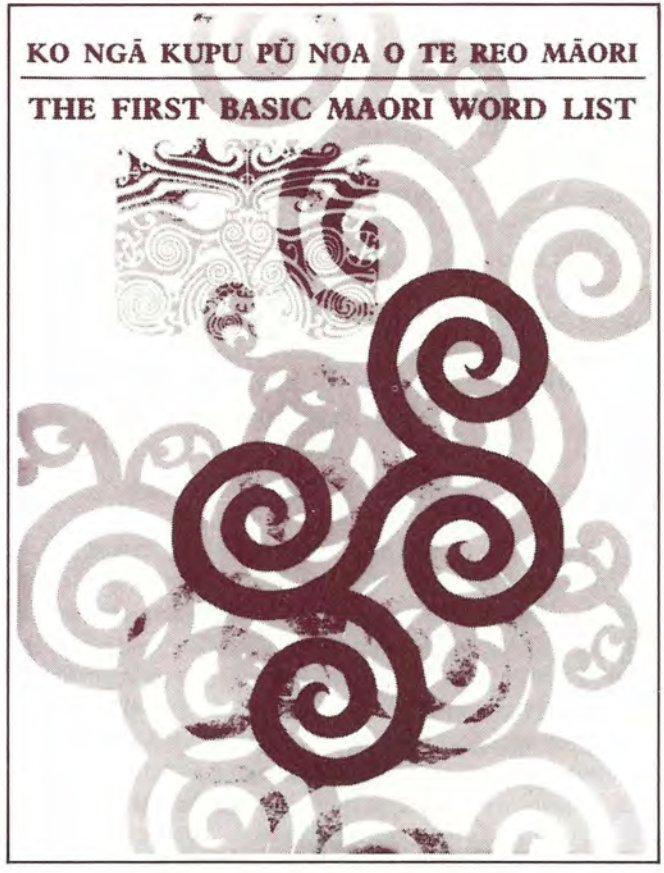
THE ANATOMY OF A WORD LIST



First Issued in *set: research information for teachers*, no. 2, 1982.

Printed by Lithoprint (NZ) Ltd.

The Anatomy of a Word List

By Richard Benton
NZCER

On June 12, 1982, a little book with a long title, *Ko Ngā Kupu Pū Noa o te Reo Māori: The First Basic Maori Word List*, was launched by the New Zealand Council for Educational Research. The 'basic-ness' of the words included in the book (there are about 800 of them) was determined by a combination of how often they occurred and how well they were distributed in the sources from which they were gathered. This article outlines the way in which the Basic Word List was developed and organized. A number of general issues relating to words and word lists are discussed, as are some similarities and differences between the content and construction of the Maori list and other word lists in common use. It also explains how the Basic Maori Word List can be used by teachers and learners of Maori.

Why have a word list at all?

One of the greatest problems which the learner of any language has to confront is that of being "stuck for words". It's not the only problem, of course, and, as we shall see, there are words *and* words (not to mention phrases and sentences of one kind and another). However, all language learners, from infants learning their first language to adults learning a second or third language, have eventually to learn many thousands of individual words in order to be able to communicate adequately with other people, and to understand the spoken and written language to which they will be exposed.

The number of words needed by an individual is astounding. A common rule-of-thumb estimate is that a child entering school will usually command a listening vocabulary of at least 5,000 words (although many hundreds of these may not be actively used by such a child). The Thorndike and Lorge *Teachers Word Book of 30,000 Words* contains very few words that an average secondary school pupil would not know – which means that a native speaker of English needs to "learn" in some degree an average of four to five new words a day from birth to age 15 or 16 in order to become a competent speaker and hearer of the language. A second-language learner has a similar task, but usually with less time to cover the same ground.

It is not surprising therefore that word lists of one kind or another have been around since the known beginnings of formal language instruction. However, since one cannot pick up 3,000, let alone 30,000, words in a day or two, language students, and their teachers, have tried to set priorities which would enable them to pick up the words that are likely to be most useful to them at first, leaving other words of less immediate relevance until later. For native speakers of a language, word-lists are most commonly associated with the teaching of reading and writing – and again the principle guiding the selection of words has been the learner's needs: sets of words included in the books a child will use first, lists of words which are difficult to spell or to recognize, and so on. The methods used to select these words have been many and varied, ranging from the whim of the teacher to the results of extensive and highly systematic research. We will briefly explore some of the research-based methods shortly.

The Basic Maori Word List

Teachers of Maori in New Zealand have never been oversupplied with resources to help them in their work. As the numbers of students of Maori have increased (from a few hundred in the 1940s to probably around the 100,000 mark in 1982, counting the various adult and after-school classes as well as children taking formal primary and secondary school courses), the need for basic reference works has become obvious. It has become particularly pressing in recent years because the vast majority of these students have English as their first language, as do quite a number of their teachers.

As part of its effort to improve the quality of Maori language teaching, and to provide better support for the teachers, the Department of Education requested the New Zealand Council for Educational Research to prepare a basic Maori word list, primarily for the use of teachers and of people writing textbooks and readers for children in the junior classes of bilingual primary schools. Work began on the project in 1979, and, after many trials and tribulations, the final computer runs were completed two years later, with the finished product ready for the printer early in 1982.

Before work started on the word-list proper, a sample list was constructed, using the entire vocabulary of the standard secondary-school text books, with a few other words thrown in for good measure, arranged according to an arbitrary system of categories. This list was circulated to a large number of teachers, in the hope that they could comment on the form of the list, the number of words, the types of information given about them, and so on, so that the "real" word-list, when produced, could take a form most closely filling the needs the teachers expressed. What actually happened was that most teachers were reluctant to part with the trial lists, and little specific information was received – a few were highly critical of the "sample", but, unfortunately, did not have a specific alternative in mind. The exercise did, however, highlight the need for a resource of this kind, whatever the form.

In the end, what may be described as a very basic list was produced, one which contains and classifies some 650 words which are the least dispensible in a wide variety of contexts. These are words which no learner can do without, but, bearing in mind the five new words a day principle, they constitute only a small fraction of the words which will be needed by anyone who hopes to be an active user of the language. However, because they are words which are likely to turn up in all sorts of situations, there are many advantages to be gained by ensuring that they are among the first to which a new learner is introduced.

What is a word?

Before a search for basic words can be mounted, it is necessary to decide just what a word might be. One easy way out (taken, for example, by the compilers of the massive, and authoritative, American Heritage *Word Frequency Book*) is to regard anything that occurs between spaces in a written text as a word, paying attention only to the way they are written down. Thus **SET**, **set**, **Set**, **sets**, **SETS**, all count as separate words, and appear at different points on an 87,000-word list. Some other lists ignore superficial differences in spelling, and combine words which differ only in the addition of a regular ending, thus treating all the variations on **set** and **sets** in the example just quoted as a single word, **set**, for purposes of analysis. The *New Zealand Basic Word List* and the *New Instant Word List* adopt this approach.

None of the lists mentioned, however, takes the acutal *meaning* of a word into account. This is partly because one of the major purposes of such lists is to give reading teachers some idea of which written forms their students most need to be able to recognize at an early stage. The Thorndike and Lorge word book does pay some attention to meaning, for example by indicating that the entry for **hurried** refers to the adjectival ("a hurried glance") rather than past-tense ("he hurried round") usage of this word-form; the same list, however, lumps together **hunter** and **Hunter**, despite the likelihood of the latter being a proper name whose bearer may never have hunted anything (although it does separtate **hunt** and **Hunt**). Michael West's *General Service List of English Words* pays very close attention to meaning – while taking the written form as the unit for each entry, it indicates the relative frequency in which different meanings or senses of each of these "words" was encountered.

To compound the problem, if we intend a word-list to be used by speakers as well as writers (or readers), we must recognize that words are not found only on paper or tablets of stone. In fact, compilers of English word lists have had their task made easier for them by the fact that the spelling system often takes account of meaning even when the spoken forms are identical, as with **two**, **too**, and **to**, for example. Because Maori words have been "reduced to writing" for only 150 years or so, when different sound-forms have converged, differences in meaning have not continued to have different spellings.

If Maori had been written for as long as English, words like **ua** *rain*, **ua** *backbone*, and **uaua** *vein* would probably reflect their varied histories, being written perhaps **'uha**, **u'a**, and **uaua** respectively. The fact that this is not the case in Maori forces us to face the question of whether there are two different *words* with the form **ua**, or whether the word **ua** has two different *meanings* – *rain* and *backbone* (among others). So also with dialect forms. Where in English variant *pronunciations* of the same "word" are usually ignored in writing (so that the Irishman asked to adjudicate between two Englishmen as to whether 'neither' should be pronounced "nee-ther" or "neye-ther" was able to inform them that "nay-ther" version was correct), Maori writing tends to reflect these differences fairly consistently – thus **rātau** and **rātou** *they* and **kei** and **kai** *at* appear as separate forms in writing as well as speech.

For the purposes of the Maori Word List, both form and meaning were taken into account in determining just what constituted a word. Where *form* was concerned, any unit of speech which had some meaning in isolation and which was normally written as a single unit was given the status of a word-form. These word-forms were then examined for their meanings (with the word-root being taken into account for reduplicated or affixed forms). A form like **tau** for example, may have any one of a large number of meanings – Williams' Dictionary lists about 30 for the simple form alone, let alone derivatives like **tatau**, **taunga**, **whakatau** and so on. In cases like this, judgements (sometimes, undoubtedly, somewhat arbitrary ones) were made as to which meanings were to be grouped together, so that any particular occurrence of **tau** could be assigned to a specific "word". Eight such "words" were encountered (or recognized) in our data:

tau 1	*come to rest, alight, ride at anchor, etc.*
tau 2	*year, season, period of time*
tau 3	*bundle*
tau 4	*spouse, lover, darling*
tau 5	*ridge, reef*
tau 6	*steeped in water*
tau 7	*bark like a dog*
tau 8	*attack*

This list was further expanded when derived forms were taken into account, so that **tatau**, *count*, was taken as being a single undivided form **tatau** 1 (probably related to **tau** 3, but sufficiently different to justify separation), while **whakatau**, *adorn, prepare, make ready*, was regarded as a combination of a ninth **tau** with the causative prefix **whaka-**. Another **tatau**, meaning *sliding door or window slab* was treated as a second simple form, **tatau** 2 in our computations, although it is listed under the heading **tau (viii)** in Williams' dictionary.

On the other hand, slightly divergent meanings were usually treated simply as "senses" of a single word. Like **tau** 1, the form **hoa** has a number of related meanings – *friend*, *spouse*, *companion*, in fact, any person in some kind of reciprocal relationship with another. Becaused of this, **hoa** was treated as a single word in our analysis. The compound form, **hoariri** *enemy* was, however, treated as a word in its own right, as were most other compounds.

Now in Maori, just as in English, and probably in every natural language, there are many groups of words which really function as single units – that is, the meaning of the group is not quite what one would expect from the normal meanings of the component parts: "see you later, alligator", "that'll be the day", "I'll say", and hundreds of others. These are normally ignored in word lists, although they are critically important to anyone wanting to understand colloquial language, particularly in spoken form. "By and large" such combinations were ignored in preparing the data for the basic Maori Word List, too. However, occurrences of a few combinations, such as the verbal particles **e** and **ana**, and the forms **kei te** and **i te** with verbs, were treated as single units, while special uses of **haere** and **noho** in greetings and farewells were separated out from the other uses of those verbs.

In the first set of computer runs over the data all forms and combinations normally written as single words (but further subdivided according to criteria of meaning as noted above) were counted separately. (Just what the data consisted of, and how it was divided up for purposes of analysis, will be revealed shortly.) All the forms in the right-hand column of the list below were computed as separate words.

Normal written form	English gloss	Computed as
tau	*come to rest, alight*	**tau** 1
tau	*attack*	**tau** 8
tau	*bundle*	**tau** 3
tau	*year, season, etc.*	**tau** 2
tau	*bark like a dog*	**tau** 7
tautau	*bark like a dog*	**tau** 7 **-tau**
whakatau	*adorn, prepare, etc.*	**whaka-tau** 9
tatau	*quarrel*	**ta-tau** 8
tatau	*count*	**tatau** 1
tatau	*door*	**tatau** 2
whakatatau	*quarrel*	**whaka-ta-tau** 8
tauranga	*anchorage*	**tau** 1 **-ranga**

After this, a second set of criteria came into play. Most content-words in Maori consist either of a word-root in isolation, or a word-root to which one or more of a few

prefixes or suffixes have been attached to modify the meaning of the root in some systematic way. Similar processes are present in English, as will be obvious from some of the translations of these examples:

Word-root	Word	English gloss
tau	**tau**	*come to anchor, etc.*
	tauranga	*anchorage*
mātau	**mātau**	*know*
	mātauranga	*knowledge, education*
	whakamātau	*make to know; test*
	whakamātautau	*try on*
iri	**iri**	*elevated*
	whakairi	*elevate*
	whakairia	*cause to be elevated*
pō	**pō**	*night*
	pōnga	*nightfall*
	pōngia	*benighted*
	whakapō	*darken*
	pōpō	*nights (archaic)*

In cases such as these, some familiarity with the general meaning of the word-root (or with its specific meaning when used as a word in itself) would give the listener or reader a very good chance of working out the meaning of the derived forms. Because the occurrences of a particular word-root could be scattered over a number of derived forms, no one of which might be particularly important, but which together might have quite a formidable presence, we decided to take the word roots themselves into account, independently of the other bits and pieces attached to them.

Similarly, we examined alternate forms of essentially the same word (as the two forms for "they" – **rātou** and **rātau** – already mentioned), and also alternate words for the same idea, such as **mangu** and **pango** *black*, **waha** and **māngai** *mouth*, and so on. The latter also have many parallels in the major dialects of standard English – **footpath** versus **sidewalk**, or, nearer home, **bach** versus **crib**. Again, the combined frequencies and distributions of all these variants were examined, as the parts taken together might in many cases form a very significant whole.

Where do the words come from?

Having decided on what could constitute a word (at least in principle), and some of the things that could be done with them, decisions still had to be made on where to

obtain the words themselves. Since the word-list was intended primarily (but not exclusively) for use by people preparing readers and other material for younger children in bilingual primary schools, it was clear that as much material as was available already prepared or suitable for such children should be included. However, since preparation of the list was likely to be a time-consuming and relatively costly business, and the number of "primary" users was likely to be very small, a significant amount of additional material was also desirable, to expand the data base and make the list more obviously helpful to other writers and to more advanced learners. Similarly, since all bilingual programmes, and most other Maori language courses, contain a high oral component, transcriptions of spoken Maori should also find their way into the data.

In the event, the data was drawn from as many suitable texts and recordings as could be prepared for computer analysis given the limits of time and money available. Our original aim had been to select the words from a data base of about 200,000 running words, but, for the first word list, we have had to be content with a base of about 120,000. When the data had been edited and coded (to take care of the nine varieties of **tau**, and similar complications), they were divided into six major groupings or "worlds".

(1) *Primary school texts:* These were books and booklets, either published, or homemade by teachers, which were designed primarily for teaching the language to younger children (for example H.T. Rikihana's *Kōrero Māori* readers).

(2) *Primary school, general:* Books and stories written to be read for information or pleasure, and either designed or suitable for children of primary school age (for example, *Crayfishing with Grandmother,* the various stories in *Te Tautoko* journals).

(3) *Spoken Maori of Children:* Transcripts of taped conversations of children who were native-speakers of Maori (mostly from the Ruatoki district in the Bay of Plenty).

(4) *Secondary School Texts:* All the narrative material in the Department of Education text books for Secondary Schools (*Te Rangatahi* series and *Te Reo Rangatira*).

(5) *Secondary School, General:* Stories and informative material written or suitable for older children (e.g. the stories in various issues of *Te Wharekura*).

(6) *General:* A selection of oral and written material directed at adult or general audiences (e.g. transcriptions of news broadcasts, articles and stories from *Te Ao Hou* and *Te Kaea*).

The "worlds" were further subdivided into 27 major sub-groups, the number of which varied from world to world (there were three for secondary-school "texts", for example, but six for the "general" primary level materials). Each individual story and conversation – there were over 300 of them – was also given a code-number, but the

computer analysis of the data was confined to the six worlds and the 27 major groupings within them.

Weighing the Words

Having decided on the various forms a word might take, capturing them and classifying their habitats, we could now begin the selection of those which could be regarded as basic.

In any particular situation, every word used (or needed) can be regarded as basic to that occasion. However, there will be some words that reappear in many different situations, and the aim of basic word lists is to isolate these for the benefit of teachers and learners.

Many basic word lists have been constructed purely on the basis of the gross frequency in which particular words are encountered in the data examined. This is a straightforward method, and probably as satisfactory as any other if a general purpose list of 50 to 100 words is all that is required. However, a very high proportion of those first hundred words will be grammatical markers of one kind or another (words like **he**, **te**, **e** in Maori, **a**, **the**, **by** in English), with very few of the concrete nouns or other words needed to express specific ideas. In any large-scale frequency count, the 20 or 25 most common words are likely to account for about 40 percent of all occurrences (this is as true for English or Spanish as it is for Maori). The frequencies of the remaining words will to an increasingly greater degree be affected by the specific texts from which the language sample was drawn, so that their relative importance in everyday life may not be reflected very reliably even when a very large data base has been used.

One way of getting around the problems posed by straight-out frequency counts is to examine how well particular words are distributed over a variety of different contexts. It might reasonably be argued, for example, that a word which pops up once in each half a dozen situations is considerably more important or useful for general purposes than one which occurs ten times but in only one context. Approaches which have taken account of both frequency and distribution in assessing the general usefulness of words have been developed by the compilers of two major word lists, *The Frequency Dictionary of Spanish Words* (1964), and *The American Heritage Word Frequency Book* (1971), among others.

In assessing the usefulness of words, however, there are factors other than frequency and dispersion to be taken into account. It is important for a teacher or language learner to be aware of these, even though no single word list is likely to be able to keep them in perfect balance.

The first set of factors is related to the role of the language user: is it productive or receptive? Often, of course, it will be both, but the gross number of words which we

need to be able actually to use is likely to be much less than those we need to be able to understand, and the order of importance of particular words within the two sets is not necessarily the same. The mode in which the words are to be used is also relevant: speaking and writing (both productive in nature), make different demands, as do listening and reading. Cutting across all these, of course, are considerations which are highly relevant to specific contexts – formality, status, and topic, for example.

A parallel set of considerations is the interrelationship between the frequency, familiarity and accessibility of particular words. In this regard, we have to take into account contexts as well as texts, and time as well as space. The concept of *familiarity* is very important for vocabulary selection. There are many words which "everybody" knows, but which are neither frequent or well dispersed in written texts or in speech. Jack Richards, who argued that familiarity should be a guiding principle in the construction of basic vocabulary lists, points out that one such word in English is **toothpaste** (you won't find it in Thorndike and Lorge – although **toothbrush** is there. It occupies rank 13085, well out of range of basic vocabulary, in the American Heritage list). Another is **incinerator**, which, like **heterozygosis**, occurred only once in the 5,000,000 running words used in the American Heritage count.

One of the reasons for the apparent gap, or chasm, between frequency and familiarity is that "frequency" is used by word-list makers in a very restricted sense, taking into account only static examples of writing or speech; so also with dispersion (distribution) within subsets of these samples. A word may be familiar, however, because it occurs frequently in a temporal sense in a *socially* well distributed domain. **Toothpaste** is one such word – it is not likely to occur often in any cross-sectional sample of speech or writing, but there is a regular need to use the object so-named (and therefore the likelihood of its having a name is very high), and, in most families, enquiries as to its whereabouts, or directions to purchase new supplies, are also likely to be common in daily life. However, because time rather than space is of the essence in cases like these, the importance of such a word is almost always under-estimated in general vocabulary lists.

The concept of *accessibility*, the need to have access to certain words in certain situations, is one of which language teachers have long been aware. Most special purpose word lists aim at making the most important words in a specified situation accessible to the learner. It should be noted that such words will not necessarily be familiar in the sense we have just discussed. The particular situations in which familiar words occur are within most people's everyday experience. However, there are many words which may be of critical importance to certain individuals, which many people, perhaps the great majority of native speakers of a language, would never encounter. An example from English would be **haplology**, which didn't occur even once in the American Heritage sample, but which will be found in many texts on linguistic

change (if you pronounce the word concerned "haplogy" you will have become a haplologist). A budding linguist therefore needs access to terms like **haplology**, but most people do not.

Because accessibility is often tied to very specific situations, attempts to build up basic word lists through the use of "domains" or "notions" as primary reference points are likely to run into difficulty. This is not to say that vocabulary lists tied to particular activities, occupations, and so on are not useful and important. Far from it. However, many items in such lists will have a rather low level of importance in general experience, however critical they may be in a specific context – one could get along in life without a word for "haplology" much more easily than in ignorance of a word for "toothpaste" – and it is not at all easy to draw a line between these two extremes.

Sorting the Wheat from the Chaff

Since an investigation of the relative familiarity of Maori words would have required the mounting of a new and time-consuming project, we had to settle for the next best thing, an approach which would take into account relative frequency in the six "worlds" into which our data were divided, as well as the degree to which each was dispersed in the 27 subsets within the various worlds. After certain computations were performed (the essential details will follow shortly) each word was assigned a *value*, and this numerical value became the basis for selection for the basic word list.

The first step was to ascertain how many times each word occurred in the various "worlds". In the data we actually used, there were big differences in the numbers of running words in each world, and so we used percentage frequencies rather than the actual "raw" numbers. In the examples which follow, however, we'll assume each "world" has a roughly equal population – the principle remains the same in either case. In our examples we have also set up a universe of only five "worlds". In practice, as many "worlds" as are needed can be created.

A glance at Table 1 will reveal that some words are dispersed much more evenly than others through our hypothetical universe. To measure the evenness of dispersion we used the formula developed by the authors of the Spanish frequency dictionary. This involves four steps:

(1) calculating the mean (average) number of occurrences per world;

(2) calculating the standard deviation from the mean;

(3) dividing the standard deviation by a number equivalent to the mean multiplied by the square root of the number of worlds minus one (in our five-world sample, the mean would be multiplied by two);

(4) subtracting the result from 1.

Table 1 Hypothetical frequencies of selected word-forms

Word	World 1	World 2	World 3	World 4	World 5	Total
te *the*	80	90	75	75	90	410
poti *cat*	5	5	10	10	5	30
tori *cat*	20	0	0	0	0	20
ngeru *cat*	0	5	0	0	10	15
tereina *train*	10	0	0	0	10	20
āpōpō *tomorrow*	5	5	5	10	5	30
tohorā *whale*	5	10	5	0	0	20
tāpuke *bury*	5	5	0	5	0	15
tāpuketia *buried*	0	0	5	0	5	10

Note: these frequencies were invented for this article, they are not actual frequencies.

This will give us a coefficient of dispersion, which will vary from 0, for a word which occurs in only one world, to 1, for a word which is perfectly evenly dispersed through all the worlds. As can be seen in Table 2, **tori** (the word for "cat" in one part of Northland), would, on the basis of the figures supplied, end up with a coefficient of zero, while **te**, the singular definite article, with a fairly even distribution, would have a coefficient of 0.96 (about as close to perfect as any reasonably frequent word would be likely to get).

Table 2 Coefficients of Dispersion (hypothetical examples)

Word	Total Frequency	Mean (5 Worlds)	Standard Deviation*		Divide by (Mean × 2)	Coefficient of Dispersion
te	410	82	6.8	÷	164 = 0.04	.96
poti	35	7	2.5	÷	14 = 0.18	.82
tori	20	4	8.0	÷	8 = 1.0	.00
ngeru	15	3	4.3	÷	6 = 0.72	.28
tereina	20	4	4.9	÷	8 = 0.61	.39
āpōpō	30	6	2.4	÷	12 = 0.20	.80
tohorā	20	4	3.4	÷	8 = 0.43	.57
tāpuke	15	3	2.4	÷	6 = 0.40	.60
tāpuketia	10	2	2.4	÷	4 = 0.60	.40

*Calculated by squaring the difference between actual occurrences and the mean in each world, dividing the sum of these by the number of worlds (5), and obtaining the square root of the result.

To give us a way of ranking words taking overall frequency and dispersion equally into account, we multiplied the total frequency by the dispersion coefficient. The resulting estimate of "usefulness" was used as the basis for rank-ordering words in the Spanish frequency dictionary, and a very similar method was also employed in the American Heritage list. We introduced a further modification, however, because of the possibly excessive influence of a single dialect which dominated two of our "worlds", but was paramount in only four of the 27 major subgroupings into which the data had been further divided.

To compensate for this bias we allocated an index number to each word, according to the number of subgroups in which it was represented. This ranged from 6 for words confined to one, two, or three subgroups, to 14 for those in 25 or more subgroups. The estimate of usefulness was then multiplied by the common logarithm of the subgroup index, and the result was called the "value" of the word. Using a logarithmic scale ensured that the words best distributed among the subgroups were given increasingly greater rewards for their virtue, while the poorly distributed words were punished with increasing severity for their sluggishness. Hypothetical estimates of "usefulness" corrected to estimates of "value" for the words in our sample are given in Table 3.

Table 3 Estimates of Usefulness and Value (hypothetical)

Word	Total Frequency	Coefficient of Dispersion	Estimate of Usefulness	Major Sources	Source Index	Common Logarithm	Estimate of Value
te	410	.96	393.6	27	14	1.15	452.6
poti	35	.82	28.7	5	7	0.85	24.4
āpōpō	30	.80	24.0	18	11	1.04	25.0
tohorā	20	.57	11.4	3	6	0.78	8.9
tereina	20	.39	7.8	4	7	0.85	6.6
tori	20	.00	0.0	1	6	0.78	0.0
ngeru	15	.28	4.2	2	6	0.78	3.3
tāpuke	15	.60	9.0	7	8	0.90	8.1
tāpuketia	10	.40	4.0	4	7	0.85	3.4

The correction of total frequency for dispersion had the effect of eliminating **tori** from our hypothetical sample (for the moment at least), and resulted in a number of other re-arrangements. The final adjustment, taking into account representation in

the major subgroupings as well as evenness of dispersion through the five worlds, has had the effect of promoting **āpōpō** into second place in the rank ordering, and reversing the relative positions of **tāpuketia** and **ngeru**. The result for **tori** (and for some readers, perhaps the whole process) is summed up by Lewis Carrol in this episode from *The Hunting of the Snark*.

> The beaver brought paper, portfolio, pens,
> And ink in unfailing supplies;
> While strange creepy creatures came out of their dens,
> And watched them with wondering eyes.
>
> So engrossed was the Butcher, he heeded them not,
> As he wrote with a pen in each hand,
> And explained all the while in a popular style
> Which the beaver could well understand.
>
> "Taking Three as the subject to reason about –
> A convenient number to state –
> We add seven, and ten, and then multiply out
> By one thousand diminished by eight.
>
> The result we proceed to divide, as you see,
> By nine hundred and ninety and two
> Then subtract seventeen, and the answer must be
> Exactly and perfectly true.
>
> The method employed I would gladly explain,
> While I have it so clear in my head,
> If I had but the time and you had but the brain –
> But much yet remains to be said.
>
> In one moment I've seen what has hitherto been
> Enveloped in absolute mystery,
> And without extra charge I will give you at large
> A lesson in Natural History."

The Return of the Cat

Two more sets of data remain to be subjected to the winnowing process. These are (1) the word-roots, and (2) the pronunciation and dialect variations representing a single idea. In computing the value of a word-root, all occurrences of the root (with or without affixes) were taken into account. The dialect variations were grouped together in a

similar manner, with one member of each selected as the code-word to represent all. We called this latter kind of grouping an "alliance" in our basic Maori word list.

In our sample list, the word-root **tāpuke** occurs in two separate words – **tāpuke** and **tāpuketia**. The idea "cat" is expressed by three words: **ngeru**, **poti**, and, of course, **tori**. We can therefore add two new items to our list, with values of their own, as indicated in Table 4.(The words for "cat" have been grouped together as ***TORI**. This name is quite arbitrary – any label would do.)

When added to the other items (see Table 3), the alliance ***TORI** will displace **āpōpō** from second position in the rank ordering, while the combined occurrences of **TĀPUKE** will be seen to be considerably more important than their separate parts, and take fourth place, ahead of **tohorā**, now a very distant fifth.

The Cut-off Point for Basic Vocabulary

Of the 3,500 or so distinct words encountered in the Maori word list data, half were confined to one world only, and were thus out of the running from the beginning. Values were computed for all the rest, plus the word roots appearing in more than one word, and some fifty "alliances". Many of the words were very poorly dispersed; only 965 of them were encountered in three or more worlds.

Because the values assigned to the less frequent words would be increasingly distorted by inadequate or biassed selection of the data from which they were drawn, it was decided to include only the 500 most valuable content-words in the basic word list, plus all the "grammatical" words, word-roots, and "alliances" which had values not less than that of the 500th-ranked content-word. These items were then cut up into five levels, each consisting of 100 content-words and the other items that correspond to them in value.

This ensured that neither the basic list nor any level was completely swamped by "grammatical" words – the various particles, pronouns, demonstratives and so on which form the building blocks of sentences, but which do not refer directly to objects or events in the outside world. Concrete nouns, which are vital if one wants to talk

Table 4 Hypothetical values of word root **TĀPUKE** and alliance ***TORI**

	Total Frequency	Dispersion Index	Estimate of Usefulness	(Sources) Index	Common Logarithm	Value
TĀPUKE	25	1.0	25.0	(11) 9	.95	23.8
***TORI**	70	.79	55.3	(8) 8	.90	49.8

about relationships or even a variety of activities, are often the victims of frequency-based lists, because they begin making their appearance lower down the scale than most grammatical words and some of the verbs. However, by taking the content words as our reference point, we were able to ensure that there were 30 or 40 concrete nouns at each of the five levels.

Altogether 649 actual words were included in the level lists – 193 at level one (almost half of them in the "grammatical" category), down to 111 at Level five. In addition, some 116 word-roots were included (30 of which would not otherwise have been represented at any level), and 40 "alliances". Taken together, these words and groupings accounted for about 92 percent of all occurrences in our data.

The Final Product

The published version of the list consists of three parts:

(1) the levels lists (already mentioned), organized alphabetically;

(2) lists of the basic words grouped by functions and notions; and

(3) an index, which brings together all the material in the previous two sections, and includes as well information on major dialect variants of all the words listed. There are a further 106 words in the index which do not appear in the levels lists, so all-in-all the basic vocabulary gives the user direct access to 785 Maori words, and indirect access (taking into account the process of affixation) to many hundreds more.

For example, the list for level three includes these entries:

ngahere
ngaru
NGAU
*NGERU

The first two are actual word-forms; the third represents a word-root; and the fourth is the alliance which includes our old friends 'ngeru', 'poti', and 'tori'. The corresponding entries in the index, to which the user of the list should refer next, are:

ngahere (forest): (3). 8. (34). S19, S20
ngaru (waves): (3). 6. (17). S19
NGAU (bite, gnaw): (3). 8. (23). ...
***NGERU** (cat): (3). 7. (34). ...
poti [*Eng]: (5). 3. (27). S23
ngeru [*W]: ... 0. (7). ...
tori [*Rarawa]: ... 0. (0). ...
pūihi [*Au]: ... 0. (0). ...

Each word is accompanied by a brief English gloss. This is followed, where appropriate, by information on derivation or dialect – **poti**, for example, is derived from an English word, **ngeru** is common in the Waikato area, **tori** among the Rarawa people and **pūihi** among the Aupouri.

For the main entries, the first figure in parentheses refers to the level to which the word or grouping has been assigned. All of these 4 main entries are of course tagged for level three. However, the members of an alliance, which are listed as sub-entries, will often be from a lower level than the alliance itself. In this case, although the idea "cat" is of level 3 value, only one of the specific words for "cat", **poti**, is a basic word at all – it is found at level 5. One dialect variant, **ngeru** occurred in our data, but was not sufficiently well distributed to be a basic word in its own right, and the other two, **tori** and **pūihi**, were known to us from other sources, but did not actually occur in our coded data. For word-roots, like **NGAU**, sub-entries with their meanings and other information are included only for actual word-forms which are included in the levels lists. Otherwise, just the word-root, with a generalized meaning, is included. All sub-entries (i.e. words like **tori**, **poti** and **pūihi**) are also given separate entries in the index, with cross-references to the major heading under which they are listed.

The next set of figures represents the computed value of the word or group (actually the value to 1 decimal place, represented as a whole number). This is followed by the actual raw frequency, in parentheses. Thus **ngahere**, with a computed value of 8, actually occurrred 34 times in our data. The final piece of information, provided only for actual words included in the levels lists, is a cross-reference to the lists of words in notional categories, which makes up the second major portion of the book.

The notional lists are actually organized in four sections; (1) word classes, etc; (2) language functions; (3) general notions; and (4) specific notions. The latter three sections are organized in the same way as those in Dr J.A. van Ek's book *The Threshold Level for Modern Language Learning in Schools*. Each section is divided into a number of subcategories, with codes in the margin to facilitate easy cross-reference from the index. Information on the words in our examples is contained in a number of categories in the 'specific notions' section, including:

2. House and home

S-19. *2.7 Region, environment*
kāinga, rākau, wai, whenua, tāone, moana, marae, awa, tuna, ahi, maunga, rori, rohe, toka, rangi (*sky*), ngahere, huarahi, rua, kōhatu, ngaru, moutere, taiapa, one, tāwāhi, motu, puna, oneone, mahinga, tai, puehu, pāmu, tāhuna, pātiki, puihi.

3. Life at home

S-23. *3.5 Pets*
kurī, poti.

The index and the notional lists can be used together in the same way as a thesaurus, although the user must bear in mind that only the most basic words are listed.

Quirks and limitations

Basic word lists can be very useful indeed as guides to which words a learner (or teacher) simply cannot afford to ignore any longer than absolutely necessary. They will not, however, provide all the words that a learner will need, even at the very earliest stages. And, of course they will give little if any help when it comes to putting these words into sentences, or using and understanding them in real communicative situations. Communication will be difficult without these words, but the words alone will not be sufficient for effective communication.

A very useful rule-of-thumb, which teachers and writers should keep in mind when using basic vocabulary in the practice of their art, is that the equivalent of about one running word in twenty should come from outside the controlled range. In a story 300 words in length, therefore, 10 or 15 words could be drawn from other sources, if the rest were basic vocabulary items (whether from one level or several). These 10 or 15 words do not all have to be different, of course – in fact, if they are likely to be new to the audience, it is a good idea to see that they occur several times in the course of the story or exercise; 15 running words could therefore mean only 3 or 4 different words. The important thing is to ensure that a high level of interest is maintained, and that learners have constant opportunities to acquire some of the thousands of "non-basic" words they will eventually need.

In this respect, it is interesting to note some of the facts about the frequency and distribution of particular kinds of words which frequency counts reveal. We have already seen that certain words, like **toothpaste**, which are extremely important in everyday life, fare very badly in most attempts to determine usefulness by counting. Another common phenomenon is the wide separation of members of sets of words – like numerals, the names of days of the week, and so on – in lists based on rank order.

For days of the week, only one – **Mane** *Monday* made the basic Maori word-list. None would have featured in a parallel list based on English or Spanish word-counts; in the American Heritage list, for example, the most important day of the week is **Saturday**, but it does not feature among the first 2,000 words in the usage list. In terms of gross frequency, the Maori words followed exactly the same order as their English counterparts, but **Mane**, fourth in frequency, was very much better distributed than any of the others, while **Hātarei** *Saturday* was very poorly distributed, and ended up fifth in value. Among the numerals, all the Maori numbers from **tahi/kotahi** *one* to **tekau** *ten* rated as basic words, but in a rather jumbled order, with **rua** *two* the most valuable and the most frequent, and **waru** *eight* at the bottom of the list. In terms

of general importance of numerals, and non-sequential ordering, the Maori list parallels that for the Spanish numerals, rather than the English list. (The different orderings discussed are illustrated in Table 5.)

Just what *use* teachers should make of such information is a moot point. It could be argued, however, that these sets of terms need not necessarily be taught all together – **Sunday** and **Monday** may be more important at the beginning than **Tuesday** and **Wednesday**, and, certainly, counting to four may be a more important exercise initially than counting to ten. In fact, most of the days of the week, and many numerals, will probably have to be brought in fairly early in the piece among the necessary "extras" we have referred to. Nevertheless, it may still be useful to remember that in Maori, English and Spanish, **āpōpō/tomorrow/mañana** is probably a lot more important than **Wenerei/Mīereoles/Wednesday**, even on Tuesday!

Table 5 Actual rank ordering for days of the week and numerals

	Days of the Week						
Term	"Sunday"	"Monday"	"Tuesday"	"Wednesday"	"Thursday"	"Friday"	"Saturday"
English Usage	2	4	5	6	7	3	1
Maori Frequency	2	4	5	6	7	3	1
Maori Value	2	1	6	4	7	3	5

	Numerals									
Term	"1"	"2"	"3"	"4"	"5"	"6"	"7"	"8"	"9"	"10"
English Usage	1	2	3	4	5	6	8	9	10	7
Spanish Usage	2	1	3	5	4	7	9	6	8	10
Maori Value	2	1	5	4	6	8	7	10	9	3

What Next

Already, the research on word value has been used as the basis for the selection of items for a picture vocabulary test of Maori word knowledge, for eventual use in bi - lingual schools. The data base for the Maori word list is being steadily expanded, and it is planned to produce a much larger and more comprehensive word list – drawn from a 300,000–500,000 word corpus – in two or three years. At the same time, work is proceeding on two children's dictionaries – one monolingual and one bilingual – with about 4,000 entries. The words in these dictionaries will be defined – i.e. explained fully, not simply paired with single word equivalents (as is the case with most currently available bilingual "dictionaries" – the one partial exception is Williams; no monolingual Maori dictionary has yet been published). It is also planned to include common idiomatic expressions in the new comprehensive word list. But these are all scheduled for **āpōpō** – not for next Wednesday.

References

Richard Benton, Hiria Tumoana, and Andrew Robb, *Ko Ngā Kupu Pū Noa o te Reo Maori: The First Basic Maori Word List,* Wellington: NZCER, 1982.

John B. Carroll, Peter Davies, and Barry Richman, *The American Heritage Word Frequency Book,* Boston: Houghton Mifflin, 1971.

Warwick Elley, Cedric Croft, and Colin Cowie, *A New Zealand Basic Word List,* Wellington: NZCER, 1977.

Edward Fry, 'The New Instant Word List', *The Reading Teacher,* Vol. 34, No. 3, 1980, pp 284–289.

Alphonse Juilland and E. Chang-Rodriguez, *Frequency Dictionary of Spanish Words,* The Hague: Mouton, 1964.

Jack C. Richards, *A Psycholinguistic Measure of Vocabulary Selection*, Quebec: ICRB, 1969.

Edward L. Thorndike and Irving Lorge, *The Teacher's Word Book of 30,000 Words,* New York: Teachers College Press, 1944.

J.A. van Ek, *The Threshold Level for Modern Language Learning in Schools*, London: Longman, 1976.

Michael West, *A General Service List of English Words*, London: Longman, 1962.

H.W. Williams, *A Dictionary of the Maori Language*, Wellington: Government Printer, 1971.

Books on Kindred Topics by NZCER

MAORI

Ko Nga Kupu Pū Noa o te Reo Maori
The First Basic Maori Word List
By Richard Benton, Hiria Tumoana, Andrew Robb (1982) 64 pages, card cover, saddle stapled, 21×17 cm. $4.50.
The 679 most valuable words in Maori arranged in five levels; over 100 cross-references to dialect variants, synonymns, etc.; a special section with all words grouped according to the notions they contain or functions they fulfil; a comprehensive index which doubles as a mini-dictionary and theasaurus.

He Pioke No Rangaunu
Exercises and Games for Practice in Maori
By Janet McCallum (1975), 64 pages, card cover, saddle stapled, 21×17 cm, $3.00.
A source of practical ideas for the teaching of Maori – language games and exercises are described with clear examples drawing on Te Rangatahi and other texts.

OCEANIC LANGUAGES

The Flight of the Amokura. Oceanic Languages and Formal Education in the South Pacific
By Richard Benton (1981) 236 pages, case bound, 22×14 cm, $18.00.
Many of the languages of Oceana are threatened with extinction. Official policies are examined and contemporary moves towards bilingual education are discussed. A 'runner-up' for the 1982 New Zealand Book Award.

ENGLISH

A New Zealand Basic Word List
By Warwick Elley, Colin Cowie and Cedric Croft (1977) 24 pages, card cover, saddle stapled, 21×17 cm, $1.50.
The 315 most basic words in the teaching of reading in New Zealand. These are arranged to be of most use to teachers: to help check the suitability of the vocabularies of books and other early reading material; to help make supplementary reading and remedial material.

Assessing the Difficulty of Reading Materials: The Noun Frequency Method
Popularly known as the 'Elley Noun Count' this method has proved as accurate as seemingly more sophisticated methods, and easier to apply. It gives teachers and writers a quick, simple, and accurate guide to the difficulty levels of books they may wish to give to children to read, and of material they are writing for children.

All these books may be obtained from leading bookshops, or directly from NZCER Book Sales Service, Box 3237, Wellington, (no charge for postage).