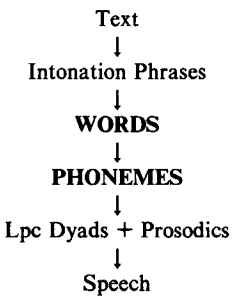
Morphological Decomposition and Stress Assignment for Speech Synthesis

Kenneth Church
Bell Laboratories
600 Mountain Ave.
Murray Hill, N.J.
research!alice!kwc
kwc@mit-mc.arpa

1. Background

A speech synthesizer is a machine that inputs a stream of text and outputs a speech signal. This paper will discuss a small piece of how words are converted to phonemes.



Typically words are converted to phonemes in one of two ways: either by looking the words up in a dictionary (with possibly some limited morphological analysis), or by sounding the words out from their spelling using basic principles.

- Dictionary Lookup
- Letter to Sound

Both approaches have their advantages and disadvantages; dictionary lookup fails for unknown words (e.g., proper nouns) and letter to sound rules fail for irregular words, which are all too common in English. Most speech synthesizers adopt a hybrid strategy, using the dictionary when possible and turning to letter to sound rules for the rest. I discussed letter to sound rules at the last meeting of the ACL [Church]; this paper will report on some new dictionary lookup approaches, with an emphasis on morphology.

Morphological decomposition is used to reduce the size of the dictionary and to increase coverage. Instead of storing all possible words, the system can store just a lexicon of morphemes and save a factor of 10 [Jon Allen (personal communication)] in storage. Now when the system is given a word and asked to determine is pronunciation, the system decomposes the word into known morphemes, looks up the pronunciation of each of the pieces and combines the results.

2. MITalk Decomp

The best known morphological decomposition system is the Decomp module in the MITalk sysnthesizer [Allen et. al.]. This system attempted to parse an input word such as *formally* into morphemes: *form*, *-al* and *-ly*. It was assumed that morphemes are concatenated together (like "beads on a string") according to the finite state grammar shown below:

The types of morphemes were:

1. Prefixes (pref): *UNtie, PERmit, REduce*
2. Suffixes
   a. Derivational (derv): *laxITY, existENCE, softNESS, kingDOM*
   b. Inflectional (infl): *boatING, toastED, coatS, manS'*
3. Roots
   a. Free (root): *sray, squeeze, large*
   b. Absolute (absl): *the, than, but*
   c. Left-Bound (lbrt): *rePEL, conCEIVE*
   d. Right-Bound (rbrt): *CRIMINal, TOLERance*
   e. Strong (root): *women, rang*

Costs were placed on the arcs to alleviate overgeneration. Note that the grammar produces quite a number of spurious analyses. For example, not only would *formally* be analyzed as *form-al-ly* but it would also be analyzed as *form-ally* and *for-mal-ly*. The cost mechanism blocks these spurious analyses by assigning compounding a higher cost than suffixation and therefore favoring the desired analysis. Although the cost mechanism handles a large number of cases, it would be better to aim toward a tighter grammar of morphology which did not overgenerate so badly.

State	Arc	Cost
word-final:	cat infl word-final	64
	cat derv right-sida-a	35
	cat root left-side-a	101
	cat lbrt middle	1091
	cat absl word-initial	1221
right-side-a:	cat derv right-side-a	35
	cat infl word-final	35
	cat rbrt left-side-a	66
	cat root left-side-a	101
	cat lbrt middle	1091
right-side-b:	cat derv right-side-a	963
	cat lbrt middle	2019
	cat infl word-final	992
	cat root left-side-a	1029
	cat rbrt left-side-a	66
middle:	cat pref left-side-a	34
left-side-a:	cat root left-side-a	133
	cat derv right-side-b	67
	cat hyph word-final	1024
	cat infl word-final	1056
	cat lbrt middle	1155
	cat pref left-side-b	34
word-initial:	cat hyph word-final	1024
left-side-b:	cat pref left-side-b	34
	cat derv right-side-a	1027
	cat lbrt middle 2083	
	cat root left-side-a	1093
	cat hyph word-final	1024
	cat infl word-final	1056

The MITalk Decomp program performed its task quite well; it could analyze 95% of running text [Allen (personal communication)]. In order to achieve this level of performance, the authors of Decomp made a conscious decision not to deal with stress alternations (*festive / festivity*), vowel shift and tensing (*divine / divinity*), and other phonological rules associated with latinate morphology. Basically, there was only one rule for combining the pronunciations of morphological pieces: simple concatenation with a few simple rules to account for spelling alternations at the juncture:

- Silent e deletes before a vocalic suffix: *observe* + *ance* → *observance*
- Consonant doubles before a vocalic suffix: *red* + *est* → *reddest*
- y → i before a suffix: *glory* + *ous* → *glorious*
- y deletes before a suffix starting with i: *harmony* + *ize* → *harmonize*

All affixes were assumed to be stress neutral. Words like *festivity* and *divinity* which require a richer understanding of the interaction of morphology and phonology were entered into the lexicon as exceptions.

The decision not to handle more complicated morphological and phonological rules was based on the belief that it is hard to do an adequate job and that it wasn't necessary to do so because the rules are not very productive and hence it is possible (and practical) to list all of the derived forms in the lexicon. I'd like to believe that morphology and phonology have progressed enough over the past ten years that this argument does not have as much force as it did. Nevertheless, I have to admit that the payoff may be marginal, especially if measured in short term savings in the size of the lexicon and memory costs. The real value in the enterprise is more long term; I am betting that pushing the theoretical linguistic understanding with a demanding application such as speech synthesis will uncover some new insights.

3. Types of Morphological Combination

It has long been recognized that "stress-shifting" morphology (e.g., *divin+ity*) differs in quite a number of respects from "stress neutral" morphology (e.g., *divine#ness*). It is a well-established convention to mark the "stress-shifting" morpheme boundary with a "+" symbol and to mark the "stress-neutral" boundary with a "#" symbol. (Scare quotes are placed around "stress-shifting" and "stress-neutral" because these terms are probably not quite right.) This paper will also use the terms *Level 1* and *Level 2* to refer to the two types of morphological combination, respectively. This terminology is taken from the literature on Level Ordered Morphology and Phonology (e.g., [Mohanan]) which argues that "+" boundary (level 1) morphology is ordered before "#" boundary (level 2) morphology and that this ordering dependency has important theoretical implications.

It is worthwhile to review some of the well-known differences between "+" boundaries and "#" boundaries. Informally "+" morphemes such as *in+*, *ad+*, *ab+*, *+al*, *+ity* are (generally) derived from Latin whereas "#" morphemes such as *#ness*, *#ly* come from Greek and German. This historical trend is only a rough correlation and has numerious counter-examples (e.g., the German suffix *-ist* behaves like "+"). The program uses the following set of prefixes and suffixes:

- Level 1 "+" Prefixes: *a, ab, ac, ad, af, ag, al, am, an, ap, ar, as, at, bi, col, com, con, cor, de, dif, dis, e, ec, ef, eg, el, em, en, er, es, ex, im, in, ir, is, ob, oc, of, per, pre, pro, re, suf, sup, sur, sus, trans*
- Level 1 "+" Suffixes: *ability, able, aceous, acious, acity, acy, age, al, ality, ament, an, ance, ancy, ant, ar, arity, ary, ate, ation, ational, ative, ator, atorial, atory, ature, bile, bility, ble, bly, e, ea, ean, ear, edge, ee, ence, ency, ent, ential, eous, ia, iac, ial, ian, iance, iant, iary, iate, iative, ibility, ible, ic, ical, ican, icate, ication, icative, icatory, ician, icity, icize, ide, ident, ience, iency, ient, ificate, ification, ificative, ify, ion, ional, ionary, ious, isation, ish, ist, istic, itarian, ite, ity, ium, ival, ive, ivity, ization, ize, le, ment, mental, mentary, on, or, ory, osity, ous, ular, ularity, ure, ute, utive, y*
- Level 2 "#" Prefixes: *anti, co, de, for, mal, non, pre, sub, supra, tri, ultra, un*

- Level 2 "#" Suffixes: *able, bee, berry, blast, bodies, body, copy, culture, fish, ful, fulling, head, herd, hood, ism, ist, ite, land, less, line, ly, man, ment, mental, mentarian, most, ness, phile, phyte, ship, shire, some, tree, type, ward, way, wise*

There is also a well-known precedence relation between + and #. With very few exceptions, # morphemes nest outside of + morphemes. Thus, we have *non* # [*in* + *moral*] but not **in* + [*non* # *moral*]. The precedence relation yields some subtle (but correct) predictions. Observe that *-able* can be a level 1 affix in some cases (e.g., *cómparable*) and a level 2 affix in others (e.g., *emplóyable*). Notice the contrast between *INcomparable* and *UNexmployable*; the + marked *comparable* takes the + marked prefix *in*+ whereas, in contrast, the # marked *employable* takes the # marked prefix *un#*. This same contrast is brought out by the famous pair: *indivisible / undividable*. (This argument is no longer considered to be as convincining as it once was because of so-called bracketting paradoxes which will be discussed shortly.)

Word formation rules are also sensitive to the difference between + and #. Note that + morphemes can attach to bound morphemes (e.g., *crimin* + *al*), but # morphemes cannot (e.g., **crimin # ness, *crimin # ly, *crimin # hood*). In addition, # morphemes attach more productively than + morphemes.

> "It is clear that *#ness* attaches more productively to bases of the form *Xous* than does *+ity*: *fabulousness* is much "better" than *fabulosity*, and similarly for other pairs (*dubiousness / dubiety, dubiosity*). There are even cases where the *+ity* derivative is not merely worse, but impossible *acrimonious / *acrinoniosity, euphonious / *euphonosity, famous / *famosity*. There is also the simple list test, which is still a good indicator. Walker (1936) lists fewer *+ity* derivatives than *#ness* derivatives of words of the form *Xous*." [Aronoff, pp. 37-38].

Aronoff continues to point out that the semantics of # boundaries tend to be more predictable and compositional than + boundaries. The meaning of *callousness*, for example, is more predictable from the meanings of *callous* and *ness* than the meanings of *variety, notoriety* and *curiosity* are from the meanings of their parts.

The following list summarizes some of the differences between + and #:

- + morphemes are (often) historically correlated with Latin; # with German and Greek
- + morphemes feed certain phonological rules (stress assignment, vowel shift); # do not.
- + morphemes take precedence over #
- + morphemes can attach to bound morphemes; # cannot
- + morphemes are less productive than #
- + morphemes have less predictable semantics than #

The remainder of the paper will be divided into two sections, the first will be concerned with level 1 morphology and the second with level 2 morphology and compounding. Level 1 morphology has been studied more heavily in the lingusitics literature; level 2 is perhaps more important for practical applications, at least in the short term.

4. Morphological Decomposition of Level 1 Affixes

A number of the differences between + and # ought to be relevant in decomposing level 1 affixes and reducing the posibility of spurious derivations. Consider how the first difference mentioned above, historical correlation, could be used to improve a decomposition program. It is very easy, for example, for a decomposition program to decide erroneously that *acclamation* is derived from *clam*, meaning roughly *the result of having been clammed up*. If the program could somehow split the Latinate and non-Latinate vocabularies, then the program could know that *-ation* cannot be attached to *clam* because *clam* is not Latinate. The program accomplishes this by maintaining a short list of words marked with an ad hoc feature [−Latinate].

The program might perform even better if the Latinate vocabulary were split still further. Consider, for example, the split between words ending with *-ent* and those ending with *-ant*. The first class are likely to have variants ending with *-ence* and *-ency* and the second are likely to have variants ending with *-ance* and *-ancy*. It seems extremely implausible for an *-ent* word such as *president* to take an *-ant* suffix: **presidant, *presidance, *presidancy*. Thus, it would be desirable to partition the Latinate vocabulary into quite a number of subsets, each with different possibilities for suffixation. But how do we do this without assigning *ad hoc* features such as [+Latinate], [+ent], [+ant], [+Declension 1], [+Declension 2], etc.?

Not only is the feature approach *ad hoc*, but it also missing an important asymmetry. Note that most words ending with *-ency* (e.g., *presidency*) are derived from words ending with *-ent* (e.g., *president*), and crucially not the other way around. The intuition that the relation "derived from" is asymmetric has some distributional support: notice that the percentage of words ending in *-ency* which are morphologically related to words ending in *-ent* is much larger than the percentage of words ending in *-ent* which are related to words ending in *-ency*. (The program estimates these percentages to be 73% (36/49) and 5% (36/710), respectively, using a procedure described below.) This asymmetry is problematic for a concatenation model like MITalk's Decomp, which would place *presidency* and *president* on equal footing, deriving both from *preside*.

Aronoff-style [Aronoff] truncation rules provide an attractive mechanism for accounting for the asymmetry. Recall that Aronoff proposed that *nominee* be derived from *nominate* by truncating the *-ate* suffix and attaching *-ee* in a single step. These truncation rules were necessary for him so that he could maintain his *Word Based Hypothesis*. The Word Based Hypothesis claims that words are formed from other words (possibly via truncation) and not from bound morphemes. Thus, in Aronoff's theory, there is no bound morpheme *nomin-*; there are only words (e.g., *nominate* and *nominee*). The generalizations that would be attributed to *nomin-* in other

theories are captured in Aronoff's system by his truncation rules.

The program uses truncation rules to capture the symmetry in the 'derived from' relation by permitting *-ent* to be truncated before *-ency*, but not the other way around. Thus, *presidency* is derived from *president* − *-ent* + *-ency*, and *president* is not derived from *presidency* because does not truncate *-ency* before *-ent*. Truncation rules are subject to a number of constraints. In particular, truncation is only found at level 1; truncation cannot apply at level 2 because, as mentioned above, level 2 affixes attach to words, not bound (= truncated) morphemes.

How does the program decide which suffixes can be truncated and when? Let me introduce the notation *-ency* > *-ent* to mean (roughly) that words ending with *-ency* are likely to be derived from words ending with *-ent*. The precise status of the '>' relation should be to be explored more fully. In some cases, the relation is a necessary condition; if *presidency* is derived from an English word then it must be derived from *president*. In other cases, the relationship expresses a possibility but not a necessity. For example, words ending in *-ation* may be related to words ending in *-ate*, but not necessarily. Marchand describes the relation as follows:

> "The English vocabulary has been greatly enriched by borrowings, chiefly from Latin and French. In course of time, many related words which had come in as separate loans developed a derivational relation to each other, giving rise to derivative alternations. Such derivative alternations fall into three main groups.
>
> Group A is represented by the pairs 1) *-acy* / 2) *-ate* (as *piracy* ~ *pirate*), 1) *-ancy, -ency* / 2) *-ant, ent* (as *militancy* ~ *militant, decency* ~ *decent*), 1) *-ization* / 2) *-ize* (as *civilization* ~ *civilize*), 1) *-ification* / 2) *-ify* (as *identification* ~ *identify*), 1) *-ability* / 2) *-able* (as *respectibility* ~ *respectible*), 1) *-ibility* /2) *-ible* as (*convertibility* ~ *convertible*), 1) *-ician* / 2) *-ic(s)* (as *statistician* ~ *statistics*), 1) *-icity* / 2) *-ic* (as *catholicity* ~ *catholic*), 1) *-inity* / 2) *-ine* (*salinity* ~ *saline*).
>
> If 1) is a derivation from an English word, the only possible word is 2), ie., if *piracy* is a derivative from an English word, only *pirate* is possible. The statement does not imply that for every 1) there must be a 2). 1) may be a loan, or it may be formed on a Latin basis without any regard to the existence of an English word at all (*enormity*, for instance, is so coined). Nor does the derivational principle involve the existence of a 1) for every 2) (many words in *-able* or *-ine* are not matched by words in *-ability* resp. *-inity*).
>
> Group B is represented by the pairs 1) *-ation* / 2) *-ate* (as *creation* ~ *create*), 1) *-(e)ry* / 2) *-er* (as *carpentry* ~ *carpenter*), 1) *-eress* / 2) *-erer* (as *murderess* ~ *murderer*), 1) *-ious* / 2) *-ion* (as *ambitious* ~ *ambition*, 1) *-atious* / 2) *-ation* (as *vexatious* ~ *vexation*).
>
> If 1) is a derivative from another English word, the derivational pattern 1) from 2) is possible, but not necessary. A derivative in *-ation* such as *reforestation* is connected with *reforest*, a derivative such as *swannery* is connected with *swan*, *archeress* is connected with *archer*, *robustious* is extended from *robust* (but otherwise an adj in *-tious* derived from a sb points to the sb ending in *-tion*, i.e. we have really type A).
>
> Group C is nothing but a variant of A and concerns adjs in *-atious* as *flirtatious*. Originally deriving from sbs in *-ation*, the type is now equally connected with the unextended radical, i.e. *flirt* (the older derivation *ostentatious* 1658 has not entered this latter derivational connection)." [Marchand, pp. 165-166]

For pragmatic purposes, the program assumes that there is only one '>' relation, not three as Marchand suggests, and that the relation can be estimated statistically as follows:

$$Probability(suffix_1 > suffix_2) = \frac{number\ of\ words\ ending\ with\ both\ suffix_1\ and\ suffix_2}{number\ of\ words\ ending\ with\ suffix_1}$$

The program estimates, for example, that *-ency* > *-ent* with a probability of 73% (36/49) and that *-ent* > *-ency* with a probability of 5% (36/710). The 36 words ending in *ency* which have a variant ending in *-ent* are: *incumbency, complacency, indecency, excrescency, residency, presidency ascendency, dependency, independency, superintendency, despondency, exigency contingency, emergency, detergency, insurgency, deficiency, efficiency sufficiency, proficiency, expediency, clemency, permanency, transparency vicegerency, belligerency, currency, competency, prepotency, consistency inconsistency, frequency, delinquency, constituency, solvency* and *fervency*. The estimate should be almost 100%; the program believes that *decency, cadency, tendency, ambitendency, pudency, agency, regency, urgency, counterinsurgency, valency, patency, potency, and fluency* are not derived from *-ent*. Most of the errors can be attributed to a heuristic which excludes short stems (e.g., *ag-*) on the grounds that these stems are often spurious. These errors could be fixed by ammending the heuristic to check a 'winners list' of one, two and three letter stems. Some of the other errors are due to accidental gaps in the dictionary.

The results of this statistical estimation are shown in the figure below (where -0 denotes the null suffix):

-ability	-able (43%), -ate (29%)
-able	-0 (24%), -ation (18%), -ate (17%), -e (14%), -al (6%), -y (3%), -ion (2%), -ity (2%), -ous (2%), -ent (1%), -ive (1%)
-aceous	-0 (19%), -e (7%), -ate (7%), -ation (4%), -y (4%), -ous (4%), -al (3%), -ary (3%), -ic (3%)
-acity	-acious (38%)
-acy	-ate (42%), -ation (18%), -al (13%), -e (8%)
-age	-0 (51%), -y (13%), -e (12%), -al (5%), -ate (4%), -ation (4%), -able (4%), -on (4%), -ion (3%), -le (3%), -ic (3%), -ar (2%), -or (2%), -ial (2%)
-al	-0 (17%), -e (7%), -ic (2%), -y (2%), -on (1%), -le (1%)
-ality	-al (76%), -0 (19%), -ate (13%), -e (9%), -ation (7%), -ary (5%), -ous (5%), -able (4%), -ative (4%)
-ament	-0 (38%), -ate (29%)
-an	-0 (6%), -e (2%), -al (2%), -ous (1%), -y (1%), -on (1%), -ate (1%), -ation (1%)
-ance	-ant (30%), -0 (26%), -e (15%), -ate (10%), -able (9%), -ation (9%), -or (7%), -al (4%), -ous (4%), -ion (4%), -ative (3%), -ive (3%), -y (3%)
-ancy	-ant (40%), -0 (19%), -ation (12%)

-ant	-ate (27%), -ation (21%), -0 (21%), -e (11%), -able (9%), -y (5%), -al (5%), -ous (5%), -ion (4%), -ent (3%), -ity (3%), -or (3%), -ive (2%), -an (1%), -ar (1%), -ic (1%), -ize (1%), -on (1%)
-ar	-ate (13%), -e (9%), -ation (7%), -0 (6%), -ous (2%), -y (2%), -able (1%), -al (1%), -ite (1%)
-arity	-ar (63%), -ate (26%), -ation (22%), -0 (13%)
-ary	-0 (25%), -al (13%), -ate (10%), -e (8%), -ation (8%), -ar (6%), -ous (4%), -y (4%), -able (3%), -ion (3%), -ic (2%), -ity (2%), -ize (2%), -ant (2%), -or (2%)
-ate	-0 (13%), -e (9%), -al (8%), -ic (4%), -y (3%), -on (1%), -le (1%), -ion (0%)
-ation	-ate (42%), -e (21%), -0 (18%), -al (9%), -y (3%), -ous (3%), -ion (1%), -ic (1%), -on (1%)
-ational	-ation (40%), -e (25%)
-ative	-ation (56%), -ate (42%), -e (19%), -0 (17%), -able (17%), -ant (12%), -al (9%), -y (5%), -ity (4%), -ous (3%), -ance (3%)
-ator	-ate (61%), -ation (48%), -ant (18%), -ative (18%), -able (18%), -e (15%), -al (9%), -0 (7%), -ar (6%), -ity (5%), -ous (4%), -ary (4%), -on (4%)
-atorial	-ation (37%), -ator (26%), -atory (26%)
-atory	-ation (63%), -ate (46%), -e (21%), -ative (20%), -ator (16%), -able (15%), -0 (13%), -ant (11%), -al (7%), -ar (4%)
-ature	-ate (26%), -0 (21%), -ation (18%)
-bility	-ble (62%), -on (14%)
-ble	-on (5%), -0 (3%), -le (1%)
-bly	-ble (73%)
-e	-0 (4%)
-ee	-0 (28%), -e (13%), -or (11%), -y (6%), -ation (6%), -ment (5%), -ate (5%), -ant (3%), -al (3%), -ion (3%), -able (3%)
-ence	-ent (54%), -e (18%), -0 (15%), -ment (3%)
-ency	-ent (73%), -ence (24%), -e (14%), -0 (12%)
-ent	-0 (6%), -e (6%), -y (1%), -ate (1%), -al (1%), -ation (1%)
-ential	-ence (59%), -ent (59%), -0 (26%), -e (20%)
-eous	-e (5%), -y (4%), -0 (3%), -ic (3%), -ous (3%), -ate (3%), -on (2%)
-ia	-ic (14%), -0 (7%), -y (7%), -e (4%), -ous (2%), -al (1%), -ate (1%)
-iac	-ia (44%), -ic (19%)
-ial	-0 (26%), -y (15%), -e (5%), -ate (3%), -al (2%), -ic (2%), -ize (2%)
-ian	-0 (23%), -y (14%), -ic (7%), -al (6%), -e (4%), -ize (3%), -ia (3%), -ity (3%), -ium (3%)
-iant	-iate (27%)
-iary	-ial (25%), -0 (22%), -e (22%)
-iate	-ial (13%), -e (9%), -0 (7%), -ate (6%), -ium (6%), -ia (5%), -ious (5%)
-iative	-iate (70%)
-ibility	-ible (73%), -ive (45%)
-ible	-ion (25%), -ive (22%), -0 (20%), -e (12%), -or (10%), -ent (7%), -able (5%), -ory (5%), -ence (4%), -al (4%), -y (4%)
-ic	-e (18%), -y (14%), -0 (12%)
-ical	-y (55%), -ic (11%), -0 (8%), -ize (8%), -e (6%), -ist (6%), -al (2%), -ate (2%)
-icate	-ication (26%), -ic (17%), -icity (15%), -e (14%), -y (11%), -0 (7%), -ical (7%)
-ication	-y (66%), -ic (14%), -e (9%)
-icative	-ication (50%), -icate (38%), -y (38%)
-icatory	-ication (50%), -y (43%), -icate (36%)
-ician	-ic (61%), -ical (32%), -0 (16%), -e (13%), -y (13%)
-icity	-ic (63%), -e (18%), -0 (16%), -y (12%), -ical (10%), -ize (8%), -al (7%), -ication (7%)
-icize	-ic (71%)
-ide	-ate (8%), -ic (8%), -0 (7%), -ite (6%), -e (4%), -on (3%), -ous (3%), -al (3%), -ize (3%), -age (2%), -ium (2%)
-ience	-ient (40%)
-iency	-ient (100%)
-ient	-e (11%), -0 (10%)
-ification	-ify (71%), -0 (22%), -e (18%), -ity (16%), -y (16%), -ic (11%)
-ify	-0 (25%), -e (15%), -ic (15%), -y (15%), -ity (13%), -al (11%), -ate (9%), -ion (7%), -ite (6%), -ize (5%), -or (5%), -ar (4%), -ary (4%), -ical (4%)
-ion	-e (31%), -0 (15%), -ic (1%), -y (1%), -al (1%)
-ional	-ion (57%), -ive (21%), -0 (18%), -e (18%), -or (11%)
-ionary	-ion (87%), -e (30%), -0 (26%), -ive (26%)
-ious	-y (15%), -ity (13%), -ion (10%), -0 (9%), -e (9%), -ial (6%), -ium (5%), -ic (4%), -ate (3%), -ive (3%), -ist (2%)
-isation	-ization (93%), -ize (70%), -0 (53%), -ity (33%), -ist (27%), -ic (20%), -e (17%)
-ish	-0 (27%), -e (11%), -y (7%), -le (2%), -ic (2%)
-ist	-0 (40%), -ic (19%), -ize (18%), -y (18%), -e (14%), -al (6%), -ity (5%), -ation (3%), -ate (2%), -able (1%), -ion (1%)
-istic	-ist (46%), -ize (29%), -0 (27%), -e (17%), -ic (15%), -ity (13%), -y (13%), -al (10%)
-itarian	-ity (57%), -ize (43%), -0 (36%), -e (36%)
-ite	-0 (13%), -ic (11%), -e (6%), -ate (6%), -ous (6%), -y (2%), -ia (2%), -on (2%), -al (1%), -able (1%), -ity (1%), -ation (1%), -ion (1%), -or (1%)
-ity	-0 (37%), -e (24%), -ous (6%), -ate (5%), -al (4%), -ation (3%), -y (2%), -ion (1%), -ic (1%)
-ium	-ic (11%), -0 (8%), -ial (6%), -y (6%), -ia (6%), -e (6%), -ite (5%), -ate (4%), -ous (4%), -al (2%), -on (2%), -ion (2%), -ize (2%), -ist (2%)
-ival	-ive (47%)
-ive	-ion (59%), -e (26%), -0 (22%), -al (1%), -y (1%), -ation (1%)
-ivity	-ive (66%), -ion (61%), -0 (39%), -or (32%), -ance (14%), -e (14%), -ible (11%)
-ization	-ize (75%), -0 (59%), -ity (31%), -ist (25%), -ic (22%)
-ize	-0 (47%), -ic (17%), -ity (17%), -y (14%), -e (12%), -ous (6%), -ate (4%), -al (4%), -ite (2%), -ation (1%), -ia (1%)
-le	-0 (11%), -y (3%), -e (3%), -on (2%), -ic (1%)
-ment	-0 (63%), -able (6%), -e (4%), -ation (4%), -or (3%), -ant (2%), -ate (2%), -ble (2%)
-mental	-ment (77%), -0 (20%)

-mentary	-ment (56%)
-on	-0 (4%), -e (2%), -ic (2%), -y (1%)
-or	-ion (30%), -e (27%), -0 (22%), -ive (16%), -ation (3%), -able (3%), -y (2%), -al (2%), -ate (2%), -ent (1%), -le (1%)
-ory	-ion (56%), -e (34%), -ive (21%), -or (20%), -0 (11%)
-osity	-ous (65%), -0 (15%), -al (12%), -ate (11%), -e (11%)
-ous	-0 (13%), -ic (7%), -ate (6%), -e (6%), -y (4%), -al (4%), -on (2%)
-ular	-le (31%), -0 (4%), -e (4%), -ate (4%)
-ularity	-ular (67%), -le (28%)
-ure	-0 (21%), -e (15%), -ion (11%), -or (8%), -ive (4%), -al (2%)
-ute	-e (8%)
-utive	-ute (67%)
-y	-0 (19%), -e (6%)

The decomposition program uses the table above to decide which suffixes can be truncated and when. Consider the word *presidency*. The program notices that this word ends in *-ency* so it looks in the table and discovers that *-ency* alternates with *-ent* (73%), *-ence* (24%), *-e* (14%) and *-0* (12%). The program tries to replace the *-ency* with each of these sequentially until it finds a word in the dictionary. In this case, it will succeed on the first try when it replaces *-ency* with *-ent* and finds that the result *president* is a word in the dictionary.

Level 1 prefixes are processed through an analogous procedure, so that *effect*, for example, is derived from *defect* by truncating the *ef-* prefix and adding the prefix *de-*. The truncation mechanism is not generally employed by most authors for prefixing, and it may be a mistake to do so, but I used it anyways, mostly because it was available and filled a practical need.

The resulting decomposition program has been used to construct a forest of related words as illustrated below:

```
(38 port
    (aport)
    (comport (comportment))
    (deport (deportation) (deportee)
            (deportment))
    (disport)
    (export (exportation) (reexport))
    (import (important (importance))
            (importation) (reimport))
    (portable)
    (portage)
    (portal)
    (portative)
    (portent (portentous))
    (portion
     (apportion (apportionment)
                (reapportion (reapportionment)))
     (proportion (disproportion
                  (disproportionate
                   (disproportionation)))
                 (proportional)
                 (proportionate)))
    (report (reportage))
    (transport (transportation)))
```

```
(36 infect
    (affect (affectation)
            (affection (affectionate))
            (affective (affectivity))
            (disaffect))
    (confect (confection) (confectionary))
    (defect (defection) (defective)
            (effect (effective (ineffective))))
    (disinfect (disinfectant))
    (infection)
    (infectious)
    (infective)
    (refect (perfect (imperfect (imperfection)
                                (imperfective))
                     (perfection (perfectionist))
                     (perfective (perfectible)))
            (prefect (prefecture))
            (refection)
            (refectory (prefectorial))))
```

The forest was constructed by applying the decomposition procedure to every word in the dictionary and then indexing the results to show which forms were derived from which stems. Thus 38 words were found to be related to the stem *port* and 36 words were found to be related to *infect*. These results seems extremely promising; most of the relations appear to agree very closely with intuition.

Now that we have a fairly accurate method of decomposing words at level 1, how can this be put to practical use? For assigning stress, it would be useful to know the weight of the syllables in the stem. This is particularly necessary before so-called weak retraction suffixes (e.g., *-ent, -ant, -ence, -able, ance, al, ous, ary*). General principles of stress retraction (e.g., [Liberman and Prince]), predict strong retractors (e.g., *-ate, -ation*) always back the stress up regardless of syllable weight (*degrâde / dègradation*), whereas weak retractors do so only if the preceding syllable is light (*refer / référent* with a light syllable before *-ent*, as opposed to (*cohére / cohérent* with a heavy syllable before *-ent*).

Given syllable weight, it is relatively well-understood how to assign stress. A large number of phonological studies (e.g., [Chomsky and Halle], [Liberman and Prince], [Hayes]) outline a deterministic procedure for assigning stress from the weight representation and the number of extrametrical syllables (1 for nouns, 0 for verbs). A version of this procedure was implemented by Richard Sproat last summer, and was discussed at the last ACL meeting [Church].

It it generally believed that syllable weight is derivable from underlying vowel length and the number of consonants, but if one is trying to assign stress from the spelling, it can be difficult to know the vowel length and the number of consonants. The fact that *inhence* has a heavy penultimate syllable and that *inference* has a light penultimate syllable is extremely difficult to determine from the spelling. It would be considerably easier if syllable weight (or some correlate thereof such as vowel length) were marked in a lexicon of stems, so that the program could determine syllable weight by decomposing a word into its peices, look them up in a morpheme lexicon, and then re-combine the results appropriately.

Not only is it convenient for practical application to assume that stems are marked in the lexicon for syllable weight, but it may be necessary for linguistic reasons as well. Consider the stress alternation *confide / confidence*. This alternation is problematic because the *i* in *confide* seems to be underlyingly long whereas the *i* in *confidence* seems to be underlyingly short, and yet, the

two stems ought to share the same underlying form since the two words are morphologically related to one another. The solution to the confidence puzzle, I believe, is to say that the stem *-fide* is marked in the lexicon as underlyingly light at least with respect to stress retraction (and to account for the tense vowel in *confide* in some other way [Church (forthcoming)]).

The table below is presented as evidence that the confidence alternation is determined, at least in part, by some sort of lexical marking on stems. Note, for example, that *-fer, -cel, -side,* and *-fide* words display the confidence alternation, but *-here, -pel,* and *-pose* words do not.

alternation	refer	reference		
	confer	conference		
	infer	inference		
	defer	deference		
	excel	excellent	excellence	excellency
	reside	resident	residency	
	preside	president	presidency	
	confide	confident	confidence	confidency
no alternation	adhere	adherent	adherence	adhesive
	cohere	coherent	coherence	cohesive
	inhere	inherent	inherence	inhesion
	expel	expellent	expellant	
	repel	repellent		
	propel	propellent	propellant	
	expose	exposal	exposure	expository
	dispose	disposal	disposure	dispository
	propose	proposal		
	compose		composure	

Assume the lexicon divides stems into at least two classes:

- Retraction Class I Stems (light): *-fer, -cel, -side, -fide, -main, -vail, -note, -cede, -pete, -pair, -pare*
- Retraction Class II Stems (heavy): *-here, -pel, -pose, -hale, -pale, -grade, -vade, -flame, -suade, -place, -plore, -void, -clude, -prove, -sume, -fuse, -duce*

where class I stems show stress alternations before weak retracting suffixes and class II stems do not.

This concludes what I wanted to say about level 1 decomposition. In summary, this section presented Aronoff-style truncation rules as an alternative to MITalk-style concatenation rules. Truncation rules have the advantage that they preserve the asymmetry in the 'derived from' relation, and that they correctly partition the lexicon into classes such as [+ent] and [+ant] without introducing unnecessary *ad hoc* features such as [+ent] and [+ant]. Some results of the new decomposition procedure were presented, and they seem to agree very closely with intuition. It was suggested that the decomposition procedure could be used in stress assignment, by decomposing words into morphemes, look up the syllable weight of the pieces in a morpheme lexicon, and then recombine the results appropriately. This last suggestion has not yet been fully implemented.

5. Level 2 and Compounding

Most of the linguistic literature deals with level 1 where we find extremely interesting stress alternations and vowel shifts and so forth. Generally speaking, the phonology of level 2 and compounding is believed to be relatively fairly straightforward. Something iike the simple concatenation model in decomp is not a bad first approximation. In fact, I believe the stress of level 2 and compounding is more interesting than has generally been thought. In particular, I am beginning to believe that level 2 affixes are not stress neutral at all, but rather they stress as if they were parts of compounds. Note that *under-, anti-* and *super-* follow the general compound pattern where stress is assigned the to the left member in nouns and to the right in verbs and adjectives.

Noun	Verb	Adjective
únderdog	undergó	underáge
ántifreeze		antisócial
súpermarket	superimpóse	supersónic

6. Are Level 2 Affixes Really Stress Neutral?

It might be possible to extend this position to its logical extreme and say that all level 2 affixes stress like compounds, and thus completely do away with the concept of stress neutral affixes.

- *Compound Theory*: (All) Level 2 affixes are stressed just like compounds; they receive main stress on the left in nouns and main stress on the right in verbs and adjectives.
- *Stress Neutral Theory*: (At least some) Level 2 affixes are stress neutral; they are simply concatenated onto the stem (a là MITalk's Decomp).

The compound theory has much to recommend it. Indeed most level 2 prefixes are like *under-, anti-* and *super-* and show the compound stress pattern (stress on the left when nominal and on the right when verbal/adjectival). These prefixes cannot be accounted for easily under the stress neutral theory. The main support for the stress neutral theory seems to come from prefixes like *un-* which (almost) never take the main stress. However, *un-* can also be accounted for under the compound theory by noting that *un-* forms adjectives and verbs, and therefore main stress would fall on the right.

Admittedly, there are a number of nominal compounds like *pro-life* and *anti-abortion* which take right stress, presumably because the semantics of the left member takes on a semi-adjectival status. Notice, for example, that the word *antimatter*

has two stress patterns, one with main stress on the left and one with main stress on the right, just like well-known compound *blackboard*. With left stress, the compound takes non-compositional semantics and with right stress the compound has a more compositional meaning. These facts suggest that the compound theory can be maintained to acocunt for cases like *pro-life*, but only if the compound stress rules are refined take the semantic facts into account.

Level 2 suffixes provide additional support for the compound theory. Consider suffixes like *ment, hood, ship* and *ness* which appear to support the the stress neutral theory because they never receive main stress. But, they can also be accounted for under the compound theory because they form nouns, and therefore the main stress would be expected to fall on the left. Moreover, consider the level 2 adjectival suffixes *-istic* and *-mental*.[1] These suffixes refute the stress neutral theory because they take the main stress, but they are no problem for the compound stress theory which predicts that adjectivial compounds should receive main stress on the right.

7. The Super-Puzzle and Compound Stress

In attempting to include prefixes as a subcase of compound stress, I did stumble over a very interesting problem in the theory of compound stress. Consider the contrast between *súperconductor* and *sùperconductívity*. Although both compounds are nominal, the first takes primary stress on the left member and the second takes stress on the right member. Upon further investigation, it appears than many compounds ending with level 1 suffixes (e.g., *-ity, -ation*) take primary stress on the right member. For example, here is a breakdown of compounds ending with the letters *ion*. Note the strong tendency for primary stress to end up on the right member.[2]

- Left-Dominant: *intersession, outstation, midsection*
- Right-Dominant: *intercommunion, supervision, anteversion, intercession, supersession, intermission, echolocation, intercolumniation, contravallation, overpopulation, interlunation, intermigration, overcompensation, aftersensation, superfetation, superelevation, interaction, intersection, contradistinction, superinduction, superconduction, underproduction, contraposition, superposition, interposition, postposition, interlocution, counterrevolution*
- Neither: *tourbillion, interrogation, foreordination, redintegration forestation, electrodeposition*[3]

Thus, it appears that compounds ending with a level 1 suffix take right stress. If correct, however, the generalization is a puzzle for the level ordering hypothesis, which assumes that the stress rules of level 1 are opaque to the stress rules of level 2 and compounding. In other words, level ordering suggests a structure like *super[conductivity]* where level 1 takes precedence over level 2 and compounding, but stress assignment requires a different structure *[superconductive]ity* where the compound stress rule applies before the level 1 suffix is analyzed.

In this sense, words like *superconductivity* are very much like the well-known bracketing paradox *ungrammaticality*, where level ordering suggests one structure *un[grammaticality]* (*un#* is a level 2 prefix which must scope outside of *+ity* with is a level 1 prefix) and syntactic/semantic interpretation (LF) requires another *[ungrammatical]ity* (*un#* attaches to adjectives and not to nouns). Note that stress assignment seems to side with the syntactic/semantic arguments in suggesting a left branching structure that violates level ordering.

A solution to these bracketting paradoxes becomes apparent when we consider nominal Greek compounds like *psychobiology* with three or more morphemes. Notice that these compounds systematically take main stress on the middle morpheme.

> aeroneurosis, aerothermodynamics, astrobiology, astrogeology, astrophotography, autobiography, autohypnosis, autoradiograph autoradiography, biogeography, biophysicist, biotechnology, chromolithograph, chromolithography, chronobiology, cryobiology, diageotropism, electroanalysis, electrocardiogram, electrocardiograph, electrodialysis, electrodynamometer, electroencephalogram, electroencephalograph, electroencephalography, electrophysiology, endoparasite, epidiascope, geochronology, geomorphology, heterochromatin, heterochromosome, histopathology, hypnoanalysis, magnetohydrodynamics, metaphysicist, metapsychology, microanalysis, microbarograph, microbiology, micrometeorology, micropaleontology, microparasite, microphotograph, microphotography, multivibrator, myocardiograph, neoorthodoxy, neuropathology, neurophysiology, orthohydrogen, otolaryngology, paleoethnobotany, parahydrogen, parapsychology, photochronograph, photoelectrotype, photogeology, photolithograph, photolithography, photomicrograph, photopolymer, phototelegraphy, phototypography, photozincograph, photozincography, pneumoencephalogram, pneumoencephalography, psychoanalyse, psychoanalysis, psychoanalyze, psychobiology, psychoneurosis, psychopathology, psychopharmacology, psychophysiology, radioautograph, radiobiology, radiomicrometer, radiotelegram, radiotelegraph, radiotelegraphy, radiotelemetry, radiotelephone, radiotelephony, semidiameter, semiparasite, spectroheliograph, spectrophotometer, stereoisomer, stereoisomerism, telephotography, telespectroscope, telestereoscope, teletypewriter, thermobarograph, thermobarometer, ultramicrometer, ultramicroscope, ultramicroscopy

Assume that compounds take stress on the right member when it is branching (bi-morphemic). Thus, *psycho[biology]* takes main stress on the *biology* because it is branching.

Let me suggest further that this same sort of explanation might carry over to explain the stress in the bracketting paradoxes such as *superconductivity* and *ungrammaticality* where I claim that the right piece is 'branching' in order to account for the fact that main stress ends up on the right half.[4] Note that I am

1. These suffixes cannot be level 1, because they don't force the secondary stress to fall two syllables before the main stress: **dèpartméntal* (cf., *dègradátion*).

2. None of the left dominant words above end in the suffix *+ion*. Note, for example, the contrast between *inter=session* and *inter=céss+ion*. The left dominant case does not end in the suffix *+ion*; the right dominant case does.

3. Almost all of these exceptions are due to errors in morphological decomposition algorithm. *Tour # billion, inter # rogation, fore # station,* and *electrode # position* are all incorrect analyses. It is highly unusual for the algorithm to make this many mistakes.

using the lexical category prominance rule in order to let one bit of information [+branching] pass through the opacity imposed by level ordering.

8. Conclusion

Two new ideas in machine morphological decomposition were presented. The discussion of level 1 proposed the application of Aronoff-style truncation rules as an effective means to capture the asymmetry in the 'derived from' relation. Secondly, the discussion of level 2 proposed ideas from the literature on compound stress as an alternative to the stress neutral approach taken in MITalk's Decomp.

References

Aronoff, M., *Word Formation in Generative Grammar*, MIT Press, Cambridge, MA., 1976.

Allen, J., Carlson, R., Granstrom, B., Hunnicutt, S., Klatt, D., Pisoni, D., *Conversion of Unrestricted English Text to Speech*, incomplete draft, underground press, 1979.

Chomsky, N., and Halle, M., *The Sound Pattern of English*, Harper and Row, 1968.

Church, K., *Stress Assignment in Letter to Sound Rules for Speech Synthesis*, in Proceedings of the Association for Computational Linguistics, 1985.

Church, K., *The Confidence Puzzle and Underlying Quantity*, forthcoming.

Hayes, B., *A Metrical Theory of Stress Rules*, Ph.D. Thesis, MIT, 1980.

Liberman, M., and Prince, A., *On Stress and Linguistic Rhythm*, Linguistic Inquiry 8, pp. 249-336, 1977.

Marchand, H., *The Categories and Types of Present-Day English Word-Formation*, University of Alabama Press, 1969.

Mohanan, K., *Lexical Phonology*, MIT Doctoral Dissertation, available for the Indiana University Linguistics Club, 1982.

4. The problem is to define 'branching' so that it gets the right results. I don't want to say that *superconductor* is branching, because that would incorrectly predict main stress on *conductor*. I don't know how to define branching to achieve the desired results, though I believe that this approach is extremely promising.