letters or digraph units consisting of a lower case letter followed by an upper case letter. For example, the word *pigFa* would be divided thus: p i gF a. The default alphabetizing convention for ASCII characters orders a space before an upper case letter, and upper case letters before lower case letters. Thus, p i g a is automatically alphabetized before p i gF a.

Since many phonetic Zapotec forms to be entered into the "\-o" field contain lax vowels, [ɩ] and [ɛ], to be keyboarded for UNIX or CP/M systems with upper case letters, *E* and *I*, the computer needs to be told specifically not to consider *E* and *I* as part of digraphs, but rather to treat them just as they would lower case letters, inserting spaces this way: p I k a, and not this way: pI k a. Likewise, *&* for [æ], and *@* for [ə] need to be treated as lower case letters. Fortunately, the lax vowels can be eliminated from the "\-p" and "\-a" fields, being non-principal allophones of tense phonemes.

There are other ways of alphabetizing[10], besides the segmentation described here, so this lower case-upper case digraph convention might not be necessary. However, I propose its use, just to keep segmentation as an open option to aid in alphabetizing.

Notes

[1] Terrence Kaufman, in personal communication, proposed another option, that of entering the data into the data base only in a form which reflects considerable reconstruction analysis. For example, a stem which is pronounced [biny] would be written as k^winy since he posits no [*b] phoneme, only [*k^w]. I do not propose to implement this suggestion for two reasons: 1) it gives a misleading picture of the modern pronunciations for the words, and 2) being an abstraction, it confuses data and analysis, possibly prejudicing results toward an analysis which may or may not turn out to be correct. Such abstract forms belong in the "\-a" field (see Chapter 6).

[2] When compiling the questionnaire sent out to the Zapotec language workers, I requested both phonetic and phonemic forms, and copies of whatever phonemic analysis had been done. I specifically asked that they not give orthography, to facilitate comparison between the languages. Despite these clear requests, some people returned only phonetic forms, others only phonemic forms (sometimes without explanation of what they meant), and a few even sent only orthography. This frustrating problem would not have arisen if I had been able to gather the data myself, but that was not possible, given time and financial constraints.

My intention in asking for both forms was to give myself the necessary data to arrive at a form compromising between the two. As discussed above, I planned to eliminate phonetic detail that proved redundant after preliminary study and comparison of all involved languages. I also planned to compare the phonemic systems and determine which phonological rules (and crucial orderings, if any) are shared by all the Zapotec languages. These seem like reasonable goals to have had, but apparently asking for both forms from everyone was excessive, resulting in less cooperation than would have resulted if I had asked for less.

My recommendations for procedure are as follows: Ask for phonemic statements and data from other linguists in whatever form they already have, making clear to all of your secondary data sources what you plan to do. Using their phonemic statements, derive broad phonetic forms from their forms, whether they are practical orthography or phonemic. How much phonetic detail is desirable will depend on the languages being studied and compared. Overall speech style characteristics such as fronted tongue position or wide range of pitch need not be written in the data since it will presumably be described in the introduction. However, things like aspiration, relative vowel height (i.e., tense versus lax) should be included in these early steps, unless they are already known to be shared and therefore redundant. Then check these derived phonetic forms with the field linguists and ask them to evaluate them for accuracy, based on their much more intimate knowledge of the specific language. Then eliminate redundant phonetic detail. Then complete as much phonemic analysis as possible in each language. (The sources' phonemic statements will be invaluable here, too.) Put your findings into written form, to be included with the comparative project and with any portion of the data to be circulated separately. Finally, implement a notation system, for use in the "\-p" data field, that reflects only the phonetic details that are not predictable by the phonological rules. For example, lax vowels occur phonetically in Zapotec, but they do not contrast with tense vowels. Therefore, lax vowel symbols should not be written in the "\-a" data field of the ZDB. However, if the data originally came written with lax vowels, then lax vowel symbols should be used in the "\-o" field. It should be noted, too, that the forms in the "\-a" field can be updated as the phonemic analysis matures. Of course, eventually, when the reconstruction of the proto-language has been completed, proto-forms (possibly created with the aid of computer programs to "undo" the phonological changes that took place over time) should also be included, in the "\re" field.

[3] ASCII is an acronym for American Standard Code for Information Interchange.

[4] Every letter on a computer is represented by a number between 1 and 255. The first half of those numbers (through 127) represent the characters available on practically every computer keyboard, with the shift key and some special control keys. These 127 character possibilities are called the "standard ASCII characters". MS-DOS computers have the capability of doubling the character potential, using a special <ALT> key and additional numbers up to 255. These additional characters are called the "upper ASCII characters". IBM has standardized an upper ASCII character set called the "IBM extended ASCII character set", utilizing these numbers, 128-255, assigning one number to each of 128 letters and letter/diacritic combinations used in the practical orthographies of major European languages (e.g., French, Danish and Spanish), graphics symbols for drawing lines on the screen, commonly-used scientific and mathematical symbols, and symbols representing different national currencies. This IBM extended ASCII character set does not contain every symbol needed to supplement the standard ASCII character set for all practical and technical orthographies, including Zapotec. When additional characters are needed, as for the ZDB, and the archiving of texts in indigenous languages of Mexico being undertaken by the Summer

Institute of Linguistics, special software is needed to make additional letters or letter/diacritic combinations available on the screen. Furthermore, most printers do not automatically print even the IBM extended ASCII set, let alone any set custom-designed for a particular project, so special software is also required for printing of upper ASCII characters.

[5] It should be noted that the mnemonicity of these abbreviations is specific to English, which may appear to be inconsistent with the decision to use Spanish in preference to English in most aspects of the ZDB. However, I view English abbreviations for tone as being better than Spanish ones (e.g., "AL" for alto, "B" for bajo, "D" for decendiente and "A" for asendiente) since the Spanish ones necessitate using one two-letter abbreviation to distinguish between the two tones that begin with the letter *a*. It seems that anyone planning to use a linguistic data base such as the ZDB, regardless of his command of English, should be able to learn four English words for tones, or at least the four abbreviations for them, even if they seem arbitrary.

[6] Chichicapan (HR, HF, LR, LF), Quiegolani (HR, LR) and Quioquitani (MR, LR).

[7] The question arises, how should syllables be identified for purposes of interpreting stress or tone markings? For example, in (6), is the stressed syllable *si*, *si?* or *si?i*? Glottal stop does not occur in Zapotec except in checked and laryngealized vowels, i.e., there is no independent glottal stop consonant in Zapotec. Thus a consonant-vowel sequence like *?i* will not occur in Zapotec. The sequence *si?i* cannot be interpreted as two consonant-vowel syllables, *si-?i*, but can only be interpreted as one single syllable whose peak is the laryngealized vowel

i?i. Likewise, if there is a Zapotec syllable *si?*, it cannot be interpreted as a consonant-vowel-consonant syllable whose final consonant is a glottal stop, but can only be interpreted as an open syllable whose peak is the checked vowel *i?*.

[8] Punctuation symbols will not be used as punctuation in the data for the ZDB since all the data are planned to be single words. Thus, commas, colons and periods can be use in the phonetic data without confusion.

[9] In initial sorting, the various types of syllable nuclei would all be grouped together, i.e., the glottal modifications would be ignored by the computer, and all three kinds ([a], [a?] and [a?V]) lumped together. Then, a decision would need to be made regarding which order would be prefered. For example, just as *pas*, *pat*, *pes*, *pet*, *pis*, *pit* are alphabetized in the order given, should *pas*, *pat*, *pa?s*, *pa?t*, *pa?Vs*, *pa?Vt* go in that order as well? Alternatively, should they be ordered thus: *pas*, *pa?s*, *pa?Vs*, *pat*, *pa?t*, *pa?Vt*? The decision depends on careful analysis of the language data which will determine whether the three vowel qualities are functioning as three separate vowels, in which case the first order would be appropriate, or as three styles of one vowel, in which case the second order would be preferable. Such fine-tuned analysis is beyond the scope of this thesis. In any case, whichever order is chosen can be specified to the computer before it alphabetizes the data.

It should be noted that all the discussion of alphabetization is relevant only for indexes, individual language bilingual dictionaries that might be made, and proto-forms that might be used at some future time as entry titles in the data base. For now, all the alphabetizing in the data base itself deals with the Spanish gloss words only, for which there already are very adequate alphabetizing conventions.

[10] For example, to order Spanish *ch* after *cz*, a consistent change operation can temporarily rename every *ch* to *czz*, and then change them back again after the words have been properly alphabetized.

Chapter 6

Data Base Design

In this chapter, three main topics are addressed: 1) "input"--summary of what needs to go into the data base, and discussion of how to do that on the computer; 2) "output"--discussion of preferred format for printing the data ("hard copy"); and 3) indexing needs.

6.1 Input--the data base record

The purpose of this section is to outline and discuss the various fields that will be needed for each data record in the data base and the form of the contents for each field.

6.1.1 Introductory fields

Examine the following partial record, whose various fields will be explained throughout this section:

```
(1)  \sg mujer
     \eg woman
     \gr n
     \sd 2-mankind
     \re
     \cm
     \xm
     \xs
     \Ate-o 131 niu̲la'
     \Ate-p ni'ula M-M-H 1
     \Ate-a
     \WIx-o 131 'nuilʌ
     \WIx-p 'nuila
     \WIx-a
```

As has already been discussed earlier in the thesis, each word is to be glossed primarily in Spanish. Thus, the Spanish gloss serves as a sort of title for each record, and its field, labeled "\sg", signals the beginning of each record. (Each record must end with a blank line, which separates that record from the next one, whose beginning is again signaled by its "\sg" field.) (In (1) above, the Spanish gloss is 'mujer', which is placed in the "\sg" field.) Next comes the field containing the English gloss (in (1), 'woman'), labeled "\eg".

The third field of each record is the one containing the grammatical category of the word, labeled "\gr". For economy of effort in keyboarding, these grammatical category labels should be as brief as possible, preferrably one or two letters long. For the ZDB, to facilitate use by Spanish speakers, they could be based on the Spanish grammatical terms, with English translation and explanation to be included in the introduction to the data base. Alternatively, they could be done completely in English, or in both languages on the same line. For this thesis, they are done in English only.

In (1), the word is a noun, so this field contains the abbreviation 'n'. Had it been a verb, it would have been labeled 'vi' for an intransitive verb, or 'vt' for a transitive verb. Similar abbreviation codes, as brief as possible, for the other possibilities will need to be devised. A partial list of suggestions follows:

(2) Abbreviation codes for grammatical relation labels in data records

aj	-	adjective
av	-	adverb
cj	-	conjunction
ij	-	interjection
n	-	non-possessed noun
np	-	obligatorily possessed noun
nu	-	number
pn	-	pronoun
pp	-	preposition
q	-	question word

r	- relative pronoun
v	- verb, whose transitivity is unclear, i.e., unknown, semi-transitive,etc.
vi	- intransitive verb
vt	- transitive verb

Three forms of each verb should be included for Zapotec in each language field of verbal records, in an order convention clearly explained in the introduction.

Whether or not a noun can (or must) be possessed should also be notated. For the ZDB, first-person possessed forms (if the noun is obligatorily possessed) should be listed along with non-possessed stem forms, sharing one record in a predetermined order convention, as for verbs.

The fourth field, "\sd", contains a broad semantic domain label, in the form of a number code corresponding to the number of the file in which the record is contained, (which reflects the semantic domain chapter number from Buck 1949, whose outline I propose to follow) and a keyword from the actual file title. In example (1), the "\sd" field contains "2-mankind", indicating that the 'woman' record appears in file number 2, titled "mankind: sex, age, family relationships".

The next field, labeled "\re", is intended for the reconstructed proto-form of the word, the form that it is hypothesized to have had in the proto-language. Since I have not yet done any reconstruction, these fields must remain empty in the record samples included in this thesis. However, others have done considerable preliminary reconstruction work in Zapotec, including Terrence Kaufman, Robert MacLaury, Joseph Benton, and María Teresa Fernández de Miranda.[1] The ZDB is planned to include data from all four of the above, as well as mine. Such a pooling of data could eventually result in multiple entries available for this

"\re" field, in which case some sort of source code would need to be devised to identify whose hypothesis each proto-form was. This could easily be accomplished with initials, e.g., 'TK' for Kaufman or 'RM' for MacLaury. In the event that someone felt less certain about one of their hypothesized proto-forms than about others, there could also be a reliability code included in this field, such as blank for very sure, '1' for probable, and '2' for speculative.

(3) below demonstrates a hypothetical example of what the contents of the "\re" should eventually look like.

(3) \re *ni'ula AB2

The '*' preceding the form is a conventional symbol to designate a proto-form (which for keyboarding ease could be omitted entirely, since all entries in the '\re' fields will be assumed to be proto-forms anyway, but a program would need to re-insert it for printing in certain formats); <u>ni'ula</u> is the hypothesized proto-form itself;[2] the 'AB' indicates that the hypothesized proto-form is Anita Bickford's; and the '2' is the reliability code indicating that AB is very unsure about this hypothesis.

6.1.2 Comment and cross-referencing fields

The last three introductory fields provide space for comments and cross-references to other records. The comment field, "\cm", will be filled when pertinent comments are made about the data which do not fit into any of the other fields in the record. This field could be used for detailed glosses, pertinent sociolinguistic information, etc., or any other comments that compilers or users would want to make about a given record. There can be more than one "\cm" field per record, and users of the data base can add more "\cm" fields to suggest additions or changes. (See Section 6.1.4 for discussion.)

The "\xm" and "\xs" fields are for cross-references to records that are related in either of two ways to the record at hand. Words can bear a morphological relationship to each other (e.g., non-causative verbs and their causative counterparts); a record should contain cross references to all other records so related to it, in the "\xm" field. Words can also bear a semantic relationship to each other (e.g., synonyms and cognates evidencing semantic shifts); again, a record should contain cross references to all other records so related to it, this time in the "\xs" field.[3] The two types of cross-references should be kept separate to facilitate various sorting processes that might be desired.

It should be noted that none of the fields are limited in the number of characters that they can contain, so comments that are relevant to only one language rather than to the whole data set can be inserted directly into the language data fields outlined in Section 6.1.3.

6.1.3 Language data fields

The next fields in the data record contain the data forms themselves. In (1) above, only two languages' data fields are shown, as examples, but all of the languages should be included, listed consistently in the pre-determined order discussed in Section 2.1, with up to three possible fields per language. Examine again example (1) above. Each field is labeled with a three-letter code to abbreviate the name of the language whose data it contains. For example, the first three language fields in (1) above are all labeled "Ate", indicating that their data is from Atepec Zapotec. A table of these abbreviations must be included with all uses of the data base. (See Appendix A for the abbreviation table proposed for the ZDB.)

The three potential fields for each language are labeled "\-o", "\-p" and "\-a". The "\-o" stands for 'original form', i.e., the form in which the data was initially received from secondary sources[4] (which could be orthographic, phonetic, or whatever). This field should indicate the source of the data (e.g., 'FM', in this example, to indicate that it comes from María Teresa Fernández de Miranda's notes), and some sort of code number or letter, (e.g., "131" in this example), to indicate as precisely as possible where to find that particular item in that source, e.g., an item number or page number. These codes cannot be standardized since they will reflect the organization used in many sources using many different styles of organization and labeling. Data in this field should be notated in exactly the same form as it was received, except for transliteration for UNIX or CP/M systems as outlined in Chapter 5.

The "\-p" field is to contain a form of the data between broad phonetic representation and classical phonemicization of the form, as discussed in Section 5.1 of this thesis. It contains either an interpretation of the form as written in "\-o", if any, or the actual rechecked phonetic pronunciation, with redundant phonetic detail omitted. The notation chosen should be basic and standardized, and should reflect a uniform level of abstraction. This field must be present for all the languages, but it might remain temporarily empty for a given language in the case of some incomplete data sets. If it is an interpretation of the form as written in "\-o" rather than being rechecked with native speakers, it should have some kind of reliability code in it, probably the 'blank, 2 or 3' outlined above for the "\re" field, to reflect the degree of confidence felt for that interpretation. See the "\Ate-p" and "\Wes-p" fields of example (1) for samples of this reliability code, shown there as "1" and blank respectively.

Thus, the difference between the "\-p" and "\-o" fields is that the "\-o" field contains "raw data", unaltered in any way (other than perhaps transliteration for CP/M or UNIX systems, according to conventions such as those listed in Appendix C), whereas the "\-p" field contains the data, rewritten in a standardized orthography and reflecting the same level of phonemic abstraction as data in other records' "\-p" fields. See Section 5.2 for discussion of how to standardize the orthography.

Either the "\-p" field or the "\-o" field should also contain a gloss for the word in its respective individual language, if that gloss differs from the one given for the entire data set.

The "\-a" field is to contain a phonemicization of the word, which reflects analysis beyond classical phonemicization. For example, in many Mixtec dialects, Stephen Marlett (personal communication) analyzes the word meaning 'wax, soap', which is phonetically [ñũmã], to be phonemically /yuwa^{n}/ (the raised *n* indicating suprasegmental nasalization which spreads over the entire word). In this case, the "\-o" field should contain whatever form of the word was given to the compiler, the "\-p" field *ñũmã*, and the "\-a" field *yuwa*^{n}.

Since the forms in the "\-a" field might reflect controversial analyses, they should also be given source and reliability codes as outlined above (e.g., "TK", indicating the initials of the person offering the analysis, and blank, "1" or "2" again for "certain", "probable" and "speculative"). For political reasons, the reliability codes should be added by the people offering the analysis rather than by the compiler of the data base (although one user may certainly disagree with another's reliability code!).

Thus, there should be three fields per language in the data base. In the ZDB, for the twenty Zapotec languages being dealt with so far, there will be sixty language data fields in each record, although some will of course be empty, waiting to be filled when data is available. Empty fields can be omitted in printing.

6.1.4 Data entry aids

In order to make data entry easier, all of the fields discussed above should be entered first into the computer, separately from the data. This skeleton record can then be inserted into a file repeatedly, to appear on the screen as a sort of worksheet into which to keyboard the glosses, data forms, etc. (This can be done in a text editor, by reading the skeleton record from a separate file into the data file, or by making use of form capabilities of text editors like MicroSoft Word.)

6.1.5 Additions and corrections by the users

A data base can be greatly enriched and improved if all users can add comments, suggestions, new data, or changes in existing data. Of course, such changes can be made easily to printed editions of the data base, simply by writing things in with brightly colored ink and/or adding a memo, and returning everything to the compilers. However, if a user wants to add anything to a soft version (still on the computer disk), the need arises to set up a standardized way for this to be done.

Where to put such additions on the disk is quite straightforward.

The "\cm" field is the most obvious place to put comments and suggestions about individual records. Corrections or additions to the data itself should go in the designated fields in the appropriate records. Whole new records should be set up for new data sets. Alternatively, extra "\cm" fields could be added to a record, one for each comment that a user wanted to make. There is no limit to the number of "\cm" fields that could be inserted to accomodate corrections and additions, etc.

However, additions of this type will go unnoticed (and thus unadded to subsequent editions of the data base, printed or otherwise) unless they are flagged in some way so that the compilers notice them and deal with them. I suggest that any additions or corrections, in any field in any record, be flagged by writing "CHANGE", in capital letters, just before whatever is to be added. For example, suppose that a word is incorrectly written "niulo". The user who spots the error should add "CHANGE niula" (or whatever the correct form is) in that field, immediately following "niulo". If a new record is to be added, the word "CHANGE" should precede the Spanish word in the "\sg" field which marks the beginning of the new record. If a comment or cross-reference is added, its first word should be "CHANGE", etc. (If all additions are handled by "\cm" fields, the contents of each added "\cm" field should begin with "CHANGE".) Then the compiler's computer can search each file for records containing this word, and take appropriate action concerning the added material. This puts a small burden on the user to remember to flag his additions in this way, but insures that his comments are noticed and given due consideration.

Any such changes should also be identified as to contributor (name and location) so that the compilers can communicate with the users regarding the changes. This could be done right on the disk, along with the first "CHANGE" flag, or by means of a memo to accompany the disk when it is sent back to the compilers.

6.2 Output--formats for printing

Once the data is entered, it can be sorted and printed by the computer in various ways. For example, using one of the first five fields per record, it can be alphabetized by gloss or by proto-form, or sorted by grammatical relation, etc., for many assorted uses. Alternatively, using the data form fields, all forms from, for example, the "\-o" or "\-p" field of a particular language (e.g., \Ate-p) could be grouped together and alphabetized for the formation of a dialect-specific bilingual dictionary (e.g., Atepec/Spanish).

The data base, stored in these many fields and records, is very flexible. It can be printed exactly as it appears in the records--the program should specify whether you want one record per page or whether records can straddle across page-breaks--or reorganized into some other layout, perhaps a sort of chart, to facilitate comparison. One thing to remember in doing this, however, is that the fields have unspecified length, i.e., there is no limitation on the number of characters entered into each one. When a chart is being planned, its column widths will need to be specified. Thus, data characters beyond this limit would need to be wrapped.

6.2.1 Layout

There are almost unlimited options available for the layout of the data base, all of them being variations of two main possibilities. The purpose of this section is to display and discuss these two main options, namely columns and prose format, and evaluate them in terms of their legibility, flexibility and use of space. Two assumptions underlying all of these suggested options are that the gloss needs to accompany each entry and that the language of each data item must be clearly labeled.

One option for data display is to arrange the cognate lists in columns. For example, figure (4) shows the various Zapotec "\-p" forms of the word glossed 'milpa' ('cornplant'), listed in a column:

(4)

milpa	(cornplant)
\Ate	'šela M-M
\WIx	'šiele
\Ytz	yel
\Ylg	yel M
\Cho	la'kyela
\Rin	y+l
\Tex	kyel
\Lac	iña?a M-L-L
\SZm	'kelë L-L
\Chi	'kela 'L
\Glv	kel
\Mit	kæhl
\Alb	kyal L
\GvH	kiahl L
\Ist	'kela L-L
\Qgl	kyel H
\Qui	kyol H
\Xan	
\Xng	kyal
\Ama	na?a

One or more such columns could appear on each page. Note, however, that most computer software that can read standard format markers does not have the capability of moving columns around, so the flexibility of data movement could be severely reduced by using multiple columns.

Thus far in planning for the ZDB, I have avoided the use of multiple columns, instead simply listing the columns one below the other. In this way, only one list can fit on each page, leaving plenty of margin space for notes and corrections. Although much of this space on the paper seems wasted, it is beneficial to have it available in the computer data fields (i.e., not to have a limited number of characters in the field of each word) when alternative word forms are available in a language or when an explanation accompanying a data form is necessary. For example, the Albarradas Zapotec entry for 'dew' in my data base is as follows:

(5) pi'ni?i H-R (kohp: L light rain, sprinkle)

As it turns out, the secondary word kohp:, fits better into a cognate set with the other languages' forms than does pi'ni?i. Had the number of characters that could fit in that field been limited, I probably would have left out the form that was listed second, and would have had to waste time later looking back in the original questionnaire for it, or, worse yet, might not have known it was there at all and left it out entirely.

A disadvantage of this format is the need to flip through a lot of pages to find numerous cognate sets containing consistent sound correspondences. However, once they are found (and the computer can assist in this process by searching for particular segments, or strings of segments, in specified languages), the data sets can be moved around, set by set, with several sets being grouped together on consecutive pages, to facilitate comparison. Thus, this disadvantage can be fairly easily overcome.

The corollary advantage, of course, is that, with the data listed in columns, it is easy to look up and down each column, to compare segments and find those sound correspondences in their analogous environments. For example, refer back to the 'milpa' data set (example (4) above). With the data listed as it is above, in a column, it is easy to see that there is some sort of sound correspondence. Perhaps the [š] of Ate and WIx corresponds to the [y] of Rin, Ylg and Ytz, the [ky] of Cho and Alb, the [k] of Chi, Glv and Mit, the [ki] of GvH, etc. Alternatively, it could be that the WIx and GvH [i] corresponds to Rin and Cho [y], etc. As mentioned above, other words with similar segments could be grouped for comparison with 'milpa', to see which analysis is better supported. Other cognate sets could be examined to determine whether the languages were consistent in the way that they grouped together on these sound correspondences. This is a first step toward reconstructing the proto-language and determining generations of relatedness among the modern languages.

Since potentially there could be three forms for each language listed, three columns could be formed, one for each form ("\-o", "\-p", and "\-a"). This sorts the three forms and separates them, to maximize the scanning advantage described in the preceding paragraph. Example (6) below, a partial listing of the 'milpa' set, demonstrates what three columns of this sort could look like.

(6) milpa (cornplant)

	o	p	a
Ate	'yela	'šela M-M	yela M-MP
WIx	yiela	'šielə	yela
Rin	yɨl	yɨl	yɨl

Notice that the middle data column is the same as the beginning of the list in (4), with the same good potential for scanning up and down for sound correspondences.

Arranging the data in prose format (like paragraphs) has one advantage: it is far more efficient in terms of conserving paper use. Since publication costs would be smaller for a book containing fewer pages, this would be a large advantage when the final product is being prepared for publication. Prose format is adequate if the final product is a finished comparative dictionary, presenting proto-forms and full analysis, and all the reader would expect to do with the data is confirm or disconfirm the author's analysis. However, if the data is to be published as working data, simply made available for others' use in further analysis, as is the case for the ZDB, it would be of limited usefulness in prose format. For example, look again at part of the 'milpa' data set, this time arranged in paragraph form:

(7) 'milpa' (cornplant): Ate, 'šela M-M; WIx, 'šielə; Ytz,
yel; Ylg, yel M; Cho, la'kyela; Rin, yɨl...

The ease of scanning which was present in a column format is clearly lacking here.

6.2.2. Cognate set "title"

The next choice involved in output format is which field's contents to use as the title for each cognate set. The logical choices are the three major fields that introduce each data record: Spanish gloss, English gloss, or reconstructed proto-form. Whichever of those three is chosen, all the rest of the data in the record will be organized under or beside it in some subordinate way.

For a true comparative dictionary, only one of those choices is acceptable as a title for each entry--the reconstructed proto-form. The presentation of proto-forms, accompanied by a description of the phonological processes that took place over time to produce the modern language reflexes of those proto-forms, is the ultimate purpose for a comparative dictionary. Thus, the first line of an entry in a true historical dictionary would look something like the first line of example (8), as shown here:

(8) *kial milpa (cornplant) n
(cross-references would go here)

Ate	'šela M-M
WIx	'šielə
Ytz	yel
Ylg	yel M
Cho	la'kyela
Rin	yɨl
Tex	kyel
Lac	iña?a M-L-L
SZm	'kelë L-L
Chi	'kela 'L
Glv	kel
Mit	kæhl
Alb	kyal L
GvH	kiahl L
Ist	'kela L-L
Qgl	kyel H
Qui	kyol H
Xng	kyal
Ama	na?a

Here, *kial is, of course, the proto-form,[2] "milpa" the Spanish gloss, "cornplant" the English gloss (placed in parentheses since English is only the secondary gloss language), and "n" the grammatical category marker indicating "noun".[3] Cross-referencing to any related data records would also be included, as outlined in Section 6.1.3.

Note that the sources and reliability codes have been omitted here. They are still available, stored in the data base, but need not appear in every single entry in the comparative dictionary. This is because a list of sources will be included in the preface to the data base, as well as a description of the analysis which lead to the postulation of the proto-forms.

Unfortunately, there is a major problem with using the proto-form as the "title" of each entry in the ZDB: there are not yet proto-forms posited for every cognate set to be included in the data base. For this reason, I propose using the Spanish gloss as the identifying portion of the entry in this data base of cognate sets. It would be inconsistent and confusing to use proto-forms when possible and Spanish glosses elsewhere. (Spanish glosses are chosen over English for the reasons outlined in section 3.1.) Thus, my proposal for presentation of the data sets at the current stage of analysis and reconstruction is the one found in the small sample in Appendix D.

6.3 Need for separate indexes

Indexes are needed for printed editions of the data base whenever the desired information cannot be found by means of a sequential, linear search. In the case of the ZDB, "linear" means alphabetical. Since the data sets are listed alphabetically by Spanish gloss, there is no need for any Spanish indexes.[4] Spanish forms can simply be found in a sequential, alphabetical search. Thus, the potentially needed indexes are Zapotec, English and (after further work in reconstruction produces reliable proto-Zapotec forms) proto-Zapotec.

An English/Spanish index is definitely needed, to allow a person to make use of the ZDB beyond what his command of Spanish allows. One can easily be produced by computer, the computer taking information from the "\sg" and "\eg" fields and converting it into an index.

In the early stages of the ZDB, there should be a proto-Zapotec/Spanish index including any proto-Zapotec forms that are posited. If, at some future time, proto-Zapotec forms are posited for all the data sets in the data base, and the data base is reorganized using those proto-Zapotec forms as titles rather than the Spanish glosses, then English/proto-Zapotec and Spanish/proto-Zapotec indexes should replace the English/Spanish and proto-Zapotec/Spanish ones. Again, any of these indexes can easily be produced by computer.

To accompany a printed edition of the ZDB, an index for each variety of Zapotec could be made, to enable people to find the cognate set that contains a particular Zapotec word. Such indexes could be generated fairly easily by computer. Example (9) below shows what such an index would look like for the small sample data base presented in Appendix D, for Albarradas Zapotec.

(9) Albarradas Zapotec/Spanish Index

č:o?n: R, tres
kahš: L, cerca
kaw, comer
ko?ol, cantar
kon: R, donde
ko:t:im, moler
kul:a?am, tener hambre
ku'sam, andar
ku'tawim, comer
kyahs: L, olla
lič, hogar
na-k:ay L-R, oscuro
nya?s L, negro
pan R, ¿dónde?
pe?t:, moler
pi?il, cantar
pi'sa:?n L-L, hermana de hombre
pi'sa:?n R-H, hermano de mujer
'raw, comer
rilya?am, tener hambre
ri'sam, andar
'ro?olim, cantar
'ro:t:im, moler
sam, andar
tat yu? pala H R L-H, sombrio
ya?s L, negro
yu L, casa

However, since one such index would be needed for each of at least twenty varieties of Zapotec, for a projected data base containing at least a thousand entries, it seems that they would add a lot of unnecessary bulk to the size of the ZDB, and maybe to any such data base project. Instead, a whole new data record could be made for the gloss of any individual cognate which varied in gloss from the rest of its set, containing a cross-reference to the main cognate set. Examine the record in example (11) below to see how cross-referencing could be done for partial data set (10), in which the glosses do not match completely. The ‘llovizna’ record in (11) contains only cross-referencing information, to tell a user where to find data with that gloss in another record, but does not contain the data itself.

```
(10)  \sg  rocío
      \eg  dew
      \gr  n
      \sd  1-world
      \re
      \xm
      \xs  cf. lluvia
      \Ate-p up:a’rela H-M-F-M
      \WIx-p pe’tani
      \Cho-p ’kup:a tao?
      \Glv-p pni?/kup:
      \Mit-p kohp
      \Alb-p pi’ni?i H-R (kohp: L llovizna)

(11)  \sg  llovizna
      \eg  sprinkle, light rain
      \gr  n
      \xs  cf. rocío, lluvia
```

Note that both records shown (in (10) and (11) above) also contain cross-references to ‘lluvia (rain)’, whose record is not shown here, to which both ‘rocío’ and ‘llovizna’ are semantically related.

Presumably, any other indexing needs, (e.g., Spanish to Spanish gloss synonyms or morphologically related words) are met within the data base itself by the cross-referencing. See Appendix D for samples of cross-references and how they would be used.

Notes

[1] I cannot cite specific published works for most of these, since, to my knowledge, their reconstruction work has not yet been published.

[2] This is just a guess for demonstration purposes.

[3] An alternative introductory field for each entry, to replace the "\xs" field, would be one labeled "\ds". This field would contain a brief description of a detailed semantic domain for the word. In example (1), e.g., it could contain the word person, indicating that this Zapotec word falls into the human domain. This field is important for filling out incomplete data sets when glosses of cognate words do not match exactly. For example, perhaps one language has a completely different word for 'woman', a word that is clearly not a cognate to the others gathered for that set. However, a cognate for ni'ula could exist with a different gloss, perhaps 'wife' or 'daughter', etc. The language worker could be helped by having a fairly specific semantic domain listed which could steer him toward other places to look in his data for cognates.

(Note that this is not the same as the "\sd" field, which contains the large semantic domain label reflecting the data file to which the record is assigned. The "\sd" field is actually redundant as long as the records remain sorted in their respective files. However, if in a printed edition the data is all merged together alphabetically, the "\sd" provides a quick way to identify to which file a given record belongs.)

Perhaps the need for fine-tuned semantic domain labels becomes clearer in the following Atepec Zapotec example (Nellis 1983:173), involving complex meaning relationships with the word niula. To the basic word, various modifiers are added which alter the meaning:

(i)	niula	mujer (woman)
	niulii	esta mujer (this woman)
	niulata'	niña (girl)
	enne' niula	dama (lady)
	yi'ni niula	hija (daughter)
	niula cuiti'	mujercita (young woman)

(This data is in Nellis's practical orthography, with tone omitted, as irrelevant to the example.) If a linguist who was beginning work in Zapotec for the first time had elicited the last five of these six nouns in separate contexts, using second-language glosses, and had not yet analyzed where the word breaks belong, he might not yet have noticed that they all share the same basic stem. He might not yet have elicited the basic noun niula, instead having learned a synonym. It would help him to have a semantic domain label, i.e., 'person', as a guide to where to look in his data for cognates to another language's noun niula 'woman'.

This notion of including a fine-tuned semantic domain field involves considerable complication, and imposes a burden on the compiler of the data base, to think of a specific semantic domain for each record. It is unclear to me at this time whether exhaustive cross-referencing would accomplish the same end with less effort. For now, I am assuming this to be the case. What is very clear is that some good way of facilitating field linguists' search for cognates in their own data files should be included in the ZDB if at all possible.

[4]A brief discussion of data sources is in order. Should the data be solicited from field linguists or elicited directly from indigenous speakers of the language? There are two sides to this question, and they are both dependent on the situation involved. If one person eli-

cits data directly from all the indigenous speakers involved, he is more likely to be consistent in transcription, which will greatly facilitate the comparative study. However, as he does the elicitation himself, he will not be able to benefit from the expertise of field linguists, who may already speak the language well and have done considerable phonological study. In addition, the great expenditure of time and travel costs would necessarily limit the scope of a one-person project. Soliciting data from field linguists is a presumption unless there is a clear agreement regarding cooperation, compensation and giving credit for the data sources. There is also the great likelihood of inconsistency of transcription, and, in fact, it is difficult to monitor the quality of the data. In addition, it is difficult to motivate linguists to make the effort to send the data, especially if forms are asked for which they do not already have in their files, necessitating new elicitation.

Obviously, neither choice is ideal. For the ZDB, I have chosen so far to seek the data from my colleagues in the Summer Institute of Linguistics, because they had all been trained similarly in transcription, and because they had far easier access to the data than I did. Eighty per cent of the linguistic teams that I approached about the study actually shared data with me, and of course my questionnaires were filled out with varying degrees of completeness. However, I received far more data in that way than I would have been able to on my own without prohibitive costs of time and money.

Additional data will be contributed by other linguists and published sources, as has already been mentioned.

[5] For the real ZDB, the grammatical category markers and cross-

references should probably be in Spanish, eg., the "n" for "noun" would be changed to "s" for "sustantivo". However, for demonstration purposes in this thesis, I am using English to avoid needing to include English translations of the Spanish.

[8]However, if the data base is printed out with the semantic domain chapters intact rather than with all the records merged alphabetically, there is need for a Spanish index directing users to the chapter containing the desired form.

APPENDICES

APPENDIX A

Language Abbreviations

This first appendix contains a list of the abbreviations used in this thesis for Zapotec language names, and the last name of my data sources, as listed in the introduction. (The choice to use three-letter codes was somewhat arbitrary; they could just as well have been two letters instead.)

Abbreviations:

Álb	Albarradas (Kreikebaum)
Ama	Amatlán (Riggs)
Ate	Atepec, Ixtlán de Juárez (Nellis)
Ayo	Ayoquezco (MacLaury, no data in this thesis)
Chi	Chichicapan (Benton)
Cho	Choapan (Lyman)
Glv	Guelavía (Jones)
GvH	Guevea de Humboldt (Kreutz, Morse)
Ist	Isthmus (Pickett)
Lac	Lachixío (Persons)
Mit	Mitla (Stubblefield)
Qgl	Quiegolani, Western Yautepec (Regnier)
Qui	Quioquitani (Ward)
Rin	Rincón (Earl)
SZm	Sur de Zimatlán (Olson)
Tex	Texmelucan (Speck)
WIx	Western Ixtlán (Thiessen, Smith)
Xan	Xanica (Reeck, via Piper)
Xng	Xanaguía (Hopkins, Olive)
Ylg	Yalalag (Newberg)
Ytz	Yatzachi (Butler)

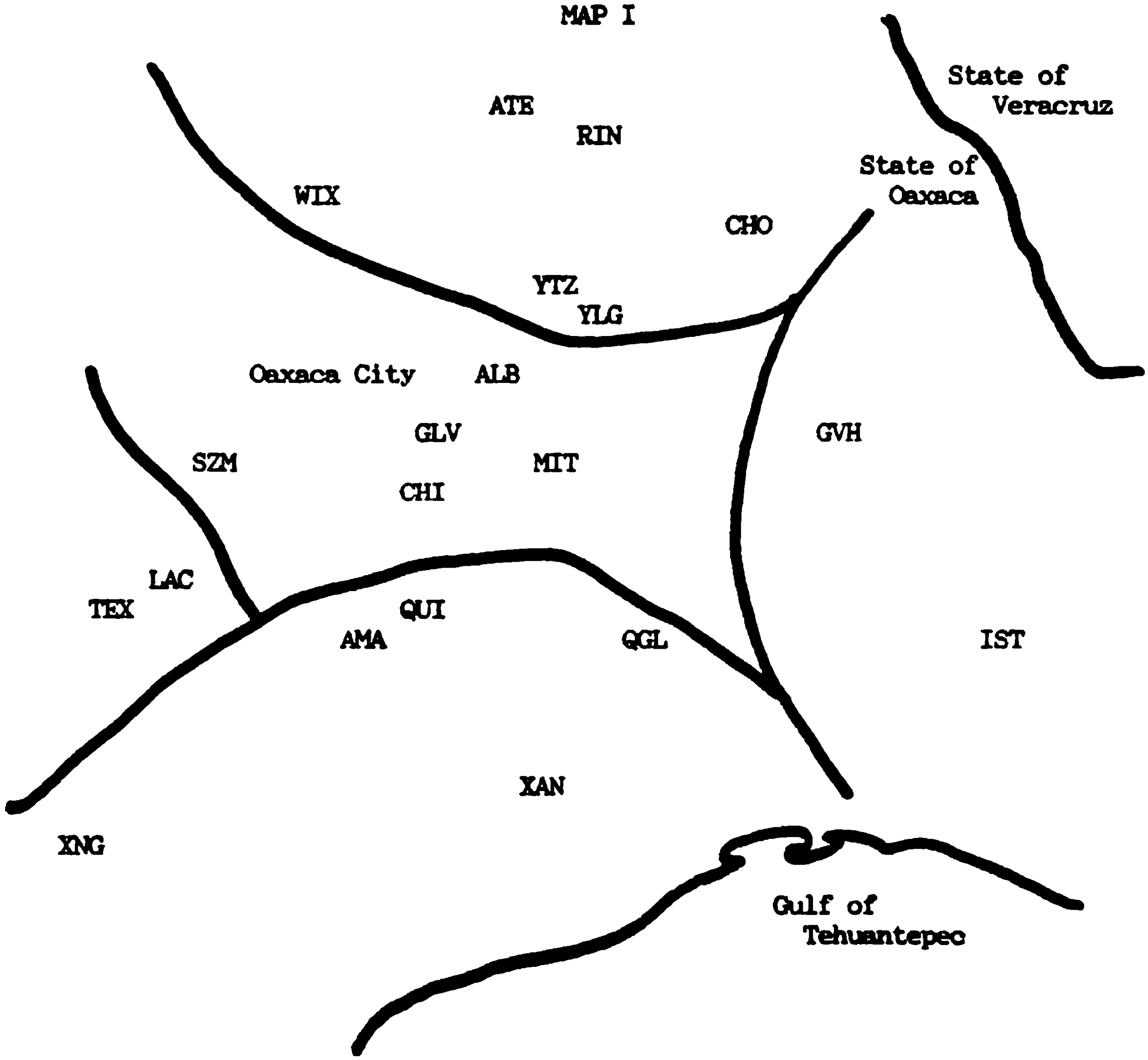
APPENDIX B

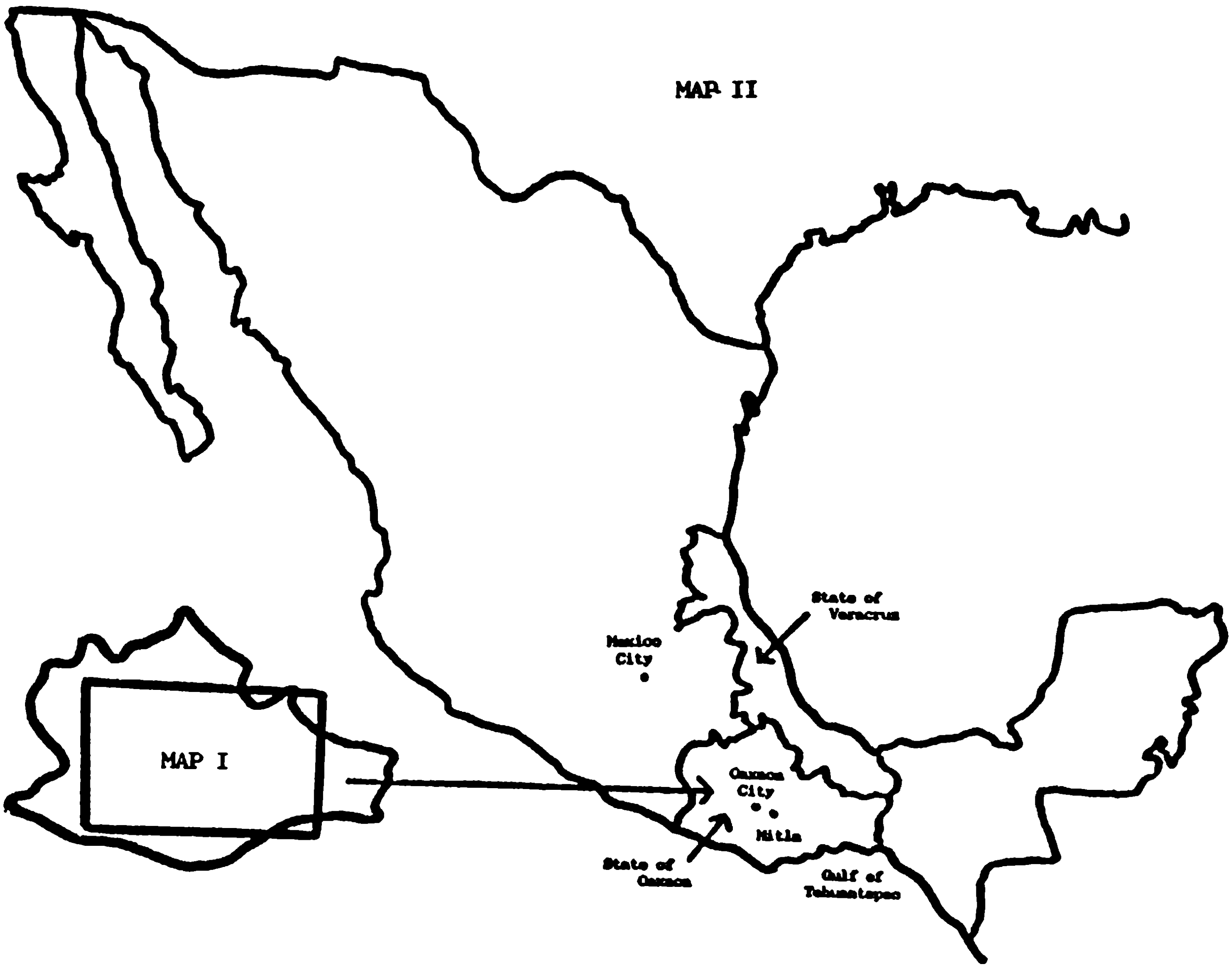
Maps of the Zapotec Region

This second appendix contains maps of the Zapotec region. Map I indicates the approximate locations of the twenty languages so far included in the ZDB. The map is adapted from a survey map found in Egland (1978:68-69). All towns and cities shown are in the state of Oaxaca.

Map II shows how Map I is oriented to the rest of Mexico.

MAP I




MAP II
MAP I
Mexico City
State of Veracruz
Oaxaca City
Mitla
State of Oaxaca
Gulf of Tehuantepec

APPENDIX C

Special Character Conventions

This third appendix summarizes the special character needs for the Zapotec data in the ZDB, and contains a summary list of proposed keyboarding conventions for non-standard ASCII characters.

Phonetic symbol	Proposed keyboarding convention for Unix or CPM
ž	zY
š	sY
ǰ	jY
č	cY
ə, ʌ	@
æ	&
i	I
e	E
ɨ	+
ë	e"
ƀ	bF
đ	dF
θ	tF
ǥ	gF
ł	lF
tʰ	tH
ñ	n˜
?	?

Phenomenon	Proposed keyboarding convention for UNIX or CPM
stress	apostrophe preceding stressed syllable ('tika)
tone	H, M, L, F, R, placed beside the word in sequence (tika H-R)
nasalization	tilde following the nasalized vowel (ti˜ka)
length	colon following the long segment (tika:); if length occurs on a segment that also has another modification, the colon should follow the other modification's symbol (ti˜:ka, tik.:a)
retroflexing	period following the retroflexed segment (t.ika)
backing	period following the backed segment (tik.a)
fortis	colon following the fortis segment (tik:a)
lenis	no colon following lenis segment (tika)
checked vowel	glottal stop following single vowel (ti?ka)
laryngealized vowel	followed by glottal stop and V (ti?Vka) OR vowel followed by two glottal stops (ti??ka) (consistently choose one of these two options)

APPENDIX D

Sample Comparative Dictionary of Zapotec

This appendix contains seventeen sample entries for the ZDB, and a sample index to accompany them. These entries were chosen for the variety of grammatical categories represented, the phonological variety displayed, and the cross-referencing potential that they have.

The first line of each entry contains the Spanish gloss in bold face, next the English gloss, next the grammatical category in italics, and finally a place for the proto-form. However, since I am not yet ready to postulate proto-forms, I have simply placed the symbols "*xxx" in the spot where the proto-forms should be placed in the completed comparative dictionary someday. The optional second line contains cross-referencing information taken from the "\xs" and "\xm" fields. The optional third line contains miscellaneous comments taken from the "\cm" fields.

The data themselves, taken from the "\-p" field of the records, are in a broad phonetic form which approaches classical phonemicization.[1] Each verb in this sample dictionary is given in three forms: completive, potential and habitual (in that order). Abbreviations used for language names are explained in Appendix A. Abbreviations used in this sample for grammatical categories are as follows:

(i) Grammatical category abbreviations

aj	adjective
av	adverb
n	non-possessed noun
np	obligatorily possessed noun
nu	number
q	question word
r	relative pronoun
v	verb

vi intransitive verb
vt transitive verb

Following the list of sample entries is the English/Spanish index for the entries. An Albarradas Zapotec/Spanish index of the words, as an example of the type of index potentially needed for a comparative dictionary, can be found in Section 6.3, along with a discussion of why I propose to use cross-referencing rather than twenty such indexes (one for each variety of Zapotec in the study).

It should be noted that the data in this sample appear as they might be printed by a formatter, without field codes, etc. See example (ii) below, to see what the actual computer record for the 'hogar' entry would look like.

```
(ii) \sg  hogar
     \eg  (someone's) home
     \gr  np
     \re
     \cm  'hogar' is obligatorily possessed, 'casa' is not
     \xm
     \xs  Cf. casa
     \Mit-o rolize my house
     \Mit-p rolis (house)
     \Mit-a
     \Alb-o li³J (someone's) home
     \Alb-p lič L (someone's home)
     \Alb-a
     \Ist-o 'liJi home
     \Ist-p liči L-L
     \Ist-a
     \Ama-o li/yo?o
     \Ama-p li
     \Ama-a
```

Sample Comparative Dictionary of Zapotec

andar   walk   vi   *xxx

Ate	’kuθa?	θ:a?	riθa?
WIx	u’ta:na	’θ:a:na	ri’ta:na
Ytz	’kwsa?a	s:a?a	čsa?a
Ylg	kwte?e	t:e?e	čte?e
Cho	uta	t:a	rta
Rin	ku’tæ?	t:æ?	ri’tæ?
Tex	psay	say	rsay
Lac	uc:e	c:e	s:e
SZm	ku’sa	s:a	rsa
Chi	ku’sa	s:a	rsa
Glv	kusa?ap	časa?ap	rsa?ap
Mit	pi’s:ah	ki’s:ah	r’s:iah
Alb	ku’sam	sam	ri’sam
GvH	kwse?	s:e?	rse?
Ist	ku’sa	s:a?	ri’sa
Qgl	u’sa?a	s:a?a	rsa?a
Qui	kws:æ	ks:æ	šs:y&
Xan	(ussame)?		
Xng	wsë	s:ë	së
Ama	nuse	yise	nšise

cantar   sing   v   *xxx

Ate	pe’l:a	ku’l:a	ru’l:a
WIx	’pil:ana	’ul:ana	rul:ana
Ytz	pil:	kol:	čol:
Ylg	pel:	kol:	čol:
Cho	pila?	kula	rula
Rin	pil	kul	rul
Tex	ptulã	tulã	rtulã
Lac	unkula?	ula?	rula?
SZm	pi?il:ë	’ko?ol:ë	’pi?il:ë
Chi	’pi?il:a	’ku?ul:a	’ru?ul:a
Glv	pi?l:	ð:aku?l:	ru?l:
Mit	pil:	kol:	rol:
Alb	pi?il	ko?ol	ro?ol
GvH	pi?il:y	ku?ul:y	ru?ul:y
Ist	’pi?inta	’ku?unta	’ru?unta
Qgl	wo?ol	ko?ol	ro?ol
Qui	pi?ily	ko?ol	č:o?ol
Xan	(ti?i)?		
Xng	wyo?ol	ko?ol	čo?ol
Ama	mpil	kol	nčol

casa house *n* *xxx

Cf. hogar

Note: 'casa' is optionally possessed, but 'hogar' is obligatorily possessed[3]

Ate	šu?u
WIx	šo?o
Ytz	yo?o
Ylg	yo?o MF
Cho	kyu?u
Rin	yu?u
Tex	yu? L
Lac	(ni?i M-L)
SZm	kiu?u
Chi	yu?u L
Glv	yu?u
Mit	yu?
Alb	yu L
GvH	yu? L
Ist	yo?o L
Qgl	yu?u L
Qui	yu L
Xng	yu
Ama	yo?o

cerca near *av* *xxx

Ate	'eṣ̌:a M-M
WIx	'kaṣ̌:a
Ytz	'kalə?esə/'pao
Ylg	au'šo
Cho	k:wet:a
Rin	'kala?
Tex	ka?ap L
Lac	ašu L-L
SZm	kësa'ka
Chi	'kaš:u 'L
Glv	kaṣ̌:
Mit	kahš:
Alb	kahṣ̌: L
GvH	kahš: H
Ist	kaš:a L-L
Qgl	keš: L
Qui	keṣ̌: L
Xng	kaš:
Ama	kaš:

comer   eat   *vt*   *xxx
Cf. moler, tener hambre

Ate	ku’t:o	ko	ro
WIx	ut:ona	kona	rona
Ytz	kwtakw	kakw	čakw
Ylg	kwtao	kao	čao
Cho	utao	kao	rao
Rin	ku’taw	kaw	raw
Tex	ptony	kony	rony
Lac	utak:u	ak:u	rak:u
Chi	ku’taw	kaw	raw
Glv	pk:i?n	č:ak:i?n	rk:i?n
Mit	ku’tahw	kaw	rahw
Alb	ku’taw	kaw	raw
GvH	kwtahkw	kahkw	rahkw
Ist	ku’to	ko	ro
Qgl	wu	ku	ru
Qui	kwt:aw	k:aw	č:aw
Ama	ntaw	kaw	nšaw

donde[1]   where   *r*   *xxx
Cf. ¿dónde?

Ate	’lat:i
WIx	ka:
Ytz	ka
Ylg	ka
Rin	ka
SZa	k:a lo
Chi	k:a H
Glv	k:un H
Mit	k:aro
GvH	pa L
Ist	ra L

¿dónde?[3]   where   *q*   *xxx
Cf. donde

Ate	’kani H-H
Cho	ka
Rin	’ckaši
Tex	k:a H
Lac	k:a H
Glv	k:a’li
Mit	ca-
Alb	pan R (where to, with verbs)/kon: R (where, with nouns)
GvH	pa (k)lo L-L
Ist	p:a’ra?a L-H
Qgl	p:a/ko H (where is)
Qui	p:a H/k:o H
Xan	p:a:na
Xng	p:aw/k:o (where is)
Ama	p:a

hermana de hombre, hermano de mujer cross-sex sibling *n* *xxx

Ate	'θana M-L
WIx	θan
Ytz	san
Cho	san
Rin	san
Tex	san M
Lac	sana M-H
SZm	pë'sã?
Chi	pi'sa?n 'L
Glv	psa?n
Mit	pi's:iahn
Alb	pi'sa:?n L-L/R-H
GvH	pysan H
Ist	pi'sa?na? L-L-L
Qgl	psa?an L
Qui	psyañ L
Xng	psan:
Ama	pson

hermano brother *n* *xxx
Cf. hermana de hombre, hermano de mujer

hogar[4] (someone's) home *np* *xxx
Cf. casa
Note: 'hogar' is obligatorily possessed, 'casa' is not[2]

Mit	rolis	home
Alb	lič L	someone's home
Ist	'liči L-L	
Ama	li	

moler grind *vt* *xxx
Cf. comer

Ate	pet:u	kut:u	rut:u
WIx	'pet:una	'ut:una	'rut:una
Ytz	pet:	kot:	čot:
Ylg	pet:	kot:	čot:
Cho	pt:u	t:u	rut:u
Rin	'pæt:uæ?	'kut:uæ?	ri'yæt:uæ?
Tex	ko?o	ko	ror
Lac	uyu:	yu:	ryu:
SZm	pe?et:u	ko?ot:ë	'ro?ot:ë
Chi	'pwe?t:u	'kut:a	'rut:a
Glv	be?et:	čakut:	rut:
Mit	pe?t:	ko?t:	ro?t:
Alb	pe?t:	ko:t:	roht:
GvH	pi?t	ku?t	ru?t
Ist	'pi?it:u	'ku?ut:u	'ru?ut:u
Qgl	wu?ut:	ku?ut:	ru?ut:
Qui	pæt:	kot:	čot:
Xng	wsui?i	ysui?i	sui?i

negro   black   *aj*   *xxx

Cf. oscuro, sombrío

Ate	šeθ:ia H-L-L
WIx	'šaθ:i
Ytz	'kas:Ṛ
Ylg	kas:Ṛ
Cho	kas:o
Rin	kas:R
Tex	k:a?as L
Lac	nac:a M-M
SZm	'nkas:ë R-L
Chi	na'kas:a 'H (ya?sa evil)
Glv	nkas:
Mit	yas:
Alb	(n)ya?s: L
GvH	nya?s:
Ist	na'ya?as:e? L-L-L
Qgl	n'kas: H (ka?as: to turn black)
Qui	nye?es:y L
Xan	nkas:
Xng	ya?s
Ama	na-kas:

olla   cooking pot   *n*   *xxx

Ate	'šeθ:u? H-H (clay)
WIx	'šeθ:u/c:a
Ytz	'yes:ə?
Ylg	ye?es:
Cho	'kyes:o
Rin	'yus:u?
Tex	kyis: L (clay)
Lac	ec:u H-L
Chi	'kis:u 'L
Glv	kes:
Mit	kehs:
Alb	kyahs: L
GvH	kiuhs: H
Ist	'kis:u L-L
Qgl	kyus: L (olla)/m'c:a?a LR (casuela)
Xng	kes:
Ama	yas:

oscuro dark *aj* *xxx

Cf. negro, sombrío

Ate	č:ul:a
WIx	č:ul:a
Ytz	šč:ol:
Cho	č:ula
Rin	č:ula
Tex	k:a?as L
Lac	k:ape M-M
Chi	na'kas:a 'H
Glv	nak:ay
Mit	na'k:ahy
Alb	na'k:ay L-R
GvH	nk:awy
Ist	na'k:awi
Qgl	nkep
Qui	nkey LR
Ama	na-ya?as:

sombrío shady *aj* *xxx

Cf. negro, oscuro

Ate	'ş̌ula F-M
WIx	'ş̌:ula
Ylg	ş̌ol
Cho	šula
Rin	'šultsa
Lac	š:k:a?la? H-L
Chi	lo pa'k:ala L 'H
Glv	kapl:a M-H
Mit	lopa'la?a
Alb	tat yu? pala H R L-H
GvH	ke?mple

tener hambre be hungry *vi* *xxx

Cf. comer

Ate	kut:uni	it:uni	rit:uni
WIx	pit:uin:a	it:uin:a	rit:uin:a
Ytz	kwton	t:on	čton
Cho	utue	t:ue	rtue
Rin	ku'tun	itun	ri'tun
Tex	pyan	kyan	ran
Lac	ulya?na	lya?na	lya?na
Chi	pi'č:i?:an	kič:i?:an	rč:i?:an
Glv	pl:ia?ana	il:ia?ana	rl:ia?ana
Mit	kulia?an	kilia?an	rlia?an
Alb	kul:a?an	kilya?an	rilya?an
GvH	pilahn	kilahn	rlahn
Ist	kun'ta?ana	kin'ta?ana	rin'ta?ana
Qgl	u'la?anen	i'la?anen	'rla?anen
Qui	kwla?añ	klya?añ	šlya?añ
Ama	---	nla?an	---

tres three *nu* *xxx

Ate	'c:un:a M-H
WIx	'c:un:a
Ytz	'š:on:ə
Ylg	č:on:
Cho	c:ona
Rin	'c:ona
Tex	č:on L
Lao	č:una L-M
SZm	'č:ona L-H
Chi	'č:on:a 'LR
Glv	č:on
Mit	č:on:
Alb	č:o?n: R
GvH	c:on:
Ist	'č:on:a L-R
Qgl	c:on-
Qui	c:on H
Xan	cona
X̣ng	c:on
Àma	č:on

English/Spanish Index

be hungry, tener hambre
black, negro
brother, hermana de hombre, hermano de mujer
cooking pot, olla
cross-sex sibling, hermana de hombre, hermano de mujer
dark, oscuro
eat, comer
grind, moler
home, hogar
house, casa
hungry, be, tener hambre
near, cerca
shady, sombrío
sing, cantar
sister, hermana de hombre, hermano de mujer
three, tres
walk, andar
where (r), donde
where (q), donde

Notes

[1] It would be preferable to present the data in both phonetic and classical phonemic forms. However, since both were not consistently available for use in this thesis, and since accurate derivation of one from the other requires results of research that is not currently available to me, I have compromised on a broad phonetic form somewhere between the two for demonstration purposes. Two partial data records in a more preferable form are included here, based on phonetic transcription and phonemicization by Grace Thiessen and Joan Smith (Western Ixtlán), Inez Butler (Yatzachi), Ronald Newberg (Yalalag), David Persons (Lachixío), Donald Olson (Sur de Zimatlán), and Joseph and Mary Benton (Chichicapan). In each entry, the phonemic data are listed first, between slashes (//), followed by the phonetic data, between brackets ([]). The language fields displayed in (i) and (ii) correspond to the "\-p" fields in the sample comparative dictionary offered in this Appendix.

(i) cerca near *av* *xxx

WIx	/gaşa/	['gaşʌ]
Ytz		['galeʔezə]
Ylg	/aužoʔ/	[awžuʔ]
Lac	/ašu L-L/	[ašu L-L]
SZm	/gëža'ga/	[gëža'ga]
Chi	/'gašu/	['gaš:U]

(ii) moler grind *vt* *xxx

WIx	/betuna/	/utuna/	/rutuna/
	['bet:unʌ]	['ut:unʌ]	['rut:unʌ]
Ytz	/'beteʔ/	/'goteʔ/	/'JoteʔJ/
	['betɛʔ]	['hotɛʔ]	['Jotɛʔ]
Ylg	/beteʔ/	/goteʔ/	/Joteʔ/
	[be't:eʔE]	[go't:eʔE]	[dzo't:eʔE]
SZm	/beʔetu/	/goʔotë/	/'řoʔotë/
	[beʔetu]	[goʔotë]	['řoʔotë]

Again, the verbs are given in three forms, completive, potential and habitual.

[2] The glosses 'casa' and 'hogar' were provided by my data sources. To my knowledge, optional versus obligatory possession is not a component of the meaning difference between those two Spanish words. Perhaps better glosses would be 'casa (opt. possessed)' and 'casa (oblig. possessed)'. It is also unfortunate that the Zapotec forms glossed 'casa' are not apparently morphologically related to those glossed 'hogar'. I included the 'casa' and 'hogar' pair for demonstration purposes because it is the only pair in my data base so far that manifests any notion of possession.

[3] Some of the forms in these two sets may be in the wrong set. In my questionnaire, I neglected to specify whether I wanted a question word or a relative pronoun for 'where', and not all of the responses specified which they gave. I have tried to sort based on which words looked most likely to be cognates to the Ate, Ist, Qui and Rin forms, of whose placement I am sure.

[4] I should have two forms of the words in this record, the third person stem and the first person form, but I have only collected one form in each language so far. See Section 4.2.6 for discussion.

REFERENCES

REFERENCES

Anttila, Raimo. 1972. An Introduction to Historical and Comparative Linguistics. New York: MacMillan Publishing Co., Inc.

Arlotto, Anthony. 1972. Introduction to Historical Linguistics. Lanham, Maryland: University Press of America, Inc.

Bartholomew, Doris A. and Louise C. Schoenhals. 1983. Bilingual Dictionaries for Indigenous Languages. México, D.F.: Instituto Lingüístico de Verano.

Bloomfield, Leonard. 1933. Language. New York: Holt, Rinehart and Winston.

Buck, Carl Darling. 1949. A Dictionary of Selected Synonyms in the Principal Indo-European Languages. Chicago: University of Chicago Press.

Butler, Inez M. 1980. Gramática Zapoteca. México, D.F.: Instituto Lingüístico de Verano.

Button, Ella Marie. 1984. Abbreviated Guide to Dictionary Format and Punctuation. México, D.F.: Instituto Lingüístico de Verano.

Crystal, David. 1985. A Dictionary of Linguistics and Phonetics. Oxford, England: Basil Blackwell, Ltd.

Egland, Steven. 1983. La Inteligibilidad Interdialectal en México: Resultados de Algunos Sondeos. México, D.F.: Instituto Lingüístico de Verano.

Felger, Richard Stephen and Mary Beck Moser. 1985. People of the Desert and Sea: Ethnobotany of the Seri Indians. Tucson, Arizona: University of Arizona Press.

García-Pelayo y Gross, Ramón, editor. 1964. Pequeño Larousse Ilustrado. Paris: Ediciones Larousse.

Grimes, Barbara F., editor. 1988. Ethnologue: Languages of the World. Dallas, Texas: Summer Institute of Linguistics, Inc.

Hockett, Charles F. 1958. A Course in Modern Linguistics. New York: The MacMillan Company.

Johnson, Mark. 1985. Computer Aids for Comparative Dictionaries. Linguistics 23:285-302.

Josserand, Judy Kathryn. 1983. Mixtec Dialect History. Ann Arbor, Michigan: University Microfilms.

Kaufman, Terrence. (unpublished manuscript, 1987) The Phonology and Morphology of Zapotec Verbs.

MacLaury, Robert E. 1970. Ayoquesco Zapotec: Ethnography, Phonology and Lexicon. México, D.F.: M.A. thesis, Universidad de las Américas.

MacLaury, Robert E. To appear. Zapotec Body-Part Locatives. (will be in IJAL; written October 1987, U. C. Berkeley)

Marks, Donna Louise. 1976. Zapoteo Verb Morphology: Categories and tonomechanics with special attention to Sierra Juárez Zapotec. M.A. thesis, University of Texas at Arlington.

Merrifield, William R. 1981. Proto Otomanguean Kinship. Dallas, Texas: International Museum of Cultures.

Miller, Wick R. 1988. Computerized Data Base for Uto-Aztecan Cognate Sets. (unpublished manuscript, University of Utah)

Morris, William, editor. 1969. The American Heritage Dictionary of the English Language. New York: American Heritage Publishing Co., Inc., and Houghton Mifflin Company.

Nellis, Donald G., and Barbara E. Hollenbach. 1980. Fortis versus Lenis in Cajonos Zapotec Phonology. IJAL, 46-2:92-105.

Nellis, Neil, and Jane Goodner Nellis. 1983. Diccionario Zapoteco de Juárez. México, D.F.: Instituto Lingüístico de Verano.

Pickett, Velma. 1955. Isthmus Zapotec Verb Analysis II. IJAL. 21:217-232.

Pickett, Velma. 1980. Vocabulario Zapoteco del Istmo. México, D.F.: Instituto Lingüístico de Verano.

Pickett, Velma. Prepublication ms. 1985. Comparación de dos 'dialectos' Zapotecos.

Pullum, Geoffrey K., and William A. Ladusaw. 1986. Phonetic Symbol Guide. Chicago: University of Chicago Press.

Speck, Charles H. 1978. The Phonology of Texmelucan Verb Irregularity. University of North Dakota: unpublished M. A. thesis.

Williams, Edwin B., editor. 1978. The Williams Spanish and English Dictionary. New York: McGraw-Hill International Book Company.