

COMPUTER METHODS FOR MORPHOLOGICAL ANALYSIS

Roy J. Byrd, Judith L. Klavans
I.B.M. Thomas J. Watson Research Center
Yorktown Heights, New York 10598

Mark Aronoff, Frank Anshen
SUNY / Stony Brook
Stony Brook, New York 11794

1. Introduction

This paper describes our current research on the properties of derivational affixation in English. Our research arises from a more general research project, the Lexical Systems project at the IBM Thomas J. Watson Research laboratories, the goal for which is to build a variety of computerized dictionary systems for use both by people and by computer programs. An important sub-goal is to build reliable and robust word recognition mechanisms for these dictionaries. One of the more important issues in word recognition for all morphologically complex languages involves mechanisms for dealing with affixes.

Two complementary motivations underlie our research on derivational morphology. On the one hand, our goal is to discover linguistically significant generalizations and principles governing the attachment of affixes to English words to form other words. If we can find such generalizations, then we can use them to build our improved word recognizer. We will be better able to correctly recognize and analyse well-formed words and, on the other hand, to reject ill-formed words. On the other hand, we want to use our existing word-recognition and analysis programs as tools for gathering further information about English affixation. This circular process allows us to test and refine our emerging word recognition logic while at the same time providing a large amount of data for linguistic analysis.

It is important to note that, while doing derivational morphology is not the only way to deal with complex words in a computerized dictionary, it offers certain advantages. It allows systems to deal with coinages, a possibility which is not open to most systems. Systems which do no morphology and even those which handle primarily inflectional affixation (such as Winograd

(1971) and Koskenniemi (1983)) are limited by the fixed size of their lists of stored words. Koskenniemi claims that his two-level morphology framework can handle derivational affixation, although his examples are all of inflectional processes. It is not clear how that framework accounts for the variety of phenomena that we observe in English derivational morphology. Morphological analysis also provides an additional source of lexical information about words, since a word's properties can often be predicted from its structure. In this respect, our dictionaries are distinguished from those of Allen (1976) where complex words are merely analysed as concatenations of word-parts and Cercone (1974) where word structure is not exploited, even though derivational affixes are analysed.

Our morphological analysis system was conceived within the linguistic framework of word-based morphology, as described in Aronoff (1976). In our dictionaries, we store a large number of words, together with associated idiosyncratic information. The retrieval mechanism contains a grammar of derivational (and inflectional) affixation which is used to analyse input strings in terms of the stored words. The mechanism handles both prefixes and suffixes. The framework and mechanism are described in Byrd (1983a). Crucially, in our system, the attachment of an affix to a base word is conditioned on the properties of the base word. The purpose of our research is to determine the precise nature of those conditions. These conditions may refer to syntactic, semantic, etymological, morphological or phonological properties. (See Byrd (1983b)).

Our research is of interest to two related audiences: both computational linguists and theoretical linguists. Computational linguists will find here a powerful set of pro-

grams for processing natural language material. Furthermore, they should welcome the improvements to those programs' capabilities offered by our linguistic results. Theoretical linguists, on the other hand, will find a novel set of tools and data sources for morphological research. The generalizations that result from our analyses should be welcome additions to linguistic theory.

2. Approach and Tools

Our approach to computer-aided morphological research is to analyse a large number of English words in terms of a somewhat smaller list of monomorphemic base words. For each morphologically complex word on the original list which can be analysed down to one of our bases, we obtain a structure which shows the affixes and marks the parts-of-speech of the components. Thus, for *beautification*, we obtain the structure

```
<<<beauty>N +ify>V +ion>N.
```

In this structure, the noun *beauty* is the ultimate base and *+ify* and *+ion* are the affixes.

After analysis, we obtain, for each base, a list of all words derived from it, together with their morphological structures. We then study these lists and the patterns of affixation they exemplify, seeking generalizations. Section 3 will give an expanded description of the approach together with a detailed account of one of the studies.

We have two classes of tools: word lists and computer programs. There are basically four word lists.

1. The Kucera and Francis (K&F) word list, from Kucera and Francis (1967), contains 50,000 words listed in order of frequency of occurrence.
2. The BASE WORD LIST consists of approximately 3,000 monomorphemic words. It was drawn from the top of the K&F list by the GETBASES procedure described below.
3. The UDICT word list consists of about 63,000 words, drawn mainly from Merriam (1963). The UDICT program, described below, uses this list in conjunction with our word grammar to produce morphological analyses of input words. The UDICT word list is a superset of the base word list; for each word, it contains the major category as well as other grammatical information.

4. The "complete" word list consists of approximately one quarter million words drawn from an international-sized dictionary. Each entry on this list is a single orthographic word, with no additional information. These are the words which are morphologically analysed down to the bases on our base list.
5. We have prepared reverse spelling word lists based on each of the other lists. A particularly useful tool has been a group of reverse lists derived from Merriam(1963) and separated by major category. These lists provide ready access to sets of words having the same suffix.

Our computer programs include the following.

1. UDICT. This is a general purpose dictionary access system intended for use by computer programs. (The UDICT program was originally developed for the EPISTLE text-critiquing system, as described in Heidorn, et al. (1982).) It contains, among other things, the morphological analysis logic and the word grammar that we use to produce the word structures previously described.
2. GETBASES. This program produces a list of monomorphemic words from the original K&F frequency lists. Basically, it operates by invoking UDICT for each word. The output consists of words which are morphologically simple, and the bases of morphologically complex words. (Among other things, this allows us to handle the fact that the original K&F lists are not lemmatised.) The resulting list, with duplicates removed, is our "base list".
3. ANALYSE. ANALYSE takes each entry from the complete word list. It invokes the UDICT program to give a morphological analysis for that word. Any word whose ultimate base is in the base list is considered a derived word. For each word from the base list, the final result is a list of pairs consisting of [derived-word, structure] The data produced by ANALYSE is further processed by the next four programs.
4. ANALYSES. This program allows us to inspect the set of [derived-word,structure] pairs associated with any word in the base list. For example, its output for the word *beauty* is shown in Figure 1. In the

```

beautied <<*>N +ed>A
beautification <<<*>N +ify>V +ion>N
beautifier <<<*>N +ify>V #er>N
beautiful <<*>N #ful>A
beautifully <<<*>N #ful>A -ly>D
beautiffulness <<<*>N #ful>A #ness>N
beautify <<*>N +ify>V
unbeautified <un# <<<*>N +ify>V +ed>A>A
unbeautified <un# <<<*>N +ify>V -ed1>V>V
unbeautiful <un# <<*>N #ful>A>A
unbeautifully <<un# <<*>N #ful>A>A -ly>D
unbeautiffulness <<un# <<*>N #ful>A>A #ness>N
unbeautify <un# <<*>N +ify>V>V
rebeautify <re# <<*>N +ify>V>V

```

Figure 1. ANALYSES Output.

structures, an asterisk represents the ultimate base *beauty*.

5. SASDS. This program produces 3 binary matrices indicating which bases take which single affixes to form another word. One matrix is produced for each of the major categories: nouns, adjectives, and verbs. More detail on the contents and use of these matrices is given in Section 3.
6. MORPH. This program uses the matrices created by SASDS to list bases that accept one or more given affixes.
7. SAS. (SAS is a trademark of the SAS Institute, Inc., Cary, North Carolina.) This is a set of statistical analysis programs which can be used to analyse the matrices produced by SASDS.
8. WordSmith. This is an on-line dictionary system, developed at IBM, that provides fast and convenient reference to a variety of types of dictionary information. The WordSmith functions of most use in our current research are the REVERSE dimension (for listing words that end the same way), the WEBSTER7 application (for checking the definitions of words we don't know), and the UDED application (for checking and revising the contents of the UDICT word list).

3. Detailed Methods

Our research can be conveniently described as a two stage process. During the first stage, we endeavored to produce a list of morphologically active base words from which other English words can be derived by affixation. The term "morphologically active" means that a word can potentially serve as the base of a large number of affixed derivatives. Having such words is important for

stage two, where patterns of affixation become more obvious when we have more instances of bases that exhibit them. We conjectured that words which were frequent in the language have a higher likelihood of participating in word-formation processes, so we began our search with the 6,000 most frequent words in the K&F word list.

The GETBASES program segregated these words into two categories: morphologically simple words (i.e., those for which UDICT produced a structure containing no affixes) and morphologically complex words. At the same time, GETBASES discarded words that were not morphologically interesting; these included proper nouns, words not belonging to the major categories, and non-lemma forms of irregular words. (For example, the past participle *done* does not take affixes, although its lemma *do* will accept *#able* as in *doable*)

GETBASES next considered the ultimate bases of the morphologically complex words. Any base which did not also appear in the K&F word list was discarded. The remaining bases were added to the original list of morphologically simple words. After removing duplicates, we obtained a list of approximately 3,000 very frequent bases which we conjectured were morphologically active.

Development of the GETBASES program was an iterative process. The primary type of change made at each iteration was to correct and improve the UDICT grammar and morphological analysis mechanism. Because the constraints on the output of GETBASES were clear (and because it was obvious when we failed to meet them), the creation of GETBASES proved to be a very effective way to guide improvements to UDICT. The more important of these improvements are discussed in Section 4.3.

For stage two of our project, we used ANALYSE to process the "complete" word list, as described in Section 2. That is, for each word, UDICT was asked to produce a morphological analysis. Whenever the ultimate base for one of the (morphologically complex) words appeared on our list of 3,000 bases, the derived word and its structure were added to the list of such pairs associated with that base. ANALYSE yielded, therefore, a list of 3,000 sublists of [word,structure] pairs, with each sublist named by one of our base words. We called this result BASELIST.

Our first in-depth study of this material involved the process of adding a single affix to a base word to form another word. By applying SASDS to BASELIST, we obtained 3 matrices showing for each base which affixes it did and did not accept. The noun matrix contained 1900 bases; the adjective matrix contained 850 bases; and the verb matrix contained 1600 bases. (Since the original list of bases contained words belonging to multiple major categories, these counts add up to more than 3,000. The ANALYSE program used the part-of-speech assignments from UDICT to disambiguate such homographs.)

Figure 2 contains samples taken from the noun, adjective, and verb matrices. For each matrix, the horizontal axis shows the complete list of affixes (for that part-of-speech) covered in our study. The vertical axes give contiguous samples of our ultimate bases.

Our results are by no means perfect. Some of our mis-analyses come about because of missing constraints in our grammar. The process of correcting these errors is discussed in Section 4. Sometimes there are genuine ambiguities, as with the words *refuse* (<re# <fuse>V>V) and *preserve* (<pre# <serve>V>V). In the absence of information about how an input word is pronounced or what it means, it is difficult to imagine how our analyser can avoid producing the structures shown.

Some of our problems are caused by the fact that the complete word list is alternately too large and not large enough. It includes the word *artal*, (plural of *rotl*, a Middle Eastern unit of weight) which our rules dutifully, if incorrectly, analyse as <<art>N +al>A. Yet it fails to include *angelhood*, even though *angel* bears the [+human] feature that #hood seems to require.

Despite such errors, however, most of the analyses in these matrices are correct and provide a useful basis for our analytical work. We employed a variety of techniques to examine these matrices, and the BASELIST. Our primary approach was to use SAS, MORPH, and ANALYSES to suggest hypotheses about affix attachment. We then used MORPH, WordSmith, and UDICT (via changes to the grammar) to test and verify those hypotheses. Hypotheses which have so far survived our tests and our skepticism are given in Section 4.

4. Results

Using the methods described, we have produced results which enhance our understanding of morphological

processes, and have produced improvements in the morphological analysis system. We present here some of what we have already learned. Continued research using our approach and data will yield further results.

4.1 Methodological Results

It is significant that we were able to perform this research with generally available materials. With the exception of the K&F word frequency list, our word lists were obtained from commercially available dictionaries. This work forms a natural accompaniment to another Lexical Systems project, reported in Chodorow, *et al.* (1985), in which semantic information is extracted from commercial dictionaries. As the morphology project identifies lexical information that is relevant, variations of the semantic extraction methods may be used to populate the dictionary with that information.

As has already been pointed out, our rules leave a residue of mis-analysed words, which shows up (for example) as errors in our matrices. Although we can never eliminate this residue, we can reduce its size by introducing additional constraints into our grammar as we discover them. For example, *chicken* was mis-analysed as <<chic>A +en>V. As we show in greater detail below, we now know that the +en suffix requires a [+Germanic] base; since *chic* is [-Germanic], we can avoid the mis-analysis. Similarly we can avoid analysing *legal* as <<leg>N +al>A by observing that +al requires a [-Germanic] base while *leg* is [+Germanic]. Finally, we now have several ways to avoid the mis-analysis of *maize* as <<ma>N +ize>V, including the observation that +ize does not accept monosyllabic bases. We don't expect, however, to find a constraint that will deal correctly with words like *artal*.

In the introduction, we pointed out that one of our goals was to build a system which can handle coinages. With respect to the 63,000-word UDICT word list, the quarter-million-word complete word list can be viewed as consisting mostly of coinages. The fact that our analyser has been largely successful at analysing the words on the complete word list means that we are close to meeting our goal. What remains is to exploit our research results in order to reduce our mis-analysed residue as much as possible.

4.2 Linguistic Results

Linguistically significant generalizations that have resulted so far can be encoded in the form of conditions and assertions in our word formation rule grammar (see Byrd (1983a)). They typically constrain interactions between specific affixes and particular groups of words. The linguistic constraints fall into at least three categories: (1) syllabic structure of the base word; (2) phonemic nature of the final segment of the base word; and (3) etymology of the base word, both derived and underived. Each of these is covered below. Some of these constraints have been informally observed by other researchers, but some have not.

Constraints on the Syllabic structure of the base word. It is commonly known that the length of a base word can affect an inflectional process such as comparative formation in English. One can distinguish between short and long words where [+short] indicates two or fewer syllables and [+long] indicates two or more syllables. For example, a word such as *big* which is [+short] can take the affixes *-er* and *-est*. In contrast, words which are [-short] cannot, cf. *possible*, **possibler*, **possiblest*. (There are additional constraints on comparative formation, which we will not go into here. We give here only the simplified version.) We have found that other suffixes appear to require the feature [+short]. For example, nouns that take the suffix *#ish* tend to be [+short]. The actual results of our analysis show that *no* words of four syllables took *#ish* and only seven words of three syllables took *#ish*. In contrast, a total of 221 one and two syllable words took this suffix. The suffix thus preferred one syllable words over two syllable words by a factor of four (178 one syllable words over 43 two syllable words). Compare *boy/boyish* with *mimeograph/mimeographish*. This is not to say that a word like *mimeographish* is necessarily ill-formed, but that it is less likely to occur, and in fact did not occur in a list like Merriam (1963).

Two other suffixes also appear to select for number of syllables in the base word. In this case the denominal verb suffixes *+ize* and *+ify* are nearly in complementary distribution. Our data show that of the approximately 200 bases which take *+ize*, only seven are monosyllabic. Compare this with the suffix *+ify* which selects for about 100 bases, of which only one is trisyllabic and 17 are disyllabic. Thus, *+ify* tends to select for [+short]

bases while *+ize* tends to select for [+long] ones. As with *#ish*, there appears to be motivation for syllabic structure constraints on morphological rules.

In the case of *+ize* and *+ify* it appears that the syllabic structure of the suffix interacts with the syllabic structure of the base. Informally, the longer suffix selects for a [+short] base, and the shorter suffix selects for a [+long] base. Our speculation is that this may be related to the notion of optimal target metrical structure as discussed in Hayes (1984). This notion, however, is the subject of future research.

The Final Segment of the Base Word. The phonemic nature of the final segment appears to affect the propensity of a base to take an affix. Consider the fact that there occurred some 48 *+ary* adjectives derived from nouns in our data. Of these, 46 are formed from bases ending with alveolars. The category alveolar includes the phonemes /t/, /d/, /n/, /s/, /z/, and /l/. The two exceptions are *customary* and *palmary*. Again, in a word recognizer, if a base does not end in one of these phonemes, then it is not likely to be able to serve as the base of *+ary*. We have also found that the *ual* spelling of the *+al* suffix prefers a preceding alveolar, such as *gradual*, *sexual*, *habitual*.

Another result related to the alveolar requirement is an even more stringent requirement of the nominalizing suffix *+ity*. Of the approximately 150 nouns taking *+ity*, only three end in the phoneme /t/ (*chastity*, *sacrosanctity*, and *vastity*). In addition the adjectivizer *+cy* seems also to attach primarily to bases ending in /t/. The exceptions are *normalcy* and *supremacy*.

Etymology of the Base Word. The feature [+Germanic] is said to be of critical importance in the analysis of English morphology (Chomsky and Halle 1968, Marchand 1969). In two cases our data show this to be true. The suffix *+en*, which creates verbs from adjectives, as in *moist/moisten*, yielded a total of fifty-five correct analyses. Of these, forty-three appear in Merriam (1963), and of these forty-one are of Germanic origin. The remaining two are *quieten* and *neaten*. The former is found only in some dialects. It is clear that *+en* verbs are overwhelmingly formed on [+Germanic] bases.

The feature [Germanic] is also significant with *+al* adjectives. In contrast to the *+en* suffix, *+al* selects for the feature [-Germanic]. In our data, there were some

two hundred and seventy two words analysed as adjectives derived from nouns by *+al* suffixation. Of the base words which appear in Merriam (1963), only one, *bridal*, is of Germanic origin. However, interestingly, it turns out that the analysis <<bride>N +al>A is spurious, since *bridal* is the reflex of an Old English form *brydealu*, a noun referring to the wedding feast. The adjective *bridal* is not derived from *bride*. Rather it was zero-derived historically from the nominal form.

Finally, other findings from our analysis show that no words formed with the Anglo-Saxon prefixes *a+*, *be+* or *for+* will negate with the Latinate prefixes *non#* or *in#*. This supports the findings of Marchand (1969). Observe that in these examples, the constraint applies between affixes, rather than between an affix and a base. The addition of an affix thus creates a new complex lexical item, complete with additional properties which can constrain further affixation.

In sum, our sample findings suggest a number of new constraints on morphological rules. In addition we provide evidence and support for the observations of others.

4.3 Improvements to the Implementation

In addition to using our linguistic results to change the grammar, we have also made a variety of improvements to UDICT's morphological analyser which interprets that grammar. Some have been for our own convenience, such as streamlining the procedures for changing and compiling the grammar. Two of the improvements, however, result directly from the analysis of our word lists and files. These improvements represent generalizations over classes of affixes.

First, we observed that, with the exception of *be*, *do*, and *go*, no base spelled with fewer than three characters ever takes an affix. Adding code to the analyser to restrict the size of bases has had an important effect in avoiding spurious analyses.

A more substantial result is that we have added to UDICT a comprehensive set of English spelling rules which make the right spelling adjustments to the base of a suffix virtually all of the time. These rules, for example, know when and when not to double final consonants, when to retain silent *e* preceding a suffix beginning with a vowel, and when to add *k* to a base ending in *c*. These rules are a critical aspect of UDICT's ability to robustly handle normal English input and to avoid misanalyses.

5. Further Analyses and Plans

When we have modified our grammar to incorporate results we have obtained, and added the necessary supporting features and attributes to the words in UDICT's word list, we will re-run our programs to produce files based on the corrected analyses that we will obtain. These files will, in turn, be used for further analysis in the Lexical Systems project, and by other researchers.

We plan to continue our work by looking for more constraints on affixation. A reasonable, if ambitious, goal is to achieve a word formation rule grammar which is "tight" enough to allow us to reliably generate words using derivational affixation. Such a capability would be important, for example, in a translation application where idiomaticness often requires that a translated concept appear with a different part-of-speech than in the source language.

Further research will investigate patterns of multiple affixation. Are there any interdependencies among affixes when more than one appear in a given word? If so, what are they? One important question in this area has to do with violations of the Affix Ordering Generalization (Siegel (1974)), sometimes known as "bracketing paradoxes".

A related issue which emerged during our work concerns prefixes, such as *pre#* and *over#*, which apparently ignore the category of their bases. It may be that recursive application of prefixes and suffixes is not the best way to account for such prefixes. We would like to use our data to address this question.

Our data can also be used to investigate the morphological behavior of words which are "zero-derived" or "drifted" from a different major category. Such words are the nouns *considerable*, *accused*, and *beyond* listed in Merriam(1967). Contrary to our goal for GETBASES (to produce a list of morphologically active bases), these words never served as the base for derivational affixation in our data. We conjecture that some mechanism in the grammar prevents them from doing so, and plan to investigate the nature of that mechanism.

Obtaining results from investigations of this type will not only be important for producing a robust word analysis system, it will also significantly contribute to our theoretical understanding of morphological phenomena.

Acknowledgments

We are grateful to Mary Neff and Martin Chodorow, both members of the Lexical Systems project, for ongoing comments on this research. We also thank Paul Cohen for advice on general lexicographic matters and Paul Tukey for advice on statistical analysis methods.

References.

- Allen, J. (1976) "Synthesis of Speech from Unrestricted Text," *Proceedings of the IEEE* 64, 433-442.
- Aronoff, M. (1976) *Word Formation in Generative Grammar*, Linguistic Inquiry Monograph 1, MIT Press, Cambridge, Massachusetts.
- Byrd, R. J. (1983a) "Word formation in natural language processing systems," *Proceedings of IJCAI-VIII*, 704-706.
- Byrd, R. J. (1983b) "On Restricting Word Formation Rules," unpublished paper, New York University.
- Cercone, N. (1974) "Computer Analysis of English Word Formation," Technical Report TR74-6, Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada.
- Chodorow, M. S., R. J. Byrd, and G. E. Heidorn (1985) "Extracting Semantic Hierarchies from a Large On-line Dictionary," *Proceedings of the Association for Computational Linguistics*, 299-304.
- Chomsky, N. and M. Halle (1968) *The Sound Pattern of English*, MIT Press. Cambridge, Massachusetts.
- Hayes, B. (1983) "A Grid-based Theory of English Meter," *Linguistic Inquiry* 14:3:357-393.
- Heidorn, G. E., K. Jensen, L. A. Miller, R. J. Byrd, and M. S. Chodorow (1982) "The EPISTLE Text-Critiquing System," *IBM Systems Journal* 21, 305-326.
- Koskenniemi, K. (1983) *Two-level Morphology: A General Computational Model for Word-form Recognition and Production*, University of Helsinki, Department of General Linguistics.
- Kucera, H. and W. N. Francis (1967) *Computational Analysis of Present-Day American English*, Brown University Press, Providence, Rhode Island.
- Marchand, H. (1969) *The Categories and Types of Present-Day English Word-Formation*, C.H.Beck'sche Verlagsbuchhandlung, Munich.
- Merriam (1963) *Websters Seventh New Collegiate Dictionary*, Merriam, Springfield, Massachusetts.
- Siegel, D. (1974) *Topics in English Morphology*, Doctoral Dissertation, MIT, Cambridge, Massachusetts.
- Winograd, T. (1971) "An A. I. Approach to English Morphemic Analysis," A. I. Memo No. 241, A. I. Laboratory, MIT, Cambridge, Massachusetts.