

Computer Processing of Texts for Lexical Analysis

Leonard Newell

*Summer Institute of Linguistics
Philippines Branch*

1. Introduction

The use of computers for processing texts for lexicographic descriptions is not new. However, until recently, computers and software programs with specifications necessary to process texts large enough to be useful were accessible only to large projects which could afford mainframe computers. With the rapid development of computer technology, however, desktop and even laptop computers are now becoming available with the capacity for processing multimillion-word texts.

Working from texts with a computer has many advantages, providing the database is sufficiently large to generate a true sample of the lexicon. Dictionary projects, other than those which aim to compile a more or less complete lexical description (i.e., an unabridged dictionary), must be selective in the vocabulary chosen for inclusion. A good dictionary project will adopt carefully prepared guidelines for selecting the lexicon rather than making random choices. One of these involves the frequency of occurrence of lexical forms¹. A frequency list, based on a sufficiently large and representative text base, arranged according to frequency of occurrence, allows the compiler to choose those items that most commonly occur within the language.

Concordances for lexical forms within sentence contexts, generated by computer from a multi-word corpus, provide valuable data for lexicographic analysis. The advantages of using natural sentences thus prepared, rather than texts fabricated specifically for lexical analysis, are obvious. A concordance provides the grammatical context of any given lexical form, aiding in establishing such features as parts of speech, grammatical restrictions, etc. It is also a major source for identifying semantic features such as senses of given lexical forms, range of meaning, homonymy, idioms, set phrases and collocation, all of which will be included in a good dictionary.

Lexicographic projects are presently in progress on several Philippine languages of small ethnic communities². A primary aim of most of these small projects is to prepare learners' dictionaries to be used for cross-cultural communication between these ethnic communities and speakers of Filipino and/or English. This presentation describes some features of this lexicographic work and demonstrates how computers are being used to assist in accomplishing it. Example language materials are from the Romblomanon Visayan project³.

2. Corpus development

A dictionary is primarily a description of the meaning of lexemes⁴ within texts. The analyst typically follows one of two procedures in analysing texts - he either bases the analysis mainly

¹ In agglutinative languages, a lexical form (i.e., the form of a lexeme) is commonly identical to that of a bound morpheme, see Sec. 3.

² Researchers of the Summer Institute of Linguistics are presently engaged in lexicographic work on some forty minority languages of the Philippines.

³ My wife, Johanna, and I are presently conducting lexicographic research on the Romblomanon language.

⁴ The term 'lexeme', as used in this paper, is a semantic unit composed of a unique bundle of distinctive semantic features necessary and sufficient to allow it to denote when it occurs in various lexical and situational contexts.

on artificial texts or on natural texts drawn from the language in use. The Romblomanon project follows the second procedure.

2.1 Why develop a Corpus?

There is a significant contrast between a dictionary based on a corpus of the living language such as the *Collins Cobuild English Language Dictionary*, and one that is not. Entry descriptions of lexical meaning, based on a good corpus, strike a native speaker as obviously correct, but not necessarily how a native speaker might describe it through introspection. This is because such descriptions are based on how the language actually functions rather than on how the analyst or a language informant might think it does. There is an inevitable degree of artificiality in the semantic description of lexemes occurring in made-up sentences (Sinclair 1984). In addition, it is virtually impossible for a native speaker, through introspection, to consistently recall the major functions of most lexical forms.

Dictionaries of lesser-known languages often include a helpful grammar sketch to accompany and support the lexical description. There is a major compatibility advantage in basing the grammatical description on the same text base as is used for the lexical description.

2.2 Spoken or Written Texts

A large project of a major language usually has access to a sizeable amount of written material either in electronic form or convertible to that form by an optical character reader (scanner). A typical large project might develop a corpus of up to 90% written texts and 10% oral texts.

A small project, however, and especially one which studies a language of a preliterate society or one in the initial stages of developing a literature, does not have written materials available to use in developing a corpus. This, then, requires developing a corpus primarily by recording and keyboarding oral texts. Although there are problems involved in this, they can be at least partially overcome as discussed in Sections 2.6 and 3.

2.3 A Corpus Restricted to a Single Dialect

For the Romblomanon project, it was decided that the corpus, and eventually the dictionary, should represent a single dialect. A learners' dictionary should, presumably, aid a learner in acquiring a facility in a single dialect rather than a "blend" of multiple dialects which no one speaks naturally. The Romblomanon dialect chosen for lexicographic research and description is spoken by people living in the geographic center of both the government and dominant religion (the Roman Catholic religion), in and around the town of Romblon. This is a prestige dialect, presumably acceptable to most speakers of other closely-related dialects and sub-dialects.

2.4 People Chosen to Contribute Texts

Contributors in the age range of about 21 to 65 are chosen. Very few are either younger or older than this. Care is taken to select those who have lived in the dialect area most of their lives. No immigrants from other languages, and no recent immigrants from other Romblomanon dialects are chosen. An attempt is made to select contributors from varied social, economic and educational sub-groups.

Included with each text is a record of information indicated above about each contributor.

2.5 Size of the Corpus

The larger the corpus, the more chance there is to assure a representative sample of the language, and thus more reliable lexicographic statements about it. The Cobuild project, for example, based its dictionary on a corpus of over 20 million words. This, however, is a goal beyond the reach of a modest project. Resource and time constraints on small projects dictate a severe limitation on size.

Experience to date indicates that it requires about 4000 man-hours to collect and keyboard about a million words of text. This is equivalent to one person working 8 hours a day, 5 days a week for about two years. The project has employed six people to collect text, each spending an average of about 4 months over a period of two years. The result to date is somewhat more than one million words of text.

Based on the experience of this project, a corpus yielding about three million morphemes is considered both attainable and adequate to meet the needs of a modest lexicographic project on a language such as Romblomanon. (See also Sec. 3.2.)

2.6 The Nature of the Corpus

Since it is practical for a small project to collect a corpus of only about two to three million words (Sec.2.5), it was necessary to carefully set priorities and guidelines for text selection.

A major consideration is to gather texts from a wide range of representative speakers, and to cover all major genres and cultural areas represented by the language of the speakers. This will assure a representation of the lexicon in as broad a cultural range as possible.

Also, by careful text selection, the aim is to cover major lexical areas with a smaller corpus than would be necessary if corpus development depended primarily on whatever printed or electronically prepared materials were available, as is the case with some large projects.

2.6.1 A Wide Range of Common Subjects

In general, texts represent a broad spectrum of usage of subjects about which people most commonly communicate. This is considered especially important for a learners' dictionary. Cultural areas in which a learner might be involved or about which a learner might want to communicate are selected. Texts are collected about specialized subjects such as the marble industry of Romblon; however, an overbalance in specialized areas is avoided. In addition, most texts on specialized subjects are contributed by lay people, rather than by specialists, to avoid highly technical vocabulary.

2.6.2 Cultural Considerations

Cultural classifications based on George Murdock's *Outline of Cultural Materials* (1982) are used as an etic starting point. These are being revised to meet requirements of Romblomanon culture. Numbering of major categories and subcategories follow Murdock. (See also Sec. 2.7.) Of Murdock's 79 categories, the Romblomanon project presently has 58, with 176 sub-categories out of a potential 637 (see Appendix). The text database will continue to grow for the length of the project.

2.6.3 Complete Texts

Collected texts are not edited except to eliminate such features as stumbles and false starts where these are obvious. Excerpts from larger texts and text fragments are avoided. Since some

lexical forms tend to occur most frequently or exclusively at paragraph boundaries or at the beginning or end of discourse units, it is considered necessary to include whole discourse units in order to get a proper representation of such features.

2.6.4 Natural Texts

Many texts are recordings of naturally occurring events such as weddings, litigation sessions, political meetings and the like. Others are discussions of cultural features suggested by the one making the recording. However, guidance of the communication event is avoided as much as possible. Texts do not include poetry, riddles, dramatized events, songs, chants or proverbs.

2.7 Text Input

Each text is assigned a file name consisting of two letters representing the language, followed by five numbers.

RO22609.it8	255K	--/--	F1 Help F2 RO Lex F3 RO Affix F4 RO GRAM F5 Box D F6 RO Parse F7 Box F F8 Box G F9 Text F0 Quit
<p>'Ang Nagapahubas Sa Dāgat. Kung 'ang tā'ub hay kadaku' 'ang manga tāwu nga 'igway manga pūkut nga 'iyugpahūbas kung 'ang tā'ub hay lantung na 'imaw 'ang pagtaktak nang pūkut pāra 'ang 'isda' hay masarhan 'indi' na sinda kapalawud. 'ang nagapahubas hay kung 'ang kadaku' kag timprānu pa nagahunas kadāmu' nga manga binanāta nang pūkut 'ang ginasugdungsugdung pāra malangkuban 'ang masigkapiliw kay kung matag'ud lang 'ang pūkut nga 'ipahubas waya' da 'adtu puyus kay malāpus da gihāpun sa punta nga wayay pūkut. Ginataktak 'ang pūkut hasta sa piliw kay 'ang 'isda' kung makakīta' ning lapūsan didtu na sinda tanan malāpus gāni' dāpat gid nga langkūyan 'ang pagtaktak. Ginabutangan ning 'ūsuk 'ang pūkut pāra 'indi' maglagmak sa hunāsan kag kung may nasākup nga bayānak 'indi' kalumpat kay matā'as 'ang pūkut. 'ang bayānak lang da nga 'isda' 'ang ma'āyu nga lumuksu sapūkut. Kung nakakīta' sinday pūkut nga sagang mintras nagapahigkat 'ang dāgat 'indi' pa pwīdi nga turagun 'ang pahubas. Mahuyat ka gid nga ma'isut na 'ang dāgat. Kung tagabatī'is na lang 'ang tūbi' ginaturag na 'ina' nang manga tāwu pāra manglabu' ning 'isa</p>			

F9-File Editor

Table 1 Sample Romblomanon Text

Table 1 illustrates a text portion. The file name in the top left corner gives the following information: RO signals Romblomanon as the language; 22 indicates a main classification “food quest”; the number 6, a subclassification “fishing” and 09, the ninth text on this subject (see Appendix).

Keyboarding follows a simple and unvaried format. The text consists of strings of words using a single typeface, separated by spaces. Each sentence begins with a capital and ends with a period.

For ease in computational operations, the orthographic transcription of texts from recorded voice and written materials is strictly phonemic. This eliminates variation of forms due

to orthographic conventions which would cause problems in such operations as spell checking, frequency counts and concordance operations. All phonemes, including length and glottal stop are symbolized. These latter are especially necessary since in Romblomanon, as in many languages, these phonemes contrast pairs of forms that would otherwise be identical and which also figure in morphophonemic variation of forms.

Each sentence within a text is assigned a unique number. This allows easy reference to any lexical form in the corpus within a discourse context. Sentence numbers are inserted automatically by computer as illustrated in Table 2.

RO22609.its	F1 Help F2 RO Lex F3 RO Affix F4 RO GRAM F5 Box D F6 RO Parse F7 Box F F8 Box G F9 Text F0 Quit
<p>\tn RO22609 001 \ex 'Ang Nagapahubas Sa Dāgat.</p> <p>\tn RO22609 002 \ex Kung 'ang tā'ub hay kadaku' 'ang manga tāwu nga 'igway manga pūkut nga 'iyugpahūbas kung 'ang tā'ub hay lantung na 'imaw 'ang pagtaktak nang pūkut pāra 'ang 'isda' hay masarhan 'indi' na sinda kapalawud.</p> <p>\tn RO22609 003 \ex 'ang nagapahubas hay kung 'ang kadaku' kag timprānu pa nagahunas kadāmu' nga manga binanāta nang pūkut 'ang ginasugdungsugdung pāra malangkuban 'ang masigkapiliw kay kung matag'ud lang 'ang pūkut nga 'ipahubas waya' da 'adtu puyus kay malāpus da gihāpun sa punta nga wayay pūkut.</p> <p>\tn RO22609 004 \ex Ginataktak 'ang pūkut hasta sa piliw kay 'ang 'isda' kung makakīta' ning lapūsan didtu na sinda tanan malāpus gāni' dāpat gid nga langkūyan 'ang pagtaktak.</p>	
	F9-File Editor

Table 2 Romblomanon Text with Sentence Numbers

Sentence numbers are the last three numbers on the text number (\tn) line.

3. Computational Operations Involve Morphemes

Since Philippine languages are agglutinative, it is more practical to work with roots, affixes and clitics than with words. For non-agglutinative languages such as English, a phonological word is most commonly identical in form to that of a lexeme. For this reason, it is practical for such languages to do computational operations on words without the necessity of segmenting texts into morphemes. For an agglutinative language, however, the unit commonly identical in form to that of a lexeme is a morpheme consisting of a bound root, affix or clitic, all smaller than a word⁵. For this reason, for agglutinative languages, it is necessary to do computational operations on morphemes. This prepares the material for lexical analysis and description based on a linguistic unit most closely related in form to that of a lexeme.

⁵ The form of some lexemes are multi-morphemic. These include lexemes occurring as idioms, set phrases or derived by derivational affixation or compounding. It is relatively easy to identify these units when concordancing is done for morphemes within sentence contexts.

3.1 Segmentation of Text into Morphemes

Words constructed of two or more morphemes are segmented into roots, affixes and clitics. Words which are single morphemes remain as is. This results in a string of morphemes. At this point, no consideration is given as to whether affixes are derivational or inflectional. The determination of such features is part of the analytical process once concordances for morphemes are provided.

For morphemes with one or more variant, a standard form is chosen and indicated on the morpheme line of the text. Morphophonemic variation is not indicated on this line so that frequency counts and concordances will be for morphemes rather than for morpheme variants.

It would be entirely impractical to undertake manual segmentation of several million words of text into constituent morphemes. Segmentation is accomplished by computer assistance. This significantly speeds up the segmentation process.

RO22609.tmp	{}	ES
\id 22609		
\tn ro22609 001		
\ex 'Ang Nagapahubas	Sa Dāgat.	
\mo 'Ang hubas=pa-=ga-=na-	Sa Dāgat	
\tn ro22609 002		
\ex Kung 'ang tā'ub hay kadaku'	'ang manga tāwu nga 'igway manga	
\mo Kung 'ang tā'ub hay daku'=ka-	'ang manga tāwu nga 'igwa=-y manga	
\ex pūkut 'iyugpahūbas kung 'ang tā'ub hay lantung na 'imaw 'ang		
\mo pūkut		
\ex pagtaktak nang pūkut pāra 'ang 'isda' hay masarhan 'indi' na		
\mo		
\ex sinda kapalawud.		
\mo		

F9-text editor

Table 3 Segmenting Text into Morphemes

Table 3 shows the cursor waiting for the word *'iyugpahūbas* to be segmented, a word not previously encountered. The operator will type in *hūbas=pa='iyug-*. This will be stored in the memory of the computer and any further occurrences of this word will be automatically segmented by the computer.

3.2 Morpheme Lists Indicating Frequency of Occurrence

With the corpus segmented into strings of morphemes, the text is prepared for performing various computer operations. One is to prepare lists of morphemes that occur in the corpus.

Two kinds of lists are prepared. One list arranges morphemes alphabetically and includes the frequency of occurrence of each morpheme, preceding the form. Table 4 illustrates a small portion of a morpheme list for Romblomanon.

1	gunamila	1	gusla'	97	gustu
6	guwa'	1	guyuguyu	1	gūlay
1	gūma	8	gūna	1	gūpuk
8	gūtum	74	gūyang	2	gūyud
1	gūyus	14	gwāpa	4	gwāpu
1	gyīra	7	ha	11	-ha
1	habhab	19	habun	4	habut
33	hadluk	1	hagashas	2	hagdan
5	hagkus	2	hagkut	5	hakay
1	hakid	1	haknit	2	hakuwat
7	hala	1	halata'	6	hali
1	halimbāwa	44	halimbāwa'	44	halin
4	haligi	1	hamadhamad	407	hambay
5	hambāwan	4	hambug	2	hamham
6	hampig	1	hamu	1	hamug
2	hana'	2	hanāgub	6	handa'
8	handum	1	hangad	1	hanggang
1	hanggid	1	hanghang	15	hangit
1	hapak	2	haphap	1	hapin
2	haplit	1	hapslip	4	hapūhap
2	hapyā'	1	hapyus	17	harāna
2	haru'	52	hasta	9	hatud
1	haw'as	4	hawhaw	2425	hay
1	hayag	3	hayakhak	1	hayakhay

ITTEXT.WDL

Table 4 Alphabetical Listing of Romblomanon Morphemes

This list is used for reference purposes to check on the frequency of occurrence of any given morpheme. Thus, for example, if a given lexeme is considered for inclusion in the dictionary, its form can easily be found and its frequency of occurrence readily checked.

Another list begins with the morpheme which occurs most frequently and ends with those occurring only once. Two or more morphemes with the same frequency number are further sorted alphabetically. Table 5 illustrates morphemes at the top of the list.

Unique: 10,679		Total in text: 1,302,665						
(Tally is based on morphemes occurring 1 or more times in the text.)								
↘	44637	'ang	↘	9709	'ini	↘	5401	gāni'
↘	43461	nga	↘	9476	'aku	↘	5376	may
↘	35583	sa	↘	9473	gid	↘	5083	si
↘	33210	hay	↘	9264	nag-	↘	5041	naman
↘	19491	nang	↘	8888	pag-	↘	4933	sinda
↘	19362	na	↘	8807	pa	↘	4857	'indi'
↘	16139	-in-	↘	8806	'iya	↘	4756	'inda
↘	15332	ning	↘	8622	kunu	↘	4412	mu
↘	14846	kay	↘	8381	ku	↘	4102	'āyu
↘	13713	'ina'	↘	7341	waya'	↘	4018	niya
↘	13667	-an	↘	7273	ka	↘	3989	'unga'
↘	12879	kung	↘	7030	mag-	↘	3971	'isa
↘	12836	kag	↘	6780	'akun	↘	3831	ni
↘	11216	'adtu	↘	6670	siya	↘	3538	'imaw
↘	11090	manga	↘	5992	'Imu	↘	3521	pIru
↘	10152	lang	↘	5450	hambay	↘	3464	kami

ITTEXT.1

Table 5 Romblomanon Morphemes Listed by Frequency of Occurrence

3.3 Selection of Morphemes for Lexical Analysis

A modest project must be selective in the morphemes it chooses for analysis and description as lexemes. Frequency of occurrence is a major consideration in accomplishing this.

A computer program run on morphemes listed by frequency of occurrence for the entire corpus, generates various counts of unique morphemes when the morphemes with low occurrence are eliminated. Table 6 indicates the number of unique morphemes which result when morphemes occurring less than 3, 5, and 10 times, respectively, in varying sizes of Romblomanon texts are eliminated⁶.

Total Morpheme units in running text	Unique Morphemes 1 x more	Unique Morphemes 3 x more	Unique Morphemes 5 x more	Unique Morphemes 10 x more
200,000	5,000	2,500	2,000	1,200
400,000	6,800	3,700	2,700	1,700
600,000	8,100	4,800	3,400	2,400
800,000	9,300	5,400	3,950	2,900
1,000,000	10,400	5,800	4,400	3,300

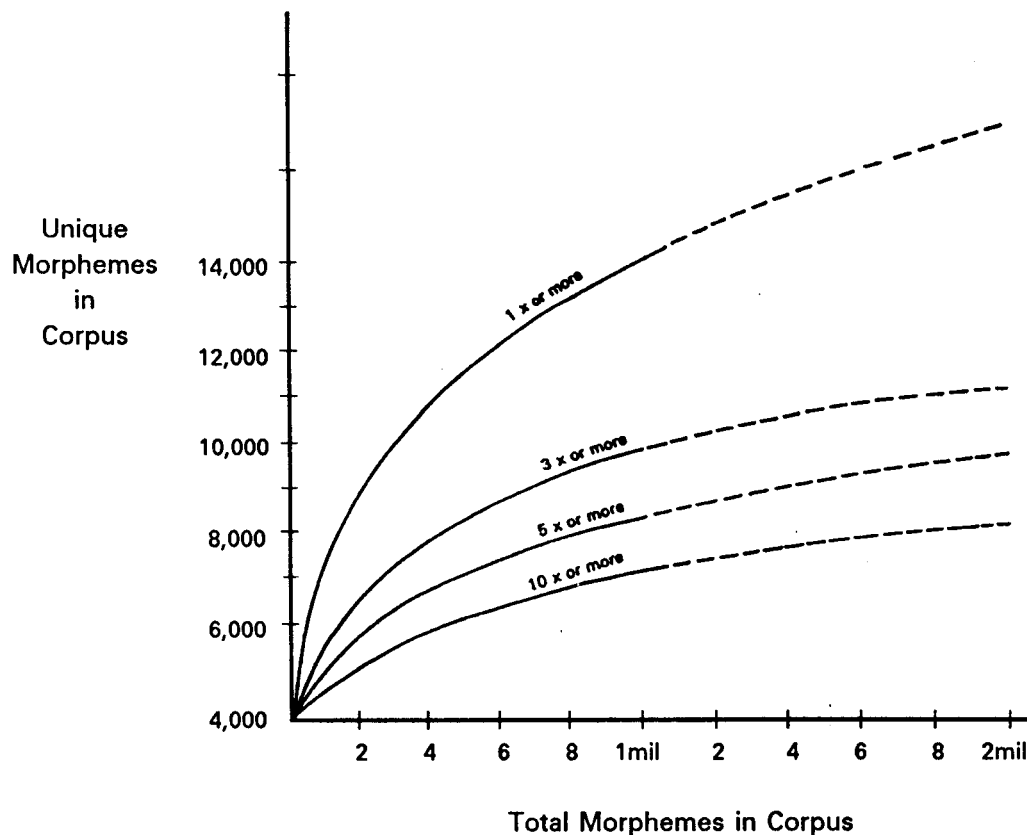


Table 6 Unique Morphemes Occurring in Various Corpus Sizes

Presumably if a morpheme occurs only once or twice in say a two million morpheme corpus, its lexeme equivalent would not be considered for inclusion within the dictionary. The only exception to this would be if it is judged to have some specific value to a language learner

⁶ This information highlights the fact that a large number of Romblomanon morphemes occur proportionately only a few times in text. This is true for other languages as well (McFarland 1989:24).

or if it were a member of a lexical set which, by its inclusion, would help to define the semantic composition of other lexemes otherwise included, by contrasting with them.

Table 6 indicates that if morphemes with one and two occurrences in the corpus were eliminated, a corpus of a million morphemes would yield about 6,000 unique morphemes. If morphemes with four or less occurrences were eliminated the result would be about 4,500 unique morphemes.

The aim set for the Romblomanon project is for a dictionary of from 8 to 10 thousand lexemes. In order to reach this goal, it is estimated that it will be necessary to select a minimum of about 6,000 morphemes. It is expected that features such as derivation, compounding, idioms and set phrases, along with senses of lexical forms will result in a considerably larger number of lexemes than morphemes⁷. Thus if 6 thousand morphemes were chosen it is expected that this goal would be attained.

Projecting for a two million morpheme corpus, eliminating morphemes of four or less occurrences would result in slightly less than 6,000 unique morphemes. On the basis of this information, a corpus of about two million morphemes or a little more will probably be a minimal requirement. The goal for the Romblomanon project is for a three million morpheme corpus.

In making lexical choices a good dictionary, and especially a learners' dictionary, will take into consideration not only frequency of occurrence but also the feature of utility. That is, the experience or interest of, for example, a beginner in a language will not be distributed equally across an entire range of cultural fields. Features such as speech styles, discussion of concrete subjects including culturally-related artifacts, kinship, cooking, eating, sleeping, time segments and similar common lexical sets need to be given special attention.

4. A Concordance of Morphemes

An important step in preparing a corpus for lexical analysis is to make a concordance for each morpheme chosen for analysis. It is impractical and unnecessary to create concordances for all morphemes in a single operation. Concordances are made for either single morphemes as they are analysed or for two or more morphemes whose lexeme equivalents are suspected of being closely related semantically, thus requiring contrastive analysis.

Concordancing is performed only on the morpheme line. (This is the line marked \no in Table 3.)

```
\lw tanda' = -i
\tn RO53975 038
\ex Tanda'i nga 'ang 'indu lang gid buy'un hay 'ang hinug na nga bunga.

\lw tanda' = -i
\tn RO57708 099
\ex Kag tanda'i kunu 'ini nga kung 'anu 'ang 'imu ginapinusung hay 'imu.

\lw tanda' = -in- = -an
\tn ro15965 100
\ex Tinanda'an 'adtu ni manung Lyüni 'ang manga tügen sa 'iya.
```

⁷ The proportion of morphemes to lexemes for the Batad Ifugao Dictionary project (Newell 1993), compiled without the benefit of computer, was about 1:1.4. Experience to date indicates that the use of a computer for lexical analysis will increase the ratio to 1:1.5 or higher.

```

\lw tanda' =ka-
\tn ro15969 045
\ex Waya' gid 'aku katanda' nga naga'away kamu magmanghud.

\lw tanda' =ka-
\tn RO75807 054
\ex Hambay ku 'ilam waya' 'aku katanda' basta 'aku gapānaw.

\lw tanda' =ka-
\tn ro78806 009
\ex Sinda hay 'indi'katanda' nang 'inda pwistu kung sa di'in sinda malinya.      TANDA'.X

```

Table 7 Portion of Concordance for tanda'

The concorded material, illustrated in Table 7, does not preserve the morpheme line. The text number line (\tn) is preserved for consulting the text source, and lexical analysis is of the boldfaced morpheme within the context of sentence texts on the text line (\ex).

4.1 Concordancing Within the Context of a Sentence

It is common for concordancing to be done on a single line of text with the unit for which a concordance is provided occurring more or less in the center of the line. The result of several lines concorded in this way is a display of the focused unit in a column running down the center of the page. This has the advantage of being able to quickly identify the focused unit. However, for the purpose of lexicographic analysis, it has the distinct disadvantage of providing only a random context, often of one or more sentence fragment rather than a complete sentence.

For adequate analysis in preparation for dictionary compilation, both semantic and grammatical, the minimal unit is considered to be the sentence. For this reason, in the Romblomanon project, concordancing is displayed within a sentence context.

Another reason the sentence context is preserved is to allow for the selection of example sentences for the dictionary directly from concorded materials. These materials, containing several examples of a given lexeme, is an excellent source for making such a selection.

We have overcome some of the difficulty in locating the morpheme for which the concordance is provided within a sentence context by having the computer mark it by boldfacing during the concordance operation, as illustrated in Table 7.

4.2 A Root is Grouped with Co-occurring Affixes

When two or more morphemes co-occur within a word (for example a root and one or more affixes) the computer program groups the sentences according to morphemes which co-occur with the morpheme for which the concordance is made. Table 7 is a list of sentences concorded for the verb root morpheme *tanda'* and grouped in this way⁸. The affix or affix combination which co-occurs with *tanda'* in each sentence is indicated on the line marked \lw.

The concordancing operation also generates a file listing the number of occurrences within the corpus of the root morpheme occurring alone and a listing of co-occurring affixes with an indication of how many times each occurs in the corpus.

⁸ This list is abbreviated for illustration purposes. Actual lists are usually much longer.

tanda' 29	tanda' = mag- 3
tanda' = CV = na- = -an 17	tanda' = ka- = pa- = an 2
tanda' = -an 16	tanda' = ma- = -i 2
tanda' = na- = -an 16	tanda' = D = ga- = -in- = -an 1
tanda' = pala- = -an 16	tanda' = D = na- = -an 1
tanda' = ma- = -an 11	tanda' = ga- 1
tanda' = -i 8	tanda' = ga- = -in- = -i 1
tanda' = ga- = -in- = -an 5	tanda' = ga- = na- 1
tanda' = ka- 5	tanda' = -in- = -an 1
tanda' = giN- = -an 4	tanda' = ka- = ma- 1
tanda' = ka- = na- 3	tanda' = nag- 1

TANDA'.SUM

Table 8 Root morpheme tanda' with co-occurring affixes

The importance of this information cannot be overstated. Verbs in Philippine and other languages are commonly inflected by affixes to indicate such features as tense, aspect, mode and role (case) relationships between verbs and co-occurring substantives of actor, instrument, goal, etc. Thus, grammar statements about verbs are commonly arranged in paradigmatic sets and verbs are classified with respect to the particular sets to which they belong. This is acceptable for grammar statements where generalized statements are made. However, this information for a language learner is misleading because not all verbs within a set are equally inflected by the affixes that constitute the paradigm. That is, for any given verb, some verb-affix combinations might occur with high frequency, some with less frequency and still others might rarely or never occur in natural text. In part, restrictions on co-occurrence is semantically controlled. Experience indicates that native intuition is unreliable, even by a trained consultant, for such co-occurrence information. Unfortunately, based on native intuition, uncommon and even aberrant constructions find their way into dictionaries and, worse yet, examples are fabricated to support them. A reliable source is concorded text using a large corpus as recommended here.

This information is crucial in guiding the lexicographer in deciding which inflected forms to include in a lexical description. Many morphemes (especially verb roots) occurring 100 or more times within the corpus are commonly found inflected by a given affix only once or twice⁹. In other words, a word consisting of a given affixed root may occur only once or twice, even though that root occurs 100 or more times within its total context. We are, in general, including a description of inflected forms in the dictionary provided they occur five or more times. We include forms occurring four times or less only if they are judged to be well-formed by trained native-language consultants and if they are necessary to fill gaps in a patterned description of a given lexeme (usually a verb). We have avoided the procedure of eliciting affixed forms not found in the corpus of natural text materials.

4.3 A Concordance Provides the Basis for Semantic Analysis

There are several major semantic areas that can be observed and sorted out when morphemes are arranged within a concordance. One is the area of homonymy. Homonyms will occur together in a single concordance since grouping by morpheme is by form rather than by meaning. Thus, homonym contrasts are easily detected.

⁹ The root morpheme *tanda'* occurred 153 times in the Romblomanon corpus of about one and a third million morphemes when the last count was taken. Nevertheless 8 co-occurring affixes or affix combinations occurred only once within the corpus, as indicated in Table 8.

Polysemy¹⁰ is also identified. This involves distinct lexemes with identical forms, related in some semantic way but which contrast in the bundles of semantic components by which they are defined (see Footnote 4).

If the corpus is sufficiently large, example sentences should also help determine the range of meaning of a given lexeme. This is one area in which a certain amount of filling in of holes is sometimes necessary through careful elicitation.

Sentences containing derivatives formed from a root for which the concordance is made will be grouped together, since segmenting by morpheme does not distinguish between derivational and inflectional affixation. In the same way, compounds (i.e. two or more root morphemes not following phrase structure rules and contrasting in meaning with all other lexemes) will also be grouped together.

A lexicographic description should indicate what other words commonly co-occur with the lexeme being described and what the grammatical and semantic relationships are. For example, a description of the English word *delectable* should include the information that it collocates only with things that can be sensed by taste or smell¹¹. Thus it would be acceptable to refer to a *delectable pastry* or *delectable aroma*, but not a *delectable ballet*, since the latter is sensed by sight and hearing.

Alphabetizing words, in turn, to the left and right of the morpheme for which the concordance is made brings together significant phrases to facilitate identification of collocational, idiomatic and set phrase information.

Recurrent phrases are identified as idioms and set phrases by alphabetizing concorded material.

```

\lw 'unu=na-
\tn ro75205 122
\ex Pag'abut nāmun sa dispinsaryu pinangutāna naman kami kung na'umu
naman 'ang 'iya pilas hay nabu'uy na 'ang tahi' kag nagkIput na 'ang
tinahi'an.

\lw 'unu=na-
\tn RO15905 097
\ex 'isa pa waya' gid sinda sa 'ākun nagapangimusta nga kung na'umu na
'aku 'u kamusta 'ang pangabūhi' ku diri labi na gid 'ang 'inda manga
tudlu' sa 'ākun.

\lw 'unu=na-
\tn RO58602 075
\ex Lunuwas 'aku 'ina' niyan kag 'ang hambay ku sa 'iya nang
pagkakIta' niya sa 'ākun siru'a 'aku ma'āyu kung 'anu 'ang 'ākun
kamutāngan kag pangutan'un mu 'aku kung na'umu 'aku.

\lw 'unu=na-
\tn RO53504 061
\ex Hambay ku sa 'ākun 'asāwa hay bāngun 'ānay dira' kag kadtu'un si
pari 'Iprin bāsi' kung na'umu na 'adtu.

\lw 'unu=na-

```

'UNU.X-

Table 9 Set Phrase Identified by Alphabetizing to the Left

¹⁰ A single lexical form with multiple significations is said to be polysemous.

¹¹ This analysis is based on my intuitive understanding of the word *delectable* in Canadian English. It would need to be verified by a concordance for this word, as described in this paper.

Table 9 illustrates a portion of a concordance for the morpheme 'unu, with the word to the left of the morpheme alphabetized. The set phrase *kung na'unu* has the meaning 'what the situation is'.

```

\lw bag'u=-ng
\tn R046108 026
\ex Magkaddu 'ina' niyan 'ang nanay kay nakabati' ning labunuk kay
bag'ung halin pa lang sa wikkadu.

\lw bag'u=-ng
\tn R059103 036
\ex Si Rimi naman nga bag'ung halin sa hospital hay dyagan naman kay
Nunuy, gani' 'aba?

\lw bag'u=-ng
\tn R058608 035
\ex 'ang hambay nang 'akun nanay kay 'aku 'adtu hay 'ang hambay nang
'akun nanay kay 'aku 'adtu hay bag'ung halin sa 'ilu'ilu kay yadtu
didtu 'ang pangita' nang 'asawa ku.

\lw bag'u=-ng
\tn R058606 078
\ex Hay yari da kunu si Dyun bag'ung halin sa Kwa'it.

\lw bag'u=-ng
\tn R015901 027
\ex Mang 'aku hay bag'ung halin sa Manila' tuyu katug 'akuy 'istar
sa Manila' piru waya' pa 'akuy 'asawa.

```

BAG'U.X-

Table 10 Set Phrase Identified by Alphabetizing to the Right

Alphabetizing the word to the right of the morpheme *bag'u* (Table 10), results in sentences containing the set phrase *bag'ung halin* 'newly-left.' Other idioms and set phrases grouped in this operation are:

<i>bag'ung abut</i> 'newly-arrived'	(11 times)
<i>bag'ung kasay</i> 'newlyweds'	(12 times)
<i>bag'ung tawu</i> 'newborn baby'	(40 times)
<i>bag'ung tu'ig</i> 'new year'	(28 times)
<i>bag'ung unga</i> 'one who has just delivered a baby'	(21 times)

The lexicographer will include these expressions in the dictionary and will illustrate them with the best choices of concorded sentences.

5. Compiling a Dictionary Entry

Once semantic and grammatical features of a given lexical form have been analyzed, this information is entered in the dictionary. Choice examples from the concorded material are transferred to the dictionary by computer.

This process of transferring example sentences continues until all semantic and grammatical features of the lexical form have been adequately illustrated.

6. Grammar Companion for the Dictionary

A grammar statement, usually a brief sketch, is an extremely helpful companion to a dictionary, especially a learners' dictionary of a lesser-known language. The grammar is ideally based on the same principles that guided the development of the lexicon. That is, the grammar

makes generalized functional statements using concordances from the same corpus as for lexical analysis. The actual grammatical behavior of lexemes is observed and, from these specific observations, generalized grammatical statements of features such as parts of speech, tense, aspect, case, degree, number, etc., are made¹². One function of the lexicon is to serve as a place where generalizations made in the grammar section are extensively illustrated. Another is to cite exceptions to these generalized statements.

7. Conclusion

Much of the drudgery of dictionary making in the past has involved collecting, sorting, arranging and filing lexical materials with the use of unending stacks and files of paper. The computer has made this process obsolete. Using the computer to perform routine tasks in preparing lexical materials is allowing the lexicographer to concentrate on the more interesting work of analyzing and describing the lexicon. And the resulting dictionary represents a more comprehensive, accurate and natural description of the lexicon.

¹² A grammar will usually include generalized descriptions of other tactical units such as phrases, clauses, sentences, paragraphs and discourse units. Restrictions imposed by specific lexemes on the composition or distribution of these tactical units, especially if they are unpatterned, will be found in the lexicon.