# The role of selectional restrictions, phonotactics and parsing in constraining suffix ordering in English

INGO PLAG

## 1. INTRODUCTION[1]

At least since Siegel (1974) there has been a debate about the principles and mechanisms that constrain the combinatorial properties of affixes, in particular of English suffixes. A classic example to illustrate the problem of combinatorial restrictions is given in (1):

(1) a. átom, atóm-ic, àtom-íc-ity
    b. átom, átom-less, átom-less-ness
    c. *atom-less-ity

While in (1a) and (1b) -*ic* and -*ity*, and -*less* and -*ness* happily combine, the combination -*less*-*ity* is impossible. Similar sets of examples can be found in abundance and pose the question what exactly is responsible for such patterns. Basically three types of answers have been given in the pertinent literature. One group of scholars argues for the existence of lexical strata (e.g. Siegel 1974, Allen 1978, Selkirk 1982, Kiparsky 1982, Mohanan 1986, Giegerich 1999) with strong restrictions holding between the different strata. This view is disputed, for example, in Fabb (1988) and Plag (1996, 1999), who claim that it is selectional restrictions of individual suffixes that are responsible for the combinatorial properties of suffixes. Finally, and most recently, Hay (2000) has proposed an attractive psycholinguistic

[1] I am grateful to Harald Baayen, Geert Booij and Rochelle Lieber for their feedback on an earlier version of this paper. Special thanks go to Jennifer Hay for critical discussion and constructive comments.

model of morphological complexity which also makes interesting predictions on stacking restrictions.

In this paper I want to test a number of these predictions that emerge from Hay's model of morphological complexity in order to see how far a purely psycholinguistically oriented account can take us. Although Hay's work must be commended for offering intriguing insights into the nature of affixation, it will be shown that with regard to suffix combinations the model's predictions are not always born out by the facts and that further, strictly grammatical, restrictions need to be taken into account. Thus, combinatorial restrictions cannot be reduced to parsing restrictions, as claimed by Hay (2000: 23, 236).

The article is structured as follows. Section 2 ('Earlier models') will briefly review earlier approaches to stacking restrictions, section 3 ('Complexity-based ordering: Hay (2000)') summarizes the main points of Hay (2000), which will be under close scrutiny in section 4 ('Testing complexity-based ordering'). Section 5 summarizes and discusses the results.

## 2. EARLIER MODELS

Until recently the debate on stacking restrictions was characterized by two opposing views. Proponents of stratum-oriented models (e.g. Siegel 1974, Allen 1978, Selkirk 1982, Kiparsky 1982, Mohanan 1986) assume that most, if not all combinatorial restrictions among English suffixes can be explained by the fact that these suffixes belong to different lexical strata and that these strata interact phonologically and morphologically in intricate ways. This is known as level-ordering, which in turn is part of most models of Lexical Phonology.[2] According to the level-ordering hypothesis, English suffixes and prefixes belong to the following classes or strata:

[2] As pointed out by Booij (1994), the main insight of Lexical Phonology is that phonology and morphology work in tandem. This is logically independent of the idea of level ordering. What concerns us here is the level ordering hypothesis. In what follows the discussion of level ordering will focus on suffixes, since this is the topic of the present article.

(2) Class I suffixes: +ion, +ity, +y, +al, +ic, +ate, +ous, +ive, +able, +ize
Class I prefixes: re+, con+, de+, sub+, pre+, in+, en+, be+
Class II suffixes: #ness, #less, #hood, #ful, #ly, #y, #like, #ist, #able, #ize
Class II prefixes: re#, sub#, un#, non#, de#, semi#, anti#

(from Spencer 1991:79)

The suffixes belonging to one stratum share a number of properties that distinguish them from the suffixes of the other stratum. Stratum 1 suffixes tend to be of foreign origin ('Latinate'), while stratum 2 suffixes are mostly Germanic. Stratum 1 suffixes frequently attach to bound roots and tend to be phonologically and semantically less transparent than stratum 2 suffixes. Stratum 1 suffixes cause stress shifts, resyllabification, and other morphonological alternations, stratum 2 suffixes do not. Stratum 1 suffixes are less productive and less semantically compositional than stratum 2 suffixes, and, crucially, stratum 1 suffixes do not occur outside stratum 2 suffixes. Thus, suffixes can only combine in such a way that they attach to suffixes of the same stratum or of a lower stratum. This is perhaps the most important generalization concerning suffix combinations that emerges from stratum models, since impossible combinations such as those in (1c) are ruled out on principled grounds.

However, there are serious problems with this approach. One major theoretical weakness of level ordering is that the two strata are not justified on independent grounds. In other words, it is unclear what is behind the distinction between the two strata, and which property makes a suffix end up on a given stratum. Originally, it has been suggested that the underlying distinction is one of etymology (borrowed vs. native, e.g. Saciuk 1969), but this does not explain why speakers can and do master English morphology without etymological knowledge. Others have argued that the stratum problem is in fact a phonological one, with differences between different etymological strata being paralleled by phonological differences (Booij 2002, van Heuven et al. 1993 for Dutch). This approach has the advantage that it would allow speakers to distinguish between the strata on the basis of the segmental and prosodic behavior of derivatives. However, explaining the nature of the strata as an epiphenomenon of underlying phonological properties of suffixes does in fact weaken the idea of strata considerably, because, as shown by Raffelsiefen (1999), not even two of the many suffixes of English trigger exactly the same type of morpho-phonological alternations, so that we would need as many sub-strata as we have suffixes that trigger morphonological alternations. Thus we end up with a continuum, rather than with a discrete dipartite or tripartite system.

Another serious problem is that a stratum can neither be defined by the set of suffixes it contains, because many suffixes must belong to more than one stratum, given that in certain derivatives they show stratum 1 behavior, whereas in other derivatives they display stratum 2 behavior, with sometimes even doublets occurring (cf. *compárable* vs. *cómparable*). Futhermore, there are a number of unexpected suffix combinations. Thus stress-neutral *-ist* appears inside stress-shifting *-ic*, or stress-neutral *-ize* appears inside stress-shifting *-(at)ion*. In order for the model not to make wrong predictions, dual membership of affixes (or some other device weakening the overall model) becomes a necessity.

Giegerich (1999) discusses cases of apparent dual membership of affixes in great detail and - as a consequence - proposes a thoroughly revised stratal model, in which the strata are no longer defined by the affixes of that stratum, but by the bases. This base-driven stratification model, which is enriched by many suffix-particular base-driven restrictions, can overcome some inadequacies of earlier stratal models, but at the cost of significantly reducing the overall predictive power of the model. These restrictions are a well-taken step towards eliminating the weakness of not making any predictions concerning suffix order within strata, which characterized earlier Lexical Phonology models. Certain problems remain, however.

For example, Fabb (1988) and Plag (1996, 1999) point out that there are numerous other important (phonological, morphological, semantic, syntactic) restrictions operative in English suffixation. About these restrictions level ordering does not say anything. For example, Fabb finds that the 43 suffixes he investigates are attested in only 50 combinations, although stratum restrictions would allow 459 out of the 1849 possible ones. He replaces stratal restrictions by individual selectional restrictions and proposes four classes of suffixes:

(3) Fabb (1988): 4 classes of suffixes

a. Group 1: suffixes that do not attach to already suffixed words (28 out of 43)

b. Group 2: suffixes that attach outside one other suffix (6 out of 43)

c. Group 3: suffixes that attach freely (3 out of 43)

d. Group 4: problematic suffixes (6 out of 43)

As pointed out in Plag (1996, 1999), this classification has also serious shortcomings. Firstly, there are numerous counterexamples to the above generalizations, secondly, the classes of suffixes are arbitrary and it is not clear why a given suffix should belong to a certain class and not to a different one, and thirdly, the classification again makes no predictions on many other restrictions. The latter point is crucial, as we will shortly see.

For any given affix, its phonological, morphological, semantic and syntactic properties (or the properties of its derivatives, i.e. of the morphological category[3]) must be stated in its lexical entry. Plag (1996, 1999) shows that these diverse properties together are responsible for the possible and impossible combinations of a given affix with stems and with other affixes. What has been analyzed as would-be stratal behavior falls out from the phonological, morphological and semantic properties of the affix. Since these properties must be stated anyway to account for the particular behavior of a given affix, no further stratal apparatus is necessary.

Plag (1996, 1999) also incorporates the idea of base-driven suffixation to explain apparent idiosyncrasies in suffix combinations. For illustration of what is meant by base-drivenness, consider the deverbal suffixes of Fabb's 'Group 1', which are said not to attach to any suffixed word.

(4) deverbal nominal suffixes not attaching to an already suffixed word

*-age* (as in *steerage*)

*-al* (as in *betrayal*)

---

[3] For those who do not believe in lexical entries for affixes, this type of information can be conceptualized as information on a given morphological category. Either theoretical preference has no bearing on the argument being made here.

*-ance* (as in *annoyance*)
*-ment* (as in *containment*)
*-y* (as in *assembly*)

Why should these suffixes behave in this way? And is this a property that has to be stated in the lexical entry of each of the nominal suffixes? In an affix-driven approach this would be essential. In a base-driven approach, however, this is not necessary, because it follows from independently needed specifications of the pertinent base words. The argument goes as follows: the only suffixed words that could in principle appear before deverbal *-age, -al ,-ance, -ment* and *-y* are verbs ending in *-ify, -ize, -ate,* and *-en*. However, *-ify, -ize,* and *-ate* require (a suffix-particular allomorph of) the nominalizer *-(at)ion*:

| | | | |
|---|---|---|---|
| (5) | magnification | verbalization | concentration |
| | *magnify-ation | *verbalize-ification | *concentrate-ation |
| | *magnify-ion | *verbalize-ion | *concentrate-ification |
| | *magnify-ance | *verbalize-ance | *concentrate-ance |
| | *magnify-al | *verbalize-al | *concentrate-al |
| | *magnify-age | *verbalize-age | *concentrate-age |
| | *magnify-y | *verbalize-y | *concentrate-y |
| | *magnify-ment | *verbalize-ment | *concentrate-ment |

These facts suggest that the behavior of verbalizing and nominalizing suffixes is best analyzed as base-driven: combinations of the verbal suffixes *-ify, -ize, -ate* with *-age, -al, -ance, -ment* and *-y* are ruled out because it is the verbal suffix (or the verbal base with this suffix) which selects the nominalizing suffix *-ion*, and crucially not the nominal suffix which selects its base.

To summarize, we can say that level ordering has serious weaknesses, including the one that its generalizations are too sweeping. Models that focus on suffix-particular restrictions are empirically more adequate, but they could be criticized for their lack of generalizations across suffixes. After all, linguists want to believe that language in general and derivational morphology in particular is not just an accumulation of idiosyncrasies. This is the point where Hay's psycholinguistic model enters the scene in an attempt to cut the gordian knot.

### 3. A NEW ALTERNATIVE: COMPLEXITY-BASED ORDERING (HAY 2000 )

In her model, Hay construes morphological complexity as a psycholinguistically real notion which heavily relies on the parsability of affixes. The basic claim concerning the problem of affix ordering is that "an affix which can be easily parsed out should not occur inside an affix which can not" (p. 23, 240). For reasons that will shortly become clear, I will refer to this approach as COMPLEXITY-BASED ORDERING.

What does it mean for an affix to be "easily parsed out"? As is well known, there are words that are clearly composed of two or more morphemes (e.g. *concrete-ness*), there are words that are clearly monomorphemic (e.g. *table*), and there are words whose status as complex words is not so clear (e.g. *rehearse*, *interview*, *perceive*). Discussions of this problem can be found in any good morphology textbook and are the daily bread of theoretical morphologists working on word-formation. Hay now shows that morphological complexity is a function of the psycholinguistic notion of morphological parsability, which in turn is largely influenced by two factors, frequency and phonotactics. The investigation of these factors leads to two conclusions. First, that morphological complexity is psychologically real, and second that morphological complexity is not a discrete, but a gradual notion. Thus some suffixes create words that are less morphologically complex than the words derived with other kinds of suffixes. Note that this distinction is reminiscent of the '+' and '#' boundaries in SPE, but that the dichotomy is replaced by a gradual notion of complexity. In Hay's words "[a]ffixes polarize along a continuum, ranging from affixes which are always parsed out during processing (displaying prototypical level 2 [i.e. stratum 2, IP] characteristics), to affixes which are never parsed out (displaying textbook level 1 [i.e. stratum 1, IP] characteristics)" (p. 22). In other words, it is hypothesized that only less complex base words can occur inside more complex derivatives, so that morphological complexity increases from the innermost to the outermost morphological layer. This is my reason for calling this approach 'complexity-based ordering'.

Before we can evaluate the idea of complexity-based ordering, let us explain how such a gradual notion of complexity can emerge psycholinguistically and what its linguistic correlates are. Hay proposes two levels in morphological processing, pre-lexical and lexical. In pre-lexical processing, speakers use probabilistic phonotactics to segment speech into potential morphemes by positing boundaries inside phoneme transitions which are unlikely to occur word-internally. For example, the combination [pf] (as in *pipeful*) is unattested morpheme-internally in English, with the consequence that wherever this combination occurs, it provides a cue to morphological juncture. This cue from pre-lexical processing is exploited in speech perception to facilitate morphological decomposition in lexical processing. Hay proves this point with a number of psycholinguistic experiments, which show that the phonotactics plays an important role in the parsing or non-parsing of complex words. It is also shown that the described effect is much stronger with prefixes than with suffixes (p. 94-97).

Hay's experiments involve consonant-intial suffixes, but the results are suggestive also for vowel-initial suffixes. Thus, consonant-initial suffixes will create phoneme transitions that are more likely to be illegal morpheme-internally than the phoneme transitions created by vowel-initial suffixes. The reason for this is that the combination of consonant-final base words with consonant-initial suffixes leads to consonant-clusters, which in general are under severe phonotactic restrictions. Violations of these restrictions provide cues for morphological juncture. Vowel-inital suffixes, on the other hand, integrate more easily into the syllabic structure of the base word, thereby providing no cue for a morphological boundary in pre-lexical processing. Hence, the model predicts that C-initial suffixes should favor decomposition to a much greater extent than V-initial suffixes. Note also that stress shifts and other morphophonological alternations are interpreted by Hay as potentially blurring the morphological juncture (p. 224), so that words with non-neutral suffixes are less easily decomposed than word with neutral suffixes. On the basis of the role of phonotactics in pre-lexical morphological processing Hay sets up the following three classes of suffixes, whose phonotactics is supposed to allow predictions on their respective parsability (pp. 224, 239[4]):

(6) three main phonotactic classes of suffixes and their respective decomposability
- V-initial non-neutral suffixes, strong whole word bias
- V-initial neutral suffixes, weak whole word bias
- C-initial, strong parsing bias

From this, further predictions concerning stacking restrictions become possible. If C-initial suffixes have a strong parsing bias and should allow attachment to decomposable bases easily, whereas V-initial suffixes, which should have a whole word bias, should prefer non-decomposable bases to decomposable bases.

Moving on from pre-lexical to lexical processing, frequency comes into play. In most current models of morphological processing access of morphologically complex words works in two ways: by direct access to the whole word representation ('whole word route') or by access to the decomposed elements ('decomposed route'). Given that frequency plays a role in determining the resting activation of lexical items, it is clear that every access via the whole word route strengthens the whole word representation, whereas access on the decomposed route reinforces the representation of the decomposed morphemes and the decomposability of the complex word. How do we know which representation will be strengthened with a given word? It is usually assumed that the absolute frequency of a word correlates with its resting activation level. Hay suggests that, with regard to the storage of complex words, the relative frequency of the derived word and its base is significant. Relative frequency is defined as the ratio of the frequency of the derived word to the frequency of the base:

(7) relative frequency:

---

[4] Note that p. 239 sets up the three classes differently. The subsequent discussion in Hay (2000) shows, however, that this is an error in the presentation. Hay (personal communication, January 2001) has confirmed my reading of the relevant passages.

frequency of derived word divided by the frequency of the base

$f_{relative} = f_{derivative} / f_{base}$

With most complex words, the base is more frequent than the derived word, so that the relative frequency is smaller than unity. In psychoplinguistic terms, the base has a higher resting activation than the derived word. This leads to preponderance of the decomposed route, since due to its high resting activation, the base will be accessed each time the derivative enters the system. In the opposite case, when the derived word is more frequent than the base, there is a whole word bias in parsing, because the resting activation of the base is lower than the resting activation of the derivative. For example, *insane* is more frequent than its base *sane*, so that *insane* will have a whole word bias in access. Conversely, *infirm* has a base that is much more frequent than the derived form, so that there will be a strong advantage for the decomposed route. Hay shows that relative frequency correlates with three other properties of complex words, summarized in (8):

(8) a. Low relative frequency correlates with high productivity
b. Low relative frequency correlates with bad phonotactics
c. Low relative frequency correlates with high semantic transparency

After the above discussion, the correlations in (8) do not come as a surprise. We know that productive morphological processes are characterized by a high number of low frequency words. The lower the frequencies of derived words the lower their relative frequencies (holding the frequency of the base constant). Thus productive processes should show a preponderance of low relative frequencies, whereas less productive morphological categories should be characterized by a preponderance of words with higher relative frequencies. We also know that productive categories are semantically transparent. That this is so can be seen as a consequence of processing, since productive processes favor the decomposed route, and decomposed storage strengthens the individual semantic representations of the elements. Decomposition leaves little room for semantic drift and opacity, which arise easily under whole word access, because the meanings of the parts are less likely to be actived. We have said above that bad phonotactics favors decomposition. It is therefore expectable that those processes that favor decomposition, i.e. the productive, semantically transparent ones, are those that are also phonologically transparent. Thus, semantic opacity and low productivity go hand in hand with a kind of phonotactics that disfavors parsing.

To summarize, Hay argues that stacking restrictions are a function of the parsability of suffixes. Parsability in turn is argued to be a function of phonotactics and relative frequency. An easily decomposable suffix inside a non-decomposable suffix would lead to difficulties in processing, whereas a less easily decomposable inside a more easily decomposable suffix is easy to process. Based on these considerations, Hay proposes that "an affix which can be easily parsed out should not occur inside an affix which can not" (p. 23, 240). This hypothesis will be thoroughly tested in the next section.

## 4. TESTING COMPLEXITY-BASED ORDERING

In this section we will look at a number of phenomena in Englich derivation and see whether complexity-based ordering can really account for the data. It will become clear that there are a number of problems that complexity-based ordering cannot solve properly.

### 4.1. Problem 1: base-driven restrictions (*-al-ize-ion, -able-ity, -al-ist*)

The first problem concerns phenomena that can be subsumed under the heading of base-driven restrictions of the kind discussed in section 2 above. In English derivation, the following suffix combinations are best analyzed as involving base-driven selection: *-able-ity, -al-ize-ion,* and *-al-ist*. For example, all words ending in adjectival *-al* productively take *-ize* as the only possible verbal suffix. In turn, *-ize* obligatorily takes -*ation* as nominalizing suffix. Analogous arguments hold for -*able-ity* and -*al-ist*. But which order would complexity-based ordering predict?

For the combination -*able-ity*, complexity-based parsing predicts that -*able,* which (in its productive variant) is a stress-neutral, V-initial suffix, is supposedly more easily parsed out than -*ity*, because -*ity* is non-neutral. Thus -*able-ity* should be ruled out as a possible combination under the complexity-based ordering approach in its present formulation. The facts show, however, the most extreme opposite: -*able* not only allows attachment of -*ity* but even base-selects -*ity* productively, i.e. every -*able* derivative can take -*ity*.

A similar problem occurs with -*al-ize-ation*. On p. 263, Hay states that -*ion-al* (e.g. *sensational*) is a possible combination, because "-*ion* heavily biases the whole word route in access" and -*al* is "tolerant of *some* degree of internal structure" (p. 263). This raises first the - perhaps minor - question why there should be a difference between -*ion* and -*al* in the first place, given that both suffixes are V-initial, and both suffixes cause stress shifts. A more serious problem for Hay's account is that -*al-ize-ion* is a productive combination: -*al* base-selects -*ize*, -*ize* base-selects -*ion*. According to complexity based parsing, this would mean that -*ize* is more easily parsed out than -*al* (cf. *coloni-al-ize*), and that -*ion* is more easily parsed out than -*ize* (cf. *coloni-al-iz-ation*). This however, contradicts both Hay's statement that "-*ion* heavily biases the whole word route in access", and the idea that non-stress-shifting suffixes (like -*ize*) should occur only outside stress-shifting suffixes (such as -*ion*). In other words, we have the problem that -*al* must be at the same time more easily parsable than -*ion* (cf. *sensational*) and less easily parsable than -*ion* (cf. *colonialization*). This is impossible, unless the claim would be seriously weakened in such a way that it would no longer hold for the suffixes as such, but only for individual words. That even such an item-based approach to suffix-ordering does not work with suffix orders determined by base-driven restrictions is shown below for the combination -*al-ist*.

With regard to -*ist*, Hay claims that this suffix only allows unparsable or weakly parsable bases (2000:243-246, 251f). If it is true that -*al* is "tolerant of *some* degree of internal structure" (2000: 263), those two claims converge on the prediction that the combination -*al-ist* should be either impossible or, where attested, -*ist* should prefer bases in -*al* that have high relative frequencies. Although this may be true for the few examples Hay cites, the prediction turns out to be wrong when tested systematically against larger amounts of data are investigated. Table 1 gives all 20th century neologisms involving *-al-ist* listed in the *OED*, accompanied by the relative frequencies of the *-al* derivative on which the *-al-ist* forms are based. For example, the first word *accentualist* is based on the *-al* derivative *accentual*. The relative frequency of *accentual* is 0.006, which means that statistically for every six occurences of the derivative *accentual* we have 1000 occurrences of the base word *accent*. Such a frequency pattern strongly favors decomposition for *accentual*. Whole word access, on the other hand, is favored with frequency patterns where the derived word is more frequent than the base word, i.e. in those cases where the relative frequency exceeds unity.[5] A similar reasoning holds for derivatives based on bound roots whose frequency is zero. In table 1, all words favoring whole word access are printed in bold.

---

[5] Note that in Hay (2000) unity is used as the parsing threshold. Hay and Baayen (this volume) propose a different parsing threshold. The following remarks are based on Hay's original proposal.

**Table 1: All OED 20th century neologisms ending in *-al-ist***

[Relative frequency of bases in *-al* = f (*-al* derivative)/f (base of *-al* derivative), Relative frequencies are computed on the basis of the BNC]

| ***-al-ist* derivative** | **rel. freq. of *-al* base** | ***-al-ist* derivative** | **rel. freq. of *-al* base** |
|---|---|---|---|
| accentualist | 0.006 | bidialectalist | 0.008 |
| anecdotalist | 0.4 | contextualist | 0.04 |
| decentralist | 0.00005 | documentalist | 0.0001 |
| environmentalist | 0.6 | existentialist | 0.03 |
| factionalist | 0.08 | factualist | 0.2 |
| functionalist | 0.14 | **fundamentalist** | **bound root** |
| incidentalist | 0.07 | **integralist** | **bound root** |
| **intensionalist** | **15** | **maximalist** | **bound root** |
| **minimalist** | **bound root** | multiracialist | 0.04 |
| **navalist** | **bound root** | nutritionalist | 0.9 |
| operationalist | 0.1 | **paternalist** | **bound root** |
| **patriarchalist** | **1.4** | pentecostalist | 0.3 |
| pro-natalist | 0.005 | racialist | (0.1) |
| situationalist | 0.006 | spatialist | 0.09 |
| stratificationalist | 0.04 | structuralist | 0.1 |
| substitutionalist | 0.009 | triumphalist | 0.05 |
| unilateralist | 0.5 | unrealist | 0.01 |

The following table summarizes the most important finding emerging from table 1, by grouping the derivatives according to whether the relative frequencies of the *-al* bases is above or below unity. Bases ending in *-al* that have bound roots as bases are listed separately.

**Table 2: Relative frequency of *-al* derivatives inside OED *-ist* neologisms**

| $f_{rel}$ **< 1** | $f_{rel}$ **³ 1** | **bound root** |
|---|---|---|
| 26 | 2 | 6 |

Table 2 shows that in their vast majority, *-al-ist* neologisms are based on parsable bases ending in *-al*. In sum, the systematic investigation of relative frequencies of *-al-ist* derivatives clearly shows that these forms, contra to the prediction of complexity-based ordering, are possible, numerous, and do not show the expected frequency effects. The reason for this is a purely grammatical restriction: *-al* base-selects *-ist* productively.

That base-driven restrictions take precedence over parsing is also evidenced by the figures presented in the appendix of Hay and Baayen (this volume). Complexity-based ordering would predict that it is only possible for suffixes with a lower proportion of decomposed derivatives to occur inside suffixes with a higher proportion of decomposed derivatives. With the combinations *-able-ity* and *-ize-ion* we find the clear opposite: applying Hay and Baayen's parsing model, 72 % of all *-able* derivatives are decomposed, whereas only 17 % of the *-ity* derivatives are above the parsing line, and while 44 % of all *-ize* derivatives are decomposed, only 18 % of all *-ation* words are decomposed. With the combination *-al-ist*, the proportions of decomposed derivatives are in accordance with the predictions of complexity-based ordering, but we have seen that the attested combinations do not behave as predicted by complexity-based ordering.

Concluding the discussion of suffix-combinations involving base-selection, we can say that in all three cases under scrutiny in this section, base-driven selectional restrictions take precedence over possible parsing restrictions. In the next section we turn to the problem of productivity.

### 4.2. Problem 2: predictions concerning productivity

As mentioned in section 3, complexity-based ordering predicts that C-initial suffixes be more productive than V-initial ones. We will see in what follows that this prediction is not correct.

In order to test the hypothesis it is necessary to find an operationalized measurement of productivity. In recent work on productivity a number of different measures have been proposed, each of which emphasizes a different aspect of the notion of productivity. In order to test the hypothesis I have systematically computed four different measurements of productivity for a wide range of C-initial and V-initial English suffixes. The four measurements are summarized in (9). The first three measures are corpus-based (see e.g. Baayen 1989, 1993), the fourth is a more traditional one, and is dictionary-based.

(9) a. Extent of use V
the number of different words (types) derived with a given suffix
b. Productivity in the narrow sense P
the number of hapax legomena divided by the number of tokens derived with a given suffix
c. Global productivity P*
the number of hapaxes with a given suffix
d. Number of neologisms N for a given period *t*
as listed in a comprehensive dictionary (*OED*)

The extent of use shows how many different words are derived by adding the suffix in question. Productivity in the narrow sense quantifies the probability of encountering a newly formed derivative of the given category among all the words (tokens) of that category. This is possible because the proportion of neologisms is highest among the hapaxes (e.g. Baayen and Renouf 1996). Global productivity simply quantifies the number of hapaxes, which in turn is indicative of how many new words are formed with a given suffix (independent of how many other derivatives of that category exist or are used). Finally, the possibility to form new words with a given suffix can be estimated on the basis of the number of neologisms listed in the *OED*. For a detailed critical discussion of these measurements the reader is referred to Plag (1999) and Bauer (2001). In principle, these measurements can (and often do) yield contradictory results, because they each highlight different aspects of productivity.

For the intended comparison I have used the set of 12 at least marginally productive English suffixes that have been investigated in Plag et al. (1999).[6] Using suffixes from a different study has two advantages, one practical, one methodological. First, it is not necessary to compute all measurements anew. Second, and more importantly, the selection of suffixes was done by a different team of authors for different purposes, which precludes the danger that the results of the present investigation are influenced by a potential bias of this author.

Let us first look at the corpus-based measures. For each suffix the different measures have been computed. The suffixes were then listed in tables 3 through 5, with the most productive at the top, the least productive at the bottom of the list. The second column in each table gives the suffix-initial segment, classified according to V or C, the third column indicates whether the suffix has an impact on the stress pattern (abbreviated as 'stress-shift, yes/no'). According to complexity-based ordering, we would expect the C-initial, non-stress-shifting suffixes clustering at the top of each table, and the V-initial clustering at the bottom. In technical terms, the mean rank of the C-initial suffixes should be lower than the mean rank of the V-initial suffixes. This is, however, not the case. As can be seen from the tables, depending on the type of measure used, the ranks of C-initial and V-initial suffixes vary a great deal. Concerning extent of use, the vowel-initial suffixes are overall more productive, concerning productivity in the narrow sense the consonant-initial suffixes are more productive, concerning global productivity the vowel-initial suffixes are again more productive.

---

[6] There are three bound forms that are included in the study by Plag et al. (1999), but excluded here, i.e. *-like*, *-type* and *-free*. As argued in detail in Dalton-Puffer and Plag (in press), complex words involving *-type* are compounds, and this also holds for complex words featuring *-like* or *-free* as their head. These elements have therefore been excluded from the present investigation. For an account of the methodological problems involved in quantifying the suffixes as they occur in the BNC the reader is referred to Plag et al. (1999).

**Table 3: Extent of Use V** (BNC written corpus, N=82 mio., figures from Plag et al. 1999)

| rank | suffix | initial segment | stress-shift | V |
|---|---|---|---|---|
| 1 | -ness | C | no | 2466 |
| 2 | -ion | V | yes | 2392 |
| 3 | -er | V | no | 1823 |
| 4 | -ity | V | yes | 1372 |
| 5 | -ist | V | no | 1207 |
| 6 | -able | V | no | 933 |
| 7 | -less | C | no | 681 |
| 8 | -ize | V | no | 658 |
| 9 | -ish | V | no | 491 |
| 10 | -wise | C | no | 183 |
| 11 | -ful (property) | C | no | 154 |
| 12 | -ful (measure) | C | no | 136 |

**Table 4: Productivity in the Narrow Sense P** (P = $n_1$ / $N_{affix}$, BNC written corpus, figures from Plag et al. 1999)

| **rank** | **suffix** | **initial segment** | **stress-shift** | **P=$n_1$/$N_{affix}$** |
|---|---|---|---|---|
| 1 | -wise | C | no | 0,0612 |
| 2 | -ish | V | no | 0,0338 |
| 3 | -ful (‘measure’) | C | no | 0,0229 |
| 4 | -er | V | no | 0,0195 |
| 5 | -less | C | no | 0,0096 |
| 6 | -ness | C | no | 0,0088 |
| 7 | -ist | V | no | 0,0036 |
| 8 | -able | V | no | 0,0022 |
| 9 | -ize | V | no | 0,0021 |
| 10 | -ity | V | yes | 0,00092 |
| 11 | -ion | V | yes | 0,00038 |
| 12 | -ful (‘property’) | C | no | 0,00029 |

**Table 5: Global Productivity P*** (P* = $n_1$, BNC written corpus, figures from Plag et al. 1999)

| rank | suffix | initial segment | stress-shift | P* |
|---|---|---|---|---|
| 1 | -ness | C | no | 943 |
| 2 | -er | V | no | 792 |
| 3 | -ion | V | yes | 524 |
| 4 | -ist | V | no | 354 |
| 5 | -ity | V | yes | 341 |
| 6 | -able | V | no | 311 |
| 7 | -less | C | no | 272 |
| 8 | -ish | V | no | 262 |
| 9 | -ize | V | no | 212 |
| 10 | -wise | C | no | 128 |
| 11 | -ful (‘measure’) | C | no | 60 |
| 12 | -ful (‘property’) | C | no | 22 |

For computing the dictionary measure the same set of suffixes was used, with the addition of the three C-initial suffixes *-ling, -ment* and *-ship*. The three suffixes were added to make up for the fact that the BNC-suffixes contained a majority of V-initial suffixes (7 out of 12). Table 6 shows that, according to the listing in the *OED*[7], the C-

[7] Note that the computation of the *OED* figures is also not without methodological problems. For example, the raw data for *-ion* contain 894 items, a considerable number of which do not belong to the morphological category in question. Exactly how many is, however, sometimes hard or impossible to decide. The largest number of forms that are problematic in this respect are those that allow more than one bracketing. For example, *decompression* can be analyzed as *decompress-ion* or as *de-compression*. Only in the former case would the word belong to the morphological category of *-ion*. Fortunately, such cases are not too numerous, so that either decision how to treat such forms would not influence the overall number significantly. My general policy was to include a form if it was possible to analyze the suffix in question as being the outermost suffix. In the above case, for example, *decompression* remained on the list. A number of uncertain decisions remain, however, so that anyone trying to verify my figures is likely to arrive at slightly different figures.

initial suffixes are generally less productive than the V-initial ones, with most of the C-initial suffixes clustering at the end of the list.

**Table 6: 20th century neologisms** (OED)

| rank | suffix | initial segment | stress-shift | N |
|---|---|---|---|---|
| 1 | -ion | V | yes | 625 |
| 2 | -ist | V | no | 552 |
| 3 | -ity | V | yes | 487 |
| 4 | -er | V | no | 564 |
| 5 | -ness | C | no | 279 |
| 6 | -ize | V | no | 273 |
| 7 | -able | V | no | 185 |
| 8 | -less | C | no | 103 |
| 9 | -ish | V | no | 101 |
| 10 | -ship | C | no | 23 |
| 11 | -ful (‘measure’) | C | no | 22 |
| 12 | -ment | C | no | 20 |
| 13 | -ful (‘property’) | C | no | 14 |
| 14 | -wise | C | no | 11 |
| 15 | -ling | C | no | 3 |

To summarize our investigation of the productivity of V- and C-initial suffixes, we can state that different measurements yield different results. However, overall it is clear that, contra to the prediction of complexity-based ordering, C-initial suffixes cannot be said to be generally more productive than V-initial ones. If anything, the opposite seems to be the case. V-initial suffixes are more productive than C-initial suffixes, with the former having an overall mean rank of 5.95, the latter of 7.31. From this result it can be concluded that the segmental make-up of suffixes ALONE is not a good predictor for the productivity and parsability of suffixes and can be easily overruled by other mechanisms.

### 4.3. Problem 3: Predictions concerning the correlation between productivity, phonotactics and relative frequency

According to complexity-based ordering as formulated in Hay (2000:224,239), V-initial suffixes should strongly favor whole word access, i.e. their relative frequencies should show a preponderance of high relative frequencies. This tendency should be most remarkable with V-initials that induce stress-shift or phonological changes of the base. In order to test this prediction, I have calculated the relative frequencies of the two productive V-initial suffixes *-able* and *-ize*, using frequencies from the BNC. The frequencies were calculated for a random sample of about 100 derivatives drawn from the word lists of the BNC written corpus (80 million words).[8] For technical reasons the relative frequencies had to be calculated using the base frequencies from the whole BNC corpus (100 million words). The relative frequencies are listed in tables 8 and 9 in the appendix. Table 7 summarizes how many of the derivatives have a relative frequency equal or above unity, below unity, or are based on unattested bases. According to Hay, frequencies above unity and forms based on unattested bases strongly favor whole word access.

**Table 7: Relative frequency of -*able* and -*ize* derivatives (BNC written corpus) -*able*: sample of 92 out of 934 total, -*ize*: sample of 101 of 658**

| suffix | $f_{rel} < 1$ | $f_{rel} \geq 1$ | no base attested |
|---|---|---|---|
| *-able* | 90 % | 4 % | 7 % |
| *-ize* | 85 % | 11 % | 3 % |

Table 7 shows that, contrary to Hay's predictions concerning V-initial suffixes, there is a strong preponderance of low relative frequencies for both suffixes. This means that productive V-initial suffixes pattern like any other productive processes in that

[8] The sample was created by taking every tenth derivative from the *-able* word list, and every sixth derivative from the *-ize* word list.

they strongly favor low relative frequencies. Contrary to the prediction, these two V-initial suffixes do not show a whole word bias, although *-ize* even belongs to those suffixes that inflict severe phonological changes on their base words.[9]

4.5. Other problem areas: conversion, the monosuffix constraint and morphological family effects

In this section I would like to discuss three other problem areas where it is not yet clear how complexity-based ordering can deal with the phenomena to be discussed. These areas are conversion, the recently proposed 'monosuffix constraint' (Aronoff and Fuhrhop 2001), and morphological family effects. I am aware of the fact that my remarks will be rather speculative, but they may nevertheless be useful in pointing out areas where further research is called for, before the potential of a theory of complexity-based ordering can be fully estimated.

Concerning conversion, complexity-based ordering would make the following predictions. If conversion is considered a rather opaque morphological process, we should not expect complex bases to undergo conversion. If on the contrary, conversion is considered transparent, we should expect to find many complex bases undergoing conversion. What are the facts? In Plag (1999) it is documented that in English noun to verb conversion, suffixed, prefixed and compounded base words are frequent, which means that conversion must be considered an extremely transparent process. I am not aware of any explicit treatment of this question, but it seems that it is not obvious that conversion should be regarded as particularly transparent.

Recently, Aronoff and Fuhrhop (2001) have proposed a new kind of constraint operative in English morphology, the MONOSUFFIX CONSTRAINT. According to their empirical findings, native suffixes only attach to suffixed bases if these bases are Latinate.[10] Since most Latinate suffixes create words of lower complexity, and the

---

[9] See Plag (1999:145-194) for a detailed analysis of the phonology of *-ize* derivatives.

[10] The only exception to this generalization which is mentioned by Aronoff and Fuhrhop is *-ness*. Another two can be added here, adverb-forming *-wise*, and nominal and adjectival *-ful*, for which

native suffixes have a tendency to create words of higher complexity, it does not come as a surprise that Latinate suffixes should be possible inside native suffixes. However, it is much less clear why native suffixes should be principally ruled out inside other native suffixes. Under the assumption of complexity-based ordering, we should not only expect that Latinate suffixes occur inside native ones, but also that native suffixes happily combine with native suffixes whenever the phoneme transition created by the outer native suffix is more likely to create illegal phonotactics than the phoneme transition created by the inner native suffix. Obviously, this is not the case. Instead, there seems to be this general prohibition of native suffixes inside native suffixes, irrespective of their respective likelihood of creating illegal phoneme transitions. Whatever the deeper source of this general prohibition may be, it does not have its source in phonotactics.

A final problem for a frequency-based account of suffix-ordering are morphological family effects. Several studies have shown that word access and representation is also determined by the size of the morphological family (e.g. Schreuder/Baayen 1997, Baayen et al. 1997, Bertram et al. 2000, Krott et al. 2001), which means that the relative frequency of base and derivative cannot be the only factor in play. Hence, frequency-based assessments of decomposability (and stacking) would need to incorporate morphological family effects as well as effects of relative frequency and absolute frequency.

## 5. Discussion and Conclusion

In the foregoing section we have discussed a number of cases where complexity-based ordering does not make the right predictions or encounters other conceptual or empirical problems.

First, base-driven restrictions were shown to override effects expected on the basis of phonotactic pre-processing. Second, it was shown that there is no principled

---

derivatives like *tumblerful, drawerful, truthful,* and *meaningful* are attested (see Dalton-Puffer and Plag (in press) for a detailed discussion of these suffixes).

difference between consonant- and vowel-initial suffixes in terms of parsing. Contrary to the predictions, consonant-initial suffixes are not generally more productive than vowel-initial ones, and vowel-initial suffixes do not generally have a strong whole word bias. In sum, the present investigation has shown that combinatorial restrictions among suffixes cannot be reduced to phonotactically based parsability.

How can these findings be explained, especially in the light of Hay's experimental evidence for phonotactic effects in the parsing of suffixes? Let us first look at the differences between prefixes and suffixes. Hay shows that the effect of phonotactics in parsing "is much stronger on prefixes than suffixes"(p. 97). In fact, her own evidence for a phonotactic effect with suffixes is not always conclusive. Thus, in some experiments Hay finds evidence for phonotactic effects, but fails to do so in other cases. For example, one experiment shows that nonsense words (of a phantasy language) are more likely to be perceived as containing a suffix if they involve an illegal (in English) phoneme transition. Another experiment shows that listeners perceive forms with the suffix *-ful* as more complex if the phoneme transition at the morpheme boundary is illegal (recall the example *pipeful* from section 3). In an experimental investigation of the suffix combination *-ment-al* subjects preferred forms that were in accordance with complexity-based ordering. However, Hay's study of the hypothesized correlation of phonotactics with semantic transparency, polysemy and relative frequency using the suffixes *-ful, -less,-ly, -ness,* and *-ment* did not yield any significant result, so that with these suffixes there was no reliable effect observable of junctureal phonotactics upon decomposability. Two questions emerge from these sometimes contradictory results. First, why do we find a much more robust phonotactic effect with prefixes? Second, why do suffixes not behave in the uniform way predicted on the basis of their phonotactics?

The difference in the parsing of prefixes and suffixes could be due to the fact that in online processing, prefixes enter the processor before the base, whereas with suffixes, the base comes in first and has thus an advantage of being recognized and accessed, even though the morpheme boundary may not readily induce parsing due to seemingly morpheme-internal phonotactics. Note also that there is abundant evidence that words are often already recognized before online-processing has reached the end of the word. This is perhaps decisive for the differences in behavior between prefixes and bases of suffixed words, because bases are usually longer than prefixes and therefore provide more cues for word recognition than short prefixes (or even shorter suffixes).[11] Thus, already due to the position of the suffix after the base, we can expect a much weaker effect of the phonotactic cue.

Another explanation for the variability in the presence/absence of phontactic effects on parsing may lie in the fact that morpho-phonological alternations of the base do not only have a negative effect on decomposition. Predictable alternations such as stress-shift, or regular stem allomorphy are highly indicative of the membership of a word in a certain morphological category. As already briefly mentioned in section 2, Raffelsiefen (1999) has shown that among the many derivational processes in English which exhibit morpho-phonological alternations one cannot find even two which show exactly the same morphophonological patterns. In other words, the kind of alternation a given word shows indicates its belonging to a specific morphological category. Apart from the phonetic material of the suffix and the meaning, morphological categories can thus be uniquely characterized by their morpho-phonological alternations, so that a derivative with a certain alternation can be easily identified as belonging to a specific category solely on the basis of the alternation. Hence, phonological opacity effects can be easily tolerated by productive morphological processes if these effects are regular and predictable. One could hypothesize that the potential phonotactically induced parsing disadvantage of vowel-initial suffixes can be made up by the fact that the coherence of the category (and thus the representation of the suffix) is strengthened by the suffix-particular alternation. If this turned out to be true it could also account for the non-existence of a clear parsing distinction between consonant- and vowel-initial suffixes.

The essential problem of any account seeking an explanation for suffix ordering is that there is no *ceteris paribus* in suffixation. Each morphological category comes with its own particular phonological, morphological, and semantic

---

[11] See Laudanna and Burani (1995) and Baayen (to appear) on the role of length in the processing of affixes.

restrictions, so that no account can work which completely abstracts away from these inherent and to a great extent idiosyncratic properties of any morphological category. While complexity-based parsing may make good predictions in terms of general tendencies in parsing, individual affixes or individual words may pattern differently due to various other reasons. Thus there might be parsing differences between certain consonant-initial suffixes and other consonant-initial suffixes, and there might also be parsing differences between certain consonant-initial suffixes and certain vowel-initial suffixes, but there is obviously no robust evidence for a general difference between consonant-initial and vowel-initial suffixes. The phonotactic difference between consonant- and vowel-initial suffixes should in principle affect (pre-)processing but it seems that it is possible for grammatical constraints to overrule constraints demanding easy processability.

The explanatory power of complexity-based parsing is also limited in another important respect. Even if we could exclude all suffix-combinations that violate complexity-based ordering, there is a considerable range of potential combinations left where parsing considerations do not make any prediction whatsoever. We can illustrate this with a hypothetical example of three suffixes A, B, and C. If complexity-based parsing allows only the order ABC, this rules out all other conceivable combinations such as CBA, BCA, BAC etc., but it does not predict whether the parsing-wise legal combinations AB, AC, BC or ABC do actually occur or are at all possible. In order to account for the existing and non-existing combinations within the subset of parsable suffix combinations, suffix-particular restrictions still need to be stated and these restrictions then rule out large quantities of parsing-wise legal combinations.

In this respect complexity-based ordering is similar to level ordering in Lexical Phonology, since in both frameworks certain combinations are ruled out in principle while for many other combinations no predictions are made. Complexity-based parsing is however empirically more accurate and conceptually more interesting than level-ordering in its prediction of possible and impossible combinations. In any case, suffix-particular restrictions on suffix combinations still need to be stated and can not be replaced by parsing restrictions. In fact, parsing restrictions and suffix-particular restrictions can be conceived of as working hand in hand: parsability determines a broad range of possible combinations, which are then further constrained by suffix-particular restrictions.

It is therefore not surprising that suffix-particular restrictions are often in line with processing necessities, but the present paper has shown that there are apparently also many cases where rather idiosyncratic grammatical restrictions work against easy parsing. It should be the goal of future studies to tease apart processing effects from purely grammatical mechanisms (such as base-driven restrictions). This kind of problem is perhaps new to morphologists but certainly not to students of syntax, in particular word order. From studies such as Hawkins (1994) we know that certain word orders are more easily processable than others. However, not all attested or grammaticalized word orders are those that have processing advantages over potential rival word orders. Sometimes less than optimally parsable word orders are possible because they have advantages in other respects. Good parsability is only one (though a major one) of many conflicting constraints that language is subject to. With regard to morphology, for example, it is easy to see that what might be good for the parser (e.g. bad phonotactics) might be bad for the articulators or vice versa. It will be the future task of morphologist to further explore both the structural and the psycholinguistic side of the matter in order to gain a deeper understanding of the organization and interaction of morphological units.

## REFERENCES

Allen, Margaret. 1979. *Morphological investigations*. Ph. D. dissertation. University of Connecticut. Ann Arbor: University Microfilms.

Aronoff, Mark, and Nanna Fuhrhop. 2000. Restricting suffix combinations in English. Closing suffixes and the monosuffix constraint. Ms. SUNY / Universität Konstanz.

Baayen, Harald. To appear. Probability in morphology. In Bod, R., Hay, J, and S. Jannedy (eds.) *Probability Theory in Linguistics*. Cambridge: MIT press.

Baayen, Harald, Rochelle Lieber, and Robert Schreuder. 1997. The morphological complexity of simplex nouns, *Linguistics* 35, 861-877.

Bauer, Laurie. 2001. *Morphological Productivity*. Cambridge: Cambridge University Press.

Bertram, R., Baayen, Harald, and Robert Schreuder. 2000. Effects of familiy size for complex words, *Journal of Memory and Language* 42, 390-405.

Booij, Geert E. 1994. Lexical Phonology: A Review. In Richard Wiese (ed.) Recent Developments in Lexical Phonology (Arbeiten des Sonderforschungsbereichs 'Theorie des Lexikons' Nr. 56), 3-29. Düsseldorf: Heinrich-Heine-Universität.

Booij, Geert E. 2002. Prosodic restrictions on affixation in Dutch. This volume.

Dalton-Puffer, Christiane, and Ingo Plag. In press. Categorywise, some compound-type morphemes seem to be rather suffix-like. On the status of *-ful, -type*, and *-wise* in Present Day English. *Folia Linguistica*.

Fabb, Nigel. 1988. English Suffixation is Constrained Only by Selectional Restrictions. *Natural Language and Linguistic Theory* 6, 527-539.

Giegerich, Heinz J. 1999. *Lexical Strata in English. Morphological causes, phonological effects*. Cambridge: CUP.

Hay, Jennifer. 2000. *Causes and consequences of word structure*. Ph.D. dissertation. Northwestern University.

Hay, Jennifer and Harald Baayen. This volume. Parsing and Productivity.

Hawkins, John. 1994. *A performance theory of order and constituency*. Cambridge: CUP.

Krott, Andrea, Baayen, Harald, and Rob Schreuder (2001) Analogy in Morphology: Modeling the Choice of Linking Morphemes in Dutch, *Linguistics* 39, 51-93.

Laudanna, Alessandro and Cristina Burani. 1985. Distributional properties of derivational affixes. In L. B. Feldman (ed.) *Morphological aspects of language processing*, 345-364. Hillsdale: Lawrence Erlbaum.

Kiparsky, Paul. 1982. Lexical Morphology and Phonology, in The Linguistic Society of Korea (ed.) *Linguistics in the Morning Calm*, 1-91. Seoul: Hanshin Publishing Co.

Mohanan, Karuvannur P. 1986. *The theory of lexical phonology*. Dordrecht: Reidel.

OED2 on CD. 1994. *The Oxford English Dictionary. 2nd Edition, on Compact Disc*. Oxford: Oxford University Press.

Plag, Ingo. 1996. Selectional restrictions in English suffixation revisted. A reply to Fabb (1988). *Linguistics* 34.4, 769-798.

Plag, Ingo. 1999. *Morphological Productivity. Structural Constraints in English Derivation*. Berlin / New York: Mouton de Gruyter.

Plag, Ingo, Christiane Dalton-Puffer, and Harald Baayen. 1999. Morphological productivity across speech and writing. *English Language and Linguistics* 3.2, 209-228.

Raffelsiefen, Renate. 1999. Phonological constraints on English word-formation. In Geert Booij and Jaap van Marle (eds.) *Yearbook of Morphology 1998*, 225-287.

Saciuk, Bogdan. 1969. The stratal division of the lexicon. *Papers in Linguistics* 1: 464-532.

Schreuder, Robert, and Harald Baayen. 1997. How complex simplex words can be. *Journal of Memory and Language* 36, 118-139.

Selkirk, Elisabeth. 1982. *The syntax of words*. Cambridge: MIT Press.

Siegel, Dorothy. 1974. *Topics in English morphology*. Ph. D. thesis. MIT.

Spencer, Andrew. 1991. *Morphological theory*. Oxford: Blackwell.

Van Heuven, Vincent J., Neijt, Anneke and Maarten Hijzeldoorn. 1993. Defining etymological sublexicons for Dutch using linguistic filters. Unpublished paper, Universities of Leiden and Nijmegen. (published version: Automatische indeling van Nederlandse worden op basis van etymologische filters, *Spektator* 23 (1994), 279-291)

Ingo Plag
English Linguistics, Fachbereich 3
Universität Siegen
Adolf-Reichwein-Str. 2,
D-57068 Siegen
e-mail: plag@anglistik.uni-siegen.de, http://www.uni-siegen.de/~engspra/

## APPENDIX

**Table 8: Relative frequency of *-able* derivatives (BNC w.c., sample of 92 out of 934 total)**

| | Derivative | Freq | Base Freq | Relative Freq |
|---|---|---|---|---|
| **1** | **deplorable** | **149** | **137** | **1.0876** |
| **2** | **comfortable** | **3977** | **3801** | **1.046** |
| **3** | **pop-marketable** | **1** | **1** | **1** |
| **4** | **resizable** | **5** | **5** | **1** |
| 5 | indictable | 57 | 117 | 0.487 |
| 6 | interconnectable | 5 | 11 | 0.46 |
| 7 | predictable | 940 | 3402 | 0.2763 |
| 8 | redeemable | 92 | 340 | 0.271 |
| 9 | smash-and-grabbable | 1 | 4 | 0.25 |
| 10 | admirable | 480 | 1993 | 0.2408 |
| 11 | detachable | 99 | 585 | 0.1692 |
| 12 | severable | 28 | 184 | 0.1522 |
| 13 | superannuable | 1 | 8 | 0.125 |
| 14 | recyclable | 82 | 683 | 0.1201 |

| | | | | |
|---|---|---|---|---|
| 15 | hydrolysable | 3 | 30 | 0.1 |
| 16 | treasonable | 32 | 330 | 0.097 |
| 17 | adapt | 238 | 2542 | 0.0936 |
| 18 | detectable | 281 | 3083 | 0.0911 |
| 19 | respirable | 2 | 22 | 0.0909 |
| 20 | throw-backable | 1 | 14 | 0.071428 |
| 21 | payable | 1681 | 27123 | 0.062 |
| 22 | tractable | 49 | 821 | 0.0597 |
| 23 | recognisable | 365 | 9215 | 0.0396 |
| 24 | iterable | 1 | 33 | 0.0303 |
| 25 | dowsable | 2 | 73 | 0.02739 |
| 26 | patentable | 26 | 986 | 0.0264 |
| 27 | footswitchable | 2 | 95 | 0.02105 |
| 28 | programmable | 87 | 4171 | 0.021 |
| 29 | taxable | 364 | 19735 | 0.018 |
| 30 | endurable | 15 | 932 | 0.0161 |
| 31 | expropiable | 1 | 64 | 0.01562 |
| 32 | consumable | 23 | 1470 | 0.0156 |
| 33 | cancellable | 11 | 1103 | 0.00997 |
| 34 | excisable | 7 | 742 | 0.0094 |
| 35 | tunable | 18 | 2157 | 0.0084 |
| 36 | loadable | 36 | 5797 | 0.0062 |
| 37 | puttable | 3 | 486 | 0.00617 |
| 38 | immunizable | 1 | 163 | 0.00613 |
| 39 | repeatable | 44 | 7228 | 0.0061 |
| 40 | allowable | 196 | 33217 | 0.0059 |
| 41 | redistributable | 1 | 189 | 0.00529 |
| 42 | lovable | 151 | 28793 | 0.00524 |
| 43 | disallowable | 1 | 196 | 0.0051 |
| 44 | editable | 8 | 1674 | 0.0048 |
| 45 | procurable | 2 | 439 | 0.00455 |
| 46 | browsable | 2 | 446 | 0.00448 |
| 47 | characterisable | 5 | 1116 | 0.00448 |
| 48 | maximisable | 3 | 721 | 0.00416 |
| 49 | rateable | 122 | 31330 | 0.0039 |
| 50 | cascadable | 1 | 269 | 0.00371 |

| | | | | |
|---|---|---|---|---|
| 51 | tillable | 2 | 628 | 0.00318 |
| 52 | inhalable | 1 | 319 | 0.00316 |
| 53 | disputable | 12 | 4825 | 0.0025 |
| 54 | countable | 18 | 7493 | 0.0024 |
| 55 | ransomable | 1 | 428 | 0.00233 |
| 56 | serviceable | 124 | 55680 | 0.0023 |
| 57 | climbable | 10 | 4742 | 0.0021 |
| 58 | blendable | 3 | 1561 | 0.00192 |
| 59 | roundable | 6 | 4058 | 0.0015 |
| 60 | playable | 67 | 45833 | 0.00146 |
| 61 | reviewable | 16 | 12076 | 0.0013 |
| 62 | harnessable | 1 | 862 | 0.00116 |
| 63 | eatable | 12 | 10466 | 0.00115 |
| 64 | trimmable | 1 | 1285 | 0.00078 |
| 65 | formulatable | 1 | 1295 | 0.00077 |
| 66 | absorbable | 1 | 1407 | 0.00071 |
| 67 | defendable | 3 | 4340 | 0.000691 |
| 68 | relatable | 8 | 12643 | 0.00063 |
| 69 | overlookable | 1 | 1664 | 0.0006 |
| 70 | conveyable | 1 | 1838 | 0.00054 |
| 71 | trainable | 3 | 6200 | 0.000484 |
| 72 | biteable | 1 | 2177 | 0.00046 |
| 73 | mountable | 2 | 4965 | 0.000403 |
| 74 | dateable | 7 | 20955 | 0.0003 |
| 75 | earnable | 1 | 3712 | 0.00026 |
| 76 | shareable | 7 | 26883 | 0.00026 |
| 77 | listable | 4 | 18316 | 0.00022 |
| 78 | codeable | 1 | 6811 | 0.00014 |
| 79 | pressable | 1 | 17922 | 0.00006 |
| 80 | dealable | 1 | 22708 | 0.00004 |
| 81 | helpable | 2 | 53085 | 0.000038 |
| 82 | researchable | 1 | 27748 | 0.000036 |
| 83 | companiable | 2 | 57953 | 0.000035 |
| 84 | meetable | 1 | 36326 | 0.00003 |
| 85 | moveable | 1 | 47029 | 0.00002 |
| 86 | statable | 1 | 62544 | 0.000016 |

| | | | | |
|---|---|---|---|---|
| 87 | sayable | 1 | 335802 | 0.0000029 |
| **88** | **air-droppable** | **1** | **0** | **?** |
| **89** | **fecundable** | **1** | **0** | **?** |
| **90** | **photoactivatable** | **1** | **0** | **?** |
| **91** | **admissable** | **2** | **0** | **?** |
| **92** | **commensurable** | **4** | **0** | **?** |

**Table 9: Relative frequency of *-ize* derivatives (BNC w.c., sample of 102 out of 658 total)**

| | **Derivative** | **freq** | **base freq** | **rel. freq** |
|---|---|---|---|---|
| **1** | **actualize** | **66** | **2** | **33** |
| **2** | **bureaucratize** | **24** | **2** | **12** |
| **3** | **fraternize** | **31** | **4** | **7,75** |
| **4** | **aluminize** | **4** | **1** | **4** |
| **5** | **vaporize** | **7** | **2** | **3,5** |
| **6** | **dolomitize** | **10** | **5** | **2** |
| **7** | **mathematize** | **4** | **2** | **2** |
| **8** | **sub-vocalize** | **2** | **1** | **2** |
| **9** | **barbarize** | **1** | **1** | **1** |
| **10** | **isomerize** | **1** | **1** | **1** |
| **11** | **laicize** | **1** | **1** | **1** |
| 12 | emphasize | 4887 | 5349 | 0,91 |
| 13 | stabilize | 3 | 4 | 0,75 |
| 14 | anthropomorphize | 2 | 3 | 0,67 |
| 15 | bitumenize | 1 | 2 | 0,5 |
| 16 | ikonize | 3 | 6 | 0,5 |
| 17 | scrutinize | 516 | 1212 | 0,45 |
| 18 | tectonize | 7 | 157 | 0,45 |
| 19 | micronize | 2 | 5 | 0,4 |
| 20 | stigmatize | 106 | 276 | 0,38 |
| 21 | synthesize | 399 | 1199 | 0,33 |
| 22 | characterize | 2683 | 12511 | 0,21 |
| 23 | sympathize | 384 | 2154 | 0,18 |
| 24 | effeminize | 1 | 6 | 0,17 |
| 25 | gnosticize | 1 | 6 | 0,17 |

| | | | | |
|---|---|---|---|---|
| 26 | legitimize | 216 | 1577 | 0,14 |
| 27 | specialize | 3025 | 22114 | 0,14 |
| 28 | crystallize | 341 | 2677 | 0,13 |
| 29 | hypothesize | 182 | 1642 | 0,11 |
| 30 | notarize | 4 | 41 | 0,098 |
| 31 | catheterize | 7 | 116 | 0,06 |
| 32 | commercialize | 80 | 1616 | 0,05 |
| 33 | equalize | 346 | 7207 | 0,048 |
| 34 | computerize | 774 | 17308 | 0,045 |
| 35 | satrapize | 1 | 24 | 0,042 |
| 36 | vocalize | 29 | 761 | 0,038 |
| 37 | mythologize | 15 | 414 | 0,036 |
| 38 | victimize | 126 | 3933 | 0,032 |
| 39 | internalize | 185 | 6722 | 0,028 |
| 40 | generalize | 937 | 34825 | 0,027 |
| 41 | pronominalize | 1 | 41 | 0,024 |
| 42 | atomize | 33 | 1548 | 0,0213 |
| 43 | theorize | 277 | 13129 | 0,021 |
| 44 | solemnize | 9 | 474 | 0,019 |
| 45 | serpentinize | 2 | 113 | 0,018 |
| 46 | popularize | 183 | 10600 | 0,017 |
| 47 | transcendentalize | 2 | 131 | 0,015 |
| 48 | habitualize | 3 | 289 | 0,010 |
| 49 | problematize | 26 | 28993 | 0,0089 |
| 50 | robotize | 4 | 450 | 0,0089 |
| 51 | naturalize | 123 | 14293 | 0,0086 |
| 52 | carnivalize | 1 | 117 | 0,0085 |
| 53 | phonemicize | 1 | 121 | 0,0083 |
| 54 | liquidize | 20 | 2749 | 0,0073 |
| 55 | anathemize | 1 | 147 | 0,0068 |
| 56 | parallelize | 22 | 3409 | 0,0065 |
| 57 | hospitalize | 107 | 17977 | 0,006 |
| 58 | annuitize | 1 | 174 | 0,0058 |
| 59 | tropicalize | 1 | 1791 | 0,0056 |
| 60 | insolubilize | 1 | 191 | 0,0052 |
| 61 | dillonize | 1 | 243 | 0,0041 |
| 62 | pearlize | 3 | 716 | 0,0041 |

| | | | | |
|---|---|---|---|---|
| 63 | siberianize | 1 | 244 | 0,0041 |
| 64 | missionize | 10 | 2623 | 0,0038 |
| 65 | clericalize | 1 | 269 | 0,0037 |
| 66 | hemispherectomize | 3 | 963 | 0,0031 |
| 67 | skeletonize | 2 | 643 | 0,0031 |
| 68 | ottomanize | 1 | 333 | 0,003 |
| 69 | monetize | 8 | 2833 | 0,0028 |
| 70 | indonesianize | 1 | 366 | 0,0027 |
| 71 | textualize | 1 | 389 | 0,0026 |
| 72 | hierarchize | 4 | 1646 | 0,0024 |
| 73 | peripheralize | 2 | 913 | 0,0022 |
| 74 | lyricize | 1 | 579 | 0,0017 |
| 75 | federalize | 6 | 3747 | 0,0016 |
| 76 | finlandize | 1 | 673 | 0,0015 |
| 77 | residualize | 1 | 700 | 0,0014 |
| 78 | poeticize | 1 | 749 | 0,0013 |
| 79 | womanize | 30 | 23213 | 0,0013 |
| 80 | memorandize | 1 | 853 | 0,0012 |
| 81 | plasterize | 1 | 897 | 0,0011 |
| 82 | exoticize | 1 | 1061 | 0,00094 |
| 83 | ethicize | 1 | 1129 | 0,00089 |
| 84 | automize | 2 | 2335 | 0,00086 |
| 85 | incentivize | 2 | 2321 | 0,00086 |
| 86 | apologize | 1 | 1197 | 0,00084 |
| 87 | securitize | 11 | 13837 | 0,00079 |
| 88 | consumerize | 1 | 1639 | 0,00061 |
| 89 | marinize | 1 | 2125 | 0,00047 |
| 90 | corporatize | 2 | 4579 | 0,00044 |
| 91 | canalize | 1 | 2606 | 0,00038 |
| 92 | alcoholize | 1 | 3033 | 0,00033 |
| 93 | ruralize | 2 | 6258 | 0,00032 |
| 94 | journalize | 1 | 3402 | 0,00029 |
| 95 | criticize | 1 | 3861 | 0,00026 |
| 96 | metalize | 2 | 4621 | 0,00022 |
| 97 | derivatize | 1 | 4956 | 0,0002 |
| 98 | opinionize | 1 | 7541 | 0,00013 |
| 99 | cinetropolize | 1 | - | - |

| | | | | |
|---|---|---|---|---|
| 100 | multiparasitize | 1 | - | - |
| 101 | uralitize | 1 | - | - |