Essentials of Language Documentation

edited by

Jost Gippert
Nikolaus P. Himmelmann
Ulrike Mosel

Mouton de Gruyter
Berlin · New York

Documenting lexical knowledge

John B. Haviland

Introduction

Lexicography, the praetice of documenting the meanings and uses of "words" (literally by "writing" them down), is, through its products, perhaps the most familiar branch of linguistics to the general public. It is also an ancient and much theorized activity. In the Boasian trilogy for language description of grammar, wordlist, and text, it is surely the dictionary whose compilation is most daunting. The process begins with a learner's first encounters with a language, and it ends, seemingly, never. Worse, it is an endeavor fraught with doubt, centrally about when enough is enough both for the whole – when one should assume that the basic or most common words of a linguistic variety have been captured and characterized – but also for any single putative dictionary entry, given the apparent endless variety of nuance and scope for words and forms, not to mention the idiosyncrasies of compound or derived expressions. Moreover, despite bounteous speculation, from many disparate linguistic traditions, on what metasemantic devices one might employ to capture meanings, despite multiple models and examples of the results of dictionary-making, and despite ample experience, for most of us, in the ordinary business of "explaining the meanings of words," doubt is likely to assail us on every single effort: have we said enough? have we forgotten something? did we get even this single word right?

This chapter introduces techniques and concepts relevant to producing a lexical database as part of a language documentation project. I concentrate on a series of doubt-producing obstacles for the field lexicographer, with some suggestions about how at least to address, if not to overcome them. My coverage is deliberately partial. I draw heavily on my own fieldwork in Mexico and Australia, to consider three general issues. First, I review familiar morals about the nature of word meaning – concepts from linguistic philosophy that are easy to forget in the heat of the lexicographic moment. Second I consider semantic metalanguages proposed to deal with different kinds of meaningful elements, from "functional" to lexical and from roots

to stems. Third, and most centrally, I review techniques for systematically extracting lexical knowledge. I largely ignore several related and important topics: lexical variation and how to represent it (see Chapter 5), ideological issues inescapably involved in promulgating any dictionary (see again Chapter 5, and the discussions in Frawley et al. 2002), and wider issues in lexical semantic theory (about sense relations, problems of extension vs. intension, etc.), which underlie all lexicographic practice but are beyond the present scope. I begin with a highly selective review of published materials on lexical knowledge, especially as relevant to documenting endangered languages.

1. Lexicography and its products

In addition to a large theoretical literature on meaning, there is a practical tradition of dictionary-making that has spawned handbooks and histories, as well as essays on the lexicographer's craft. These rarely provide solace for the field worker.

The lexicon, in modern linguistics, has come to mean a repository for otherwise anarchic facts, an inventory of arbitrary pairings of pronunciations with bundles of features. It is where language stores its idiosyncrasies and irregularities. What systematicity there is to the lexicon so conceived derives from feature systems themselves, taken to represent syntactic and semantic patterning underlying surface lexical forms. Studying such patterning is the usual province of lexical semantics, which catalogues various relations between the senses of members of different subsets of lexical forms (Cruse 1986), systematic properties of surface word classes or "parts of speech," facts of argument structure, diathesis, and the like. The main contribution to linguistic theory of much empirical lexicography has been in elucidating semantic and syntactic interrelationships at the level of the surface word (Levin 1993).

Field linguistics, once the province of anthropological linguists, gave rise to much of the underlying conceptual apparatus of lexical semantics. Early theories pursued an analogy between phonological features and the "components" of meaning in structured sets of "folk terminology," from kinship to ethnobotany, from pronoun systems to verbal typologies. The classic studies of "ethnoscience" investigated culturally elaborated lexical systems, particularly in "natural" domains like ethnobotany. Further empirical inspiration for semantic theorizing came, for example, from the languages of

Aboriginal Australians, celebrated for their linguistic acuity and creative genius. Dyirbal verb semantics and the properties of special Dyirbal "mother-in-law" vocabulary for affinal avoidance led Dixon (1971) to postulate a fundamental difference between semantically basic or "nuclear" words, requiring some sort of decomposition into sublexical meaningful dimensions, and non-nuclear words which could be *defined* in terms of the nuclear words plus other devices of the grammar. Verbal play in ritual language games learned by Warlpiri and Lardil initiates suggested that Aboriginal ethnolinguists had developed sophisticated semantic analyses of ordinary vocabulary (Hale 1971, 1982).

The classic referenee manual on lexicography is Zgusta (1971).[1] Of spccial interest to the field lexicographer is Frawley et al. (2002), a collection of essays by practieing lexicographers working on American Indian languages, which also considers problems in *creating* a lexicographic practice in communities without one.[2] These range over theoretical issues in lexical semantics (the nature of definition, the range of lexical knowledge that speakers possess or a dictionary might include, and the interplay between diachronic and synchronie lexical facts); to questions of representational form, to sociopolitical issues in dictionary making (for whom is a dictionary compiled and for what purposes; or, what kinds of sociolinguistic categories – speeialized speech genres, gender or class specific lexical forms, for example – are to be distinguished). These works go well beyond the limited selection of topics addressed here.

The field linguist need not be a semanticist, except "for practical purposes," and lexicography in the service of documentation needs to strike a balance between opposing desiderata. For example, in what sense is "completeness" – however that might be defined for an endangered language – something to strive for? What about the mix of theoretically versus practically motivated metalanguages for representing lcxical information? In the field one should avail oneself of all possible tricks: bilingual dictionaries, for example, can often start with existing word lists, in either the source or the target language, and there is no reason to stand behind strict methodological principles or purism in generating lexemes for incorporation into a lexical database.

Different lexicographic products reflect different starting points and goals for compilers of lexical databases. Zgusta (1971) dedicates separate chapters to the distinct issues involved in compiling polylingual (usually bilingual) versus monolingual dictionaries. The contrast, and the choice of

which languages to include in a multilingual dictionary, raise obvious questions. For what sort of use is a lexical database produced? What knowledge on the part of the user is presupposed in its design? Why did its compiler produce it in the first place? Let me review several different kinds of field dictionaries, related to my own research in Mexico and Australia. Especially useful to me have been the introductions to two Tzotzil dictionaries by Robert M. Laughlin (1975, 1988), one modern and the other based on a sixteenth-century work.

In what I call the Colonial tradition, collecting vocabularies was always a vocation of imperialists, often an accidental byproduct of exploration and conquest. Explorers collected flora and fauna, and often they also collected words. Somewhat less innocent were the wordlists ereated explicitly to *aid* in conversion, conquest, and control. The friars' dictionaries of Indian languages in the New World, or vernacular vocabularies destined for colonial bureaucrats in Africa and India, represented unabashedly instrumental "documentation," often of languages whose eventual endangerment was a byproduct of colonial expansion in the first place. Such wordlists were plainly not made "for" the speakers of the languages so documented.

The missionary tradition continues to produce many field dictionaries, and reading them gives some flavor of the purposes and populations served by this particular lexicographic practice. In Chiapas, Mexico, the Summer Institute of Linguistics – a Protestant Bible-translating organization – has published many dictionaries of Indian languages from the region (Delgaty and Ruiz [1978] for Tzotzil, Aulie and Aulie [1978] for Chol, to mention just two), and they are widely used even by speakers who do not share the religious beliefs of the translators. Such dictionaries are subtly infused with cultural metacomment and religious ideology.

Here, for example, is a translation of the entry in Aulie and Aulie (1978) for the Chol word *ajaw*, reflex of a root which means "lord, master, God" in other Mayan languages. According to the Aulies, the Chol word means "*espiritu malo de la tierra*," and they go on to comment:

> They call it *lak tat* 'our father.' It is believed that a person ean make a pact with it. Sueh a person can make requests of the spirit for or against another. The person who establishes such relations with the *ajaw* is called a "sacristán." If a man or woman offends the sacristán, the latter appeals to the spirit to curse the other, and in a short time the other person will die.

Herc both the lexicographers' voice and its underlying ideological accent are plainly on display. Thus, for the Aulies there is no apparent dissonance

between their proposed gloss, "evil spirit of the earth" and the alternate locution "our father" (with a first-person plural inclusive prefix). Furthermore, the 'they' of the comment is clearly someone other than the dictionary writers (though perhaps not different from the dictionary users). Note finally an interesting voicing contrast. Although the possibility of "making a pact" with *ajaw* is cited as something "believed" (presumably by 'them'), the consequences of the appeal on the part of the hypothetical *sacristán* (the term itself a Spanish loan introduced into Chol during the Catholic conversion of Chol speakers following the Conquest) are given a different epistemological status: "in a short time the other person will die." The dictionary thus incorporates different, perhaps mutually contradictory stances towards Chol beliefs and practices into the lexical entries themselves.

Slightly different is the "ethnolinguistic" lexicographic tradition, whose immediate origins are in ethnographic research. Sticking again to highland Chiapas, Laughlin's exhaustive dictionary of contemporary Zinacantec Tzotzil (1975) has the form of a traditional bilingual dictionary. The first section gives extensive glosses (in English) of Tzotzil words, both derived and simple, and arranged under their putative underlying roots. There follows an English index to the Tzotzil section. Laughlin's dictionary has over 35,000 Tzotzil to English entries, making it one of the largest dictionaries of an indigenous language of the Americas. However, it is a bilingual dictionary in Tzotzil and *English*, limiting its direct use to the handful of people who speak those two languages.[3] It is also a defiantly dialect-bound (and even gender-bound) dictionary, documenting the way middle-aged men spoke during the 1960s and 1970s in just the single municipality of Zinacantán, arguably a minority variant of what has since become a dominant Indian language in highland Chiapas with a much larger number of speakers from other dialects. Thus, the choice of language variety in the dictionary reflects accidents of the background research rather than principled lexicographic or sociolinguistic design. Moreover, grouping entries by a theoretical underlying root (a form which does not occur in speech, having only psychological rather than surface "reality"), and stripping words of all affixes – i.e. lemmatizing them – makes locating a word in this dictionary something of an analytical challenge, again, a reflection of the intellectual priorities of its producers, but with possibly inconvenient consequences for many potential Tzotzil-speaking users.

A different variant of the ethnolinguistic wordlist, from Australia, illustrates another aspect of the field lexicographer's dilemma. Many linguists have documented Australian Aboriginal languages with very few remaining

speakers, often not fully fluent. My own work on the now defunct Barrow Point language (see Haviland 1998) is a minor example. In such cases, wordlists reflect serendipitous opportunity more than systematic planning, and coverage is spotty, based on happenstance and luck. Nonetheless, even haphazardly assembled lists of words may be significant when political processes – for example, "native title" claims to traditional Aboriginal territory – use linguistic evidence to establish links between land and Aboriginal culture and society (Henderson and Nash 2002). Everything from a place name to a plant name may turn out to have unsuspected relevance. Thus the issue of coverage is less a matter of scientific "completeness" than an ideological issue of clear political import, another matter to which I return fleetingly at the end of the chapter.

There is also a *pedagogical* tradition in dictionary making, source of the most common dictionaries: those used by students to look up unfamiliar words, or by tourists to translate menus. Here the question of dimension is telling. Dictionaries of Mexican Spanish (for example, Lara Ramos 1986) are explicitly graded by size: a small version meant for schoolchildren with several thousand "basic" words, a larger intermediate version with more, and so on. All celebrate Mexican Spanish, the most widely spoken variety of the language, but one relegated to a subsidiary status by the language academy of the colonial home country. The lexicon chosen and the facts of usage are drawn from a huge corpus of Mexican textual material, from letters, to newspaper articles, to popular songs. In Chiapas, the government has similarly commissioned a variety of "*diccionarios de bolsa*" or pocket dictionaries for the Indian languages of the state. These, along with a series of grammatical sketches, are meant as both pedagogical tools and political trophies, evidence of government concern for Indians in the wake of the Zapatista uprising of 1994. Of a similar design but with an opposite ideological thrust are the illustrated school primers, or basic wordlists, designed as literacy aids by Zapatista community schools which resist all government aid and standardized school materials.

2. Referential indeterminacy and other pitfalls of fieldwork

What sorts of creatures are the "meanings" of words we wish to set down in a lexical database? It is hard to escape the weight of many centuries of Western philosophizing on the subject (although there are useful antidotes in J. L. Austin's early essay "The meaning of a word" in Austin 1961).

Following Frege (1892) it is customary to begin with the notion that words (characteristically nouns) can typically be used by speakers to pick out entities in the world – the words' "referents" – by virtue of their "sense" or "denotation" independent of any instance of their use for referring or predicating about a specific state of affairs. Words, on this view, are a kind of instruction from speaker to hearer, grounded in some shared understanding of the "meanings" of expressions, and typically designed to achieve common reference.

Even with apparently simple cases, of course, the conundrums of reference as a theory of meaning immediately surface. Suppose someone wants to refer to me as I am lecturing. Consider the following expressions she might use:

(1) Expressions referring to the same referent

a. That guy (with a pointing gesture)
b. The linguistics professor from Oregon.
c. The tall guy with a black moustache at the front of the room.
d. The Mexican with a black moustachc at the front of the room.

The speaker's "instruetions" if successful – that is, if they induce the interlocutor to pick me out as the person to whom she refers – rely on quite diffcrent sorts of relations to the "meanings" of the words she uses. The first relies on some sort of categorial understanding of what we can use 'guy' to refer to, combined with two direct indexical devices, the deictic *that* and the pointing gesture. At the other extreme, (b) picks out a presupposably identifiable individual from the intersection of sets of denotata generated compositionally from the constituent words (along perhaps with presuppositions of existence and uniqueness built into the definite article *the*). Expression (c) combines such a compositional strategy with some implied deixis (calculating *which* room and where its *front* is), and (d) paradoxically is likely to succeed as well as (c) despite the fact that, though I live and teaeh in Mexico and possibly even look Mexican, I am *not* a Mexiean at all – therefore, the "meanings" of the constituent words cannot add up to a true denotation.

So reference, although it is where we start in field linguistics, cannot be where we want to end up. Quine's famous *gavagai* example (Quine 1960) – in which a hypothetical and ontologically challenged linguist, in a parodied setting of monolingual fieldwork, hears the word *gavagai* in the presenee of rabbits, but eannot decide whether the word means 'rabbit' or 'rabbit part'

or 'rabbit essence,' etc. – underscores the profound referential indeterminacy of linguistic behavior. Perhaps more to the point is Zgusta's analogy (Zgusta 1971: 25–26) with trying to discover the meanings of traffic signs (in a system like the European one), but only on the basis of observing the regularities in drivers' behavior. Perhaps, speculates Zgusta, one could in time decipher the meanings of, say, the red, yellow, and green signals of a traffic light by direct observation; but the meaning of a "great capital H on a rectangular shield (which means in many countries that there is a hospital not far away)" would be much harder to divine, since such signs stand in many different kinds of locations and "a uniform effect on the behaviour of other drivers is hardly observable."

Here is a less fanciful example from the annals of real field lexicography. In 1770, Lt. James Cook and his crew collected wordlists from the Guugu Yimithirr language, spoken near what is now called Cooktown, in northeastern Australia. (One word was *gangurru*, the name for a partieular species of what the world now calls kangaroos). Collating the shared entries of different observers, one can see precisely that referential indeterminacy of the *gavagai* variety plagued these early lexicographers. Thus, under the gloss 'branch (with buds or stalk)', the ship's illustrator Parkinson has *maiye*, Banks the botanist writes *maye butai* (adding the annotation 'with leaves') or *mayi bambier*. Based on the modern language, I assume that these expressions are based on the word *mayi* 'edible plant' – so not just any old branch is involved – and more specifically *mayi bambiir* 'the (edible) fruit of the mangrove species called *bambiir*'. The other "name" Banks reeords is plainly the expression *mayi buday* which is really an entire sentence that means "the edible part has been eaten" or "someone ate the fruit."[4] Cook's journal entry shows he was painfully aware of sueh Quinean problems of lexical elicitation.

> ...the list of words I have given eould be got by no other manner than by signs enquiring of them what in their Language signified such a thing, a method obnoxious to many mistakes: for instance a man holds in his hand a stone and asks the name of [it]: the Indian may return him for answer either the real name of a stone, one of the properties of it as hardness, roughness, smoothness &c, one of its uses or the name peculiar to some partieular species of stone, which name the enquirer immediately sets down as that of a stone. (Cook's journal, see Cook 1955)

Part of the problem, clearly, is in a primitive model of both reference and ostension: what you can pick out by pointing, or what you can show "the Indian."

A very different model of "exemplification" is advocated by J. L. Austin in "A plea for excuses" (Austin 1961). Faced with a pair of expressions (famously, in Austin's case, the apparently similar *by mistake* vs. *by accident*) one elucidates the difference in their meanings by constructing a careful example of when you would use the first expression but not the second, and vice versa. In such a method one points not at *things* but at contexts of use.

Contexts themselves can be crucial in accessing lexical knowledge. In trying to recover words from the native Barrow Point language of the late Roger Hart, he and I worked largely through Guugu Yimithirr, a second language for both of us (see Haviland 1998). We would often search – sometimes quite naively – for the Barrow Point equivalent of a Guugu Yimithirr word. Even looking for the names of plant or animal species, however, we were often stymied, partly because the flora and fauna of Barrow Point were frequently different from those of Cape Bedford, more than a hundred kilometers to the south, but partly because the environment in general was just wrong. Roger had learned his tribal language before he was removed from his family around the age of six. I first heard him speak the language without hesitation, however, sixty years later. After a long trek back overland, he and I stumbled out onto the beach where he had been born. The country he had not seen for sixty years, its trees, rocks, and animals, seemed to speak to him in his childhood tongue, and he was only there able to respond fluently.

Referenee – or more precisely those aspects of linguistic expressions that render them useful for achieving reference – though the staple of most modern formal semantics, is of course an inadequate basis for understanding meaning in an ordinary sense. The traditional notion of "connotation," for example, is based on the intuition that different words can in some sense "refer to the same thing" without, thereby, "having the same meaning." This is not the same as Frege's elassic distinction between sense (what an expression means) and reference (what it just happens to refer to, as a function of what it means) where two different expressions, with different senses, can happen to refer to the same individual. Zgusta's somewhat quaint example is the lexical triad 'decease', 'die', 'peg out' (the last in my own dialect of English would be something like 'cheek out' or perhaps 'go belly up'). Zgusta (1971: 39–40) cites Armenian as a language which has exact counterparts (*vačxanvel, mernel, satkel*) for these English

words, and Chinese as another with a considerably more elaborated set covering the same referential territory. (We could, of course, add more English expressions, changing thereby the dimensions of "connotation" evoked: 'pass [away]', 'go [to a better place] or [to meet his/her maker]', 'croak', etc.) The way to capture the difference between the terms in question, presumably, is to specify not truth conditions on the states of affairs they are used to describe (which are stipulated to be identical) but appropriateness conditions[5] on the indexical circumstances of their use: who can use which expression, to whom, speaking about which sorts of deceased entities, and in what sorts of situations, among other things.

Zgusta likens the lexicographer's problem with connotation to others related to ranges of meaning, selectional restrictions, and collocational specificity. One of Quine's examples was 'addled': "used only of eggs and brains" (see McIntosh 1961). Zgusta cites Černý on two Georgian words meaning 'to have': *makvs* (applied to things one has) vs. *mqavs* (applied to persons and animals), "but motorcars are treated not as things but as animals because one says *mankana mqavs* 'I have a motorcar'" (Zgusta 1971: 44).[6] Berlin's (1967) study of Tseltal[7] verbs of eating in which different kinds of foods require one of six different verbs of eating exemplifies a parallel phenomenon. There are conceptual muddles here which there is no space in this chapter to untangle: whereas words with different connotations seem to be appropriate to different contexts of use, or different speaker attitudes, can we distinguish selectional restrietions from denotational limitations? Perhaps *makvs* denotes a different state of affairs from *mqavs*, not merely 'the samc eoncept' applied to different kinds of objects. Perhaps Tseltal *we`* 'eat (tortillas, for example)' is "really" a different action from *k'ux* 'eat (crunchy things, for example).' Whatever our semiotic theory, such systematic meaning distinctions elearly belong in a documenting lexicon: recording them is part of the lexicographer's "duty" and a task to which methodologieal attention must be directed.

Here, the problem of negative evidence (or rather the lack of it in naturally occurring talk) is critical in compiling a lexical database for an imperiled language. Evidence about limits on the range of meaning of a word or phrase, or about restrictions on its use or appropriateness in different intertextual and cultural contexts, may simply be non existent in a textual corpus, and systematic elicitation of specific lexicographic intuitions may be impossible. In the Colonial Tzotzil dictionary, for *pesar el negocio con cordura o diligencia* "treat a matter prudently or diligently" (Laughlin 1988), the friars gave an inflected version of the Tzotzil expression *-a`i ta-olonton*,

literally "hear (or feel, or understand) in the heart." The Tzotzil phrase requires morpho-syntactic completion: the transitive verb *-a`i* needs both a syntactic subject (the one who presumably "treats" some matter) and object (the "matter" treated). Moreover, the word *-olonton* 'heart' also requires an obligatory possessor, which judging by the modern language must be coreferential with the subject of the verb, thus "x hears with his/her OWN heart" – not, with someone else's. These morphosyntactic restrictions are not obvious from the original usage. Nor is it clear that the expression is limited to the sort of referential context suggested by the English (or original Spanish) gloss: it seems instead simply to suggest careful consideration of anything, whether a "*negocio*" 'matter, business' or something less specific or concrete. Without access to fully fluent native speakers it is impossible to supply more lexical detail. More problematic, and perhaps more relevant to documenting an endangered language, is the case of an archaic word, or one in limited use in a speech community. Again, Colonial Tzotzil provides an instructive example. The ritual language of modern Tzotzil uses the expression *tza-uk*, evidently formed from a (non-attested) nominal root *tza* plus an irealis or subjunctive suffix *-uk*. Laughlin (1975) suggests as a meaning for *tzauk* 'take heed' – a translation suggested by knowledgeable modern speakers. However, somewhat arbitrarily it seems, in the modern dictionary he lists the word under the root *tzak* 'catch, grab'. Only the discovery of the Colonial dictionary (Laughlin 1988) revealed an archaic root *tza* which has entirely fallen out of existenee in Zinacantec Tzotzil except for its surviving ritual use. The Colonial lexicographers reeorded it with the meanings "cleverness, cognizance, eraftsmanship, guess, industriousness, intelligence, opinion, prudence, skill, speculation, talent, thought," but no evidence is provided by modern usage.

Perhaps the oldest chestnut of anthropological linguistics is denotational diversity in lexical mappings of "reality," captured in the slogan that "different words" imply "different worlds." One classic domain is ethnoanatomy, the lexical (and thus, perhaps, conceptual?) slicing up of the body into discrete parts. Whereas English speakers distinguish 'hands' from 'arms', Russian and Tzotzil speakers do not. Tzotzil has the single root *k'Ab*[8] which can mean either 'hand' or 'arm'. Worse, it ean also mean 'branch', 'sleeve', 'crossbar (of a cross)', 'front leg (of a cat)', and so on. Tzotzil *ni`* 'nose' denotes not only noses, but any relatively sharp-pointed protrusion, or the thin end of almost any sort of object, not necessarily a face or a head. So why privilege a 'body part' gloss like 'hand' or 'nose'? Perhaps a non-anatomical modcl is involved in such partinomies.

Another possibility is that a "basic meaning" is extended in various ways into a chain or continuum of derived meanings without well defined end-points. Cruse (1986) argues that terms like 'mouth' in English participate in "sense spectra" where each "derived" or "metaphorical" meaning leads to another.

(2) "sense spectrum" (Cruse 1986: 71 ff.)

John keeps opening and shutting his *mouth* like a fish.
This parasite attaches itself to the *mouths* of fishes, sea-squirts, etc.
The *mouth* of the sea-squirt resembles that of a bottle.
The *mouth* of the cave resembles that of a bottle.
The *mouth* of the enormous cave was also that of the underground river.

The kinds of meaningful elements one chooses for a lexical database are also inextricably linked to the whole of one's categorial analysis for a language, what "parts of speech" are postulated, and what sorts of semantic profiles are associated with them. The standard formal semantic starting point that nouns will map onto things (i.e. sets), adjectives to "properties" (i.e. subsets), and verbs to events or states of affairs (predicates over n-tuples of entities), quickly disintegrates in the face of the diverse sorts of semantic *conflation* (Talmy 1985) routinely observed in lexical items. A standard example is 'climb' in English, whose Frame Net[9] definition is: "to move vertically usually upwards, usually with effort." That is, the verb suggests, in the default case, vertical movement upward, combined with the sort of effort Fillmore called "clambering." Either of these conflated elements – upward motion, or effort – can be suspended, but not both without semantic oddness.

(3) Conflation in *climb* (Fillmore 1982)

The snake climbed (up) the tree.
The monkey climbed (up/down) the tree.
?The snake climbed down the tree.

Another commonplace of anthropological linguistics is that languages conflate semantic domains in unexpected ways, perhaps most characteristically in verbs. For example, the following Tzotzil positional predicates all might receive a similar English gloss 'stuck'.

(4) Tzotzil words for 'stuck'

Kakal 'stuck (between two surfaces)'
Ch'ikil 'stuck (into a narrow or tight crevice)'
Katz'al 'stuck (in a jaw-like orifice)'
Xojol 'stuck (inside an enelosing hole)'
Tz'apal 'stuck (a pointed thing anchored in a surface)'

As the detailed glosses show, however, each word specifies different configurations, kinds of attachment, and different shapes, in both figure and ground.[10] The exact conflation, I believe, involves such factors as the following, taking the root *tz'ap* as an illustration.

(5) Conflation in *tz'ap*

a. the "end" of the Figure is "inside" the Ground;
b. the Ground need not have a *y-ut* 'inside' (or perhaps it must not be so structured, conceived of instead as a mere surface);
c. the Figure has a "pointed" "end" (in Tzotzil, *s-ni`* 'nose');
d. typieally the Figure is "stuck" into the Ground pointed end-first, i.e., attached somehow, and self-supporting; and
e. typically it is vertically oriented.

Linguists have posited various classifications of semantie types, in different root elasses, and the field lexicographer should borrow shamelessly from such typologies: from frames, to verb types (Dixon 1972), to verb classes based on patterns of diathesis (Levin 1993), and so on.

The multiplicity of "language games" – something that cannot long remain hidden from a serious field linguist – further complicates the traditional referential view of lexical meaning. We use words to refer; but also for many other things. Here is part of Wittgenstein's list:

> Giving orders, and obeying them – Describing the appearance of an object, or giving its measurements – Constructing an object from a description (a drawing) – Reporting an event – Speculating about an event – Forming and testing a hypothesis – Presenting the results of an experiment in tables and diagrams – Making up a story; and reading it – Play-acting – Singing catches – Guessing riddles – Making a joke; and telling it – Solving a problem in practical arithmetic – Translating from one language into another – Asking, thanking, cursing, greeting, praying.
>
> (Wittgenstein 1958: sect. 23)

Cruse (1986: 270 ff.) reminds us of the differences between what he calls "semantic modes," as in the contrast between the following two utterances.

(6) "Semantic modes"

I just felt a sudden sharp pain.
Ouch!

If semantics is only about reference and predication, then it will be difficult to capture the meaning of 'ouch!' semantically, because the word involves neither reference nor predication. Instead, it will be important to understand such things as interjections (see Kockelman 2003) in terms of very different semiotic modes: indexing speaker stance, interlocutor's relationship to speaker, putative bodily and affective states, expected responses, and so on. That words like 'ouch' are hard to model in terms of denotata does not relieve us of the lexicographer's responsibility of recording them and explaining how they work – a problem which I return to below.

A broader and more appropriate conception of meaning derives from one of the well-known trichotomies of ways that signs can signify or "stand for" other things, due to C. S. Peiree (1932). The three semiotic modes are based on very different principles, although they generally co-mingle in most signs, linguistic or otherwise. Peirce pointed out that some signs stand for other things because of a resemblance between the sign vehicle and the thing signified – thus a photograph of a person can stand for that person (for example, in a directory or catalogue). The sign bears an "iconic" resemblance to what it signifies, although the nature of the "resemblance" can vary tremendously (consider diagrams, drawings, silhouettes, graphs, for example, or conventionalized but nonetheless onomatopoetic words whose sounds suggest their meanings: 'moo' or 'caw' or 'cackle', perhaps). There can also be an "indexical" relationship between sign and signified, such that physical, spatial, or direct causal relationships exist between the sign vehicle and what it signifies. A footprint, for example, may not "resemble" the person who made it (although it may, of course, "resemble" his or her foot), but it stands as an 'index' of the person by virtue of the fact that it took the person's foot to make the mark (hence, indicating, for example, that that person has been at a certain place). In language, 'ouch!' stands for (indeed, displays) sudden pain precisely because we imagine that the pain itself somehow (involuntarily?) produees the utterance. In a similar way, we know what person 'I' or 'you' refers to by observing the contextual relationship between the sign – the word – and the person uttering it or

to whom it is uttered. Such words, then, rely on an indexical relationship (in a context) to convey their meanings. Finally, there are signs whose significance is essentially unmotivated by either resemblance or context: these are Peircean 'symbols' which rely on a conventional relationship between signifier and signified – Saussure's "arbitrariness" of the linguistic sign, in which 'cat' means cat only because that is what a particular linguistic tradition has legislated.

Figure 1 shows a sign which transparently combines all three Peircean semiotic modalities: the iconic resemblance between the drawing and a (stylized) smoking cigarette; the conventional meaning (at least in much of the Western world) of the shaded circle with the diagonal bar as a "prohibition"; and finally, the location of the sign itself, whose physical position signals indexically exactly *where* smoking is prohibited.

Figure 1. A semiotically trichotomous sign

An adequate description of the meaning of linguistic elements must capture all three modes of signification, although the major lexicographic traditions limit themselves largely to "conventional" or symbolic meaning, almost exclusively in referential terms.

3. Metalanguages for meanings and units of lexical knowledge

A second major set of issues for lexical databases is how to represent the meanings of lexical items, and how to delimit such items in the first place. Bilingual definitional equivalents are often manifestly inadequate, for the reasons that have always worried translators: mismatehes in grammatical class, inexactness or lack of equivalenee between target and souree language terms, incompatible ranges of meaning, infinite regress or vieious eireles, and so forth. Mueh depends on the available metalanguages.

My colleague Matt Pearson, trying to illustrate the interdependence of different expressive modalities in language, challenges beginning linguistics students as follows: "Can you define 'spiral' without using your hands?" (You might try it yourself before reading on.)

To repeat, everything depends on the available metalanguages. Even a novice mathematician can respond by giving a formula for a 3-dimensional graph, i.e., by defining a series of values for the (x,y,z) axes. Here are some sample formulas.

(7) spiral

(eos(t), sin(t), t) [for a spring-like spiral]
(e*t*cos(t), c*t*sin(t), e*t) (where e is some constant)
[for a cone-like one]

Just to see how these formulas work, on the following page are two graphs of their results, plotted by my statistieian colleague Albyn Jones.

The beauty of the mathematical metalanguage involved is its preeision, parsimony, and presumed universality.[11] The drawback is its potential arcane incomprehensibility.[12] Moreover, though the formulas may describe quite precisely a class of geometric forms, and perhaps even would help define 'spiral,' we might still need recourse to some further (though perhaps equally general) non-mathematical devices to capture the meaning of the word in expressions like "Prices are spiraling out of control," or "We must control the insane spiral of nuclear proliferation."

One difficulty with presuming a language-independent semantic metalanguage (aside from prejudging the semiosis of words and limiting it to referential information – a worry of the previous section) is that it may do violence to the conceptual organization of particular languages. Here is the emie-etic dichotomy of classical anthropologieal linguistics: do we give

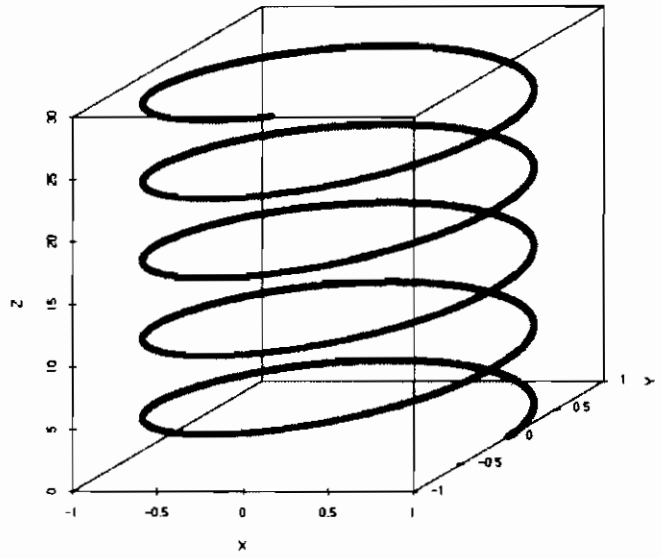


Figure 2. (cos(6t),sin(6t),t) for t in (0, pi)

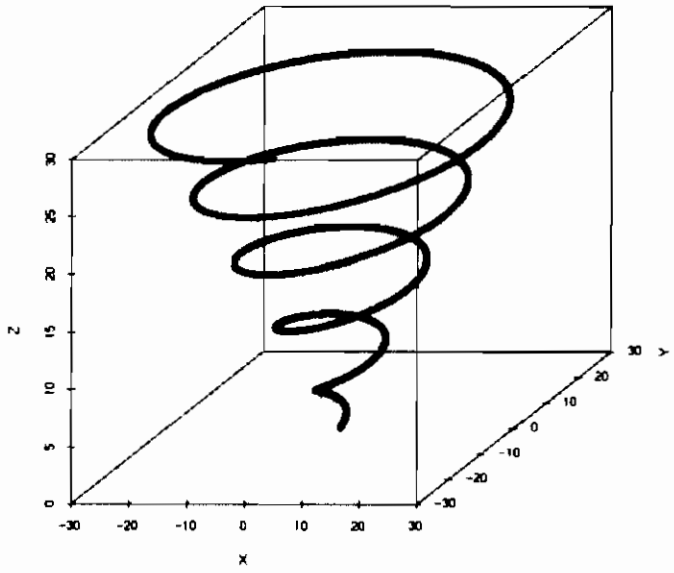


Figure 3. (t*cos(t),t*sin(t),t) for t in the same range

priority to language-specific organization of forms and meanings, or to descriptive categories derived from language-external conceptualizations. An early and instructive demonstration of the dilemma is Conklin's treatment of Hanunoo pronouns.

(8) Hanunoo pronouns (Conklin 1962)

kuh 'I' 1s
muh 'you' 2s
yah 's/he' 3s
tah 'we two' 1du
tam 'we all' 1pl INCL
yuh 'you all' 2pl
dah 'they' 3pl
mih 'we (but not you)' 1plEXCL

If we adopt the standard pronominal metalanguage, *kuh* will be glossed as "first person singular" or *tam* as "first person plural inclusive". The metalanguage thus involves a 'person' component (with possible values 1, 2, or 3), a 'number' component (with possible values, for Hanunoo, of singular, dual, or plural), and an 'inclusivity' component (with possible values inclusive or exclusive, and perhaps an unmarked value) which is defective in that it can by definition apply only to non-singular first person pronouns. Using such meaning components it should be possible to distinguish between 11–13 different pronominal forms (three different persons, with three different numbers, and an inclusive/exclusive distinction on all non-singular first-person forms). The paradigm has only eight pronouns, however. Worse, the primitive terms in the descriptive metalanguage (the number and person categories, plus the terms 'inclusive' and 'exclusive') themselves total eight, suggesting that there is little to recommend this particular metalanguage over just using the raw Hanunoo terms themselves as "primitive" or "undefinable" elements.

Conklin observed that a better analysis is possible, taking as metrics of evaluation efficiency (so that exactly three binary distinctions should be able to distinguish eight [$=2^3$] terms), and "faithfulness" to the native Hanunoo logic. His proposed three binary features are ±Speaker, ±Hearer, and ±Minimal, giving a table like Table 1, whose aesthetic symmetry inspires hope that one is discovering rather than imposing the underlying system.

Table 1. Hanunoo pronouns

	S	H	M
kuh 'I' 1s	+	–	+
muh 'you' 2s	–	+	+
yah 's/he' 3s	–	–	+
tah 'we two' 1du	+	+	+
tam 'we all' 1pl INCL	+	+	–
yuh 'you all' 2pl	–	+	–
dah 'they' 3pl	–	–	–
mih 'we (but not you)' 1plEXCL	+	–	–

Another useful descriptive paradigm widely applied to (and in fact driven by) lexicographic practice is the "frame-semantics" approach associated with Charles Fillmore (see, for example, Fillmore and Atkins 1992). Individual words, on this view, project wider, structured "frames" – configurations of elements and actions, some of which receive explicit grammatical realization and some of which remain implicit in the frame. Families of words then share frames. For example, the Framenet description of the "Commerce-buy" frame – which might be instantiated by such verbs as *buy*, *lease*, or *rent* – is

> These are words describing a basic commercial transaction involving a buyer and a seller exchanging money and goods, taking the perspective of the buyer. The words vary individually in the patterns of frame element realization they allow. For example, the typical pattern for the verb BUY: BUYER buys GOODS from SELLER for MONEY. Abby bought a car from Robin for $5,000.

Clearly, frames themselves can be interrelated. Compare the description for the "Giving" frame, which the "Commerce" frame above "inherits":

> A Donor transfers a Theme from a Donor to a Recipient.[13] This frame includes only actions that are initiated by the Donor (the one that starts out owning the Theme). Sentences (even metaphorical ones) must meet the following entailments: the Donor first has possession of the Theme. Following the transfer the Donor no longer has the Theme and the Recipient does.

In some ways related as a metasemantic device is the approach, most explicitly developed in Levin (1993), that uses various syntactic diagnostics – such as patterns of diathesis – to partition lexical sets into families or classes. Testing various diagnostic syntactic behaviors against their occurrence with specific verbs partitions the verbs into classes which can, according to this logic, be expected to display commonalities of meaning. For example, Levin proposes the following constructions as relevant tests to discover semantic classes among transitive verbs.

(9) Diathesis diagnostics

MIDDLE: The bread cuts easily.
CONATIVE: Carla hit at the door.
BODY-PART POSSESSOR ASCENSION: Terry touched Bill on the shoulder.

Applied to specific verbs (each of which may have a variety of hyponyms, thus forming meaning families), these tests reveal different syntactic classes corresponding to putative meaning families. The meaning families can, in turn, be used to group individual lexical items, and the groupings are thus justified not simply on notional but also on syntactic grounds.

(10) Diathesis diagnostics applied to different verbs (from Levin 1993: 6)

	touch	*hit*	*cut*	*break*
CONATIVE	No	Yes	Yes	No
BODY-PART POSS. ASC.	Yes	Yes	Yes	No
MIDDLE	No	No	Yes	Yes

4. Systematic extraction of lexical databases

After one has documented the basic structures of a grammar, and collected an ample corpus of texts, how does one supplement elicited examples and textually situated tokens of use to achieve a systematic compilation of lexical knowledge? Interlinear glossing of a large corpus can be used mechanically to generate a structured word list, whose analytical perspicaeity is in direct proportion to the eompiler's care and consistency in morphological and semantic tagging during the glossing procedure. Various computational tools aid lexical extraction from text corpora – not only dedicated linguistic database tools like SIL's Shoebox/Toolbox, but also both general and spe-

cialized concordance tools (written, for example, as unix shell scripts, or with programming languages like PERL or ICON[14]).

Other computer techniques can also aid in eliciting lexemes in a language, taking advantage of regular phonological patterns. A well-known example is Terry Kaufman's method for generating an exhaustive list of "potential roots" in Mayan languages, based on the observation that the root canon in Mayan is CVC or some simple variant thereof. Table 2 shows a short Icon program that begins with all the consonants and vowels[15] in the Mayan language Tseltal and produces a complete list of all permutations of the form *CV(:)(j)C*. The program produces 8820 potential roots. (The first of those beginning with *b* are shown in Table 3.) Each of these can be exhaustively (and exhaustingly) tested with native speakers to see which forms actually produce recognizable lexical items – many speakers of Mayan languages and others with similarly straightforward phonotactics have, over the years, been subjected to such a mind-numbing task.

Table 2. Tseltal root salad, in the Icon programming language

```
procedure main()
C := "`bcCjkKlmnpPrstTwxyzZ"
V := "aAeEiIoOuU"
M := "0j"
 every (c1 := !C) do {
    every (v1 := !V) do {
       every (m1 :=!M) do {
                every (c2 := !C) do {
                root := c1||v1||m1||c2
                write(root))
}}}}
end
```

Table 3. The first possible Tseltal roots beginning with *b*

```
ba' bab bach bach' baj bak bak' bal bam ban bap bap' bar bas bat
bat' baw bax bay bats bats' baj' bajb bajch bajch' bajj bajk bajk' bajl
bajm bajn bajp bajp' bajr bajs bajt bajt' bajw bajx bajy bajts bajts'
baa' baab baach ... etc.
```

Mechanically generated wordlists will inevitably reveal areas requiring further lexicographic work – phrasal lexical units, syntagmatically defined paradigms, "functional" vs. "lexical" elements, or particles, for example – and they ordinarily also expose to view especially elaborated lexical domains worthy of deeper exploration. Sueh domains may, on the other hand, emerge not from obvious gaps or hypertrophy in lexical sets revealed in text collections or elicited wordlists, but in clues from the communicative practices of a speech community itself: aesthetic judgments about "beautiful" or "eloquent" – if not "ugly" or "awkward" – speech, for example, especially marked and evaluated kinds of talk, or specialized speech genres or performances, on the one hand; and, on the other, cultural "preoccupations" with associated lexical expression: elaborated vocabularies for professions, activities, or other kinds of interests, or insistence on "getting the right word" or on "proper" and "accurate" expression.

Most methods for lexical elicitation are, for better or for worse, "extensional" and "referential" – that is, they are based on presenting exemplars of things or situations in the world to native speakers and asking for appropriate linguistic expressions which can be used to refer to or to characterize them. Such a method is perhaps inescapable for first-level lexical documentation, but it leaves largely unanswered difficult questions about the intensions of words: what they actually mean, what meaning distinctions they encode, what sorts of meaning relationships they enter into with other words and expressions, rather than simply what states of affairs they can be used truthfully to refer to. Such elieitation techniques are also often helpless to capture such non-referential aspects of meaning as politeness registers, specialized uses and contexts, and the like. Such issues can – and perhaps must – be ignored for the first stages of building lexical databases in language documentation, but they cannot be ignored forever.

Here is a single example from my own fieldwork on Guugu Yimithirr. I quickly learned that the everyday Guugu Yimithirr word *nambal* meant 'stone' but was also extended to mean 'money'. My primary teacher (and social father) in the community, who sometimes had occasion to borrow money from me, often instead used (or whispered) another word to me when he wanted to refer to money: *wambugan*. However, *wambugan* is really a polite equivalent for the ordinary word *nambal* in the respeetful vocabulary, obligatory in speech with avoided affines and referred to in the published literature as "Brother-in-law language" (Haviland 1979). Its denotative range is in fact somewhat broader than that of *nambal* – it includes stones (including specially named grinding stones, quartz, etc., which are not

normally called *nambal*) AND money. Crucially it is an over-polite word, no longer used in modern Hopevale with avoided affines nor, indeed, widely known beyond a few old men, and with them still carrying a euphemistic tone of respect. Both factors combine to make *wambugan* a perfect code word for an embarrassing task like asking one's courtesy son and pupil for a loan.

Ignoring such difficulties for the moment, let us consider techniques for supplementing the lexical information haphazardly collected through mechanical reversal of text corpora. The trick, obviously, is systematic but controlled elicitation, by presenting or simulating aspects of "external" reality so as to stimulate native speakers into using words and expressions to represent as yet unencountered states of affairs. Somewhat artificially I have divided sample methods according to what aspects of "reality" they purport to simulate: 'natural' facts, socio-cultural institutions, and in the final sections pragmatic facts of (inter)action, and ideological constructions on language and society.

4.1. 'Nature'

The tradition in anthropological linguistics, variously labeled "ethnographic semantics" or "ethnoscience," purports to display culturally specific knowledge about the natural world by detailing the semantics of lexical domains related to the corresponding natural phenomena: Hanunoo medicinal plants, Tseltal categories of firewood, ethnobotany or ethnozoology; parts of houses or bodies, taxonomies of disease, local technology, and so on. A classic example of the genre is Berlin's (1968) detailed study of Tseltal numeral classifiers, a detailed compendium of the several hundred classifiers once obligatory in Tenejapa Tseltal numeral expressions. Numeral classifiers specify countable units of different kinds of substance, often on the basis of shape. The notable feature of Berlin's study, for our purposes, is his use of carefully elaborated photographs both as stimuli (i.e. to elicit Tseltal numeral expressions from speakers) and as a vehicle for metasemantic representation: the photos accompany and illustrate his verbal characterization of the Tseltal forms so elicited. (Berlin also used Kaufman's mechanical procedure to generate potential numeral classifier roots, as described earlier.) To give an idea of both the semantic specificity of the Tseltal forms and the nature of the photographic stimuli, here are two sample pictures from Berlin's study. (Note that in Figure 5, illustrating the classifier *hiht'*, the

caption suggests that a Tseltal speaker also noticed an appropriate use for *behč'* in the same stimulus photograph – a nice example of the serendipitous consequences of using such stimuli.)

Figure 4. Tseltal /b'ehč'/: "'individual wraps of slender-flexible objects in sequential spiral around some long non-flexible objects, as a piece of wood.' Included in photo: /lahunb'ehč' laso/ 'laso in the state of ten sequential wraps around long non-flexible object'" (Berlin 1968: 39 Pl. I)

Figure 5. Tseltal /hiht'/: "'individual wraps of slender-flexible objects in sequential lash-loops around two pieces of long non-flexible objects at 90° angles to one another, as in fence making.' Included in photo: /ho?hiht' laso/ 'laso in five lash loops around two pieces of long non-flexible objects' [noted to the left of the photo, the rope in state of /?ošb'ehč'/ 'three continuous wraps']" (Berlin 1968: 39 Pl. II)

Other semantic fields with somewhat more abstract cognitive structures have been recently explored, also with the help of various artificial stimuli.

Following Talmy's typological deconstruction of motion verbs (Talmy 1985), and using a variety of "elicitation kits" involving photographs, drawings, videos, and cartoons,[16] field researchers have explored in detail linguistic systems of spatial adpositions,[17] directionals, motion verbs and other auxiliaries, and what have been called spatial "frames of reference" (Levinson 2003).

For a slightly different sort of example, just as Tzotzil speakers use a highly elaborated set of semantically specific positional roots, it is clear in practice that certain 'families' of verbs grouped by rough notional meaning categories (Dixon 1991) incorporate distinctions, often unfamiliar to speakers of other languages, that require careful lexicographic delimitation. Zgusta (1971: 89 ff.) provides a rich discussion of such families of verbs, what he calls "chains" of "near synonyms," citing as an example multiple Chinese words for 'carry'. There are many monolexemic Tzotzil transitive verbs which might most naturally be translated into English as 'carry', although it is not clear that anything justifies grouping the words together other than this fact about English translations. Thus, for example,

kuch 'to carry (a largish burden) on the back, usually with the aid of a tumpline'

pet 'to carry or hold in the arms, in front of the body (e.g. a baby)'

lik 'to carry by holding a handle from which the burden dangles (e.g. a pail)'

kach' 'to carry by gripping between two surfaces, normally in the jaws (e.g. a dog with a bone)'

jop 'to carry cupped in the hands or some other concave surface (e.g. an apron)'

tom 'to hold or carry in the hand, usually a longish thing gripped in the hand but extending above or beyond it (e.g. a torch, a rifle)'

mich' 'to carry squeezed, usually between the fingers or fist'

etc.

There is, incidentally, as far as I know no more general Tzotzil 'carry' verb that could be used to cover all of these cases.

Another such Tzotzil verb family is that of 'insert' (Haviland 1994) where – as in the case of "carry" verbs – the distinguishing criteria involve the shapes of inserted object and container, the types of contact or containment involved, the tightness of fit, the orientations of container and inserted object, etc. Both to elicit and to illustrate such distinctions I have made

small films of different kind of "inserting" actions, performed with familiar objects, which speakers can view and discuss: what is the best way to describe what they see? are there other ways to describe it? and so on.[18]

It is hard to know in advance what areas of vocabulary will enjoy lexical hypertrophy in an undocumented language. The advantage of the elicitation tools developed by the MPI and elsewhere is that they can be used to invite speakers to exploit their full repertoire of expressive resources by describing standardized stimuli. Children's cartoons such as the *Maus* series from German television[19] are both entertaining and useful for investigating domains of motion, for example. Of course the sense in which speakers of different languages, with different sorts of cultural backgrounds and life experiences, will see these stimuli as "the same" is problematic and, in fact, a central issue to be investigated in linguistic fieldwork.

4.2. Socio-cultural reality

Of obvious interest for language documentation are lexical domains that encapsulate central aspects of society. Linguistic anthropology again provides the classie example: kinship terminologies, once a central part of comparative ethnography, are for speakers of many endangered languages an area of intense personal and conceptual concern (see also Chapter 8). In societies where the central social categories are defined by family relationships, whether genealogically or otherwise construcd, the terminology denoting such eategories is essential to any characterization of social life. The asymmetry in Tzotzil sibling terminology, for example, seems suggestive about family relationships. For a male Ego, Tzotzil distinguishes older and younger brothers (*bankil, itz'in*) from older and younger sisters (*vix, ixlel*). For a female Ego, however, the gender distinction is neutralized between younger brothers and sisters. Thus, a female speaker distinguishes older brother and older sister (*xibnel, vix*) and lumps together younger siblings of both genders (*muk*). Furthermore, note that the distinction between gender of Ego is neutralized precisely in the case of the term for older sister, *vix* for both men and women speakers (see Figure 6). These asymmetries suggest that the relationship between an older sister and her younger siblings of either gender is specially marked terminologically and conceptually. A plausible explanation is the expectation in many Tzotzil speaking communities that an older sister has special mother-like responsibilities for the care of her *muk* or younger siblings, regardless of their gender. This special care

is terminologically matched by a reciprocal terminological projection for younger siblings that their *vix* or older sister is a kind of substitute mother.

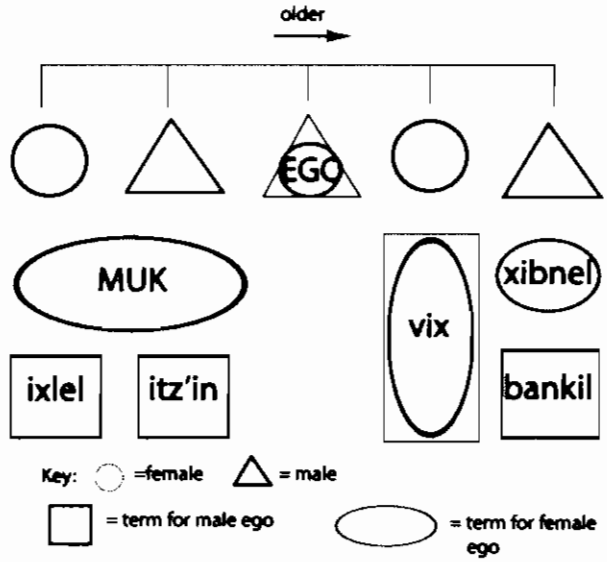


Figure 6. Tzotzil sibling terms

As the classic debates show, however, kinship "algebras" and diagrams conceal a central problem in documenting lexical knowledge, one already mentioned above: the tension between so-called "etic" metalanguages and "emic" categories. In any given language, one can justifiably question whether putatively universal descriptive terms for characterizing a particular kin relationship (in terms, say, of gender, generation, and kin-line, or with allegedly primitive relational terms like F[ather], M[other], H[usband], W[ife], or with algebraic symbols like +, –, ♀, ♂) do justice either to the meaning of a particular natural language term or to a specific relationship between two individuals. Indeed, in societies which display a clear obsession with kinship and kinship terminologies (for example, in the Australian Aboriginal communities where I have worked), a central area of dispute and conceptual wrangling is often exactly how to give the proper lexical label to a relationship, or how to explain what a particular unambiguously named relationship entails. My main Guugu Yimithirr teacher, for example,

would often point out a kinsman walking past and say, "You should call that man X; because his father was your W; but then again, he turned around and married your Y, so what does that make him? your Z?" A genealogical relationship between two individuals does not uniquely determine what the relevant kin term might be, since that, in turn, may respond to considerably more complex factors about what aspects of the relationship are most important. In modern Zinacantán, in some cases a ritual relationship of *compadrazgo* or fictive-mutual-parenthood (between the parents and the godparents of a newly baptized child, for example) may actually take precedence over an immediate genealogical relationship: brothers may become *compadres* and cease to refer to each other with sibling terms.

For purposes of systematic documentation, this domain again illustrates the tension between a "corpus" of examples and systematic eliciting. No single network of actual social/genealogical relationships and the corresponding terminological distinctions can hope to capture the systematicity of the overall terminological-conceptual complex. At the same time, no extensional metalanguage (such as the genealogical primitives of kinship algebra) will be sufficient to guarantee that all socially significant variables emerge from mechanical elicitation. An adequate lexical database must combine both kinds of information.

4.3. Pragmatic reality

Methods for enriching a lexical database to include the use of indexical linguistic units inextricably bound to context are somewhat harder to find in recent literature. All linguistic behavior is, of course, tied to context and linked with action, but some of the most intractable lexical items frequently have *inherent* links to their indexical surrounds – pronouns and other deictics being the most obvious examples, since even their referents (whom they pick out) must be computed by reference to the contexts of their use. Studies of such lexical domains suggest that the only practical approach to the description of such parts of the lexicon is a kind of exhaustive observational fieldwork. Thus, Hanks (1990) gives detailed analysis of the system of demonstratives in Yucatec Maya based on extensive fieldwork in which he recorded, in detail, situated occurrences of spontaneous deictic usage, inducing from the corpus and from the linguistic forms the theoretical components of an adequate account of deictic practice.

Another exemplary domain is that of exclamations and interjections. Kockelman's extended treatment of interjections in Q'eq'chi (Kockelman

2003) involved a field methodology much like that of Hanks. He systematically recorded the circumstances when utterances eategorized as interjections occurred in a Q'eq'chi speaking community in Guatemala. On the basis of such a corpus, he elaborated a theory of interjections which goes well beyond the received model of their "expressive" nature (part of an ancient tradition in Western linguistic thought, dating back to the Latin grammarians), to consider the multiple and bi-directional indexical properties of these expressions: exhibiting emotional and affeetive stances, explicitly inviting reciprocal exhibits from interlocutors, drawing interlocutors' attention to circumstances, requesting actions, and so on. Such studies suggest that there are few shortcuts to an adequate account of what such pragmatically charged linguistic elements mean, and that extensive ethnographic fieldwork is thus an essential part of field lexicography.

The same can be said of more prosaic vocabulary, from ordinary body part terms to specially marked polite and impolite registers, such as joking and cursing speech. I have already mentioned the residual lexical eomplexities produeed by changed use of Guugu Yimithirr respectful or "brother-in-law" vocabulary, and sueh complexities are only multiplied when several more or less well regimented speech registers are in active use in a speech community. Classic anthropological descriptions of such phenomena attest to the subtlety and nuance communicated by strategic choice between alternate lexical forms in societies from Aboriginal Australia and Samoa to Bali (Duranti 1992; Errington 1984; Geertz 1960), or between address terms and personal pronouns from Europe to Japan (Brown and Gilman 1960). Laughlin (1975) proposes a series of labels to distinguish in Zinacantee Tzotzil such things as "ritual speech, joking speech, male and female speech, baby talk, polite speech, scolding, denunciatory speech, archaic [words]," etc. Whether or not a field lexicographer can give a complete account of such facts for an entire lexical database, it is important to be aware of the sorts of metalinguistic speech categories that might be relevant in a given speech community.

For self evident reasons, systematic investigation of such genres – for example, tabooed speech – may be hard for inexperienced fieldworkers. Similarly difficult are whole systems of linguistic tropes which sometimes dominate parts of a language's expressive resources. Again, the only remedy seems to be wide ranging and systematic ethnographic attention. Here are two examples from my own fieldwork. As I learned Guugu Yimithirr, I noticed that many expressions dealing with human propensities and "inner states" were transparently metaphors, based on a small set of words which

seemed simply to name parts of the body. Whether or not, as anthropologists have sometimes suggested, these expressions represent an implieit theory of the anatomieal distribution of emotions and mental faeulties (as we might argue, for example, with English expressions like 'hard-headed' or 'hard-hearted'), or instead are simply opaque culturally eonventionalized idioms (as we might argue for 'green thumb' or 'lily-livered'[20]) it was clear that Guugu Yimithirr had a semi-produetive system for generating diverse expressions based on "body-part" tropes. (11) gives an example based on the Guugu Yimithirr word *miil* 'eye'. The only way I could document the system was to keep my ears open (as it were) for relevant expressions in conversation, and to try systematically to force new combinations of body-part words with adjectives and verbs, usually yielding only guffaws instead of new lexemes.

(11) Guugu Yimithirr expressions based on *miil* 'eye'

mülgu	= (lit., eye + EMPHATIC suffix) awake
miil warnggu	= (lit., 'eye sleep') sleepy
miil nhin-gal	= (lit., eye sit) wateh out, keep an eye out
miil biyal	= (lit. eye sinew) staring all the time
miil ngamba	= (lit. eye careless) unobservant, shutting one's eyes to something
miil waarril	= (lit., eye fly) feel faint, go crazy, faint, get drunk[21]
miil bagal	= (lit., eye poke) deceive, trick, become jealous
miil bathibay	= (lit., eye bone) sharp-eyed, always staring
miil biinii	= (lit., eye die) go blind
miil gulnggul	= (lit., eye heavy) sleepy
miilgu nhin-gal	= (lit., eye-EMPHATIC sit) stay awake ....

Consider, too, the language of Tzotzil ritual (Gossen 1974, 1985; Haviland 1987, 1996, 2000). In contexts from prayer and song to formal denuneiation, Tzotzil speakers abandon ordinary lexicon and grammar in favor of a highly structured speech style that involves parallel lines which differ in only a single word or phrase. These parallel lines are interpreted in terms of a standard "stereoscopic" image (Fox 1977) invoked by the paired expressions. Thus, to refer to the body one can use different doublets, depending on the context. One is highly literal, using *pat, xokon* 'baek, side' as a metonym for the whole. Another is considerably more opaque, and suggests an image of humility, as in the following extract from a euring prayer,

where the doublet *lumal, ach'elal* 'earth, mud' (both in possessed form) refers to the patient's body or self.[22]

(12) From a Zinacantec curing prayer

ja' me ta jmala lalumale
I am waiting for your earth.

ta jmala lavach'elale
I am waiting for your mud.

A further example is the doublet in Zinacantec ritual speech to refer to liquor: *xi`obil, sk'exobil,* literally 'cause for fear, cause for shame'. Such expressions share properties with euphemism, always a problematic phenomenon for lexicography that requires careful ethnographic fieldwork. Systematic elicitation reveals little about the overall system of imagery in ritual language, although it is an essential part of the language's expressive power. Laughlin's (1975) dictionary of modern Zinacantec Tzotzil annotates and illustrates words that participate in parallel constructions under the rubric 'ritual speech'. In my own work, I have relied on exhaustive recording and transcription of prayer and other genres that employ parallelism to expand on the list of doublets.

5. Conclusion

When does documentation of the lexicon end? While the lexicon is a repository for the exceptional and the chaotic in language, it is also a site of considerable regularity and productivity. Nonetheless, field lexicographers like Laughlin express doubts about how well structured or widely-shared lexical knowledge is across a speech community, basing his skepticism on elicitation with both Zinacantec peasants and Washington D.C. university students. Notoriously difficult even for well-studied languages is distinguishing between 'literal' and 'figurative' or tropic uses of words: older Tzotzil speakers describe airplanes as *xulem k'ok'*, literally (as we say) 'buzzard fire' or telephones as *ch'ojon tak'in* 'wire of metal' – enduring the giggles of younger speakers (who simply use a Spanish loan instead). Even more difficult is distinguishing obscure polysemy from simple (but formally unpalatable) homonymy. Laughlin's Tzotzil dictionary posits two homonymous roots, *jav*(2) – a positional root meaning 'belly (or face) up' –

and *jav*(1), a transitive verb root meaning 'to chop in half' because the two meanings seem divergent enough to warrant separate entries. However, Zinacantec folk etymology conjures a succinct image that connects the senses: when you split, say, a log in two (using a verb based on *jav(1)*), the two halves fall "belly up" (*jav(2)*). This is thus a case of covert polysemy,[23] or perhaps of underlying monosemy of a single root with different grammatical costumes. Such phenomena may remain intractable throughout a lexical documentation project.

Similarly, how much ought the lexicographer to include of what might be labeled "erroneous usage" – malapropisms, puns, or nonce creations? Zgusta (1971: 56–57) distinguishes "systemic" from "occasional" uses of words. An author may use 'bondage' occasionally to mean 'marriage,' without thereby changing the systemic meaning of either term. Zinacantec men, during several weeks of ribald gossip sessions in 1970, coined what was at the time a highly creative Tzotzil sexual euphemism using a loan *inyeksyon* from Spanish *inyección*, at a time when hypodermic injections were still a relatively novel foreign introduction. Some of these men still jokingly use the term almost 40 years later. The word is not in Laughlin's Tzotzil dictionary – but perhaps it should be.

Finally, questions already mentioned about aims and audience – for whom is a lexical database produced? to what ends will it be put? – complicate decisions about what words must be documented and how. The problems are especially vexed when a lexical database may serve as the basis for standardization or stabilization, especially in the form of a published dictionary.[24] When people can use a dictionary to look up a word, to see how it is spelled, and to read a definition, the speech community's authority over "proper" usage is irrevocably altered. How much belongs in the lexical database of a language documentation project is thus never simply a matter of "completeness" or "coverage" but also involves ideological decisions that may have far-reaching effects on the future of a language.

Building a lexical database is an expected part of any documentation project, perhaps the final most demanding analytical task of all. It can be aided by mechanical techniques applied to textual corpora and by familiarity with the great lexicographic traditions, which have already grappled with most of the problems a fieldworker is likely to encounter: lexical units, the nature of meaning, the vagaries of usage, and, finally, ideologies of language and social life. The end product is essential, but producing it relies on both drudgery and ethnographic inspiration, on systematic elicitation and serendipitous discovery. One inevitably (re)discovers that enough is

never enough, and that calling a halt by declaring the database closed is simply an arbitrary rest stop on a very long journey.

Acknowledgements

This chapter, loosely based on the lecture presented at the DoBeS summer school in Frankfurt, September 2004, owes a considerable debt to experiences in lexicography shared with my Tzotzil and Guugu Yimithirr teachers, to comments by Nikolaus Himmelmann and Jost Gippert, and to the hospitality of Elena, Renato, and Lisetta Collavin during its final drafting.

Notes

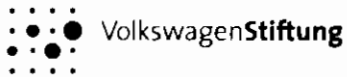
1. Especially with reference to dictionaries for literate European traditions, both Landau (1984) and Svensén (1993) are useful surveys. See also the multiple volume handbook edited by Hausmann et al. (1990–1991).
2. Although languages like Nahuatl enjoy their own centuries' old dictionary traditions (Canger 2002; Amith 2002).
3. A Tzotzil-Spanish version is currently (2005) in press, to be published by the *Centro de Investigaciones y Estudios Superiores en Antropología Social* in Mexico. As Tzotzil speakers inereasingly cross the border into the United States, the number of Tzotzil-English bilinguals will, of course, only grow.
4. See Haviland (1974). Nick Evans' (2002) remarks on misunderstandings of Aboriginal expressions, even in English, in hearings before the Australian Land Tribunal shows how such misunderstandings can have serious legal consequences.
5. See the notion of "rules of use" in Silverstein (1976).
6. Jost Gippert reports that "Georgian native speakers confirm that *mqavs* is applied to anything mobile, such as cars, bicycles, airplanes, or the like."
7. In Berlin's works the older spelling "Tzeltal" is used.
8. The symbol *A* denotes a hypothetical vowel that alternates between *a* and *o* in derived stems.
9. See http://framenet.icsi.berkeley.edu/index.php and Seetion 3 below.
10. English is interestingly different in its elaborations, as can be seen by the entries in the Framenet "being_attached" frame which include: *affixed, anchored, attached, bolted, bound, chained, fastened, fused, glued, handcuffed, lashed, manacled, moored, nailed, pasted, pinned, plastered, riveted, sewn, shackled,*

stapled, stuck, taped, tethered, tied, welded. In English the central variable seems to be the kind of material creating the attachment.

11. There are proposals from linguisties itself about a "Natural Semantic Metalanguage" through which definitions of complex notions can bc framed in terms of simpler, allegedly universal (hence 'natural') semantic primes. See http://www.une.edu.au/arts/LCL/disciplines/linguistics/nsmpage.htm, where one can find a bibliography of thc many publications of Anna Wierzbicka.
12. Faced with Pearson's challenge, Recd College senior Chris Haulk "promptly came up with, 'oh, you mean – wrap a string around a cylinder; versus, wrap a string around a cone'" (Albyn Jones, personal communication, March 1, 2005) – proving that mathematicians can bc lexicographers, too.
13. Note that "Donor" here is a single entity, defined in Framenet as "The person that begins in possession of thc Theme and causes it to be in the possession of the Recipient."
14. Visit http://www.cs.arizona.edu/icon/.
15. The program symbolizes glottalized or ejective consonants and long vowels as capital lctters, and a 0 is used to signal the absence of medial *j*.
16. See the descriptions of various stimulus kits developed by the Language and Cognition Group at the Max Planck Institute for Psycholinguistics at http://www.mpi.nl/world/data/fieldmanuals/.
17. See Levinson et al. (2003) for an unashamedly extensional, comparative approach.
18. A short video used to elicit deseriptions for Tzotzil 'inserting' actions is available on the book's website.
19. Samples of thc sort of cartoon I have found useful for such tasks are available at http://www.wdrmaus.de/lachgeschichten/mausspots in streaming video format.
20. The expression is not confined to English; both Italian *pollice verde* (according to Elena Collavin) and German *grüner Daumen* (according to Nikolaus Himmelmann) have exactly the same metaphorical and literal meanings as 'green thumb', i.e., somcone good at gardening. Similarly, Italian *senza fegato* 'without a liver' suggests a meaning similar to 'lily-livercd.'
21. I ignore basic syntaetic issucs here: for example, in the expression *miil waarril* the word *miil* 'eyc' is the syntactic subject of *waarril* 'fly.' In *miil bagal* 'eye' is syntactic object of *bagal* 'poke.'
22. In the Tzotzil of nearby Larraínzar, the equivalent ritual doublet is at once humble and literal: *ach'elal, takopal* 'mud, body.'
23. See Zgusta's discussion of polysemy (1971: esp. 77 ff.); also Evans and Wilkins (2000, 2001), Evans (1992).
24. See Janc Hill's discussion of the Hopi dictionary project in Chapter 5.

Mouton de Gruyter (formerly Mouton, The Hague)
is a Division of Walter de Gruyter GmbH & Co. KG, Berlin.

Published with support of VolkswagenStiftung, Hannover, FRG.



The hardcover was published in 2006 as volume 178
of the series *Trends in Linguistics Studies and Monographs.*

♾ Printed on acid-free paper which falls within the guidelines
of the ANSI to ensure permanence and durability.

The Library of Congress has cataloged the hardcover edition as follows:

Essentials of language documentation / edited by Jost Gippert, Nikolaus P. Himmelmann, Ulrike Mosel.
p. cm. - (Trends in linguistics. Studies and monographs ; 178)
Includes bibliographical references and index.
ISBN-13: 978-3-11-018864-6 (cloth : alk. paper)
ISBN-10: 3-11-018864-3 (cloth : alk. paper)
1. Linguistics - Documentation. 2. Language and languages - Documentation. I. Gippert, Jost. II. Himmelmann, Nikolaus P., 1959- III. Mosel, Ulrike IV. Series
P128.D63E85 2006
025.06'41 - dc22
2006001315

ISBN-13: 978-3-11-018406-8
ISBN-10: 3-11-018406-0

Bibliographic information published by Die Deutsche Bibliothek

Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the Internet at <http://dnb.ddb.de>.

Editors' preface

Language documentation is concerned with the methods, tools, and theoretical underpinnings for compiling a representative and lasting multipurpose record of a natural language or one of its varieties. It is a rapidly emerging new field in linguistics and related disciplines working with little-known speech communities. While in terms of its most recent history, language documentation has co-evolved with the increasing concern for language endangerment, it is not only of interest for work on endangered languages but for all areas of linguistics and neighboring disciplines concerned with setting new standards regarding the empirical foundations of their research. Among other things, this means that the quality of primary data is carefully and constantly monitored and documented, that the interfaces between primary data and various types of analysis are made explicit and critically reviewed, and that provisions are taken to ensure the long-term preservation of primary data so that it can be used in new theoretical ventures as well as in (re-)evaluating and testing well-established theories.

This volume presents in-depth introductions into major aspects of language documentation, including a definition of what it means to "document a language," overviews on fieldwork ethics and practicalities and data processing, discussions on how to provide a basic annotation of digitally-stored multimedia corpora of primary data, as well as long-term perspectives on the preservation and use of such corpora. It combines theoretical and practical considerations and makes specific suggestions for the most common problems encountered in language documentation.

The volume should prove to be most useful to students and researchers concerned with documenting little-known languages and language varieties. In addition to linguists and anthropologists, this includes students and researchers in various regional studies and philologies such as African Studies, Indology, Turkology, Semitic Studies, or South American Studies. The book presupposes familiarity with the basic concepts and terminology of descriptive linguistics (for example, basic units such as *phoneme* or *lexeme*), but most chapters will also be accessible and useful to non-specialists, including educators, language planners, politicians, and government officials concerned with linguistic minorities.

Acknowledgements

We gratefully acknowledge the very generous support of the Volkswagen-Stiftung (http://www.volkswagenstiftung.de) which has been instrumental in producing this book. The foundation not only funded the summer school for which most chapters were drafted, but also provided the means to distribute a substantial number of copies of this book free of charge outside of Western Europe, North America, and Japan. By granting a research fellowship for Himmelmann in 2004–2005, it has allowed him to focus his research on the issues dealt with in Chapters 7 and 10 and to engage in the editing of the book in a way which otherwise would not have been possible. Through its *DoBeS Programm* (Documentation of Endangered Languages program), which started in the year 2000, it has made a major contribution to the development of documentary linguistics as an innovative field of study and practice within the humanities.

Our sincerest thanks are due to the contributors of the volume who spent a lot of time on conceiving their chapters and have always been ready to cooperate with us in the difficult task of preparing a consistent book.

We also gratefully acknowledge much practical help we have received in putting the volume together. Marcia Schwartz checked English and style conventions; Judith Köhne compiled the combined list of bibliographical references at the end. At Mouton, Ursula Kleinhenz did a great job of seeing the book through to press. Many thanks to all of you.

Contents

Editor's preface v

Acknowledgements viii

Chapter 1 Language documentation: What is it and what is it good for? 1
Nikolaus P. Himmelmann

Chapter 2 Ethics and practicalities of cooperative fieldwork and analysis 31
Arienne M. Dwyer

Chapter 3 Fieldwork and community language work 67
Ulrike Mosel

Chapter 4 Data and language documentation 87
Peter K. Austin

Chapter 5 The ethnography of language and language documentation 113
Jane H. Hill

Chapter 6 Documenting lexical knowledge 129
John B. Haviland

Chapter 7 Prosody in language documentation 163
Nikolaus P. Himmelmann

Chapter 8 Ethnography in language documentation 183
Bruna Franchetto

Chapter 9 Linguistic annotation 213
Eva Schultze-Berndt

Chapter 10 The challenges of segmenting spoken language 253
Nikolaus P. Himmelmann