Word Families

Laurie Bauer and Paul Nation, *Victoria University of Wellington, New Zealand*

Abstract

The idea of a word family is important for a systematic approach to vocabulary teaching and for deciding the vocabulary load of texts. Inclusion of a related form of a word within a word family depends on criteria involving frequency, regularity, productivity and predictability. These criteria are applied to English affixes so that the inflectional affixes and the most useful derivational affixes are arranged into a graded set of seven levels. This set of levels and others like it have value in guiding teaching and learning, in standardising vocabulary load and vocabulary size research, in investigating lexical development and lexical storage, and in guiding dictionary making.

1. Introduction

Research on native speakers' vocabulary growth as it relates to reading has indicated the importance of morphological knowledge in determining what words will be known by developing readers (Graves 1987). Related research has also shown that this morphological knowledge develops over a considerable period of time. Nagy, Diakidoy and Anderson (1991) for example have shown that learners' ability to interpret the effect some derivational affixes have on a word is still developing well into the teenage years. There are suggestions that a useful focus for reading teachers would be to help learners increase and clarify their knowledge of the most useful affixes (Wysocki and Jenkins 1987).

Dealing with growth in morphological knowledge involves consideration of the idea of a "word family". From the point of view of reading, a word family consists of a base word and all its derived and inflected forms that can be understood by a learner without having to learn each form separately. So, *watch*, *watches*, *watched*, and *watching* may all be members of the same word family for a learner with a command of the inflectional suffixes of English. As a learner's knowledge of affixation develops, the size of the word family increases. The important principle behind the idea of a word family is that once the base word or even a derived word is known, the recognition of other members of the family requires little or no extra effort. Clearly, the meaning of the base in the derived word must be closely related to the meaning of the base when it stands alone or occurs in other derived forms, for example, *hard* and *hardly* would not be members of the same word family. Table 1 gives a brief overview of the significance of the various levels of inflection and affixation

 © 1993 Oxford University Press

Table 1. Additions to a word family at different levels of inflection and affixation

	Word families		
2	develop develops developed developing	wood wood's woods wooded	bright brighter brightest
3	developable undevelopable developer(s) undeveloped	woody woodiest woodier woodiness	brightly brightish brightness
4	development(s) developmental developmentally		
5	developmentwise semideveloped antidevelopment	wooden	brighten
6	redevelop predevelopment	anti-wood	

described in this article. At Level 1 every different form is counted as a different word, so *develop*, *develops*, etc are all different words at Level 1. At Level 2 items with the same base but different inflections are all counted as members of the same word family. So, in the case of the *develop* word family, what would be four words at Level 1 become 1 word family at Level 2. From Levels 2 to 6 the base form must be recognizable as a freely occurring word.

The following example shows how new members of a word family require little learning effort. Many readers will not be familiar with the lexeme *marmelize*, which is a British dialectal word, jocular in tone, for 'to beat up'. However, now they have met the word-form *marmelize* such readers will have no difficulty in understanding

She marmelized him

even though it contains the word-form *marmelized* which they have never met before. Similarly, they will easily understand

He fully deserved his marmelization

which contains a new word-form *marmelization*.

The aim of this article is to use the available findings on the productivity, frequency, regularity and predictability of English affixes to set up a series of levels of affixes that could provide the basis for the staged systematic teaching

and learning of these affixes for learners reading English. More optimistically, this series of levels could also be used to provide a consistent description of what should be considered to be part of a word family for readers at different levels of morphological awareness.

In setting up these levels priority is given to ease of learning for reading, and thus regularity of the word building process for the written form is most important. A series of levels could be designed for other purposes and this would result in a different ordering of word building devices.

Although this study focuses on the understanding of inflections and affixes in reading, it has significance for dictionary makers. The grouping of affixes into levels which is described later in this paper can provide a rationale for the treatment of affixes in dictionaries. Stein (1985) has pointed out the inconsistent treatment of affixation both between and within dictionaries. From one viewpoint, the levels can be viewed as an attempt to set up levels of word transparency. Moving along the levels will require more attention from the dictionary maker as irregularities of spoken form, written form, and meaning increase. The levels can be used as a guide to help answer questions like the following.

Are dictionary makers deciding in a principled and consistent way on what derived forms to include as full entries, defined sub-entries, and non-defined sub-entries, and what forms not to include?

Which prefixed derivatives should be included with their stems?

Let us now look at the criteria which are used to assign the affixes to the various levels.

2. Criteria

The criteria which are described here are not unique to this study. Thorndike (1941) in his classic study of English suffixes used many of them. Level ordering within Lexicalist Morphology (Siegel 1974; Allen 1978; Scalise 1984; Halle & Mohanan 1985; Mohanan 1986; Szpyra 1989; Bauer 1990; etc.) uses some of the same criteria, but in very different ways and with different aims. Level ordering attempts to divide affixes and other morphological processes into a small number of types based on their phonological and morphological behaviour.

Here are the eight criteria which were used to determine the level at which a particular affix should be placed.

i Frequency: The number of words in which an affix occurs. The affixes put in the early levels occur in a very large number of words. For example, *-er* is a frequent suffix, while *-ette* as in *usherette* is much less frequent. In the terminology of Bauer (1983, 1988) frequent affixes are highly generalized. This means that the ability to recognize them is going to be useful in any text.

ii Productivity: The likelihood that the affix will be used to form new words. The affixes put in the early levels are highly productive; in practical terms this means that words containing these affixes are less likely to be listed as headwords in the alphabetical order used in a particular dictionary. Correspondingly, readers cannot guarantee to find such words easily in dictionaries (if at all), and need to be able to recognize them as related to their bases. For example, *-ly* and *-ness* are frequently used to make words that have not been met before. Affixes like *en-* and *-most* are much less productive.

iii Predictability: The degree of predictability of the meaning of the affix. The meaning of most of the affixes at the early levels is predictable. *-less* for example has only two meanings, one of which is rare. For others, the meaning is predictable once the part of speech of the base to which it is added is known. For example, the meaning of an *-s* suffix is not known until you know whether the word you are adding it to is a noun or verb.

iv Regularity of the written form of the base: The predictability of change in the written form of the base when the affix is added. At the early levels, removal of the affix will leave the base orthographically intact (and hence recognizable), for example *green + ish*. At later levels, the removal of the affix leaves a form that is not exactly the same as the free form of the base, for example *sacrilegious*.

v Regularity of the spoken form of the base: The amount of change in the spoken form of the base when the affix is added. At the early levels, removal of the affix will leave the base phonologically intact and recognizable. This means that at the earlier levels the base is a potentially free form. Some words from the later levels in which the base in its spoken form is not a potentially free form include *permeable*, *dramatize*, *syllabify*.

vi Regularity of the spelling of the affix: The predictability of written forms of the affix. At the early levels, the affix has a fairly constant orthographic form, and is thus easily recognizable. The prefix *pre-* only has one written form, while *in-* ('negative') has *in-*, *im-*, *il-* and *ir-*. These are predictable but do not at first sight appear to be related.

vii Regularity of the spoken form of the affix: The predictability of spoken forms of the affix. At the early levels, the affix has a predictable phonological form, and is thus relatively easily recognizable. The *-s* and *-ed* inflections each have three spoken forms though these are predictable.

viii Regularity of function: The degree to which the affix attaches to a base of known form-class and produces a word of known form-class. For example, *-ess* always attaches to nouns and always produces nouns.

These criteria act in two ways. They determine the level at which an affix is placed, and they also place restrictions on what particular words can be included as part of a word family at a given level. For example, *-ism* is placed

at Level 4, but the word *Fascism* cannot be included in a word family with *Fascist* at this level because removing the suffix leaves a bound form.

It is clear from the criteria that learners must draw on different types and levels of knowledge in order to use the relationships between words in a word family.

1. Learners need to know word bases.
2. Learners need to be able to recognize known bases in words. To put it another way, learners need to be able to recognize that two words are related to each other because they share a common base, for example seeing the relationship between *happy* and *happiness*. Tyler and Nagy (1989) call this *relational* knowledge. This knowledge also involves not finding bases in words that do not contain that base even though the orthographic string for the base occurs, for example *me* in *mean*.

Several studies show that native speakers of English have a well developed relational knowledge of familiar bases by the time they are about 10 years old (4th grade) (Tyler and Nagy 1989; Nagy, Diakidoy and Anderson 1991).

3. Learners need to have at least implicit knowledge of the contribution that affixes make to a word. In the case of most prefixes and some suffixes this can involve knowing the lexical meaning of the affix, for example that *-less* attached to nouns means 'without'. For suffixes, the learners need to know the syntactic information carried by the suffix, for example that *-less* makes the derived word an adjective.

Native speakers' knowledge of the syntactic role of suffixes continues to develop through high school and even at that level some learners experience problems.

4. To use affixes productively in speaking and writing, learners need to be able to produce allowable base-affix combinations.

Let us now look at the levels.

3. Levels

The following levels have been set up for practical reasons and have no theoretical value. They have been designed with a focus on recognition of written words. The divisions between the levels are arbitrary and are merely intended to give easily identified steps along a cline. Ideally, others repeating our classification should end up with the same levels, and so be able to apply them to other affixes or processes not mentioned in this article and to other languages. With the exception of Level 2, data about the number of types and tokens covered by each affix and each level was gained from an analysis of the 1,000,000 token Lancaster-Oslo-Bergen (LOB) corpus. This data was not used in the selection of affixes for the levels – that was done by applying the eight criteria described above. The data from LOB was used to see the cumulative coverage each level provided.

The levels only deal with affixation. After Level 2 which contains inflectional suffixes, it would be possible to add a level dealing with transparent compound words with known parts. Bauer (1978, 1979a, 1983) looks at compounds.

Level 1: Each form is a different word

At this level it is assumed that learners will not recognize that *book* and *books* are members of the same word family. This is a very pessimistic assumption and it is unlikely that with regard to written forms, learners are ever at this level. From this point of view, this level is insignificant. From another point of view however it may be too large a step, simply because one form may have quite different meanings (*bear* = 'carry' or *bear* = 'ursine mammal'), and different functions (*to faint* and *to become faint*). For more details see Marchand (1969) and Bauer (1983). In addition there is lack of morphological marking to be considered where an item has a zero ending (*cut* as in *to cut*, and *cut* as in *to be cut* and *a cut*). Learners need to be aware of these difficulties and they must be taken into account in any estimation of vocabulary load for reading.

Level 2: Inflectional suffixes

At this level, words with the same base and inflections are considered as members of the same word family. The inflectional categories used here are – plural; third person singular present tense; past tense; past participle; *-ing*; comparative; superlative; possessive. See Appendix 2 for a discussion of the difficulties in deciding on the set of inflectional suffixes.

The number of types and tokens covered by the affixes at this level was largely calculated from data supplied in Kučera (1982) reporting on the lemmatization of the Brown corpus. The lemmatization of the Brown corpus did not include *-er* and *-est* as members of each lemma, so figures for these were added separately. Level 2 involves 24,188 types and 378,519 tokens in a 1,000,000 word corpus, and thus groups the 61,805 types into 37,617 word families – a drop of almost 40%.

Readers have to be able to undertake minimal morphographemic analysis in order to recognize regular inflections. A word-final ⟨e⟩ is regularly elided before an affix beginning with a vowel, as in *finer*, *finest*, *fining*. Depending on what the underlying form of the *-s* and *-ed* affixes is taken to be, the same might be true of *fined* and *fines*. Similarly, a word-final ⟨y⟩ becomes ⟨i⟩ before any affix that does not begin with ⟨i⟩, as in *jollier*, *jollies*, *jollied*, *jollying*. Letter doubling also applies before many affixes, as in *jotted*, *hotter*, *sinned*. These graphemic alternations do not have any phonological correlates.

Level 3: The most frequent and regular derivational affixes

At this stage, the eight criteria outlined at the beginning of this article are applied to derivational morphology. All the criteria are applied quite strictly at this level, and the strictness with which they are applied is reduced at subsequent levels. Here, only orthographic alternations, such as ⟨y⟩ becomes ⟨i⟩, which are already required for Level 2 will be permitted. The affixes included at level 3 are *-able*, *-er*, *-ish*, *-less*, *-ly*, *-ness*, *-th*, *-y*, *non-*, *un-*. They are all applied with restricted uses which are detailed in Appendix 2.

Table 2 gives frequency information on the affixes at Level 3. The number in brackets after some affixes indicates that they occur with different uses at another level. So, *-able* also occurs with a different use at Level 6.

Table 2. The number of types and tokens of Level 3 affixes in a 1,000,000 token corpus

Affix	Number of types	Number of tokens
-able (6)	155	685
-er (2)	705	4,003
-ish	54	124
-less	124	408
-ly (5)	1,227	12,789
-ness	307	1,059
-th (6)	30	356
-y (6)	265	994
non-	98	171
un-	538	1,867
Total	3,503	22,456

The addition of affixes at Level 3 reduces the 37,617 word families at Level 2 to 34,114. Appendix 2 contains a discussion of each affix at this level. It is unlikely that a change in purpose for a set of levels of affixation, such as setting up levels for productive rather than receptive use, would result in much change at this level.

Level 4: Frequent, orthographically regular affixes

At this level, the eight criteria are prioritized. In particular, the fact that an affix is frequent (widely generalized) is taken to be more important than whether it is productive or not, and orthographic criteria are taken to be more important than phonological criteria. This decision is based on the assumption that the processes recommended here are aimed at allowing comprehension of written rather than spoken forms.

One result of this prioritizing is that the levels used within Lexicalist Morphology referred to above and the ones used here fail entirely to coincide at this level, since many of the criteria for Level 1 affixes in such theories are phonological (stress shift, morphophonemic variation).

The following affixes are included at this level – *-al*, *-ation*, *-ess*, *-ful*, *-ism*, *-ist*, *-ity*, *-ize*, *-ment*, *-ous*, *in-*, all with restricted uses. See Appendix 2 for a discussion of each of the affixes.

However, since the degree of generalization of an affix is an important criterion at this level, it might be that some other affixes should be added to this level for people who are reading in specific scientific fields such as chemistry or geophysics, for example *a-*, *hypo-*, *epi-*. We restrict ourselves here to what happens in fairly general English.

Table 3 provides frequency data about the affixes at this level based on a

Table 3. The number of types and tokens of Level 4 affixes in a 1,000,000 token corpus

Affix	Number of types	Number of tokens
-al (5)	213	3,064
-ation	451	4,463
-ess	26	174
-ful	84	951
-ism	105	312
-ist (6)	143	697
-ity	243	1,788
-ize	69	158
-ment	129	2,678
-ous	121	948
in-	180	610
Total	1,764	15,843

study of the LOB corpus. The addition of the affixes at Level 4 to those of all the preceding levels results in a total of 32,350 word families in the 1,000,000 token corpus.

Level 5: Regular but infrequent affixes

This level adds a number of affixes whose behaviour is fairly regular, which may be productive, but which, because they are not widely generalized, do not individually add greatly to the number of words that can be understood. These affixes will simply be listed below. They are added to free bases, e.g. *contributory*, at this level and no new principles are involved. At Level 7 they are added to bound bases, e.g. *introductory*.

The following list of Level 5 affixes contains frequency figures in square brackets for types and then tokens based on occurrences in the 1,000,000 word LOB corpus. So, *-age* occurred in 42 different words with a total frequency of 322 tokens.

-age (*leakage*) [42:322], -al (*arrival*) [30:334], (*idiotically*) [59:152], -an (*American*) [666:1,116], -ance (*clearance*) [43:779], -ant (*consultant*) [17:145], -ary (*revolutionary*) [30:234], -atory (*confirmatory*) [11:23], -dom (*kingdom*; *officialdom*) [10:328], -eer (*black marketeer*) [5:49], -en (*wooden*) [9:20], -en (*widen*) [25:48], -ence (*emergence*) [33:773], -ent (*absorbent*) [21:663], -ery (*bakery*; *trickery*) [16:88], -ese (*Japanese*; *officialese*) [9:120], -esque (*picturesque*) [4:11], -ette (*usherette*; *roomette*) [3:56], -hood (*childhood*) [15:72], -i (*Israeli*) [3:5], -ian (*phonetician*; *Johnsonian*) [72:271], -ite (*Paisleyite*; also chemical meaning) [9:25], -let (*coverlet*) [4:18], -ling (*duckling*) [8:10], -ly (*leisurely*) [56:502], -most (*topmost*) [5:16], -ory (*contradictory*) [18:93], -ship (*studentship*) [26:302], -ward (*homeward*) [25:418], -ways (*crossways*) [3:14], -wise (*endwise*; *discussion-wise*) [5:24], ante- (*anteroom*) [5:6], anti- (*anti-inflation*) [52:71], arch- (*archbishop*) [6:30], bi- (*biplane*) [5:19], circum- (*circumnavigate*) [2:2], counter- (*counter-*

attack) [28:39], en- (*encage*; *enslave*) [90:643], ex- (*ex-president*) [32:41], fore- (*forename*) [18:38], hyper- (*hyperactive*) [3:4], inter- (*inter-African*, *interweave*) [42:173], mid- (*mid-week*) [32:74], mis- (*misfit*) [36:63], neo- (*neo-colonialism*) [3:3], post- (*post-date*) [17:37], pro- (*pro-British*) [5:9], semi- (*semi-automatic*) [34:63], sub- (*subclassify*; *subnormal*) [23:117], un- (*untie*; *unburden*) [47:93]. More details on these can be found in Marchand (1969). In total the 50 affixes occurred in 1,762 different types which were represented by 8,556 tokens. The addition of the affixes at Level 5 to those of all the preceding levels results in a total number of 30,588 word families in the 1,000,000 token corpus.

Level 6: Frequent but irregular affixes

This level includes those affixes which provide major problems of segmentation, either because they cause gross (orthographic) allomorphy in their bases (that is, parts of the base are deleted or additions besides the suffix are needed), or because there are major problems involved in segmenting them caused by homography. Because of the problems these affixes cause, only the very widely generalized ones are worth dealing with: less widely generalized affixes that cause similar problems are better treated by learning the common words which contain them. Some of the affixes dealt with here have already been mentioned. In such cases it is assumed that the transparent cases will have been dealt with at an earlier level, and the more opaque cases are dealt with here. Each affix and its problems are discussed in Appendix 2. Although the problems are dealt with in terms of particular affixes, note that many of the problems recur, and some of them recur with the less widely generalized affixes already mentioned. The affixes are *-able*, *-ee*, *-ic*, *-ify*, *-ion*, *-ist*, *-ition*, *-ive*, *-th*, *-y*, *pre-*, *re-*.

Table 4 gives the frequency figures for the affixes at this level.

Table 4. The number of types and tokens of Level 6 affixes in a 1,000,000 token corpus

Affix	Number of types	Number of tokens
-able (3)	22	53
-ee	25	136
-ic	221	1,086
-ify	14	134
-ion	475	6,748
-ist (4)	10	48
-ition	25	459
-ive	177	1,633
-th (3)	12	511
-y (3)	94	457
pre-	70	135
re-	225	881
Total	1,370	12,281

Including Level 6 affixes results in the number of word families dropping from 30,588 at Level 5 to 29,218 at Level 6 – a change of 4.5% between these levels, and a 2.2% change compared to Level 1.

Level 7: Classical roots and affixes

At this level belong all the classical roots which abound in English words and which occur not only as bound roots in English (as in *embolism*) but also as elements in neo-classical compounds (such as *photography*). Native speakers as well as L2 speakers have to be taught these explicitly, and there are books available which deal with these matters (Brown 1971), and look at the patterns of classical allomorphy and at correspondences such as those resulting from Grimm's Law. While there are undoubtedly gains in comprehension to be made in this area, it is outside the scope of this paper to discuss them.

What are commonly called combining forms e.g. *Franco-*, *Gallo-*, *Euro(po)-*, *agri-* are also included at this level.

It should be noted that this level includes many of the very frequent English prefixes (Bock 1948; Stauffer 1942), such as *ab-*, *ad-*, *com-*, *de-*, *dis-*, *ex-*, and *sub-*.

Further affixes

Not all the affixes of English have been mentioned above. In some cases the omission is deliberate, such as the omission of *cis-* because it appears in so few words (e.g. *cisalpine*) that it is scarcely worth learning. In other cases like the cases of what Marchand (1969: 356) calls 'semi-suffixes', the formatives can almost be taken to be compound elements, and *child-like* can be interpreted as a compound. Though productive, these are not particularly highly generalized. Generally, a greater proportion of suffixes than prefixes have been mentioned. If a prefix like *pseudo-* is found to be useful in a particular field of study, it can easily be slotted in to the system. A prefix like *dis-* is much harder to deal with because of its semantic range, but can be put in Level 5 or 6, depending on how useful it is found to be.

Table 5 summarizes the frequency data for the various levels.

Table 5. The number of types and tokens at six levels of affixation in a 1,000,000 token corpus

Level	Number of affixes	Number of types	Number of tokens	Number of families
1	0	61,805	1,000,000	61,805
2	8	24,188	378,519	37,617
3	10	3,503	22,456	34,114
4	11	1,764	15,843	32,350
5	50	1,762	8,556	30,588
6	12	1,370	12,281	29,218
Total	Levels 2–6	32,587	437,655	

Table 5 shows how the number of word families reduces as the affixes of each level are included in the definition of what makes up a word family. The number of word families is found by subtracting the number of types at each level from the number of families at the previous level. Thus, the figures for the number of families are probably slightly lower than they should be as they assume that the base form of each affixed item occurs in the corpus or that a family member occurs at a previous level. The greatest change comes at Level 2 with the inclusion of forms with an inflectional suffix in a word family. The next greatest change is at Level 3 with the most regular, frequent and productive of the derivational affixes. From Level 4 on, the irregularity and low frequency of the items make the return for learning and teaching effort much less.

This study has looked at only part of the knowledge that is relevant to the setting up of a description of a series of levels of what to include in a word family. It has focused on the language system aspects, making sure that the various levels meet criteria of frequency, productivity, predictability and regularity.

It is also possible to focus on the psychological and pedagogical aspects of levels of affixation. The psychological aspect looks at the learner and would deal with questions like these.

Does a native-speaker's vocabulary grow according to the content and order of the levels?

Does the storage and retrieval of derived vocabulary fit with the levels?

Evidence for the psychological reality of the levels might come from first and second language acquisition research (Gordon 1989; Wysocki and Jenkins 1987), and from the study of retrieval tasks (Cole, Beauvillain and Segui 1989; Nagy, Anderson, Schommer, Scott and Stallman 1989).

The pedagogical aspect looks at the value of a series of levels for teaching. Does learning best occur when the content and order of the levels are followed?

A well-developed series of levels would satisfy the demands of the language system, psychological and pedagogical aspects. This paper should be seen as an attempt to deal with one of these aspects.

4. Applications

This section describes a range of applications for the system of levels. These applications will undoubtedly lead to revision of the levels.

4.1 *Setting goals and stages for vocabulary teaching*

The vocabulary learning task facing learners of English as a second language is large. However, particularly for receptive uses of the language, a large amount of vocabulary learning can occur through control of the systematic word-building and affixation features of English. The levels described in this article provide the basis for a graded syllabus to help learners gain control of affixation. The figures showing the number of types and tokens in the million-

word corpus give guidance on how much time and effort should go into each level. Nagy, Diakidoy and Anderson's (1991) research indicates that native speakers of English could also benefit from attention to the later levels.

Decisions about what is included in a word family have significant effects on teaching and learning. If only a few inflected and derived forms are included in a word family, there will be a lot more word families to be taught and learned than if many inflected and derived forms are included. Teaching *govern* and then briefly pointing out the possibilities of *governor*, *government*, and *ungovernable* is much less work than teaching each of these as a separate item. Some of the questions that have been asked relating to teaching and learning include

What are feasible goals for a second language vocabulary development programme?

What are the contributions to vocabulary growth of the various sources of vocabulary learning, such as inferring from context, morphological generalization, and direct teaching (Wysocki and Jenkins 1987; White, Power and White 1989)?

How much attention should a teacher give to the word building systems such as affixation and compounding?

Is a particular reading text suitable for a particular learner?

The set of levels described in this article is a step towards answering these questions.

A wide variety of options is available for helping learners with affixes. These include direct formal analysis of isolated words, text based activities, and indirect learning through gradual introduction in simplified reading texts.

4.2 *Standardization for vocabulary load and vocabulary size research*

Over the last century there has been a continuing interest in vocabulary size and growth. One of the earliest studies to gain academic publication was Kirkpatrick's (1891) study of his own vocabulary. The motivations for this interest have been varied, ranging from the perception of vocabulary size as an indicator of intelligence to vocabulary as a major factor affecting reading.

In spite of this continuing interest, there have been major methodological issues related to the measurement of vocabulary size that have not been addressed satisfactorily. One of the most important of these is the idea of a word family. When we say that someone has a vocabulary size of ten thousand words, or say that a particular text contains seven thousand different words, what is being included in the unit we call a word?

In some studies, for example Carroll, Davies and Richman (1971), a word was regarded as a form, with any change in form such as capitalization or the presence of inflectional suffixes being sufficient to result in an item being counted as a different word. So, *govern*, *governor*, *Governor*, *governs* were all counted as different words. Other studies, for example Thorndike and Lorge (1944), consider items with an inflectional suffix to be members of the same

word family as the base. So, *govern, governs, governed, governing* were all members of the same word family and were counted as one word, but *governor* was a member of a different family. In other studies, for example West (1953), items with a common base but containing a variety of derivational and inflectional affixes were counted as the same word.

These different definitions of the members of a word family were the result of the different purposes of the research and the constraints governing it. All the researchers mentioned here were aware of the importance of explaining what their definition of a word included and they included such explanations in reports of their work. All too often however researchers have let dictionary makers make the decision for them.

If research on the vocabulary load of texts and on vocabulary size is to be cumulative and comparable, it is necessary for researchers to agree or at least to use similar criteria to decide on the vocabulary unit that is being investigated. For example, research on the vocabulary load of novels for non-native speakers of English (Hirsh and Nation 1993) sought to answer questions like the following:

- Does the ratio of known to unknown vocabulary increase as a learner proceeds through a novel?

- Does a long continuous text, such as a novel, provide opportunities for vocabulary learning through repetition of previously unknown vocabulary in a variety of contexts?

- How large a vocabulary is needed to read an unsimplified novel?

The answers to all these questions require a clear and sensible description of what is an item of vocabulary for the purposes of the research. If this description matches those used by other researchers, then useful comparison of results can occur.

Similarly, research on vocabulary size requires clear, justifiable descriptions of the unit of counting. If different researchers use the same description, then comparison of results is easy. If they use different descriptions but these are all based on a standard set of levels, then recalculation and adjustment is feasible if comparison is needed. Failure to have a standard basis for describing the unit in vocabulary size studies has contributed to the widely varying estimates of vocabulary size in numerous studies (Fries & Traver 1960).

Not including derived forms at a suitable level in the same word family has resulted in a failure of some tests of vocabulary size based on frequency lists to show implicational scaling between the various frequency levels. That is, instead of the number of words known decreasing as frequency decreases, there have been irregular and unpredictable variations. This has usually been the result of a low frequency derived form of a high frequency base word being included in a low frequency section of the test. Because the learners sitting the test can see the derived-base connection, they succeed on that item, whereas base words of the same low frequency level are not known. If the test was

based on a suitable word family description, this problem would be avoided and the expected implicational scaling would be more likely to occur.

4.3 *Acting as a reference point for developmental research*

The affixes chosen for investigation in research, mainly on native speakers of English, have been selected intuitively or because they combine with free base forms. The levels described in this article now allow differentiation between these affixes. If researchers take account of these levels in the design or even only in the reporting of their research, we can gain a more detailed picture of how control of affixation develops and of the factors that affect development.

For example, Nagy, Diakidoy and Anderson (1991) investigated knowledge of suffixes by choosing *-able*, *-ize*, *-er*, *-ful*, *-ness*, *-ist*, *-ism*, *-less*, *-ish* and *-ly*, which are all at Levels 3 and 4. It would require only minor changes to the design and reporting of the experiment to see if suffixes at Level 4 were less likely to be known than those at Level 3.

4.4 *Investigation of lexical storage*

There has been continuing debate over the way derived words are stored in the brain (Nagy, Schommer, Scott and Stallman 1989). Is each derived word stored separately? Or, is only the base word stored along with access to the regular word-building rules so that each derived form has to be reconstructed each time it is met or used? There is considerable evidence that for some affixed words this is the norm (see Tyler and Nagy 1990 for a recent review, or Stemberger and MacWhinney 1988 on the case of inflected forms).

It is most likely that there are different kinds of storage for different kinds of derived words. Some very high frequency derived forms, such as *business* and *electricity*, are probably stored in their own right. But what are the factors that determine the type of storage? Are high frequency items likely to have separate storage? Are more irregular forms likely to have separate storage? The levels described in this article are based on a set of criteria which take account of frequency, productivity, predictability and various kinds of regularity. It would be worth investigating if these same criteria affect the kind of storage of a derived item. If this is so, the levels or variations of them should correlate with the kind of storage. This in its turn has implications for teaching and learning.

4.5 *Guidelines for dictionary makers*

There is little consistency between various dictionaries in their treatment of derived forms. For example, in *Chambers English Dictionary* (1988) *cereal* is a sub-entry for *Ceres*, but it is a headword in COD8 and *Collins English Dictionary* (1979). *Cerebrate* is a headword in Collins, a sub-entry under *cerebration* in COD8, and a sub-entry under *cerebrum* in Chambers. In such a competitive market there is no need and probably no desire for consistency. However, dictionary making requires decisions to be made regarding the treatment of inflections and affixes. In particular, the following questions need to be considered. The levels described in this paper provide a step towards answering them in a principled and consistent way.

1. What word forms do not need to be listed at all? Many dictionaries do not list regularly inflected forms like *aims, aimed, aiming*. Items ending in *-ly* are often not listed. The criteria that are used to guide the listing or non-listing of items are some of the criteria that are used to set up the levels of inflection and affixation described in this paper.
2. If items are listed, how should they be included – as separate entries, defined sub-entries, or non-defined sub-entries? Words formed using inflections and affixes at Levels 2 and 3 are the most likely candidates for non-defined subentries. The further one moves through the levels, the greater the justification for listing words formed from these items as separate entries, although semantic specialisation might be seen as the most important criterion for lexicographers.
3. Which prefixed derived forms should be included under their stems in addition to their separate alphabetical listing? Because of alphabetical listing, prefixed forms such as *predetermine* are not placed near their base, *determine*. There is some value in learners being able to see in one entry, or in adjacent entries, the range of word-building devices a particular base can take. The more regular, frequent, and productive such devices are, the more justification there is for including them in the same entry.
4. Which affixes should be signalled as worth learning? *The Longman Dictionary of Contemporary English* (1987) contains an appendix on word-formation. The appendix is in two parts. The first part tabulates important prefixes and suffixes with substantial explanation of each one. The second part contains an alphabetical list of prefixes and an alphabetical list of suffixes. The first part of the appendix contains what the dictionary makers consider to be "the most common" items. When this list is compared with the items in the levels, we find that important affixes are not in the list (*-ish, -th, -al, -ess, -ism, -ist, -ous*) and that some rather irregular ones are (*dis-, de-, -ical*). Some of these differences between the list and the levels are the result of different prioritizing criteria. Some are not.
5. What word-building devices should be included in limited defining vocabularies? The LDOCE lists its defining vocabulary in an appendix and includes a list of permitted affixes. Almost all the affixes in Levels 3 and 4 are included in this list. The coverage at Levels 4 and 5 is patchy. The defining list includes several affixes which were considered too irregular to be included in Levels 3 to 6.

Although the levels are largely based on regular patterns, there are clearly unusual items within a pattern (see Bolinger 1985 for a discussion of *-less*). By being aware of the regular, the dictionary maker can more easily deal with the unusual. Indeed, because a dictionary's main concern is providing information about particular words rather than about the wider system that they fit into, information gathered from dictionary construction could provide the most stringent test of the validity of the levels, in that it would reveal items that go against the criteria used to place an affix at a particular level.

5. Some cautions

Although it is tempting to conclude on a high note, it is safer to repeat and elaborate some of the cautions expressed in this article.

Firstly, the divisions between the levels are arbitrary. This arbitrariness is most noticeable in the later levels (see for example the treatment of *-ion*). At each level it was necessary to prioritize criteria and balance them against each other. It is easily possible to subdivide various levels, and, particularly at the later levels, to add and delete items. It is also possible with more statistical data to apply the criteria for inclusion within a level more precisely.

Secondly, the levels will not make sense for very high frequency derived words, such as *direction* and *statement*. Thus, if for the purposes of analysis of a text it is assumed that derived forms at Level 3 are known for a given vocabulary, it may also be necessary to include some very high frequency derived forms from later levels as separate word families.

Thirdly, different purposes could result in a different number of levels and different items in some levels. In this article the focus has been on setting up levels for receptive knowledge of written text, with a focus on interpretation while reading. This meant that orthographic regularity was given priority over phonological regularity. A set of levels for productive use in speaking would differ considerably around Levels 4 and 6.

References

Allen, M. R. 1978. 'Morphological Investigations.' PhD thesis: University of Connecticut.

Bauer, L. 1978. *The Grammar of Nominal Compounds*. Odense: Odense University Press.

Bauer, L. 1979a. 'On the need for pragmatics in the study of nominal compounding.' *Journal of Pragmatics* **3**: 45–50.

Bauer, L. 1979b. 'Patterns of productivity in new formations denoting persons using the suffix *-er* in modern English.' *Cahiers de Lexicologie* **35, 2**: 26–31.

Bauer, L. 1983. *English Word Formation*. Cambridge: Cambridge University Press.

Bauer, L. 1988. *Introducing Linguistic Morphology*. Edinburgh: Edinburgh University Press.

Bauer, L. 1990. 'Level disorder: the case of *-er* and *-or*.' *Transactions of the Philological Society*. **88**: 97–110.

Beard, R. 1982. 'The plural as a lexical derivation.' *Glossa* **16**: 133–148.

Bock, C. 1948. 'Prefixes and suffixes.' *Classical Journal* **44**: 132–133.

Bolinger, D. 1985. 'Defining the indefinable' in R. Ilson (ed.).

Brown, J. 1971. *Programmed Vocabulary*. New York: New Century.

Carroll, J. B. 1940. 'Knowledge of English roots and affixes as related to vocabulary and Latin study.' *Journal of Educational Research* **34, 2**: 102–111.

Carroll, J. B., Davies, P., Richman, B. 1971. *The American Heritage Word Frequency Book*. New York: Houghton Mifflin, Boston: American Heritage.

Chambers English Dictionary. 1988. Cambridge: Cambridge University Press.

Cole, P., Beauvillain, C., Segui, J. 1989. 'On the representation and processing of prefixed and suffixed derived words: a differential frequency effect.' *Journal of Memory and Language* **28**: 1–13.

Collins Dictionary of the English Language. 1979. London: Collins. [*Collins English Dictionary*]

Concise Oxford Dictionary. 1990. Oxford: Oxford University Press. [8th edition: COD8]

Corson, D. J. 1985. *The Lexical Bar*. Oxford: Pergamon Press.

D'Anna, C. A., Zechmeister, E. B. In press. 'Toward a meaningful definition of vocabulary size.' *JRB: A Journal of Literacy*.

Fries, C. C., Traver, A. A. 1960. *English Word Lists*. Ann Arbor: George Wahr.

Gordon, P. 1989. 'Levels of affixation in the acquisition of English morphology.' *Journal of Memory and Language* **28**: 519–530.

Goulden, R., Nation, P., Read, J. 1990. 'How large can a receptive vocabulary be?' *Applied Linguistics* **11, 4**: 341–363.

Gove, P. B. (ed.). 1963. *Webster's Third New International Dictionary*. Massachusetts: G. & C. Merriam Co.

Graves, M. F. 1987. 'The roles of instruction in fostering vocabulary development' in M. G. McKeown and M. E. Curtis (eds.) *The Nature of Vocabulary Acquisition* Hillsdale, New Jersey: Lawrence Erlbaum.

Halle, M., Mohanan, K. P. 1985. 'Segmental phonology of modern English.' *Linguistic Inquiry* **16**: 57–116.

Harwood, F. W., Wright, A. M. 1956. 'Statistical study of English word formation.' *Language* **32**: 260–273.

Hirsh, D., Nation, I. S. P. 1993. 'What vocabulary size is needed to read unsimplified texts for pleasure?' *Reading in a Foreign Language* **8, 2** (In press).

Ilson, R. (ed.) 1985. *Dictionaries, Lexicography and Language Learning*. Oxford: Pergamon Press.

Jensen, J. T. 1990. *Morphology*. Amsterdam/Philadelphia: Benjamins.

Kirkpatrick, E. A. 1891. 'Number of words in an ordinary vocabulary.' *Science* **18, 446**: 107–108.

Kučera, H. 1982. 'The mathematics of language' in *The American Heritage Dictionary*. Second College Edition. Boston: Houghton Mifflin.

Ljung, M. 1977. 'Problems in the derivation of instrumental verbs' in H. E. Brekle and D. Kastovsky (eds.): *Perspektiven der Wortbildungsforschung*. Bonn: Bouvier 155–164.

Longman Dictionary of Contemporary English. 1987. Harlow: Longman. [LDOCE].

Marchand, H. 1969. *Categories and Types of Present-Day English Word-Formation*. Munich: Beck. 2nd edition.

Mohanan, K. P. 1986. *The Theory of Lexical Phonology*. Dordrecht: Reidel.

Mugdan, J. 1989. Review of Bauer 1988. *Yearbook of Morphology* **2**: 175–183.

Nagy, W. E., Anderson, R. C. 1984. 'How many words are there in printed school English?' *Reading Research Quarterly* **19, 3**: 304–330.

Nagy, W. E., Anderson, R. C., Schommer, M., Scott, J. A., Stallman, A. C. 1989. 'Morphological families in the internal lexicon.' *Reading Research Quarterly* **24, 3**: 263–282.

Nagy, W. E., Diakidoy, I. N., Anderson, R. C. 1991. 'The development of knowledge of derivational suffixes.' (Unpublished paper).

Oxford Advanced Learner's Dictionary of Current English. 1989. Oxford: Oxford University Press.

Scalise, S. 1984. *Generative Morphology*. Dordrecht: Foris.

Siegel, D. 1974. 'Topics in English Morphology.' PhD Thesis: MIT. Published 1979, New York: Garland.

Stauffer, R. G. 1942. 'A study of prefixes in the Thorndike list to establish a list of prefixes that should be taught in the elementary school.' *Journal of Educational Research* **35, 6**: 453–458.

Stein, G. 1985. 'Word-formation in modern English dictionaries' in R. Ilson (ed.).

Stemberger, J. P., MacWhinney, B. 1988. 'Are inflected forms stored in the lexicon?' in

M. Hammond and M. Noonan (eds.): *Theoretical Morphology*. San Diego: Academic Press.

Szpyra, J. 1989. *The Phonology-Morphology Interface*. London: Routledge.

Thorndike, E. L. 1941. *The Teaching of English Suffixes*. Teachers College, Columbia University.

Thorndike, E. L., Lorge, I. 1944. *The Teacher's Word Book of 30,000 Words*. Teachers College, Columbia University.

Tyler, A., Nagy, W. 1989. 'The acquisition of English derivational morphology.' *Journal of Memory and Language* **28**: 649–667.

Tyler, A, Nagy, W. 1990. 'Use of derivational morphology during reading.' *Cognition* **36**: 17–34.

West, M. 1953. *A General Service List of English Words*. London: Longmans, Green & Co.

White, T. G., Power, M. A., White, S. 1989. 'Morphological analysis: implications for teaching and understanding vocabulary growth.' *Reading Research Quarterly* **24, 3**: 283–304.

Williams, E. 1981. 'On the notions "lexically related" and "head of a word".' *Linguistic Inquiry* **12**: 245–274.

Wysocki, K., Jenkins, J. R. 1987. 'Deriving word meanings through morphological generalisation.' *Reading Research Quarterly* **22, 1**: 66–81.

Zwicky, A. M. 1992. 'Clitics' in W. Bright (ed.) *International Encyclopaedia of Linguistics*. Oxford: Oxford University Press.

Zwicky, A. M., Pullum, G. K. 1983. 'Cliticization vs inflection: English *-n't*.' *Language* **59**: 502–513.

Appendix 1

Summary of the levels

Level 1
A different form is a different word. Capitalization is ignored.

Level 2
Regularly inflected words are part of the same family. The inflectional categories are – plural; third person singular present tense; past tense; past participle; *-ing*; comparative; superlative; possessive.

Level 3
-able, -er, -ish, -less, -ly, -ness, -th, -y, non-, un-, all with restricted uses.

Level 4
-al, -ation, -ess, -ful, -ism, -ist, -ity, -ize, -ment, -ous, in-, all with restricted uses.

Level 5
-age (leakage), -al (arrival), -ally (idiotically), -an (American), -ance (clearance), -ant (consultant), -ary (revolutionary), -atory (confirmatory), -dom (kingdom; officialdom), -eer (black marketeer), -en (wooden), -en (widen), -ence (emergence), -ent (absorbent), -ery (bakery; trickery), -ese (Japanese; officialese), -esque (picturesque), -ette (usherette; roomette), -hood (childhood), -i (Israeli), -ian (phonetician; Johnsonian), -ite (Paisleyite; also chemical meaning), -let (coverlet), -ling (duckling), -ly (leisurely), -most (topmost), -ory (contradictory),

-ship (studentship), -ward (homeward), -ways (crossways), -wise (endwise; discussion-wise), ante- (anteroom), anti- (anti-inflation), arch- (archbishop), bi- (biplane), circum- (circumnavigate), counter- (counter-attack), en- (encage; enslave), ex- (ex-president), fore- (forename), hyper- (hyperactive), inter- (inter-African, interweave), mid- (mid-week), mis- (misfit), neo- (neo-colonialism), post- (post-date), pro- (pro-British), semi- (semi-automatic), sub- (subclassify; subterranean), un- (untie; unburden).

Level 6
-able, -ee, -ic, -ify, -ion, -ist, -ition, -ive, -th, -y, pre-, re-.

Level 7
Classical roots and affixes.

Appendix 2

Discussion of individual affixes

Level 2
There are many problems in defining the set of inflectional affixes.

Firstly, there are disagreements in the literature as to what constitute the set of inflectional categories of English. Some include plural, others do not (Beard 1982); some include comparative and superlative (Jensen 1990: 116), others do not (Mugdan 1989: 178); most do not include the negative marker *n't*, but some do (Zwicky & Pullum 1983); some include the possessive *-'s* but increasingly it is not included (Zwicky 1992). Consequently, any list of the inflectional categories of English is controversial.

Secondly, having made the decision regarding what inflectional categories to include, we have the problem that not all of the words constructed according to the principles of these categories are necessarily clearly inflectional. Consider

He is shooting clay-pigeons

I watched him shooting clay-pigeons

His shooting clay-pigeons disturbed me

His shooting of clay-pigeons was very disturbing

The shooting of clay-pigeons went on all day

Clay-pigeon shooting is an expensive pastime

Without wishing to claim that there is a definite level in this list at which the *-ing* ceases to be inflectional and becomes derivational, we note that the interpretation of the *-ing* in the first and last items in the list is not the same. Similar (though less striking) problems arise with the use of past participles as adjectives. We will conveniently ignore this problem and include all *-ed* and *-ing* forms formed from verbs.

Thirdly, there are degrees of regularity. Every non-modal verb must have a

third-person singular present tense form; not every noun must have a plural form. If a verb ends in an *-s* affix, it must be the third person singular of the present tense; if a noun ends in an *-s* affix it does not have to be a plural (consider *measles*, *news*, *Mums* (as a term of endearment), etc.). The well-known problems that native speakers and writers of English have with apostrophes mean that the presence or absence of an apostrophe cannot be taken as criterial or meaningful outside a very strictly edited text.

The possessive presents its own problems. As noted above, many do not now recognize it as an inflectional affix at all, but treat it as a clitic. This is because it is added not to nouns, but to noun phrases, as the examples below show:

The woman's hat

The woman I met yesterday's hat

The woman in green's hat

The woman who died's hat

The *-s* attaches itself to the last word in the noun phrase, rather than to the noun, as is the case with inflectional affixes. To recognize *died's* as containing a possessive despite being verbal requires a level of grammatical sophistication.

It would be possible to sub-divide the inflectional affixes according to any one or more of the types of problem mentioned above. It would be possible to have a Level 2a consisting of only those affixes which everyone agrees are inflectional, and a Level 2b consisting of the others mentioned; it would be possible to have a Level 2a consisting of only those affixes which never cause orthographic alternations (such as *'s*), and a Level 2b consisting of those which do; and so on. Where English is concerned, the number of inflectional affixes and their behaviour does not seem to merit such distinctions, though we recognize that they might be desirable or necessary in other languages.

Level 3

-able This affix is extremely productive when added to transitive verbs. The variant *-ible*, which is frequently added to bound bases should not be introduced at this stage but at Level 7. The meaning of *-able* added to nouns as in *knowledgeable*, *seasonable*, *sizeable* is not necessarily predictable and is not included at this level. Note that the rule of ⟨e⟩-deletion before a vowel is applied only inconsistently before this affix.

-er This is one of the affixes learnt earliest by English-speaking children, and it continues to have virtually automatic productivity. Children first learn the agent meaning, which is the commonest, and subsequently learn the instrument meaning. It occasionally denotes a location, as in *diner* but this is rare. It is added most commonly to verbs, but productively to other parts of speech (Bauer 1979b), where it means 'person connected with __'. There are some animal names in *-er* such as *retriever*, *warbler*. Note the spelling rules in *potter*, *diner* and (unusual, but not irregular) *trafficker*. When *-er* is added to a base

ending in *-r* the ⟨r⟩ at the end of the base is pronounced, thus /mɜ:dərə/ *murderer*. The most common words with *-er* in the LOB corpus include *leader*, *teacher*, *worker*, *farmer*, and *writer*. There could be problems with *-er*, as in for example *sewer* /su:ə/, *teller*, *summer*, and *banner*. However, its high frequency and productivity justify including it at this level. (Note *-er* as the comparative at Level 2.)

-ish This is productive, but does not give rise to quite the same numbers of new words as some of the other affixes at this level. It forms adjectives from nouns, numbers and adjectives. Common examples are *childish*, *selfish*, and *foolish*. When it forms nationality adjectives the base may not be recognizable (*English*), but such words would simply not be analyzed by students at this level, and so only items with bases which are potentially free forms (*Jewish*) are included here. Note that the ⟨y⟩ > ⟨i⟩ rule does not apply with this affix: *fortyish*, *boyish*, and final ⟨e⟩ deletion is variable: *blueish*, *whitish*. A form like *bookish* is semantically unpredictable and therefore is not included at this level. There are some deceptive forms like *rubbish* and *finish*, but these should not be a problem.

-less This suffix forms adjectives meaning 'without __' from nouns. The few familiar words where the base is not a noun can be taught individually: *weariless*, *tireless*. Instances where it is not clear from the form whether the base is a noun or not, can be dealt with as though they are nouns: *restless*, *doubtless*. The noun *wireless* is unlikely to cause great problems, since it is scarcely used these days.

-ly This is the suffix which forms adverbs from adjectives: *fortunately*, *sadly*. The only problem is that it is a homograph of a different *-ly* affix, and that when the two occur together one *-ly* is frequently dropped. Thus we find not only *a leisurely walk* but also *he strolled leisurely to the corner*. Fortunately, there are not many of this type and they are included in Level 5 along with other adjectival uses of *-ly*.

-ness This is one of the most productive derivational affixes in English. When it is added to adjectives it regularly means 'state, condition, quality of __' (Marchand 1969: 334). It is regularly found in modern texts added to bases other than adjectives, though with more or less the same meaning. Despite this regularity of meaning, there are a few cases where the meaning is not predictable, e.g. *(your) Highness*, *business*, and, from a noun, *witness*. These have to be learned as individual items. Note that ⟨y⟩ > ⟨i⟩ applies, although many authorities distinguish *busyness* from *business*. This is a modern distinction dating from the late 1800s (especially with no hyphen). The only phonological change that this affix causes is /nn/ > /n/, and even that is not obligatory. *Suddenness* may be pronounced with one or two /n/s.

-th The *-th* affix referred to here is the one which makes ordinal numbers from cardinal ones. It causes some irregular allomorphy of the base to which it is attached in *fifth*, *twelfth*, but since *first*, *second* and *third* have to be learned independently anyway, adding two more is not a great burden. A different *-th* affix also occurs at Level 6 in words like *warmth*.

-y There are at least two *-y* suffixes, the one at this level being the one which forms adjectives from nouns, with a meaning something like 'characterized by __'. The other is at Level 6. This one is extremely productive in children's language, less so in scientific prose. There are several semantically irregular forms here, including *fishy*, *funny*, *mousey*, and *nosey*, and these would not be included at this level. In some cases, confusion with the *-y/-ie* diminutive ending is possible in the form (though not usually in context). Thus *hors(e)y* (however spelt) could be 'addicted to horses' or 'a dear little horse'.

non- This affix is added productively to adjectives and nouns, is virtually always written with a hyphen (but *nonsense*) and has a regular meaning, which, however, needs to be distinguished from that of *un-*. While *un-* tends to form antonyms, *non-* forms complementaries: that is, it is possible to be neither popular nor unpopular, but if one is not popular, one must be non-popular. It is one of the affixes least likely to have phonological effects, but it can be pronounced with the last nasal assimilated to the place of a following stop.

un- There are various homophonous *un-* affixes (see Marchand 1969: 201–7). The most regular one is the one added to adjectives to provide an antonym (see note under *non-*). The nasal consonant is frequently assimilated to the place of articulation of a following stop consonant in fluent speech, but keeping the /n/ is perfectly possible. No assimilation is noted in the orthography. Other uses of *un-* are at Level 5.

Level 4

-al There are at least two distinct *-al* suffixes. The one included here is the one which produces adjectives from nouns (and occasionally adjectives). The suffix does not appear to have any meaning in itself, but simply marks the change of form-class. *-ical* could be treated as a separate affix, as a sequence of two affixes, or as a variety of *-al*. Orthographic variants *-ial* and *-ual*, if treated as separate entities, are not widely generalized, but can be recognized as allomorphs of *-al* in e.g. *professorial*, *habitual*.

-ation This is a widely generalized affix, but an extremely difficult one to deal with. From a theoretical perspective, it is clear that the suffix *-ation* has a number of variants. We can recognize *-ation* as the basic form, found not only when added to verbs in *-ize* but also with forms like *flirtation*, *interpretation*. Where the base ends in *-ify* we get *-ication* as in *justification*. Where the base ends in *-ate* there is no repetition of *-at-* in derivatives: *assassination*. The variant *-ution* as in *revolution* is probably not recognizable and so is not included at this level. The distribution of *-ition* and *-ion* in words like *definition* and *dilution* is not predictable according to general principles, so that *addition*, *competition*, *definition*, *exposition*, (and other words whose base ends in *-pose*), and *repetition* (with irregular orthography) are perhaps best learned as exceptions. While the *-ation* allomorph is fairly transparent in orthographic terms, the other allomorphs give rise to considerable stem allomorphy (having variation in the form of the stem e.g. *convene–convention*), which makes them difficult to recognize consistently. For this practical reason, the various allo-

morphs are treated as separate suffixes, and only *-ation* is considered at this level.

-ess This affix is frowned upon these days for social reasons, but is still found in a wide range of words, though perhaps not very frequently. Its basic meaning is 'female', as in *heiress*, *tigress* (note the ⟨e⟩-deletion before ⟨r⟩), but it can also, occasionally, mean 'wife of a __' as in *mayoress* (which is actually ambiguous).

-ful This suffix can be viewed as a suffixal variant of the free word *full*, although that meaning is weak (at best) in words like *hopeful*, *wonderful*, especially when applied to non-animates (*hopeful results*, *wonderful weather*). This suffix potentiates (Williams 1981) *un-* prefixation (*unfruitful*, *unhopeful*) in a large number of words. The most common words with this suffix are *useful*, *successful*, and *beautiful*. *Awful* should not be considered as a member of the *awe* family. Note the use of *-ful* in words like *armful* and *mouthful*, which is also included at this level.

-ism This suffix has a number of possible meanings: 'doctrinal system of principles' as in *Marxism*; 'something viewed as such, usually disparagingly' as in *colonialism*, *fanaticism*; 'abnormal condition' as in *alcoholism*; 'an expression typical of __' as in *Irishism*, *vulgarism*; and a few unclassifiable cases such as *criticism*, *mannerism*. This affix is frequently perceived as being linked to *-ist*, such that e.g. *Marxism* and *Marxist* form a close pair.

-ist This suffix is widely generalized, but not always analyzable because it can be added to non-word bases (as in *deist*). In many cases it is related to *-ism* (q.v.), denoting the adherent of the system of principles; in other cases it denotes a scientific profession, or something viewed as such (*economist*, *abortionist*); this meaning is also found when the base ends in *-ology*, as in *morphologist* (note the ⟨y⟩-deletion); in other cases it marks an artistic profession, or something viewed as such (*harpist*, *novelist*, *humorist*); in other cases it is perhaps best glossed as 'person connected with __' as in *cyclist*, *motorist*. All these words are – or started out as being – fairly formal in style. At this level *-ist* is only included when it is added to free bases. The most frequent examples include *artist*, *socialist*, and *specialist*. See Level 6 for other uses of *-ist*.

-ity While this suffix causes large numbers of phonological changes, it causes far fewer orthographic ones. The alternation between *-able/-ible* and *-ability /-ibility* is irregular, and needs special notice, particularly since it is so common. The meanings of words in *-ity* are frequently specialized within particular scientific domains, so that the actual meaning may be hard to deduce. Thus *sensitivity* (note the ⟨e⟩-deletion) has technical uses in (at least) physiology, electronics and photography, while *sensitiveness* is far less precise in its meaning. Countable meanings like *a fatality* may also not be easily predictable.

-ize This suffix, which also appears as *-ise*, creates verbs, particularly but not exclusively transitive verbs, from adjectives and nouns. It is added mainly to Latinate bases, which is why *Englishize* (see the OED) for 'anglicize' sounds so odd. Marchand (1969: 320) distinguishes five different meanings for this

affix, illustrated by the words *legalize, itemize, hospitalize, bowdlerize* and *ionize*. Note the ⟨t⟩ in *dramatize, dogmatize*, which Marchand suggests may be related to the ⟨t⟩ in *dramatic, dogmatic*. A final ⟨y⟩ regularly elides before *-ize* as in *colonize*. Note also the loss of final ⟨a⟩ in the (rare) *propagandize*. Words like *hypnotize* are probably best left unanalyzed at this stage.

-ment This suffix, probably no longer productive (Bauer 1983), is otherwise quite regular, forming nominalizations from verbs. The problems of interpretation are the same as those facing all nominalizations. The following words need to be specifically noted as exceptional and are not included in the word family of the base at this level: *basement, betterment, devilment, merriment, wonderment*.

-ous This suffix forms adjectives from nouns. The form *sacrilegious* is orthographically irregular (though phonologically regular). The form *gorgeous* is semantically irregular and is not included at this level. Nouns ending in *-ion* regularly lose the *-on* before *-ous* is added: *ambitious, contagious*. This suffix loses its ⟨u⟩ when it precedes *-ity*: *generosity*.

in- This prefix forms negative forms of Latinate adjectives. There are also some equivalent nouns, which in most cases can be seen as derived from the adjective: *insignificance* is the property of being insignificant, rather than the opposite of *significance* (Marchand 1969: 169). The orthographic variants *im-, il-* and *ir-* can be treated as variants of *in-*. Of the three, only *im-* recurs in a wide range of words, but there are many words beginning *im-* which do not contain this prefix. The number of appropriate words beginning with *ill-* is so small that it would be better to learn them as a list. Most words beginning *irr-* are appropriate ones (see the COD, for example), but there are not very many of them.

Level 6

-able This suffix also occurs at Level 3. All instances of *-able* at Level 6 require the formative *-ate* to be truncated before *-able* suffixation, as in *attenuable, permeable*.

-ee The formative *-ate* is normally truncated before *-ee* suffixation, as in *nominee*, but there are exceptions, such as the rare *educatee*. For some discussion see Bauer (1983: 243–50). While an *-ee* derivative typically denotes a person who is the direct object of the verb (a nominee is a person who is nominated), other patterns are also common, with *-ee* being used for non-humans and also being used more for subjects (Bauer 1983). While this rarely gives rise to misunderstandings for native-speakers, it may be confusing for L2-speakers. Special attention must be drawn to the irregular *bargee, bootee, committee, settee* which should not be treated as analyzable.

-ic This suffix derives adjectives. Because many of the adjectives are actually borrowed from French, Latin or even Greek, there are patterns of base allomorphy which are not normal in English. Consider the following (in no particular order):

(i) truncation of *-ia* before suffixation (*anemic*);

(ii) truncation of *-y* before suffixation (*geographic*); sometimes this is accompanied by other changes (if we assume that the same *-y* suffix is involved) as in *heretic*, where ⟨s⟩ > ⟨t⟩ as well;

(iii) the alteration of *-itis* to *-itic* (which could be seen as truncation of *-is*), as in *bronchitic*;

(iv) the addition of ⟨n⟩, as in *Platonic*, *embryonic*;

(v) the truncation of *-ous* before suffixation as in *ferric*;

(vi) the insertion of ⟨t⟩ as in *dramatic*;

(vii) the insertion of ⟨at⟩ as in *diagrammatic* (this could be seen as the same rule as the last one).

In other cases, it seems simpler to avoid talking about truncation, but rather to say that one affix is swapped for another to provide the appropriate part of speech; thus many words in *-ism* have corresponding adjectives in *-ic*, as for example, *endomorphic*, *endomorphism* (Marchand 1969: 295). This discussion of the problems with *-ic* is not exhaustive.

-ify This suffix produces verbs. Words which end in ⟨y⟩, or ⟨e⟩ usually drop this letter before *-ify*, but sometimes a ⟨y⟩ is retained, and the ⟨i⟩ of *-ify* is dropped instead (especially if the base ends in ⟨ey⟩) as in *Cockneyfy*. Various final syllables of bases are deleted before the suffixation of *-ify*. Consider, for example, *quantify*, *mystify*, *vitrify*, *syllabify*. There are numerous irregular formations not included at this level, such as *argufy* (with variant spellings), with a verbal base, and *modify* and *amplify*, whose meaning is not predictable from that of the base.

-ion If we consider *-ion* to be a different affix from *-ation*, as was suggested at Level 4, then the amount of base allomorphy it causes clearly means it should be listed here. Consider, among others,

redeem	redemption
perceive	perception
describe	description
demolish	demolition
convene	convention
include	inclusion
concede	concession
transmit	transmission
decline	declension
propel	propulsion
unite	union

These alternations really belong to Latin morphology rather than to English, and while there may be quite large classes which fit some of these patterns, they make recognition of words in this class extremely difficult.

-ist At this level all the words ending in *-ist* have an unexplained consonant before the suffix, as in *tobacconist, lutenist; egotist, dogmatist. -ist* also occurs at Level 4.

-ition Like *-ion*, of which it can be seen as a co-allomorph, *-ition* causes synchronically irregular base allomorphy, as in *admonition, apparition, nutrition.* It can also be difficult to analyze because it may not be clear whether the affix is *-ition* or *-ion*: contrast words such as *addition* and *edition.* Fortunately, *-ition* is rather rare, so that not much would be lost, even at this level, if it were ignored, and words containing it learned as units.

-ive At this level, this suffix is frequently preceded by an unpredictable *-at-*, and also gives rise to other types of base allomorphy. Consider *representative, persuasive, permissive, impulsive.* Note also *additive*, where the *-it-* appears to be the same as in *addition.*

-th Noun-producing *-th* is easily recognized in a few words like *warmth*, but is masked in common words like *depth* and *length* and completely opaque to most native speakers in a word like *dearth, wealth* or even *truth.* The number of words in which this suffix occurs is fairly small, but it may nevertheless be a useful analytical tool for some of these fairly common words. *-th* in ordinal numbers occurs at Level 3.

-y This *-y* affix is the one forming abstract nouns. It is not clear whether a number of distinct affixes should be recognized here (cf Marchand 1969). If we recognize a single affix, it involves recognizing a fair amount of allomorphy, as in

diplomat	diplomacy
pirate	piracy
supreme	supremacy
heretic	heresy
bigot	bigotry
tyrant	tyranny

-y also occurs at Level 3 in words like *smelly.*

pre- The problem with this prefix is not allomorphy or semantic regularity, the problem is recognizing it in the first place. If we could consider only words with *pre-* followed by a genuine (not a line-end) hyphen, it would be perfectly regular. But even forms like *preexist* are written without a hyphen, despite the ⟨ee⟩ sequence, so that the hyphen rule would omit many of the regular cases. (Just over half of the types with *pre-* at this level in the LOB corpus are written with a hyphen.) On the other hand, if instances with no hyphen are considered as well, it becomes difficult to analyze words like *precede, prelate, prepare, preposition, prescribe, present, preside, pretext, prevent*, which do not mean what they would appear to mean. Note that the meaning 'earlier in time' is more common than the meaning 'in front of' and the meaning 'to the greatest extent', although all three have to be learned. They are exemplified in *preexist, premolar* and *pre-eminent.*

re- The problem with *re-* is the same as that with *pre-*, although *re-* does not have as many distinct meanings. Words that will be wrongly analyzed include *react*, *reagent*, *rebus*, *rebut*, *recap*, *recess*, *recite*, *recoil*, *recollect*, *recommend*, *record*, *recover*, *recur*, *redeem*, *redoubt*, *redress*, etc. The undoubted productivity of *re-* provides a double bind: on the one hand, learning it would save a lot of time because it occurs in so many words, on the other, there are so many lexicalized instances with *re-* and cases that are likely to be mis-analyzed, that trying to use it might be counter-productive. Just under one-third of the word types with *re-* at this level in the LOB corpus are written with a hyphen.

Although the cases of allomorphy at Level 6 which are listed above are confusing, note that a few patterns are repeated.

i *-ate* is deleted or *-(a)t(e)* is inserted.

ii ⟨n⟩ is inserted.

iii Latin inflectional endings are deleted.

iv ⟨d⟩ alternates with ⟨s⟩.

v ⟨t⟩ alternates with ⟨s⟩.

Such processes can be considered individually, and shown to apply over several types of affix. Gross base allomorphy is sometimes caused by the accumulation of such processes, with fairly automatic orthographic adjustments as a consequence of them. For example, *diagrammatic* can be seen as *diagram* + *ate* + *ic* (as in (i)), with deletion of ⟨e⟩ before a vowel, and (unnecessary) ⟨m⟩ doubling.