LANGUAGE DEVELOPMENT, LEXICAL COMPETENCE AND NUCLEAR VOCABULARY

Michael Stubbs

Kevin Durkin, ed (1986) *Language Development in the School Years*. Croom Helm.

A slightly revised version is published in: M Stubbs (1986) *Educational Linguistics*. Blackwell).

(This version was converted from an ancient computer file to an HTML file and further converted to this pdf-file. It therefore may not correspond exactly to the published versions.)

A few additional notes have been added to the text here, like this:

NOTE added February 2008. The term used in the title, "nuclear vocabulary", followed work by Dixon, Hale, and Stein and Quirk (which is referred to in the article), but has not been much used in later work. The term has however been retained here. The term "core vocabulary" is probably the most current nowadays.

This article uses the following conventions:

- italics for linguistic forms
- "double quotes" for meanings
- asterisk * for ill-formed strings

When people think of a language, they think almost inevitably of words: vocabulary. And when they think of language development, they also tend to think of vocabulary enlargement. There are obviously many other aspects of language development, and there is the danger that an attempt to 'increase someone's word power' leads to a quiz mentality. Nevertheless, the notion of extending someone's vocabulary is a perfectly plausible one in itself. It rests on a powerful, though sometimes hazy, intuition that some words are simpler, more important or more basic than others. It underlies the commonsense fury against much bureaucratic gobbledegook; and the often repeated observation that children's everyday vocabulary does not prepare them for reading the unfamiliar academic vocabulary in school textbooks (Perera, 1980).

This article sets out in detail several criteria for defining basic or nuclear vocabulary, and discusses some of the implications of the concept: for theoretical linguistic studies of lexis, for psycholinguistic studies of children's language development, and for practical educational concerns.

In some form the idea of basic vocabulary must underlie all vocabulary teaching. It certainly underlies vocabulary lists of various kinds including: Ogden's (1930) *Basic English*; Thorndike's (1921) and Thorndike and Lorge's (1944) *Teacher's Wordbook*; West's (1953) *General Service List*; Kucera and Francis' (1967) computational analysis of American English; Carroll et al's (1971) American frequency list; Hornby's (1974) *Advanced Learner's Dictionary*; Hindmarsh's (1980) *English Lexicon*; the Keyword scheme in Ladybird readers; and, in fact, lists of vocabulary in any language textbook. Historically, the distinction between basic and non-basic expressions can be traced back to seventeenth century speculations on the possibility of a logical universal language. This work exerted a powerful influence on Roget's (1852) attempt at a Thesaurus which logically classifies the whole vocabulary of English. It also influenced Ogden's (1930) *Basic English*, intended as an international auxiliary language. (Lyons, 1981: 64.)

Such lists have very different purposes, including: teaching English as a foreign language to different groups; facilitating international communication, given the position of English as a world language; and making prescriptions about the educational level expected of native English-speaking schoolchildren of different ages. Underlying some such lists is therefore a concept of the 'usefulness' or 'communicative adequacy' of different words. A clear statement of the fundamental intuitive notion involved is by Jeffery in the Foreword to West (1953: v):

> A language is so complex that selection from it is always one of the first and most difficult problems of anyone who wishes to teach it systematically. ... To find the minimum number of words that could operate together in constructions capable of entering into the greatest variety of contexts has therefore been the chief aim of those trying to simplify English for the learner.

The widespread use of such a large number of lists in teaching of different kinds illustrates how important vocabulary development is felt to be, sometimes as an end in itself, and sometimes as a way of facilitating cognitive development.

There remain problems, however. For example, later lists have generally been constructed on the basis of earlier lists, which have themselves built-in biases in their sampling. The Thorndike list, often used by later scholars, was based on a corpus of 4.5 million words, of which 3 million are from 'the Bible and the English classics', including Boswell's *Life of Johnson* and Gibbon's *Rise and Fall of the Roman Empire*. Earlier work is of course generally reinterpreted, but via a 'teacher's discretion' (Hindmarsh, 1980: ix); and some lists (eg Van Ek and Alexander, 1977) are set up with no indication at all of how they were constructed.

Frequency counts are obviously inadequate on their own, although basic frequency data cannot be entirely ignored. And totally inexplicit use of intuition is also inadequate: apart from any other reasons, intuitions about lexical frequency are often wildly inaccurate. This article therefore aims to provide a more precise concept of what might be meant by 'basic' versus 'non-basic' vocabulary, by returning to first principles and using published lists only at a later stage. I do not aim to provide a review of empirical research on vocabulary development, though some work is referred to. The aim is rather to discuss the systematic linguistic basis for a distinction which has far-reaching implications for linguists, child language researchers and teachers.

Words are idiosyncratic

It is regularly pointed out that words are idiosyncratic. Every individual word is unique in its etymology, and in its meaning and behaviour, including its collocations. Furthermore any individual speaker's vocabulary is unique: an idiosyncratic network of personal connections which do not appear to concern linguistic competence as this is usually understood.

Phonological and grammatical competence are essentially different from lexical competence in this respect. Any adult native speaker of any dialect of English (or any other language) has basically the same phonological competence, involving intuitive knowledge of the phonemes of the language, their allophonic variants, their possible phonotactic constraints, and so on. This competence is acquired by the age of around seven years: after that there is simply no more to learn. The same is true of much of the grammar of the language: in most of its main features this is learned by the age of five or six years, though some of the more complex syntactic structures may be learned later and some stylistically formal syntactic structures, largely restricted to written language, may be learned in adulthood, if at all. Lexical competence simply never approaches this kind of completeness. The learning of new vocabulary is clearly very rapid in early childhood, and then slows down. But a person's vocabulary may nevertheless keep growing throughout their whole life. New meanings can be learned for old words, and new relations between words can be formed.

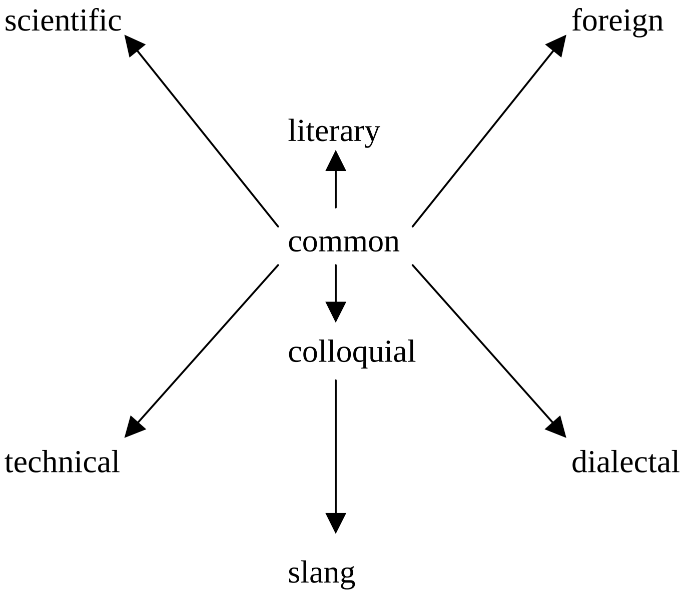
Relational lexical semantics

Despite this apparent inherently idiosyncratic aspect of lexical competence, there are, of course, systematic ways of studying vocabulary. One set of approaches could be called relational lexical semantics, and comprises: semantic field theory (expecially Roget, 1852, and Trier, 1931; but also other work by Humboldt in the 1800s, and by Meyer and Weisgerber between 1900 and 1930); structural semantics (Lyons, 1963, 1968); and componential analysis (Nida, 1975; Lehrer, 1974). The basic concept is that meaning is a relational property of language systems: words have no absolute value or meaning, but are defined in relation to other words. The sense relations involved include synonymy, antonymy and hyponymy, and these can be given formal definitions in terms of logical entailment and contradiction. Such approaches are well known and well reviewed in many standard textbooks (see especially Lyons, 1977, vol l). I will therefore not discuss them here except in so far as they can help to support a rather different way of discussing relations between words: a distinction between nuclear and non-nuclear vocabulary. Nor will I explicitly discuss the question of how children acquire such semantic relations. There are detailed analyses of children's acquisition of the hierarchical organization of vocabulary, their initial overextensions and later narrowing of word meaning, and the structure of their concepts, by Clark (1973), Livingston (1982), Nelson (1982), Palermo (1982) and Rosch (1973).

The common core

An important part of native speakers' linguistic competence is the ability to recognize that some words are 'ordinary' English words, in some sense, whilst others are rare, exotic, foreign, specialist, regional and so on. Such intuitions are by no means always accurate: for example, regional words are often not recognized as such. As a speaker of standard Scottish English, I realized only recently that *skelf* ("splinter of wood stuck in a finger") is regionally restricted to Scotland and some other northern areas of Britain.

This intuitive notion that part of the vocabulary is more basic than the rest underlies the definition of the vocabulary of a language which is discussed in detail in the introduction to the Oxford English Dictionary. It is argued there that the vocabulary of English is 'not a fixed quantity circumscribed by definite limits', but rather a nebulous mass with 'a clear and unmistakeable nucleus (which) shades off on all sides ... to a marginal film that seems to end nowhere'. The introduction also provides a helpful diagram which neatly sums up this concept:



NOTE added December 2002. A very useful discussion (and gripping narrative) of the OED and its development was published in 1998 in a quasi-novelistic form, by Simon Winchester: *The Surgeon of Crowthorne: A Tale of Murder, Madness and the Oxford English Dictionary*. (Penguin.) The book is published in the USA under the title *The Professor and the Madman*.

Dialectal and diatypic variation

The concept of 'core' evident in the position adopted by the OED is, however, not quite the concept which we require here. Comprehensive dictionaries and grammars wish to define the whole of what is 'unquestionably' English. What we require is a considerably more restricted subset of this core. In addition, the 'core' in the sense already discussed occurs naturally as the intersection of many different varieties. We require also to build in the concept of a deliberate and planned selection within this core. Stein (1978) and Quirk (1981) call such a reduced and planned English 'nuclear English', with reference to the lexical and syntactic characteristics of a restricted variety of international English. Hale (1971) and Dixon (1971, 1973) also use the term 'nuclear' in a relevant sense.

Blum and Levenston (1978) point to a related aspect of lexical competence which is closer to our requirements. An important part of native speakers' linguistic competence is the ability to do with less than their full vocabulary when required to do so. Speakers have an intuitive sense of which words to avoid when, for example, talking to younger or older children or to foreigners (cf Snow and Ferguson, eds, 1977; Bohannon and Marquis, 1977); or, conversely, which words ought to be taught first to foreign learners or used in simplified reading books for children, and, in general, which words are of maximum utility (Rosch, 1975; Cruse, 1977; Blewitt, l983; Shipley et al, 1983). Speakers have many strategies for avoiding words if they require to. One strategy is to use a paraphrase or circumlocution: instead of waddle, they might talk of a clumsy walk, and such paraphrases are constructed in systematic ways (see below). However, such intuitions have limits, hence the debates over which words should be taught in foreign language textbooks, and hence the need for criteria which are not purely intuitive.

In order to develop this sense of nuclear vocabulary, I require to develop the concepts in the OED diagram cited above. It is usual to distinguish between: regional or geographical dialects (eg Scottish versus Anglo English); social dialects (eg working class versus middle class); temporal dialects (eg Old English versus Middle English); and individual dialects (usually called idiolects). There are exceptions, but many individual speakers have full native competence in only one dialect, defined geographically, socially and temporally, and fixed in adolescence. On the other hand, any individual uses many different diatypes, according to the field of discourse (the activity going on at the time), the tenor of discourse (the social relations between the speakers), and the mode of discourse (predominantly speech versus writing). My formulation here is a Hallidayan one (see Halliday, 1978; or Gregory and Carroll, 1978, for a very simple account).

There is no implication that dialects and diatypes are separate. There was obviously regional and social variation within Old English; and there is diatypic variation within all dialects. Moving to a formal social situation may involve dialect switching as well as diatype switching. And Standard English is an intersection of dialect and diatype. It is not a geographical dialect, since it is used everywhere: 'normal' dialects are geographically restricted. It is a social dialect, used predominantly by the educated middle classes, which has particular diatypic uses, for example, in education (field), in formal settings (tenor) and in writing (mode). (See Stubbs, 1983a: 32-37, for a more detailed definition.)

The essential idea, then, is that English vocabulary has a central area 'whose Anglicity is unquestioned', which contains a smaller, naturally occurring common core. Within this it is also possible to select, for some communicative or pedagogical purpose, a planned nuclear English. The wider and the more restricted foci have fuzzy boundaries and shade off imperceptibly into marginal and peripheral forms, including obsolete words (restricted to earlier temporal dialects), regional words (restricted to particular geographical dialects), rare, specialist, technical or foreign words (restricted to certain fields of discourse), colloquial or slang words (restricted to particular tenors of discourse) and literary words (restricted to an intersection of field and mode); and so on.

Here is an initial example, before more detailed definition. The word *child* and its plural *children* are both common core and nuclear, not restricted in dialectal or diatypic usage. But there are many related words which are restricted, for example: *childe* (archaic, "young man of noble birth"); *childer* (an archaic or regional plural); *kid* or *kiddy* (colloquial); *offspring* and *progeny* (formal or technical); *paedophilia* and *paediatrics* (technical); *babe* (archaic or religious or American colloquial for "young woman"); and so on.

The tests which follow are intended to make explicit our strong, if sometimes hazy, intuitions, that some words are more basic than others.

Nuclear vocabulary: definition and tests

First, nuclear vocabulary is pragmatically neutral, in the sense that it conveys no information about the situation of utterance. (By pragmatics, I mean the study of relations between language and its contexts of use.) The nuclear vocabulary can be used by anyone, to anyone, at any time, to speak or write about anything.

A second general observation is that nuclear words are known by all normal, adult native speakers. This is a first requirement, a *sine qua non*. No user of English knows its whole vocabulary. A large unabbreviated general dictionary, such as the OED contains half a million entries, many of them unknown to most speakers. This gives us, in effect, a rough distinction between everyday and specialist words, and therefore concerns field of discourse. Words are not nuclear because they are known to all speakers. They are known to all speakers because they are nuclear: because, for example, they are pragmatically neutral and occur in a wide range of contexts. More precisely, we have to say that nuclear words are known in a particular sense. For example, speakers may know the word *frog* in its everyday sense of "small reptile", but few will know its specialist sense of "recess in a brick to save weight".

For ease of discussion, I will generally refer below simply to words, but what is really at issue is nuclear lexemes. A more detailed discussion would distinguish systematically between: word forms and lexemes; words and lexical items (phrasal verbs are a major complication here); and between different senses of homonyms. Different meanings of word forms will be left almost entirely out of account (except in test 10 below). This last point is a serious lacuna, since it begs the question of what it means to 'know' a word. Probably most words are known by most people in only some of their meanings. And are we talking about active use or passive recognition? These points also have important developmental implications. Nevertheless, they will have to be left for a more detailed discussion elsewhere, and I will assume here that it is possible simply to recognize the central or 'normal' meaning of a word.

Here then is a series of tests which elaborate these general points.

Pragmatic neutrality of nuclear vocabulary

1. Nuclear words have a purely conceptual, cognitive, logical or propositional meaning, with no necessary attitudinal, emotional or evaluative connotations. For example, to call someone *thin* could be good or bad. Consider:

   (1) She is lovely and thin. She is horribly thin.

   On the other hand, part of the meaning of *svelte* is "elegant" and the word implies a positive value judgement. This test is also an indication that nuclear words are less specialized in meaning and that they can occur in a wider range of contexts and collocations (cf test 9). This does not deny that words may have idiosyncratic connotations for individual speakers, and that they may be used with such connotations in context. However, they may be used without such connotations, and therefore be pragmatically neutral: they need not convey any information about the speaker's attitude to the referent.

2. Nuclear words are culture-free. This criterion is a development of points made above about the geographical neutrality of nuclear vocabulary. In any language variety, it is lexis which reflects culture, whereas phonology and grammar do not. For obvious reasons, languages have specialized vocabularies for local flora and fauna, and the like. Again for obvious reasons, when words are borrowed from one language into another, it is very often words which relate to new cultural artefacts, trading products, religious, cultural and artistic customs: consider the French words in English which have to do with cuisine, and the Italian musical terms. On the other hand, it is rare, but not unknown, to borrow words for the universals of human experience, including: basic bodily and biological functions, natural physical phemonena, dimensions of size and shape, words for pronouns.

   Arguably, words such as *sleep, eat, sun, earth, big, round* are culture-free in the sense intended. However, an attempt to set up a variety of a language which is 'as culture-free as calculus, with no literary, aesthetic or emotional aspirations' (Quirk, 1981: 43) may be exaggerated if carried too far. The criterion probably has to be relaxed to admit words which are culture-free relative to some geographical or cultural area (such as Western European or Anglo-American). This would admit such words as *aeroplane, upstairs, shop, school,* even though there are obviously many areas of the world which have no need of such words in everyday life.

   Dixon (1973) points out that nuclear verbs such as give have no cultural associations, and are typically easy to translate between languages. Non-nuclear *donate* and *award* have complex selectional restrictions which depend on cultural institutions. For example, one can *donate* only to a deserving cause and with no expectation of anything in return. Such non-nuclear verbs are typically difficult or impossible to translate directly.

3. Nuclear words are also pragmatically neutral in that they give no indication of the field of discourse from which a text is taken. For example, if we come across the words *port* and *starboard*, we know that the general context has something to do with ships or aircraft: the words *left* and *right* have the same logical meaning, but are not restricted in this way at all. The most obvious distinction here is between specialist and everyday terms. Thus for parts of the body, we find pairs such as

   (2) brain, cerebellum; shin bone, tibia; skin, epidermis; stomach, abdomen; teeth, dentition.

   Admitting that the technical term is often more precise in meaning, and that there are rarely if ever true total synonyms, both members of each pair convey the same logical meaning: they differ in the additional meaning they convey about the social setting of the language used.

4. Nuclear words are also neutral with respect to tenor of discourse: they are not restricted either to formal, or to casual or slang usage. This implies that nuclear words are also neutral with respect to mode of discourse: since written language is on average more formal than spoken. For example, alongside nuclear *help*, we have colloquial *give a hand*, and more formal *come to the aid of* and *render assistance*. Alongside *drunk*, we have formal *intoxicated* and *inebriated*, and a very large number of colloquial words, including *pissed, smashed, sozzled*. The last is also non-nuclear on the grounds that it is out-of-date: that is, it belongs to an earlier temporal dialect. Taboo subjects such as death and insanity attract a very large number of approximate synomyms. Thus alongside nuclear *mad*, we have formal *insane*, and many colloquial words: *crackers, nuts, loony*, and so on. *Mad* also has much wider meanings (cf also test 10).

5. Nuclear words are used in preference to non-nuclear words in summarizing original texts. This is a statement about the use of vocabulary for different purposes. For example, I performed the following experiment (reported fully in Stubbs, 1983b, chapter 10). I gave copies of Hemingway's short story *Cat in the Rain* to a hundred people, and asked them to summarize the story in their own words. A cat is an important character in the story and different words for "cat" appear with the following frequencies: *cat*, 13; *kitty*, 6; *tortoise-shell*, 1; *gatto*, 1 (Italian for "cat": the story takes place in Italy). Despite the fact that *kitty* is common in the original story, and that the story lays considerable stress on the fact that it is a small cat which a woman wants to hold and stroke, informants overwhelmingly preferred the word *cat* in their summaries. Nor did they introduce other non-nuclear words such as *kitten, pussy, moggy, feline*. This characteristic of nuclear words presumably reflects the fact that speakers intend summaries to represent propositional content, but not the style and attitudes of the original author (cf test 1).

Syntactic and semantic relations between nuclear words

Tests 1 to 5 have to do with the relation between words and social context. The next series of tests, 6 to 11, involve syntactic and semantic relations between words: the essential notion underlying them is that nuclear words are generic rather than specific. Note therefore that these two series of tests point to two rather different senses in which vocabulary may be 'basic' or 'simple'. The pragmatic neutrality tests above concern, roughly, the notion of everyday, non-technical words. The tests which follow concern the notion that words may be 'basic' in the sense that they could be used to define a greater proportion of the vocabulary, and could therefore be useful in constructing an elegant and systematic semantic description of a language (Lyons, 1981: 65). There is no logical reason why such generic terms should be everyday words: in fact, many are clearly not (eg mammal, substance, state, event). However, the extent to which the same words are defined by both series of tests is an empirical question (cf further below). Bearing in mind these points:

6. Nuclear words tend to be superordinate rather than hyponyms. Hyponymy or class inclusion is a basic sense relation. A *rose* is a kind of *flower*: if something is a *rose*, then this logically entails that it is a *flower*; but not all flowers are roses: the reverse entailment does not hold. The concept seems most obviously applicable to nouns which denote classes of objects, but it applies also to adjectives (*scarlet* is a hyponym of *red*), and to verbs. Consider the words *kill, execute*, *murder*, *assassinate*. If A *assassinates* B, then this entails that A *murders* B, and this entails that A *kills* B. But the reverse entailments do not hold. A might *kill* B by accident, and this does not count as *murder*. *Execute* is similarly more restricted in meaning than *kill*. This test is discussed by Mackey and Savard (1967).

7. Since nuclear words are generic, it follows that nuclear words can substitute for non-nuclear, but not vice-versa. The examples with *kill* above may be reconsidered from this point of view. Similarly, *give* can be substituted for any of the italicized verbs in the following examples:

   (3) I donated money to the hospital.
   (4) I awarded him the medal. (cf gave him it for services rendered.)
   (5) I lent him the car. (cf gave him it for a short period.)

   Conversely, the non-nuclear *donate, award, lend* cannot occur in sentences such as:

   (6) I gave him a book for Christmas.
   (7) I gave him a lift.

   The above examples are adapted from Dixon (1973), who argues further that nuclear verbs have all the syntactic and semantic properties of non-nuclear verbs, but not vice-versa. Consider:

   (8) I gave it to him. I gave it him. I gave him it. I donated it to him. *I donated it him. *I donated him it.

   Mackey and Savard (1967) propose that it is possible to calculate the replacement value of a word by using a dictionary of synonyms or a thesaurus. We would find, for example, that *seat* can replace more words than *chair*.

8. It follows from test 7 that nuclear words (which are known to everyone) are used to define non-nuclear words, but the reverse is difficult or impossible. The following types of definition are typical:

   non-nuclear verb = nuclear verb + adverb
   chuckle = laugh softly

   non-nuclear noun = adjective + nuclear noun
   drudgery = tedious work

   non-nuclear adj = adverb + nuclear adjective
   svelte = elegantly thin

   This is yet another way of saying that the meaning of nuclear words is more general and less specialized than non-nuclear words. Versions of this test are discussed by Dixon (1971), Hale (1971) and Carter (1982). Mackey and Savard (1967) propose further that the defining power of a word can be measured by calculating how often it is used in the definitions in a chosen dictionary. For example *young* would be useful in defining *calf, lamb, puppy* and many other words. Ogden's (1930) dictionary of *Basic English* is constructed in just such a way by using a self-imposed restricted vocabulary of 850 words to define, and therefore replace, other words. Less radically, the definitions in the *Longman's Dictionary of Contemporary English* are written in a 'controlled vocabulary of approximately 2,000 words'. A study such as Mackey and Savard propose could to some extent be circular, since the Longman editors selected their controlled vocabulary from published frequency and pedagogical lists. Nevertheless, even a study of the Longman dictionary would show what words it is possible to use for such a purpose: some selections would not have worked.

9. Words vary enormously with respect to the freedom with which they can combine syntagmatically with other words, and this provides another test. Nuclear words have a wide collocational range. Collocation refers to the relation between a word and its co-text. For example, *good* can collocate with almost any noun. In some contexts, it is a near synonym for *mild* (*good/mild weather*). *Mild* and *lukewarm* are almost exact conceptual synonyms, but they have very narrow and very different collocational possibilities:

   (9) mild weather; *mild liquid; ?mild reception; *lukewarm weather; lukewarm liquid; lukewarm reception.

   In the following examples, based on Carter (1982), a plus indicates a possible collocation, and a minus indicates an impossible collocation and therefore an ill-formed string:

	man	baby	belly	animal	lie	paycheque
fat	+	+	+	+	+	+
stout	+	-	+	-	-	-
obese	+	-	?	-	-	-

Fat is shown to be nuclear on the basis of its wider collocations. Rudzka et al (1981) give a large number of observations on such collocations for English.

This test is a consequence of the pragmatic neutrality criterion (no restriction on diatypic occurrence), and of the generic criterion (wide uses). It leads directly to the next test.

10. Since nuclear words are generic, it follows that they have the property of extension: the power to create new meanings (Mackey and Savard, 1967). It is commonly observed that everyday words have wide general meanings, and are consequently often more difficult to define than specialist words. A simple measure of extension is the number of dictionary entries which a word (lexeme) has for related, but different senses. This obviously depends on the unexplicated intuition of the lexicographer, but the figures are usually striking enough. The following figures are from the *Collins English Dictionary*, which groups together related senses of a lexeme irrespective of part of speech:

- run 83, sprint 3;
- walk 24, saunter 3, stroll 3;
- strong 20, potent 5, powerful 4;
- give 29, award 4, donate 1;
- fat 19, stout 5, obese 1;
- kill 19, murder 8, execute 8, assassinate 2;
- thin 9, slim 3, svelte 2, emaciate 1;
- house 28, mansion 5, villa 3, bungalow 2;
- father 14, paternal 3;
- child 9, kid 5, paediatrics 1.

The following words all have relatively high figures and are therefore candidates for the nuclear vocabulary:

- blind 31, block 39, key 31, pair 14, raise 34, stop 39, time 60.

11. A final measure of the nuclearity of a word is the number of compound lexical items it can help form. Again (as proposed by Mackey and Savard, 1967), this can be studied in published dictionaries. For example, Collins lists about 150 combinations starting with *well*, and 32 for *run* (eg *runabout, runner, run-of-the-mill* and phrasal verbs such as *run up* (*debts*)).

The structure of the nuclear vocabulary

A third and final characteristic of nuclear vocabulary is that it is not simply an unstructured list of words but a unified whole. This can probably best be tested by experimental methods. In general, the structure of semantic relations between words can be studied by word association tests. It is well known that, especially for common, unemotive words (cf test 1), people's responses to stimulus words are not original, but follow predictable paths. English has particularly high levels of associational stereotypy. (Meara, 1980, discusses such data in the context of foreign language vocabulary acquisition.) The following test is one reflex of this general claim.

12. Nuclear words have obvious antonyms. For example, in elicitation experiments, *good* will predicably elicit the antonym *bad*; *fat, thin*; *clean, dirty*; etc. On the other hand, responses will be much less predictable with *excellent, obese, spotless*. This criterion amounts to the claim that nuclear words are more tightly integrated into the structural organization of the vocabulary.

Tests 7 to 12 above show that the nuclear vocabulary is self-contained and communicatively adequate for some purposes in so far as nuclear words can substitute in different ways for non-nuclear words. These tests also show that the general structuralist notion of the vocabulary of a language as a single, integrated, coherent system, is not entirely adequate. The nuclear vocabulary is more tightly integrated than the rest. This is perhaps the main theoretical point of the argument of this article. Lyons' (1968, 1977) concept of sense relations has been criticized as applying only to the type of carefully chosen examples which he discusses, and not to the language as a whole. But this criticism can be turned on its head: taken together, sense relations define nuclear vocabulary. This point also has important psycholinguistic implications for the mental organization of the lexicon, and this could provide an interesting topic for research.

Nuclear words: other tests

There are other tests for nuclearity which I will mention much more briefly. For example, there is a broad split in English vocabulary between words of Germanic and Romance origin: this has many reflexes in field, tenor and mode of discourse. For well known historical reasons, much of the vocabulary of the law, religion and government is Romance. But the split is much more widespread than that, as is seen in pairs such as the following, with the nuclear Germanic word first in each case:

(11) strong, potent; mother, maternal; teach, instruct; sheep, mutton.

There is a related tendency for nuclear words to be simple rather than compound. Consider:

(12) thin, undersized; strong, powerful; work, drudgery.

Finally, for a well defined semantic field, Berlin and Kay (1969) have given a careful set of definitions for what they call basic colour terms, intended to be universal, though I will illustrate them here only from English. A basic colour term: (a) must be monolexemic (blue not bluish); (b) must not be a hyponym (*red* not *scarlet*); (c) must not be restricted to one class of objects (not *blond*); (d) must be psychologically salient and stable in meaning in all idiolects; (e) must have the same distribution as other basic terms (*reddish, greenish, *chartreusish*); (f) is suspect if it is also the name of an object (not *gold, rose, claret*); (g) is suspect if it is a recent loan (not *beige*). Several of these specific tests are obviously related to the more general tests I have given above.

Frequency, range and evenness of distribution

It may seem odd that I have not used frequency at all as a criterion of nuclearity. This is because raw frequency will clearly not do on its own, and it is best to discuss other criteria and then to interpret frequency in relation to these. First, frequency and related statistics are an empirical consequence of nuclearity, not a test for it, as such: though some frequency statistics can be used to identify nuclear words. Second, as Mackey and Savard (1967) have shown, indices of usefulness correlate only weakly with frequency. By usefulness they have in mind such indices as use in definitions (cf test 8), genericness (cf test 6), extension (cf test 10), and combination (cf test 11). Third, frequency counts go out of date rather quickly (some words are prone to fashion); and they can differ significantly for British and American English, and for adults' and children's language.

The best known word lists for pedagogic purposes are 'general' in the sense that they are not designed for any particular subject, topic, purpose or diatype. This reveals a serious limitation on such lists: for any purpose, students must know all of the first few hundred items on a general frequency list; but after that there seems little to choose between the next item and an item a few hundred or a thousand ranks down. The dilemma appears in a sharp form if one considers text coverage. The word *the* accounts for about 7 per cent of an average English text. The 100 most frequent words account for about 50 per cent of an average text. The 1,000 most frequent words account for about 70 to 75 per cent. The curve of text coverage is clearly flattening off quickly, and the next 1,000 and the 1,000 after that give little extra in terms of text coverage: around 7 and 3 per cent respectively. It is clear what is happening in general. A few words are very frequent; most words are relatively infrequent; and some words are vanishingly rare, and are unlikely to occur more than once or twice in a corpus of millions of words. Many basic lexical statistics have been calculated for corpora such as: the London-Lund corpus of about half a million words of spoken British English, and the Brown University and Lancaster-Oslo-Bergen corpora of about one million words of written English, American and British respectively.

Of the 100 most frequent words in English (as calculated, for example, for the Lund, Brown and LOB corpora: see Svartvik et al, 1982) most are grammatical words. The lexical words in the first 100 of the Lund corpus are: *know, got, see, now, just, mean, right, get, really, people, time, say, thing*. Presumably one would want to include all such words in a list of nuclear words. This provides in any case a way of including grammatical words, many of which do not get identified on the tests above. And presumably there would be little disagreement on the next 300 or 400 words. After that, however, raw frequency of occurrence is of limited direct interest.

There are, however, two related statistics which are very easy to calculate, especially with computational techniques. The first is range: the number of different texts in which a word occurs, if only once. The second is evenness of distribution: ie the fact that a word occurs with significant and relatively even frequency in a wide range of texts. In combination, these two calculations provide statistical measures of pragmatic neutrality, since they show whether a word is restricted to particular diatypic uses.

> NOTE added December 2002. This article pre-dates the first corpus-based dictionary, published by Cobuild in 1987, and it also predates the easy availability of word-frequency lists based on much larger corpora, and the easy availability of software which can check the range of occurrence of words across different texts and text-types, but the general points made still seem to hold.

It should be clear that I am not claiming that the set of nuclear words is entirely clearcut. A typical situation in linguistics is that a class of words (eg nouns or grammatical words) is defined by a series of tests: some words pass all or most of the tests and are the clear, central or focal members of the class. Other words are more or less central. (See Comrie, 1981: 100, for a sensible discussion of such multi-factor definitions which are stated in terms of prototypes, rather than in terms of necessary and sufficient conditions; and Rosch, 1975, for psycholinguistic discussion of the concept of lexical and conceptual prototypes.)

Some educational implications

This article has been concerned with some of the principles underlying the organization of the vocabulary. I have mentioned only in passing data on children's language development and pedagogical issues of how vocabulary should be taught. I have, however, discussed an issue which appears to give considerable problems to educationalists. It is obvious that the vocabulary of English is very large, and that selections have to be made from it for many educational purposes. And it is generally accepted that the vocabulary known by individual speakers is related to their educational skills: it is widely agreed (Jenkins and Dixon, 1983) that there is a significant correlation between vocabulary size and both reading comprehension and overall verbal intelligence (though there is no real agreement at all on whether vocabulary influences IQ and reading comprehension, or vice versa, or whether the relation is indirect). There are, however, major uncertainties about how or whether to try and teach vocabulary, and one major problem is: Where to start? This article has attempted to provide some principles which are directly relevant to this question.

I have discussed the question: how can we talk systematically about the dimensions of diversity along which lexical competence can develop, with or without instruction? Many ideas for teaching materials do, however, follow in fairly obvious ways from the criteria for nuclear vocabulary: the tests specify, in effect, dimensions along which vocabulary can be extended. As Meara (1980) points out, the type of argument I have put forward has to do with the management of learning, not with learning itself. The article defines a set of words which ought to be known already by native speakers, and suggests ways of structuring learning and teaching so that this vocabulary can be extended in principled ways.

In addition, these definitions can also be used to help assess the linguistic difficulty of texts for use in schools, or to help simplify existing texts for various purposes (eg language teaching, making bureaucratic documents more readable). There are many so-called readability formulae for calculating the difficulty of texts, and they generally operate on word and/or sentence length, variously calculated. Such formulae have their place, but they are open to well known problems (cf Perera, 1980), since ease of comprehension depends also on features of syntactic structure, discourse organization and subject matter. But familiarity of the vocabulary is also a major factor. Although there are legitimate purposes for simplifying texts by controlling their vocabulary, I am not, of course, recommending that textbooks ought to be written in nuclear English: it seems best to state that explicitly.

Directions for research on language development

Despite hundreds of years of interest in basic vocabulary and many relevant studies of child language in recent years, there is still very little research concerned directly with the developmental and educational implications of nuclear vocabulary. In particular, there is a lack of research which is based directly on speakers' actual usage of lexical items in conversations with children. I will therefore conclude with some specific suggestions for textual and observational research.

First, the definitions of nuclear vocabulary which I have proposed require to be developed. The following steps define, in themselves, a substantial research project. A candidate list of nuclear words (lexemes) can be provisionally established by including (a) the 500 or so most frequent words in English, and (b) words in a chosen dictionary with a large number of distinct listed senses, say six or more (cf test 10 above), and/or a large number of listed combinations (cf test 11). Check all the words on this candidate list against all the tests above. This in itself will doubtless lead to a more precise formulation of some of the tests. Check if there are any words which are, on intuitive grounds, nuclear, but which have not been captured : eg check the next 5,000 words on frequency counts for English; check published lists of 'basic' vocabulary. It is intuitively plausible that there will be a correlation between the results of the various tests. Check if this is so. Rank order the words on the list according to how many tests they pass: ie from most to least nuclear. Collect experimental data on those tests where this is appropriate: eg on the antonymy test. Investigate more generally the word associations between words on the list. Check the list against a corpus which contains as wide a diatypic range as possible: for example, there is a prediction that the nuclear words occur in a wide range of texts at least once, and in addition are evenly distributed across different samples. This can be checked easily by computational methods. Take texts which are intended to be written in a reduced vocabulary: eg texts for beginning readers or for English as a foreign language. Calculate measures of richness of vocabulary. For example, in a type:token ratio of the form 1:n, one would expect n to be relatively high. Similarly, one would expect the number of hapaxes (words which occur only once) to be low. Assuming that such texts do, as predicted, have relatively 'poor' vocabulary, check whether they have correspondingly high percentages of nuclear vocabulary. Take published lists of 'basic' vocabulary: test their adequacy against the now considerably revised definitions of nuclear vocabulary.

A research programme along these lines and the resulting list and associated detailed specifications of the words would have many applications in studies of children's language development, in teaching English as a mother tongue and as a foreign language, in studies of readability, and in the design of dictionaries.

Given a well tested list of nuclear vocabulary of this kind, many developmental questions, such as the following, are then also open to investigation. It is plausible that children acquire nuclear words first and most rapidly: is this the case? Do adults use mainly nuclear words in talking to children? This would require a study of the spontaneous speech of parents and teachers to children of different ages in different situations. How and when do children acquire non-nuclear vocabulary? How much is acquired through reading? Is the acquisition of non-nuclear vocabulary related to other developmental variables? Is it, for example, a predictor of any other measures of educational success?

Finally, it is intuitively highly plausible that nuclear vocabulary is a universal: that is, for any language, native speakers will always feel that some words are more important and basic than others. Most of the tests proposed above are applicable to any language and comparative research is therefore a possibility.

ACKNOWLEDGEMENTS

I am grateful to: Ron Carter, for getting me interested in the topic of this article, and for several points on defining nuclear vocabulary (see Carter, l982); Kevin Durkin, Joanna Channell, Walter Grauberg and Gabi Keck for helpful discussion and comments on previous drafts.

REFERENCES

Berlin, B. & Kay, P. (1969) Basic Color Terms. University of California Press.
Blewitt, P. (1983) Dog versus collie: vocabulary in speech to young children. Developmental Psychology, 19, 4: 602-09.
Blum, S. & Levenston, E. A. (1978) Universals of lexical simplification. Language Learning, 28, 2: 399-415.
Bohannon, J. N. & Marquis, A. L. (1977) Children's control of adult speech. Child Development, 48: 1002-08.
Carroll, J. B., Davies, P. & Richman, B. (1971) The American Heritage Word Frequency Book. New York: Heritage Publishing Co.
Carter, R. (1982) A note on core vocabulary. Nottingham Linguistic Circular, 11, 2: 39-50.
Clark, E. V. (1973) What's in a word? In T. Moore, ed: 65-110.
Comrie, B. (1981) Language Universals and Linguistic Theory. Oxford: Blackwell.
Cruse, D. A. (1977) The pragmatics of lexical specificity. Journal of Linguistics, 13: 153-64.
Dixon, R. M. W. (1971) A method of semantic description. In Steinberg & Jakobovits, eds: 436-70.
Dixon, R. M. W. (1973) The semantics of giving. In H. Halle & M-P. Schutzenberger, eds, The Formal Analysis of Natural Languages. The Hague: Mouton. 205-23.
Gregory, M. & Carroll, S. (1978) Language and Situation. London: Routledge & Kegan Paul.
Hale, K. (1971) A note on the Walbiri tradition of antonymy. In Steinberg & Jakobovitz, eds: 472-83.
Halliday, M. A. K. (1978) Language as Social Semiotic. London: Edward Arnold.
Hindmarsh, R. (1980) Cambridge English Lexicon. Cambridge: Cambridge University Press.

Hornby, A. S. (1974) Oxford Advanced Learner's Dictionary of Current English. London: Oxford University Press.

Jenkins, J. R. & Dixon, R. (1983) Vocabulary learning. Contemporary Educational Psychology, 18: 237-60.

Kucera, H. & Francis, W. N. (1967) Computational Analysis of Present-day American English. Rhode Island: Brown University Press.

Kuczaj, S. A., ed, (1982) Language Development, 2 vols. Hillsdale, NJ: Lawrence Erlbaum.

Lehrer, A. (1974) Semantic Fields and Lexical Structure. London: North Holland.

Livingston, K. R. (1982) Beyond the definition given: on the growth of connotation. In Kuczaj, ed: 429-44.

Lyons, J. (1963) Structural Semantics. Oxford: Blackwell.

Lyons, J. (1968) Introduction to Theoretical Linguistics. London: Cambridge University Press.

Lyons, J. (1977) Semantics. Vols. 1 & 2. Cambridge: Cambridge University Press.

Lyons, J. (1981) Language, Meaning and Context. London: Fontana.

Mackey, W. & Savard, J-G (1967) The indices of coverage: a new dimension in lexicometrics. International Review of Applied Linguistics, 2-3: 71-121.

Meara, P. (1980) Vocabulary acquisition: a neglected aspect of language learning. Language Teaching and Linguistics Abstracts, 13, 4: 221-46.

Moore, T. ed, (1973) Cognitive Development and the Acquisition of Language. New York: Academic Press.

Nelson, K. (1982) The syntagmatics and paradigmatics of conceptual development. In Kuczaj, ed: 335-64.

Nida, E. (1975) Componential Analysis of Meaning. The Hague: Mouton.

Ogden, C. K. (1930) Basic English: A General Introduction with Rules and Grammar. London: Kegan Paul.

Palermo, D. S. (1982) Theoretical issues in semantic development. In Kuczaj, ed.

Perera, K. (1980) The assessment of linguistic difficulty in reading material. Educational Review, 32, 2: 151-161. Reprinted in R. Carter, ed. (1982) Linguistics and the Teacher. London: Routledge & Kegan Paul. 101-13.

Quirk, R. (1981) International communication and the concept of nuclear English. In Smith, ed, 1981: 151-65.

Richards, J. (1971) Coverage: what it is and what it isn't. ITL, 13: 1-15.

Roget, P. M. (1852) Thesaurus of English Words and Phrases.

Rosch, E. H. (1973) On the internal structure of perceptual and semantic categories. In Moore, ed: 111-44.

Rosch, E. H. (1975) Cognitive reference points. Cognitive Psychology, 7: 532-47.

Rudzka, B., Channell, J. & Putseys, Y. (1981) The Words You Need. London: Macmillan.

Shipley, E. F., Kuhn, I. F. & Madden, E. C. (1983) Mothers' use of superordinate category terms. Journal of Child Language, 10: 571-88.

Smith, L. E. ed. (1981) English for Cross-Cultural Communication. London: Macmillan.

Snow, C. E. & Ferguson, C. A. eds (1977) Talking to Children: Language Input and Acquisition. Cambridge: Cambridge University Press.

Steinberg, D. D. & Jakobovitz, L. A. eds (1971) Semantics . London: Cambridge University Press.

Stein, G. (1978) Nuclear English: reflections on the structure of its vocabulary. Poetica, 10: 64-76.

Stubbs, M. (1980) Language and Literacy: The Sociolinguistics of Reading and Writing. London: Routledge & Kegan Paul.

Stubbs, M. (1983a) Language, Schools and Clasrooms. 2nd ed. London: Methuen.

Stubbs, M. (1983b) Discourse Analysis: the Sociolinguistic Analysis of Natural Language. Oxford: Blackwell.

Svartvik, J., Eeg-Olofsson, M., Forsheden, O. Orestrom, B. & Thavenius, C. (1982) Survey of Spoken English: Report on Research 1975-81. Lund Studies in English, 63.

Thorndike, E. (1921) The Teacher's Wordbook. New York: Columbia Teachers College.

Thorndike, E. & Lorge, I. (1944) The Teacher's Wordbook of 30,000 Words. New York: Columbia Teachers College.

Trier, J. (1931) Der Deutsche Wortschatz im Sinnbezirk des Verstandes. Heidelberg: Winter.

Van Ek, J. A. & Alexander, L. G. (1977) Threshold Level English. Oxford: Pergamon.

West, M. (1953) A General Service List of English Words. London: Longman.