Class 8: Taking stock; Models of lexical access

I. Taking stock

(1) Evidence for a domain that ≠ the syntactic word?

- Observation: There are rules/constraints that "see" the stem and some affixes together, but not other affixes. For example, in Dutch a CV will be syllabified together iff they belong to the same stem+suffix$_1$ (or in certain cases of cliticization): *on.+aar.d+ig* 'unkind' = (*on)(aard+ig)*. Similarly, in Dutch an interconsonantal *t* is obligatorily deleted if the CtC sequence belongs to the same stem+suffix$_1$—but not stem+suffix$_2$, as in *zicht-baar*.
- Counteranalysis: Attach the affixes that the rule/constraint sees before it applies, and attach the other affixes later. Or, have low output-output faithfulness between, say, *aard* and *aardig*, but high faithfulness between *aardig* and *onaardig* (where faithfulness can say things like "keep syllable-initial vowels syllable-initial").
- Problems with counteranalysis
  - Misses phonological generalizations. Except for *–achtig*, Dutch suffixes' status as suffix$_1$ or suffix$_2$ is predictable from their phonological shape.
  - Bracketing paradoxes. Affixes that are morphologically closer to the stem will sometimes have to be added later. E.g., ungrammaticality: assuming that *un-* subcategorizes for adjectives, it must be added to *grammatical*, not *grammaticality*. But *–ity* must be added earlier in the derivation, since it shifts primary stress and *un-* doesn't *(ùncógent* instead of **úncogent*).

(2) Assuming we need the p-word, what kinds of phonology refer to it?

Raffelsiefen claims that truly segmental rules—like total assimilation of *n*—don't refer to p-word structure. They can appear to, though, when they refer to syllable structure, and syllable structure is conditioned by p-word structure.

- Have we seen any plausible counterexamples yet?

(3) Assuming we need the p-word, how is it constructed?

The languages we've looked at so far have a p-word for each root (though Greek compounds are said to form a single p-word), and vary in whether prefixes and suffixes can join. In general, it seems that it's easier for prefixes to stay out than for suffixes.

We haven't seen much on clitics yet.

II. Hay 2003[1]

(4) Representation of morphologically complex words

The literature we've been reading so far assumes that affixes are sometimes included in their stems' p-words and sometimes not, leading to different behaviors.

[1] Jennifer Hay (2003). *Causes and Consequences of Word Structure*. New York & London: Routledge.

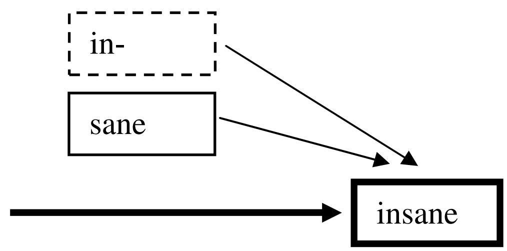
Distinctions made by the grammar
Morphological status of affix: In Italian, suffixes always join the stem's p-wd (evidence: by primary stress), whereas prefixes vary.

Phonological shape of affix: In Dutch, a suffix joins the stem's p-wd iff it is vowel-initial—except –*achtig*—or lacks a full vowel (evidence: syllabification, gapping).

Distinctions made by the lexicon
Affix by affix: In Dutch, -*achtig* always forms its own p-word, despite being vowel-initial.
Word by word: In English, some *in-* words are a single p-word and in others the stem forms its own p-word (evidence: stress). Recall also Baroni's *s*-voicing study, where some words vary even within speaker.

For Hay, any reasonably frequent complex word is represented as a whole, and any reasonably transparent complex word is connected to its subparts.
Therefore, most complex words can be accessed in two ways: directly or through their subparts.

(5) Dual-route models of lexical access—let's compare this to Raffelsiefen
Hay's Figure 4.1 (adapted): direct route is faster because *insane* has higher resting activation (because higher token frequency), shown by thicker outline on box, than *sane*. (Dashed line around *in-* because I don't know how its frequency compares.)



Not explicitly addressed by Hay is the strength of the connections between *insane* and its subparts—this ought to affect the speed of the decomposed route.

(6) Importance of relative frequency (ch. 4)
Hay argues that existing models of processing—even if they claim to predict an effect on the likelihood of direct access for complex words of word frequency only—really predict an effect of *relative* frequency, because direct and decomposed access are in competition.

(7) Importance of phonotactic boundary signals (ch. 3)
If we look at a language's monomorphemes, we'll find some sequences that are very infrequent, such as (for English) *pf*. When such a sequence is encountered (*pipeful*), it could be a signal to the hearer that a morpheme or word boundary is present.

Hay assumes a phonological pre-processing stage that, before any lexical access has occurred, attempts to segment the speech stream using only phonological cues.

Would be worth looking at languages like Dutch where C- and V-initial suffixes are supposed to behave differently, to see how much can be explained by phonotactics: C-initial suffixes might be much more likely to produce illegal sequences, and thus a boundary signal. (Hay mentions something along these lines for English, where there's also a tendency for C-initial suffixes to be separate.)

(8) Direct judgments of morphological complexity (experiments 4, 3)
Hypothesis: words accessed decomposedly will be rated as more morphologically complex than words accessed directly.

E.g. *unobtrusive* vs. *unaffected*: they have similar frequency (42 vs. 54), but *obtrusive* has lower frequency than the prefixed words (17) and *affected* has higher (169).

Result (exp. 4): Subjects rated words like *unobtrusive* as less complex than words like *unaffected*—true of both prefixed and suffixed items.

E.g. *bowlful* vs. *pipeful*: they're matched for word and base frequency, but only *pipeful* contains a low-probability sequence.

Result (exp. 3): For suffixed words, 56% of responses rated the *pipeful*-type word as more complex (as predicted). In prefixed words, however, it was 50%-50%. Hay speculates that the lack of result in prefixed words is due to the semantic opacity of some of the items she used.

(9) Pitch accent placement in a reading task (experiment 5)
Hypothesis: the prefix in a word accessed decomposedly will be more likely to bear a contrastive-focus pitch accent.

Subjects were asked to read sentences like *Sarah thought the document was* ***legible****, but I found it completely* ***illegible***.

Result: Pitch accents occurred more frequently on words like *illiberal* (less frequent than *liberal*) than on words like *illegible* (more frequent than *legible*).

- Do you think this relates at all to Raffelsiefen's claims about secondary stress on true prefixes?

- The effect is gradient (see Figure 4.4 from p. 94): the words don't just form two groups. Can this be explained within Hay's model?

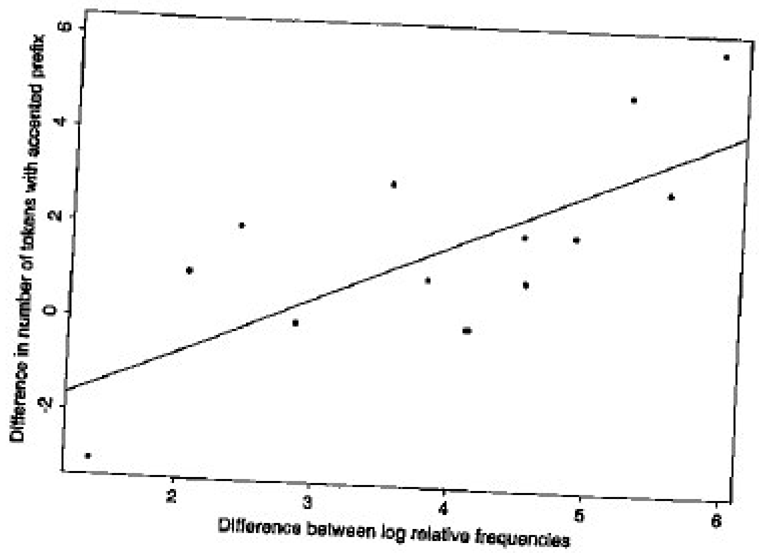


Figure 4.4: Difference in number of tokens attracting a pitch accent to the prefix, as a function of the difference between the log(base freq/derived freq) for each matched pair. (r^2=.53, p<.005)

(10) *t*-deletion in a reading task (experiment 6)
<u>Hypothesis</u>: t → Ø / C__C should apply more frequently/strongly in words accessed directly.

Subjects read sentences like

> *Sam cleaned up the mess very* ***swiftly*** (>*swift*: expect direct access and thus little *t*)
> *Fran tapped Sue's arm very* ***softly*** (<*soft*: expect decomposed access and thus much *t*)
> *John toppled onto stage very* ***daftly*** (<*daft*; low absolute frequency: expect very much *t*)
> *Chris dropped by very* ***briefly*** (control: expect no *t*)

For each speaker, each quadruple is ranked by *t* duration.

<u>Results</u>: As expected, words like *daftly* had the longest *t*s, then *softly*, then *swiftly*, then *briefly*.

As Hay points out, there's a challenge here to other theories (paradigm uniformity): frequent bases themselves should show less *t* than infrequent bases. But, holding derived-word frequency roughly constant, words <u>containing</u> frequent bases show <u>more</u> *t*.

(11) Resting activation of affixes? (e.g., p. 157)
Hay suggests that an affix's resting activation is increased only when the decomposed route wins. Thus, it's a function not simply of how frequent the affix is, but of how frequent are words containing the affix that are accessed decomposedly.

Similarly, we could expect that a base's resting activation depends on how frequent the base is in isolation, but also how frequent it is in decomposed words.

So the dynamics of this model are kind of complicated...in order to predict what a given word will do, you really need to know the whole lexicon.

(12) Toy example

word	*frequency*	*word*	*frequency*	*word*	*frequency*
taste	35	tasteless	30	tasty	50
list	5	listless	90	listy	30
mist	45	mistless	40	misty	60
fame	35	fameless	30	famy	50

Comparing words to bases, we'd expect *listless, tasty, listy, misty,* and *famy* to be accessed directly. Note that *–less* and *–y* are equally frequent.

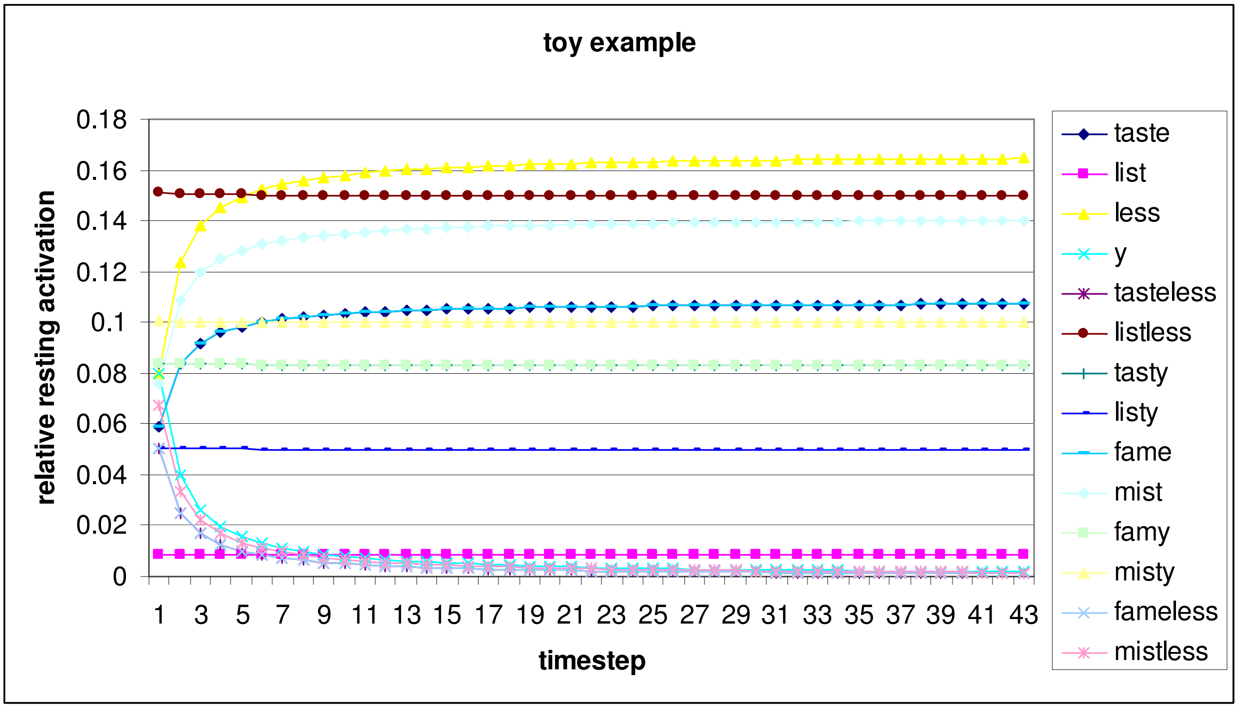
morpheme	*starting activation*	
fame	35	
list	5	
mist	45	
taste	35	
less	47.5	for suffixes, starting activation is frequency / 4 (4 is arbitrary—if it's increased to about 6.3, suddenly most of the words get accessed directly, and both prefixes end up weak)
y	47.5	
fameless	30	
listless	100	
mistless	30	
tasteless	60	
famy	60	
listy	50	
misty	60	
tasty	50	

At each timestep, update each item's resting activation (root, suffix, or word). For *taste*, add 35 (*taste* in isolation), plus 30 for *tasteless* if *taste* and *–less* both have higher activation than *tasteless* at this timestep, plus 50 for *tasty* if *taste* and *–y* both have higher activation than *tasty*.

We end up with a situation where...

- *listy* is the only word getting accessed directly
- *-less* is stronger than *–y*

morpheme	*ending activation*
fame	4390
list	340
mist	5740
taste	4390
less	**6748**
y	**47.5**
fameless	30
listless	6120
mistless	40
tasteless	30
famy	3400
listy	**2040**
misty	4080
tasty	3400



That was just playing around. See Hay & Baayen (2002)[2] for a real model taking into account the frequencies of the whole lexicon to predict which items should be decomposed.

(13) Prefixes vs. suffixes (ch. 3, ch. 5)
Hay points out that in a prefixed word, the whole word can start being accessed before the base can, giving an advantage to direct access. In suffixed words, not so (p. 103). She suggests that this should be another ingredient in the dual-route race.

This is the opposite of what we've seen in the p-word literature, where suffixes seem to be more coherent with their stems than prefixes!

But note that in a prefixed word, the prefix can also start getting accessed early (assume equal resting activation for *pre-* and *–less*, prefix and *fixless*):

prefix	heard *pre* all 400 (say) items beginning with string *pre* are activated, including *pre-*, *prefix* ⇒ relatively weak activation for *pre-*, *prefix*; no activation for *fix*	heard *prefix* now only *pre-*, *prefix*, and *fix* are activated, with *fix* lagging behind
fixless	heard *fix* all 50 (say) items beginning with string *fix* are activated, including *fix*, *fixless* ⇒ relatively strong activation for *fix*, *fixless*; no activation for *-less*	heard *fixless* now only *fix*, *fixless*, *-less* activated, with *–less* lagging behind

[2] Jennifer Hay & Harald Baayen (2002). Parsing and productivity. In G. Booij & J. van Marle (eds.) *Yearbook of Morphology 2001*. Kluwer Academic Publishers. Pp. 203-235.

In sum, in a prefixed word the base lags a bit behind the prefix and whole word, and in a suffixed word, the suffix lags a lot behind the base and whole word. So perhaps we actually predict that prefixed words are more decomposable?

To really know what's predicted, we have to implement the model.

I'm unclear on what effect boundary signals should have on the prefix-suffix difference. In ch. 3, Hay points out that a phonotactically illegal junctural sequence in a prefix word occurs earlier, and thus could have more effect—but Hay's model also assumes that boundary signals have an effect early in processing, before lexical access.

Hay notes that there are more prefixed words (10%) than suffixed words (8%) that manage to be more frequent than their bases, and that the correlation between base frequency and word frequency is stronger for suffixed words than for prefixed words (though there are also more data points for suffixed words).

(14) Semantic drift (ch. 3, 5)
Cf. Raffelsiefen p.p. 179-181.

<u>Hypothesis</u>: Words that are accessed directly are less likely to be related in meaning to their bases (here, less likely to have their base appear in their dictionary definition).

Prefixed words
<u>Results (frequency)</u>: 83% of prefixed words that are less frequent than their bases mention the base in their definition, but only 62% of prefixed words that are more frequent than their bases do. Absolute frequency shows no effect.

<u>Results (phonotactics)</u>: 88% of prefixed words with phonotactically iffy junctural sequences mention the base in their definition, but only 80% of other prefixed words do.

Suffixed words
<u>Results (frequency)</u>: The difference is much weaker than for prefixed words (though still significant): 91% of suffixed words less frequent than their bases mention the base in their definition; only 84% of suffixed words that are more frequent than their bases do. Suggestion of an effect of absolute frequency, but it's not significant.

<u>Results (phonotactics)</u>: No significant difference: 91% of suffixed words with illegal transitions mention their bases in their definitions, and 90% of words with legal transitions do.

(15) Semantic drift—back to Raffelsiefen
Raffelsiefen says something intriguing about the mechanism of semantic drift (p. 178). How do we learn word meanings? If a word doesn't get decomposed, you just guess its meaning from context. If you do decompose the word, you guess its meaning from a combination of context and what you already know about the stem.

Say that *impotent*—somehow—acquires a unitary prosodic structure (*ímpotent*). It's thus parsed as a unit, and its meaning is inferred from context (sample quotes from OED—all from before earliest quote with sexual meaning):

1390 GOWER *Conf.* III. 383 And also for my daies olde That I am feble and **impotent**.
1444 *Pol. Poems* (Rolls) II. 219, I sauh a krevys, with his klawes longe, Pursewe a snayl, poore and **impotent**.
***c*1450** LYDG. *Secrees* 482 He was feble and Oold, And **inpotent**.
1535 COVERDALE *Neh.* iv. 2 Saneballat..saide..What do the **impotent** Iewes?
1538 STARKEY *Engl.* I. i. 3 He ys by syknes or age **impotent** and not of powar to helpe hym selfe.
1568 in H. Campbell *Love-Lett. Mary Q. Scots* App. (1824) 11 When any of the persons of the said councell shall depart, or become **impotent** to serve.
1596 SPENSER *F.Q.* V. xii. 1 O sacred hunger of ambitious mindes, And **impotent** desire of men to raine!
1601 R. JOHNSON *Kingd. & Commw.* (1603) 184 Those onely who are **impotent** in their limes.
1604 SHAKES. *Oth.* II. i. 162 Oh most lame and **impotent** conclusion.

Whereas *unpleasant* for whatever reason, acquires a complex prosodic structure and gets stressed as *ùnpléasant*, which must be parsed as *un+pleasant* (to explain the stress), and so contexts for both *unpleasant* and *pleasant* are used to learn its meaning:

1487-9 J. BARBOUR *Bruce* I. 10 And suth thyngis that ar likand Tyll mannys heryng, ar **plesand**.
1509 J. FISHER *Mornynge Remembr. C'tess of Rychemonde* sig. Biv, A **pleasaunt** & a swete lyfe..a lyfe full of ioye & pleasure.
1535 COVERDALE *Ecclus.* xxii. 6 Euen so is the..doctryne of wyszdome euer **vnpleasaunt** vnto fooles.
1538 ELYOT, *Rancidus*,..vnsauery, or **vnpleasaunt**.
1545 *Primer Hen. VIII* in *Three Primers* (1848) 502 Arise, Lord.., let..the righteous and Christ's disciples make **pleasant** and merry.
1551 TURNER *Herbal* I. 109 The colour is **vnpleasanter** and blacker.
1552 ABP. J. HAMILTON *Catech.* Pref., Na thing culd be to God mair **plesand**.
1553 T. WILSON *Arte of Rhetorique* II. f. 75, Wee confute wholy his saiynges, with some **pleasaunt** iest.
1560 J. DAUS tr. J. Sleidane *Commentaries* f. cccxlvii[v], Ihon Cardinall of Lorayne..had bene all his life time a most **pleasaunt** gest and companion.
***a*1568** R. ASCHAM *Scholem.* II. (Arb.) 132 Preceptes in all Authors..without applying vnto them the Imitation of examples, be..barrayn, vnfruitfull and **vnpleasant**.
1575 GASGOIGNE *Making of Verse* §5 Wordes of many syllables do cloye a verse and make it **unpleasant**.
1581 G. PETTIE tr. S. Guazzo *Ciuile Conuersat.* (1586) I. 45 Which kinde of men, a **pleasant** writer scoffing at, sayth, That that meate is **vnpleasant** in tast, which smelleth of the smoake.
1596 W. RALEIGH *Discov. Guiana* 55 Some of our captaines garoused of his wine till they were reasonable **pleasant**, for it is very strong with pepper.
1585 T. WASHINGTON tr. *Nicholay's Voy.* III. i. 69b, An euill fauoured and **vnpleasant** harmonie.
1604 E. GRIMESTONE tr. J. de Acosta *Nat. & Morall Hist. Indies* I. xiv. 47 From our Peru..they might well bring gold, silver, and **pleasant** monkies.

(16) Polysemy (ch. 5)
Hypothesis: Words that are accessed directly should have more meanings (here, number of definitions in a dictionary). [I'm not completely clear on why this prediction is made, actually...I can see why the number of meanings should be more independent of the base's number of

meanings, but why, all else being equal, should a word that has become independent have more meanings—couldn't it lose some of the meanings of the base form and thus have fewer meanings?]

Prefixed words
Result (frequency): Of the words more frequent than their bases, 57% had an above-average (>5) number of definitions; of the words less frequent than their bases, only 36% did. The difference looks fairly consistent across the prefixes investigated (*dis-, un-, in-* "not"*, in-* "within"*, em-, up, mis-, ex-, trans-*).

But, all of the effect comes from words whose absolute frequency is below average. For high absolute-frequency words (which have above-average polysemy regardless of whether base or derived in more frequent), Hay speculates that there is a ceiling effect.

Result (phonotactics): Hay constructed a set of 24 pairs like *desalt* (legal) vs. *deice* (illegal), matched for word and base frequency and found that "legal" items tended to have more meanings than the "illegal" items.
Then, for all 515 words containing one of 9 prefixes, the number of definitions was counted. For the items with an illegal sequence, 23% had an above-average number of definitions. But for the items with a legal sequence at the juncture, 41% had an above-average number of definitions.

Suffixed words
Results (frequency): There must be an error in Table 5.12. Working from Table 5.14, 42% of the words that are more frequent than their bases have an above-average (>2) number of meanings, and only 29% of the words that are less frequent than their bases do.

Again, the effect is all from words of below-average absolute frequency—words of above-average frequency just have lots of meanings regardless of relative frequency.

Results (phonotactics): Effect is opposite of predicted. 36% of words with illegal transitions have an above-average number of definitions (>2), and 27% of words with legal transitions do.

(17) Phonotactics vs. relative frequency (ch. 3)
Prefixed words
Result: For the 515 words, 12% of the words with legal junctural phonotactics are more frequent than their bases, but only 4% of those with illegal junctural phonotactics are. Thus, the two factors that are supposed to predict direct access agree.

Suffixed words
Result: For both words with legal transitions and words with illegal transitions, 8% are more frequent than their bases—i.e., no relationship.

- Chicken or egg? Do words that are accessed directly, because of relative frequency, change to get better phonotactics, or do words with good phonotactics get directly accessed, which (somehow) changes their relative frequency? Which would be more consistent with Raffelsiefen's model?

(18) Baayen's $\wp$

For a given affix *i*, let $n_{i,k}$ be the number of word types in a corpus that contain affix *i* and occur *k* times. Thus, $n_{un\text{-},1}$ is the number of word types that contain *un-* and occur just once in the corpus. Let N_i be the number of tokens in the corpus that contain the affix *i*.

Baayen proposes a measure of productivity $\wp_i = n_{i,1} / N$.

Intuitively, the idea is that a very productive affix should produce lots of nonce formations, and thus there will be lots of singly-occurring words with that affix.

[See Lüdeling, Evert & Heid 2000[3], Evert & Lüdeling 2001[4] for useful critical discussion of this measure. They note that $\wp$ changes (decreases) as the sample size increases. Thus, if you compare two different affixes with different frequencies, the more-frequent one's $\wp$ will look artificially smaller. Baayen's way of correcting for this depends on extrapolated vocabulary-growth curves, using parameter values that are, LEH argue, highly sensitive to errors in the corpus data—and, as a model of morphological learning, maybe highly sensitive to individual differences in acquisition environment:

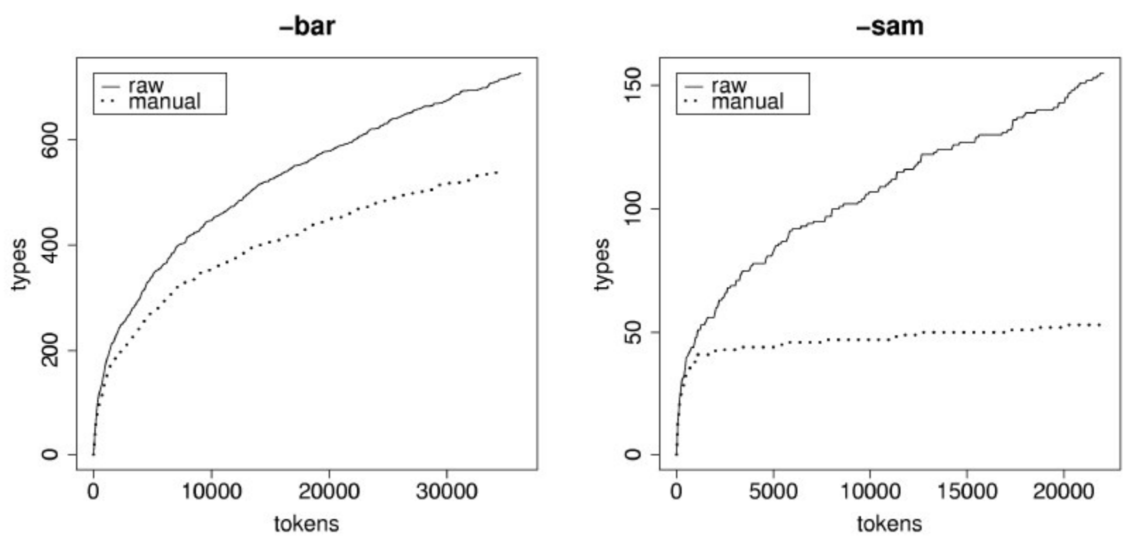


Figure 2: Raw and manually corrected vocabulary growth curves for *–sam* and *–bar*.

(Evert & Lüdeling 2001, p. 4 of ms. version: German *-bar* looks productive before and after the cleanup, but *–sam* looks unproductive after the cleanup)]

(19) Productivity (ch. 7)

Hay shows that, for a set of 12 English affixes, the correlation between type frequency and $\wp$ is poor (p. 146).

---

[3] Anke Lüdeling, Stefan Evert & Ulrich Heid (2000). On measuring morphological productivity. *Proceedings of KONVENS 2000*: 57-61.

[4] Stefan Evert & Anke Lüdeling (2001). Measuring morphological productivity: Is automatic preprocessing sufficient? In Paul Rayson, Andrew Wilson, Tony McEnery, Andrew Hardie & Shereen Khoja (eds.) *Proceedings of the Corpus Linguistics 2001 conference*: 167-175.

The correlation between ℘ and proportion of types more frequent than their bases is good (negative—the more derived>base, the less productive), however (p. 149):

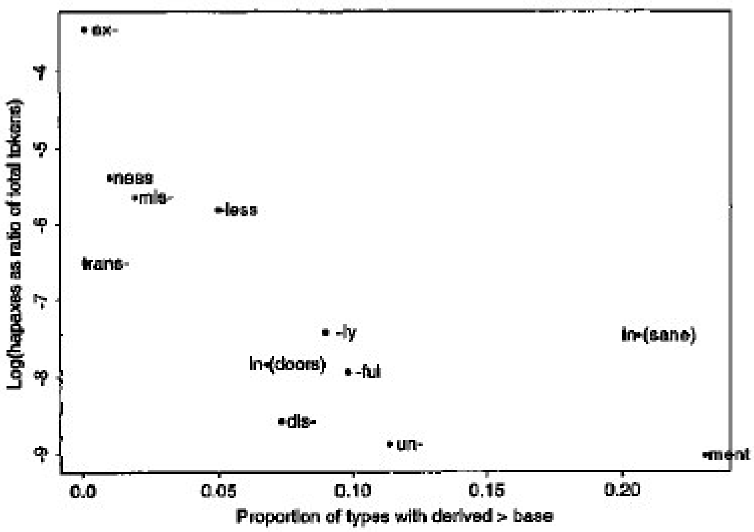


Figure 7.3: Proportion of forms for which the derived form is more frequent than the base, vs log productivity – measured as the ratio of hapaxes to total tokens occurring in corpus (cf Baayen and Lieber 1991).

And the correlation between ℘ and proportion of types with illegal phonotactics at the morpheme boundary is good (positive—the more phonotactically illegal sequences, the more productive) (p. 150):

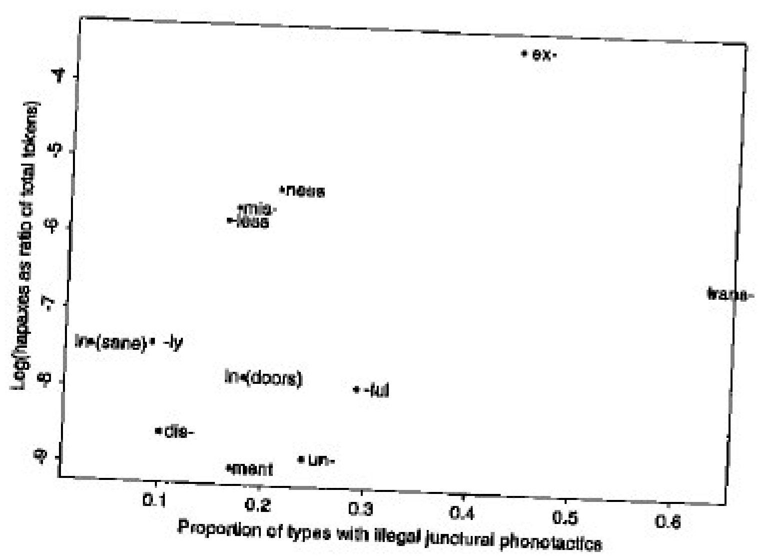


Figure 7.4: Log proportion of forms containing illegal junctural phonotactics vs log productivity – measured as the ratio of hapaxes to total tokens occurring in corpus (cf Baayen and Lieber 1991).

Hay acknowledges that if we look at a wider set of affixes (these 12 are all reasonably frequent and reasonably productive), then type frequency could conceivably be important.

(20) Affix ordering (ch. 8)
"[A]n affix which can be easily parsed out should not occur inside an affix which can not." (p. 161) E.g. **helpfulism*: *pf* sequence is boundary signal, suggesting *help+fulism*. Since *fulism* isn't an affix, hard to recover the meaning.

	"Level 1"	"Level 2"
suffix examples	*-al, -an, -ary, -ate, -ese-, -ette, -ian, -ic, -ify, -ity, -or, -ory, -ous, -th*	*-age, -dom, -en, -er, -ful, -hood, -ish, -less, -let, -like, -ling, -ly, -most, -ness, -ship, -some*
phonotactics	most begin with V → unlikely to produce illegal sequence → direct access	most begin with C → likely to produce illegal sequence → decomposed access
how many forms more frequent than base?	from 4% (*-ify*) to 32% (*-ic*); average 17% → direct access	from 0% (*-dom*, *-hood*, *-let*, *-ship*) to 12% (*-age*); average 5% → decomposed access
Baayen's ℘	average = .002	average = .030

(21) Acceptable affix combinations (ch. 8.4)
Consumerist: *-er* is more frequent than *–ist*, so in general we might not expect *–erist* words (though recall that affix parsability depends on more than affix frequency), but *consumer* is more frequent than *consume*, and so treated as a whole: *consumer+ist*.

Similarly, *–ionist* words mostly come from *Xion* words that are more frequent than their bases: *conservation+ist*. Same goes for *–ionary* and *–ioner* words.

(22) Prefixes vs. suffixes II (ch. 8.10)
Hay notes that the famous bracketing paradoxes involve a Level I suffix attaching to a word with a Level II prefix, but not vice-versa:

[[de – congest] –ant] LII root LI	but not	*[in- [care – ful]] LI root LII

Consider the timecourse of lexical retrieval (for the hearer, I guess):

heard so far	*de-* or *in-*	*decongest* or *incare*	*decongestant* or *incareful*
getting activated	*de*	*de,* ***decongest****, congest*	*de, decongest, congest, decongestant,* ***congestant***
	in	*in,* ***incare****, care*	*in, incare, care, incareful,* ***careful***

So all else being equal, *[[prefix root] suffix]* has a head start over *[prefix [root suffix]]*.
This makes *decongest-ant* relatively easy to parse correctly, and **in-careful* hard to parse.

(23) Experiment 7a, 7b: judgments of *–al* affixability
<u>Hypothesis</u>: decomposed suffixed words should be less able to be further suffixed with *–al* than directly-accessed suffixed words.

Task was to pick the better member of pairs like *arrangemental – investmental*.

Results:

- (7a) 56% of responses preferred the *Xmental* form whose *Xment* was more frequent than *X*. Since the items were roughly matched for *Xment* frequency, this means that items with lower-frequency *X* were more able to take *–al*—might be surprising under some theories.
- (7b) 67% of responses preferred the *Xmental* form whose *X* ends in a vowel, creating a highly probably V-C transition (*deploymental*) to the *Xmental* form whose *X* ends in a C and creates a low-probability C-C transition (*recruitmental*).