Chapter 10. The written record: grammars, dictionaries, and text collections

10.1. Grammars

10.1.1. Why write a grammar?

I once heard a colleague remark that no one should write an article about a language until they have written a grammar of the language. I think he was serious. It is good advice, even though it is exceedingly stringent. For one thing, writing a grammar deepens and broadens one's knowledge of a language. Also, writing a grammar helps a linguist get the 'big picture', seeing how things fit together – kind of like helping to understand each tree better by understanding its place in the forest. This is what Pike (19--) would have called the 'field' view of language.

To my way of thinking, writing an article about a single construction or rule without a good grasp of the grammar as a whole is like writing about the function and form of a puzzle piece without knowing what the whole puzzle looks like. Of course, many languages already have grammars. So one could read a grammar instead of writing one. Certainly, noone should write about a language without at least reading entire grammars of that language when available. I would, on the other hand, tend to avoid working on languages where a grammar has been written, simply because so many other languages need to have descriptive grammars written.

In any case, reading a grammar is a poor substitute for writing one, if you really want to figure out how a language works. Writing a grammar teaches you that every part of a grammar is a series of decisions on how best to bring order from the chaos of data that the linguist has collected. There is more or less security in the usefulness of these decisions in different parts of the grammar. Often only the grammar-writer knows this, even when they are not at all trying to sweep recalcitrant facts under the rug. I can think of no more challenging, rewarding, important or urgent task for individual linguists than writing a grammar of a language based on their own field research. Ultimately, a grammar is the linguist's theory of how a specific language works, atomistically and holistically, i.e. what are its 'bits' and how do these 'bits' fit together. A grammar is the result of careful methodology, lots of hard thinking, innumerable conversations with native speakers, artistic flair, boldness (to propose connections or to say that 'x' does not exist in language 'y', etc.), lots of luck, and huge amounts of reading, planning, testing, and interpreting. A grammar may even identify a 'theme' of a particular language – recurring properties, semantic or formal or cultural or societal – that are found in multiple constructions, text types, social patterns, etc. throughout the language. A grammar should ideally include an ethnography of communication, for reasons presented in chapter __ above.[56]

So how does one do fieldwork with writing a grammar in mind? First, have a basic outline of a grammar in mind from the start of your fieldwork. One such outline is the *Lingua Descriptive Questionnaire* by Norval Smith and Bernard Comrie (this can be found on the Department of Linguistics website for the Max Planck Institute for Evolutionary Anthropolgy: http://lingweb.eva.mpg.de/fieldtools/linguaQ.html). Another is my own phonology questionnaire (see appendix ___, also on the MPI site above) or the phonology outlines of one of the volumes in the Oxford *Phonology of the World's Languages* series.

Another way to do fieldwork guided by questions relevant to writing a grammar is to invest time learning the craft of grammar writing by reading grammars that are

[56] Give sources on writing ethnographies of communication.

considered exemplary by other linguists, either for a given region of the world or in general. Linguists should read grammars for fun and for professional development. If you do not like to read grammars, you may be in the wrong profession.

However, ultimately, the questions you ask will reflect your own interests and background. What attracts your attention most in this grammar or language-culture pair? What kinds of things do you believe are in most need of being said about this language? What have you learned from native speakers about their view of their own language? What would they liked highlighted? (Alternatively, would they want anything omitted from public discussion, even grammatical forms, certain kinds of texts or semantic domains, etc.?)

First I begin with a discussion of the different kinds of grammars, including the following three main types: Reference, Descriptive, and Pedagogical Grammars, as well as the usefulness of 'grammar fragments'. I will emphasize that quality is more important than quantity, but that a certain coverage is required to get a 'feel' for the 'genius' of the language and knowing how and where to 'hang' each piece in the grammar. I discuss grammar-writing as a literature genre in relation to the remarks on its fallibility in Chapter One and the implications of this for different methods and attitudes towards grammar-writing. I also discuss ways of testing the grammar as a whole with native speakers, e.g. reading it to groups of them as it is written and after the entire grammar is in draft. I next turn to consider the task of dictionary-making, comparing and contrasting different kinds of dictionaries. I then discuss the importance and methodology of compiling representative collections of texts, framing the discussion in terms of at least the following parameters:

10.1.2. Grammars are the nucleus of documentation and description

The goals of a project of documentation and description are to provide a record of the sounds and meanings of a language and how these are associated with one another and with the culture in which they are embedded. Sound files of high quality, texts, and a useful and fairly extensive dictionary form the core of such documentation. But the grammar provides not only the description (the analytical account of how the pieces of the structure of the language fit together) but also is the tie that binds together the various components of the documentary evidence of the language's sounds, texts, and words. A complete documentary and descriptive project for a language will include the following components:

(10.1)
a. Descriptive grammar
b. Sound files (all prosodies and segments exemplified)
c. Dictionary
d. Texts
e. Ethnography of communication
f. Pedagogical grammar portions

So for most fieldworkers, even if their own objectives are not a full documentation and description of a language, contributing to the construction of a grammar of the language under study is at least an important goal to which their research should contribute. In other words, if the linguist is doing research on the phonetics of a language's segments, on information questions and dislocation, or on voice alternations, etc., the research should be written up in such a way that it can be

incorporated into a grammar of the language eventually. This is done by providing careful glosses, high-quality sound files, and relative jargon-free descriptions.

10.1.3. Types of grammars

10.1.3.1. Pedagogical vs. Reference Grammars

There are at least three types of grammars – pedagogical, descriptive or reference, and theoretical. A comparison between pedagogical and reference grammars is given in Table 1 below (see also the SIL website (http://www.sil.org/linguistics/GlossaryOfLinguisticTerms/WhatIsAReferenceGrammar.htm)

***Table One**: Reference vs. Pedagogical Grammars*

Type of Grammar	Audience	Purpose	Organization	Style of Presentation
Pedagogical	Native speakers looking for a tool to preserve and teach their language	To provide lessons in how to speak and write the language	In terms of short lessons on specific topics of pronunciation, word usage, etc.	Textbook like, for learners, to inform and to develop skills.
Reference	Linguists who want specific kinds of technical information and who will not believe everything you say.	To persuade and to inform. To provide a cohesive picture of the language that allows for both comparisons with other languages and understanding the 'genius' of the language at hand.	Things are presented in the way the linguist considers most effective to understand the language, but will usually be in the order phonology-phonetics-morphology-syntax-semantics-discourse	Clear for a wide readership, but technical when necessary and with sufficient argumentation to convince well-informed readers.

A pedagogical grammar is an educational tool and as such requires different skills and training from a reference grammar. Also, since the pedagogical grammar is largely, if not exclusively developed for the needs and uses of the language community, it must be built up slowly, based on experience in using it as a teaching tool, discussing it with community teachers, getting advice from educational experts, and so on. Moreover, a pedagogical grammar should be based carefully and solidly on a comprehensive

reference grammar. It should not be developed in isolation from the larger effort of describing, arguing, and reasoning about the structures of the language.[57]

Pedagogical grammar development can require skills in motivation as well. For example, if the community leadership wants the linguist to lead an effort for language revitalization, where a pedagogical grammar is to play a major role, the linguist will have to engage the active participation of native speakers and those trying to learn the traditional language of their people, to help shape and improve the pedagogical and analytical aspects of this specialized grammar.

A reference grammar, on the other hand, is designed to provide analyses, descriptions, and data in a cohesive, relatively standardized form to a general audience of professional linguists and other linguistically sophisticated readers. These can follow rigid guidelines (e.g. the Croom Helm descriptive series, see Everett & Kern (1997)) or allow greater freedom for the linguist to describe the grammar as they think best (e.g. the new Cambridge series of descriptive grammars (see Dixon (2004)). The former model, when part of a series, has the advantage of producing a range of grammars to serve as a comparative reference set, an encyclopedia of language structures appropriate for rapid comparisons. This is also its disadvantage, however, because its detailed, meticulous, and highly constraining outline can be too confusing or force the linguist to distribute across various subpoints of a detailed outline a discussion that would have been best presented as a whole, a single portion of the grammar. Reference grammars have the general advantage of being directed towards all linguists and thus do not generally require any special jargon or theoretical background to read.

10.1.3.2. Theoretical grammars

The final type of grammar I want to discuss in this section is the theoretical grammar. The first in-depth application of generative grammar was G.H. Matthews's grammar, *Hidatsa Syntax* (Matthews 1964). Hale (1967, 341) concludes his review of this book by claiming that this grammar is "... an outstanding landmark in American Indian studies as well as a highly significant contribution to general linguistic inquiry. This work is very much a credit to modern linguistics." Chomsky further situates the theoretical grammar in modern theory:

"LSLT [Logical Structure of Linguistic Theory, Chomsky (1975), DLE] *and other detailed work of the 1950s (particularly G.H. Matthews, Hidatsa Syntax) at once revealed a tension between descriptive and explanatory adequacy. As soon as serious descriptive work was undertaken, it was discovered that available accounts of language, however extensive, barely scratched the surface; even the most comprehensive grammar provided little more than hints that sufficed for the intelligent reader; the language faculty was tacitly presupposed (without awareness, of course). The same is true of the most comprehensive dictionary. To attain descriptive adequacy, it seemed necessary to construct extremely intricate and complex grammars, radically different for different languages. On the other hand, to approach explanatory adequacy*

[57] In my Arts and Humanities Research Council project for the documentation and description of the Kisedje language, one of my research associates was an expert in indigenous education with more than fourteen years of experience in education among indigenous groups of the Xingu Park, where the Kisedje live.

it was necessary to assume that the states attained are determined to an overwhelming extent by the initial state, which is language-invariant. Thus languages must all be cast to the same mold, differing only superficially. The major research project was aimed at overcoming this tension by showing that the apparent complexity and variety of language was only superficial, the result of minor changes in a fixed and invariant system." Chomsky (1994)

Moreover, as Hale's review shows in considerable detail, Matthews's book provides detailed argumentation and explicit structures for most major aspects of Hidatsa syntax. In addition, for researchers working within Generative Grammar in the early to mid 1960s, *Hidatsa Syntax* provided a model for how to write a grammar. Judged when it appeared, therefore, this volume was a pioneering, extremely important study, almost universally accepted by theoreticians as a significant improvement over previous grammars.

The problem is that today, forty-one years on from the publication of *Hidatsa Syntax*, this grammar is almost universally lamented as being of very little use to anyone wanting to learn about either Hidatsa or Generative Grammar (the latter because the field has changed so dramatically in the intervening years). So this once proud grammar is today relatively useless, exercising almost no influence at all, whereas the so-called 'taxonomic' (i.e. reference or descriptive) grammars it was once so favorably compared to continue to be read and serve to enlighten new generations of linguists about the languages they describe. And in my opinion and experience the same fate has befallen every theoretical grammar written since.

This is *not* to say that theoretical grammars should be avoided. But we need to draw at least two lessons from the fate of *Hidatsa Syntax* and its ilk. First, if one does write a theoretical grammar, the theoretical discussion contributes to very short-term (howbeit potentially important) goals, a sort of rock in the stream that can divert the flow around it, though it itself is long forgotten downstream. Another lesson is that it is useful to reconsider the organization of a theoretical grammar. My PhD on Pirahã, for example (Everett (1990 [1983]), was divided into two major sections, a reference grammar (also published in English as Everett (1986) and a theoretical discussion. The theoretical discussion is no longer particularly relevant, but the reference grammar portion can still be used by linguists regardless of theoretical orientation. To sum up, the ideal grammar should be useful for present and future generations; it should be clear; it should anticipate reader questions and objections and deal with these by detailed argumentation (except in the case of a pedagogical grammar).

Let us conclude by considering a couple of ways to enhance the usefulness of the grammar. First, the grammar must be tested. This can be done by re-checking the entire grammar, or at least the parts that are controversial, theoretically significant, or for which the linguist feels less secure, with native speakers in groups and with individuals. Second, be sure that the grammar includes argumentation for every major analytical assertion. If the linguist says that 'Stress goes on every other syllable from right to left within the word', then they should explicitly alternative analyses (e.g. stress first picks out the rightmost syllable and then goes from left to right), by considering hypothetical or ungrammatical forms, by spectrographic analysis, etc. (see ___ above). Argumentation is at once a display of erudition and reasoning and tremendously enhances the reader's opinion of the grammar-writer, the usefulness of the grammar, and the likely reliability of the grammar as a whole. In a reference grammar, argumentation will be less theory-internal than in a theoretical article, but it is

nonetheless crucial. Finally, tell the reader something about the empirical basis for the grammar. What size is the corpus? What is the corpus like? How long was the linguist in the field? What was the role of the community in the analysis (including how language teachers were involved)? How many hours a day did the linguist work and how did they work (e.g. 'I was in the community for six months and I worked three hours per day with language teachers and five hours a day filing, analyzing, and retesting examples by walking around the village trying out my knowledge). And so on.

I turn now to consider the corpus upon which the grammar is based.

10.1.4. The corpus for the grammar

To write a grammar of any type requires a significant corpus of data. What is the nature and size of an 'adequate' corpus for a grammar? The answer is simple: an adequate corpus must be varied, natural, and big enough. Let's consider each of these terms.

First the corpus should be varied. That is, it should contain the greatest possible number of distinct form, meaning, and construction types. It should vary for ages of language teachers, emotional states, content, social class of teacher, genders, speeds of utterance, topics of discourses, genres of discourses, styles of discourses, and channels of discourse (see ___ for the concept of 'channel').

The corpus should be natural, providing data that native speakers utter (in the appropriate context) or naturally reject (That is, a good corpus contains (tested) ungrammatical examples as well). This latter point is very important theoretically and methodologically. Theoretically, the presence of ungrammatical utterances recognizes that there are many facts about speaker's knowledge of their language that cannot be learned simply by observing what they say or do not say. Methodologically, however, this raises the issue of how to collect grammaticality judgments. So before considering the final question, i.e. when is the corpus 'big enough', we need to take up more carefully the implications of ungrammatical examples in field research.

Ungrammaticality is a difficult notion, as Schutze () and Cowart () demonstrate at length. The reason that ungrammaticality is problematic is that utterances can be rejected by native speakers for a variety of reasons. Let me give an example from my fieldwork on the Wari' language. As Barbara Kern and I were working on the grammar of Wari' (Everett & Kern (1997)), a particular construction (see Everett (2005c)) type struck me as extremely weird. I could not believe that the facts were exactly as Barbara had described them to me. Even though Kern's knowledge of the language seemed impeccable and even though I had never had reason to doubt her before, I just found these facts hard to take in. When I raised this concern, Kern's response was typical of her – "Then go to the village and check them out with the Wari' for yourself." Since we were working only a day's travel from the village, I took her suggestion and departed the next morning. I met thirty or so Wari' speakers on the banks of the Mamoré river on the Brazil-Bolivia border and we sat down to consider the examples.

As I read a couple of examples out loud, the people seemed mystified. "That makes no sense at all", said the man on my right. "Who told you we say that?" asked the man on my left. So just as I thought, these example are ungrammatical! Then I answered the man on my left, "Barbara told me you say these things." Surprised looks all around. Someone then asked, "Barbara said that? Hmm. Then we must say that. Let's think." So people began discussing the example and then they began to smile. "Oh, yeah, we *do* say that. If we are talking about something and want to say something about it like this, then we say it just like you said there. Just like Barbara told you." The

utterances were now perfectly fine, we discussed them, they corrected my pronunciation (segmental and prosodic) and I was satisfied that they were indeed grammatical utterances. But this was clearly a nonstandard way of getting at their grammaticality. In the normal course of events in, say, English, linguists simply ask individual language teachers if this or that utterance is grammatical. Yet the Wari' example shows that speaker's initial reactions can be quite misleading.

Why and when is an utterance accepted or rejected? The Chomskyan distinctions between competence (grammatical factors) and performance (cognitive factors, among others), as well as the Saussurean distinction of *parole* vs. *language* (social factors), when applied to such judgments predict that speakers can accept or reject utterances for a variety of reasons. For example, the sentences in () and () are grammatical but unacceptable, both famous examples in the literature:

(10.2) Buffalo buffaloes Buffalo buffaloes buffalo buffalo Buffalo buffaloes.
Meaning: Buffaloes from Buffalo, NY who hoodwink buffaloes from Buffalo, NY hoodwink buffaloes from Buffalo, NY – i.e. this is a 'center-embedded' tautology.

(10.3) Oysters oysters eat eat oysters.
Meaning: Oysters that oysters eat (also) eat Oysters.

Both examples are hard to process but are grammatical, as linguists have known for decades. The traditional distinction, based on the competence-performance dichotomy, is to say that both are grammatical but unacceptable. Sometimes one hears too that an utterance can be acceptable but ungrammatical, as in the fairly trivial example I have constructed in ():

(10.4) Q. Who would do such a childish thing? (Wife to husband)
A. Me do it. (Husband replies sarcastically to wife.)

The answer in () is not grammatical, but it is an acceptable utterance in that circumstance for the purpose of expressing sarcasm. Although we can thus make the distinction between 'acceptable' and 'grammatical' conceptually clear, separating them practically in fieldwork is nontrivial (see also Schutze ()). Moreover, the difficulty of distinguishing these vital concepts, in a Whorfian twist, becomes even more intense in the lack of any terms for these concepts. That is, there are relatively few languages in the world that actually have metalinguistic terms like this such that the linguist can actually ask "Is this grammatical?" or "Is this acceptable?" And circumlocutions such as "Is this 'pretty'?", "Is this OK?", "Can I say this?", "Is this good English?", etc. are all extremely unreliable.

Consider "Can I say this?" (or "Can you/one/a Pirahã/a speaker of your language say this?"). In my experience language teachers often say "Yes, you can." without the slightest regard for the acceptability or grammaticality of the utterance. This is partially because the linguist is paying to talk so, to some language teachers, the linguist can say anything they damn well please. None of these expressions in fact can be relied upon to give clear evidence of either the acceptability of an utterance or, a fortiori, its grammaticality. There are simply too many variables for the fieldworker to know why a particular answer has been given – does the speaker's judgment reflect pragmatics, semantics, syntax, phonology, lack of concentration, desire to please, or feelings of different status? There is no foolproof way to know the answer to these questions.

So what can the fieldworker do? Well, one thing is to get the speaker to confirm his or her broad assessment (e.g. "Yeah, that is OK.") by uttering the sentence themselves. Than the linguist can repeat it slowly and ask for yet another repetition. In my experience, the native speaker will not repeatedly utter an ungrammatical sentence. They will 'edit' it, i.e. change it slightly to make it grammatical. Or they will refuse to say it. (If the utterance is grammatical there often will be no problem for the speaker to repeat it.) The kind of correction made should identify what was wrong with the example structurally. On the other hand, if the example is *pragmatically* bad (roughly, if it is inappropriate in the environment of the lab session), it is quite likely that the speaker will still not say it. In this case, the linguist cannot tease apart the ungrammatical from the inappropriate.

One way around this latter problem is to work with multiple language teachers simultaneously when checking for grammaticality. Recall from the earlier example from Wari' that I only discovered that a particular kind of construction was possible by working with a group of speakers who discussed the example among themselves. Usually language teachers attempt to be helpful and when several are working together they tend not to accept snap judgments by one of their group. They discuss and refine. Therefore, checking grammaticality judgments with a group of speakers has benefits. On the other hand, none of this guarantees the field linguist that they know exactly, precisely why a given utterance is rejected. Moreover, even utterances that are acceptable could be ungrammatical, as we have seen. So what is a poor linguist to do?

They must carry out grammaticality tests with multiple speakers (serial language teachers and groups, see ___ above). Where the results of these various interviews produce a unanimous judgment, one way or the other, the linguist can have *moderate* confidence in the result. But it is still necessary to test and interview for *truth conditions* (see ___ above), for naturalness (in speakers' opinions), look for related examples in natural texts, etc. But in fact it is likely that the results of all these interviews will not produce a unanimous judgment but, rather, a mixed judgment. Speakers are quite likely to disagree, especially when they are not in the room together as they offer their judgments. In this case, the linguist could go with his or her best judgment (using their theoretical predictions for example), advising the reader of their grammar or article that this is what they have done and why. (Always, always be honest with your reader.) Or a more scientifically respectable approach can be adopted and the linguist can employ a statistical approach to grammaticality, following suggestions in Cowart (). After all, other social scientists deal with conflicting judgments and accounts of behavior all the time. Their reports, therefore, present statistical analysis of these different judgments, etc. and distinguish statistically significant vs. insignificant results. But using statistics and standard social science methodology requires more and different training of linguists in general and fieldworkers in particular.

The upshot is that determining grammaticality is never a straightforward matter. Judgments, testing, statistical analysis, text tracking, eavesdropping, etc. are all crucial. As a consumer of grammars, the linguist reader must always take all crucial grammaticality judgments with a huge grain of salt, in the absence of the corroborating studies just mentioned. All grammars ever done, past, present, or future, need retesting and can never be understood as a 'God's eye view' of the language in question. Once again, grammars are stories we tell about a language as we see it.

Let us turn now to the final question concerning the corpus for writing a grammar, namely, 'When will I have enough data?' Samarin (1967, ---) gives a number of useful

suggestions in this regard. I have borrowed some of Samarin's suggestions and added some of my own to come up with the list in ():

(10.5) A complete corpus is obtained when:
a. All the closed classes of linguistic elements are fully accounted for
b. When there are no 'holes' yet in the data needed for analysis (partially, therefore, the answer depends on theory)
c. When there are multiple tokens of all types.
d. When it is maximally useful for other disciplines, as well as linguistics
e. All new material collected only contains structures and meanings already found in the corpus collected.

Let's consider each of these points in more detail. First, what does it mean to say that all the closed classes are fully accounted for? This simply means that when you have all the prepositions, all of the adjectives, all of the verbs, i.e. classes with a small number of members that do not expand in membership. In some languages verbs will be in the open classes of lexical items (e.g. English), while in others they will be among the closed classes (e.g. Mosetén, Sakel ()). How do you next determine that there are no 'holes' in the data? Well, you have to have a view of how language works, partially based on general principles shared by most linguists and partially based on the particular theory that you are most influenced by. And you must be able to argue for your conclusions. On the latter, see ___.
Table –
The phonetic segments of hypothetica

p	t	k
b	d	
m	n	ŋ

There is a missing segment in Table -, i.e. a voiced velar, [g]. Is this an accidental gap or an actual asymmetry in the segmental inventory? The linguist will need to look for examples of [g]. At some point, they might conclude that the system is indeed asymmetrical, certainly not all that uncommon. But until they can say with confidence that this is the case, the corpus is incomplete.

It is also important to ensure that for every segment, prosodic pattern, syntactic construction, suffix, etc. in the language, that the corpus includes multiple tokens of each. And the linguist's analysis must be the guide as to when there are enough tokens of each. One useful criterion in answering this question is 'Are all tokens I am now recording simply repeating the patterns that I already have?' If so, then there are probably sufficient tokens in the corpus. However, one cannot simply rely on texts to magically produce all the tokens and their distributions that are necessary for a complete corpus. The fieldworker must *think*, based on his or her analysis and ask questions like the following: 'If my analysis is correct, then there will be forms of interpretation/shape **x** but never forms of interpretation/shape **y**.' Then the linguist must look for the missing forms, both those they predict to be missing (no matter how long they search) and those which they predict to be found eventually, but which are currently absent (i.e. accidentally) from the corpus. The linguist must be able to assure the readers of the grammar that the corpus is complete by this metric.

The corpus should also be maximally useful for other disciplines. The fieldworker may be working on a rare language that few people are likely to have access to. In this case especially, but in all cases ultimately, the linguist should collect texts and data relevant to other disciplines insofar as they have time and knowledge to do so. Text collections should include all important cultural values, to the degree that the community is willing to allow access to these. Claims about numerals or counting should be accompanied by experimental evidence corroborating the claims (even if this means bringing in an expert consultant). And so on. Finally, once all new data appears to contain no new structures, etc. then the linguist can consider that, with respect to his or her current working hypotheses and purposes, the corpus is complete. But, as we have been saying, the 'complete corpus' is a relative, never an absolute concept.

We move now to another core component of language documentation, the dictionary.

10.2. Dictionaries

The traditional view of the dictionary in linguistic theory up until twenty-five years ago, and still widespread, is that the dictionary is an asylum for the misbehaved, i.e. where we put forms that are not derivable by regular rules of syntax or phonology. People who work under this view may be tempted to produce trivial dictionaries that are little more than lists of words, idioms, and morphemes. But this would be a mistake, even for those with the 'asylum' (or 'jail' – see Williams and DiSciullo ()) view of the dictionary, because it renders the dictionary less useful. A dictionary is formed by a view of its potential users, not merely by a particular theoretical perspective.

In their volume, *Making Dictionaries: Preserving Indigenous Languages of the Americas*, the editors address the purpose of making a dictionary:

> "*A reasonable person might ... ask, Why do it? One way to read the contributions to this volume is as personal answers to this question. But a more general response can be discerned in all the chapters and, indeed, in the work of every lexicographer. There is something at once both marvelous and practical about producing a guide to the mind, world, and behavior of a group of people. The benefits that accrue from such a handbook – literacy, preservation, history, discovery – only add to the excitement of seeing the published dictionary standing upright on the shelf.*" (Frawley, Hill, and Munro (2002, 2-3)).

The editors go on to suggest that the ten most important issues in compiling a dictionary, with immediate application to languages of the Americas, but with a clear pan-geographic relevance are as follows:

(10.6) Ten crucial issues in dictionary compilation:

a. Entries
b. Theory
c. Literacy
d. Graphics
e. Role of the community
f. Types and numbers of dictionaries
g. Historical information
h. Technology
i. Presence or absence of dictionary-making tradition
j. Handling exceptions

Let's consider these in turn:

Entries: how is the *headword* (the objects of definitions, possibly with subentries) determined? Should the headword be a 'basic form' (e.g. citation form)? How can the fieldworker decide on a basic form?

According to the editors, ultimately the choice of headword for a dictionary will result from a "... tradeoff between the pressures for maximal explicitness and the desire to match the users' minds to facilitate their inferences as they fill in what must be left implicit." And further, "In the end, entries are a wager that the tension between the way the dictionary ought to look to the compiler and the way it feels to the user will not be too great." (Frawley, Hill, and Munro (2002, 5)).

Theory: How much information in each entry should be there for theoretical vs. applied reasons? How much should linguistic theory affect the overall form of the dictionary? Each entry should contain as a minimum sufficient phonetic, phonological, morphological, syntactic, semantic, and pragmatic information for the reader to know how to use and pronounce the entry and where it fits in the grammar and culture of the language. Some linguists build large amounts of additional theory into entries, while others see the dictionary more as a service to the language community and prioritize its utility rather than writing a lexicographic treatise on each entry. I favor the user-friendly view, though it is conceivable that linguist-only dictionaries can be done, in addition, if the linguist has inclination and time to do so.

Literacy: The dictionary may be the first or one of the first documents ever produced in the language under study. In this case its impact on literacy and discussions of representations of the language will be massive. But regardless of when the dictionary appears in the literary history of the the people, it will be or can be an important part of their self-identification and 'represents' their language to themselves and to people outside the community, the latter especially if the dictionary is bilingual (see ___ below).

An issue that arises in this regard is the extent to which the national or other major language(s) should influence the orthographic representation of the vernacular. For example, in Romance languages, as in many other languages, vestiges of history are included in the national orthography. So, consider the words in () from Brazilian Portuguese:

(10.7) a. casa 'house'
b. cicatriz 'scar'

As is well-known, in Romance languages, the old palatalization rule from Latin, whereby the voiceless velar occlusive and the voiceless coronal affricate are in complementary distribution such that the latter occurs before high front vowels and the latter elsewhere, continues to be encoded in the orthography by using a single letter, 'c', for both. (Also imported to some degree into the representation for loan words in English, e.g. 'electric' vs. 'electricity'.) So in some cases 'c' is [ts], in others it is [k]. And to complicate matters further, [k] is represented in still other cases, i.e. to represent the voiceless velar before a high vowel, by [qu], as in ():

(10.8) a. quinze 'fifteen'

This representational inefficiency is OK for a language with a long history of literacy and in which it is apparently desirable to continue to maintain some of the history of the language (and avoid 'imperialist' symbols, e.g. 'k'!). But should it be imported into a language that does not share its history, a minority indigenous language, say? Should the language of field study be obliged to share the inefficiencies of the national language? The questions become more radical and urgent when considering whether to use national scripts, such as Chinese, Devanagari, etc. or Western symbols. These are just some of the reasons why the dictionary is at once an important milestone in the literacy of a community and why it must be subject to community discussion and approval. The linguist cannot simply present a dictionary as a 'gift' to the community, but must develop it in conjunction with them.

Graphics: Will your dictionary include illustrations, photos, different colors, etc? These will add to its expense (in print media in particular), but funding for well-planned dictionaries is readily available from a number of sources, so the expense should not be allowed to become overly discouraging. Well-designed and utilized graphics can enhance the dictionary's usefulness as a tool for literacy since beginning readers can use different graphics to figure out on their own the meanings and form of certain words. Of course, graphics can get out of hand and create complications out of all proportion to their benefits. So once again the use of graphics has to be constrained via discussions between the linguist and the speech community.

Expense of graphics can be tremendously reduced of course if the dictionary is electronic, rather than (or complementary to) print media. See __ below for more on this.

Role of Community: This is discussed throughout this section.

Number of dictionaries: By and large the average documentation project could be rightly proud of producing a single high-quality dictionary. But other types of dictionary are possible, e.g. dictionaries of place names, flora and fauna, thesauri, cultural concepts and artifacts, etc. More specialized additional dictionaries will not be something the average fieldworker expects to produce, though communities could change the linguist's mind and plans by requesting these. In such a case, the linguist needs to determine what the community wants in terms of specialized dictionaries and figure out how to produce it.

History: How many dictionaries and how much linguistic and, particularly, lexicographic work has been done? To what extent should dictionaries track changes in lexical meaning? How much, if any, space should be dedicated to etymologies in the dictionary? The answers to such questions will depend partially on the linguist's training and interests, partially on community wishes, and partially on the view of the readership most likely to use the dictionary outside of the community. It is obvious that the more historical information the better. This enriches the dictionary and its role in the cultural heritage of the people. However, at the same time, it is nontrivially complex to provide etymologies and such an endeavor can get out of hand rapidly in the hands of someone with little training in etymology.

Technology: There are a number of advantages to constructing an electronic dictionary, whether as a complement to a print dictionary (recommended) or as the sole dictionary (not recommended). Just a few of the things that an electronic dictionary can do better and more cheaply than a print dictionary are: better graphics, audio files (providing pronunciations of each entry and many examples), easier dissemination, easier connections and cross-references from entries to the text or portion of field data from which they come, much greater and easier searchability, permanence (if done properly, an electronic dictionary can last forever), accessibility (more people have access to the dictionary), lower cost, easier to look for and predict trends in semantic change of lexical entries, and much greater interactivity – changes, corrections, and updating the dictionary become much easier when it is in electronic form. Anyone who has consulted an on-line dictionary source, e.g. one finds on internet sites such as Yourdictionary.com (http://www.yourdictionary.com/) knows that being able to listen to the pronunciation of entries, being able to find entries almost instantly, etc. are tremendous advantages of electronic media dictionaries.

Ultimately, however, the crucial thing to remember once again is akin to Postal's Maxims – the linguist and the community must see dictionary making as a fluid process, subject to negotiation, constrained only by imagination and budget.

We conclude the chapter now with a discussion of the nature and importance of text collections.

10.3. Text Collections

What is the purpose of collecting texts? Well, for some linguists texts should be the core component of the corpus upon which analysis and the grammar are based. To some degree, I agree with this view, although as has been pointed out numerous times in this and other chapters, texts are but one type of data that is essential for a good corpus. They are important because they are revealing with respect to the interaction of language and culture and they provide natural data, often providing examples and structures that the linguist would never have discovered on his or her own. It ought not to be difficult for any fieldworker to understand why text collections are important to the enterprise of linguistics in general and fieldwork in particular. Accepting the importance of texts, then we come to questions about the nature of textual documentation. First, what kinds of texts should be collected? Second how many texts are needed? As to the kinds of texts, my suggestions are in ():

(10.9) Kinds of texts needed:

a. All genres (narrative, procedural, hortatory, expository, and any other type that the fieldworker identifies as relevant for the language).

b. Texts that cover all significant cultural topics, e.g. life, death, harvest, hunting, dealing with the outside world, creation, fiction, history, etc.

Different genres will reveal different kinds of linguistic information. Different moods, aspects, tenses, participant coding and tracking, different kinds of logical connectives, etc. all are associated to a greater or lesser degree with different text genres. And anthropologists, ethnographers of communication, and linguists will all benefit from a rich array of text topics, linked to the culture (which implies that the linguist must understand the culture more than superficially, in order to know which topics to ask for texts on).

Next how many texts does the fieldworker need to collect? This gets us back to the issue of the 'complete corpus' and is answered by the considerations given in the previous section. However, as a rule of thumb, I would suggest at least two hundred pages of transcribed texts with morpheme by morpheme glosses, single spaced, no greater than 11 point font.