LSA.360 The Phonological Status of Morphologically Complex Words

Class 3: Inner vs. outer affixes (vs. compounding); productivity

1 Review of last time

(1) Baroni's study of Northern Italian [s]-voicing

Italian prefixed words: some behave as really prefixed

- re+**s**uscitare 'to solve' (pronounced with [s] in 55/60 of Baroni's experimental tokens)

...some behave as though monomorphemic

- de**s**inenza 'desinence' (58/60 with [z])

...and some vary, within and across speakers

- co(+)seno 'cosine' (36/60 [s], with 6/12 subjects showing internal variation)

Best predictors were stem transparency (subjects' ratings of the degree to which the whole word contains the meaning of the stem) and prefix length, in syllables (longer prefix → more [s]).

(2) Morpheme separability

- Can be seen as a property of an individual word (is it *reform* or *re+form*, or both, or *re+form* (the '+' is supposed to be gray there))?
- Can be seen as a property of an affix overall (how much is *ana-* a prefix? *hyper-*?)
- Can be seen as a property of a stem overall (how much of a stem is *-ject*?)

(3) Preview

- Inner vs. outer affixes: derivational approaches, problem of phonological correlates
- Measures of productivity
- Flash forward: basics of Hay's model
- Prefixes vs. suffixes
- Affix ordering
- Affixation vs. compounding
- From properties of words to affixes overall

2 Inner vs. outer affixes

(4) Classic English example: stress-shifting vs. stress-neutral suffixes

Based on discussion in Raffelsiefen 2005.

	stress-shifting	*stress-neutral*	*what's weird about the stress-neutral*
rádical	ràdicál-ity	rádical-ness	i.cal.ness: 3 unstressed syllables
phónème	phoném-ic		
devélop		devélop-ment	lop: unstressed, closed penult
sólid	solíd-ity		
Sùdán	Sùdan-ése		
hóspital		hóspital-ìze	pi.tal: word-internal unstressed-unstressed sequence in verb (internal lapse)
seléct		selèct-ée	lèc.tée: adjacent stressed syllables (stress clash)

(pp. 218-219)

(5) Lexical-phonology perspective (e.g., Kiparskyan)

- Stress-shifting affixes attach earlier, and result is subject to normal stress system.
- Stress-neutral affixes attach later, when some stress is already in place.

	/radical/	/radical/	/Sudan/	/select/
attach earlier affix (e.g., "stem-level")	radicality	--	Sudanese	--
apply stress (rules or OT)	ràdicálity	rádical	Sùdanése	seléct
attach later affix ("word-level")	--	rádicalness	--	seléctée
further stress adjustments, with some faithfulness to stress already in place		--	--	selèctée

(6) Prosodic-domain perspective (Aronoff & Sridhar, Szpyra, Hammond)

Word-formation rules associated with stress-neutral affixes place them outside the stem's prosodic word:

stress-shifting	*stress-neutral*	
(ràdicál-ity)$_\omega$	(rádical)$_\omega$-ness	
(phoném-ic)$_\omega$	(devélop)$_\omega$-ment	
(solíd-ity)$_\omega$	(hóspital)$_\omega$-ìze	Raffelsiefen point out that in cases like this, either the syllabification is at odds with the p-word structure, or resyllabification can let one segment (*l*, *t*) leave the p-word without otherwise altering the structure: *(hóspita)$_\omega$l-ìze*, *(selèc)$_\omega$t-ée*
(Sùdan-ése)$_\omega$	(selèct)$_\omega$-ée	

(7) Phonological correlates of class membership: do you think these are problematic?

- V- versus C-initial suffixes in English (Raffelsiefen)

The stress-shifting suffixes are all V-initial (converse not true)

- How could we analyze this in the prosodic view?
  - What do we do about non-stress-shifting V-initial suffixes?
- What role could this play in the derivational view?

- Full vowels vs. schwa in Dutch (Booij)

Except for *–achtig*, suffixes that are vowel-initial or lacking a full vowel behave as more cohering, at least with respect to gapping:

Gapping:

(zicht)-(**baar**) en (tast)-(baar)	→	(zicht)__ en (tast)-(baar)	'visible and tangible'
(christen)-(-**dom**) en (heiden)-(dom)	→	(christen)__ en (heiden)-(dom)	'Christianity and heathendom'
(rod-**ig**) of (groen-ig)		*rod-__ of groen-ig	'reddish or greenish'

(Booij 2002, pp. 171-172)

- Same questions as above.
- Extra fact: *vijf.-ling* 'quintuplet', not **vij.f-ling* as would normally be expected by Dutch syllabification rules (van Oostendorp 1994). Ideas?

- Number of syllables in Italian prefixes

As we saw earlier, Italian disyllabic prefixes can bear the signs of primary stress (long vowels, lower-mid vowels), but monosyllabic prefixes (except maybe *ex-*) can't.

  - Same questions as above.

3 Distributional correlates: productivity

(8) Basics

Productivity of an affix ≈ its ability to form new words

English example

- *–ness* is pretty productive—I feel comfortable attaching it to adjectives to form words that I don't think I've heard before: *kludginess*, *dyspepticness* (no Google hits!)
- *–th* is not productive—I can make jokes like *coolth*, but that's it

Productivity probably has something to do with phonological behavior! Less-productive affixes tend to be more phonologically integrated with their stems.

- e.g., vowel changes with *–th*: d[i]p – d[ɛ]pth, l[ɔ]ng – l[ɛ]ngth

(9) Correlates of productivity

"Taking into account the way complex words are stored and processed in the mental lexicon, it is argued that the productivity of a given process emerges as a syndrome of properties, with parsability, relative frequency, semantic and phonological transparency as important factors." (Plag 2004, pp. 1-2 of ms.)

(10) Why look at correlates of productivity

- How do we learn whether an affix can be used productively or not? Or, to think of it another way, how does a language user's ability to coin new words with some affix emerge from the user's experience with that affix? To investigate this, we need to look at what information is available to the learner/what the learner's experience is like.
- Are there convenient proxies for productivity that we can find in, say, a corpus, instead of having to rely on vague intuitions?

(11) Coinages over time

Use a dictionary, like the OED, to trace the dates at which words with some affix have entered the language. This information is not available for many languages!

And we have to trust that the criteria for getting into the dictionary reflect something (like what) we want to measure.

(12) 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, $n_{\text{-ness}1}$ is the number of word types that contain *–ness* and occur just once in the corpus.

If the corpus is big enough, these single-occurrence items ("hapax legomena") should tend to be new coinages, or otherwise unfamiliar to readers/listeners—and yet presumably understandable, or they wouldn't have been used.

Let N_i be the number of tokens in the corpus that contain the affix *i*.

Baayen proposes a measure of productivity $\wp_i = n_{i,1} / N_i$: the proportion of word tokens with affix *i* that are hapax legomena.

As Plag explains, the relationship between $\wp$ and productivity should go both ways: if you're always encountering new words with the affix, you're forced to parse it out, so the affix or word-formation rule gets reinforced in your grammar. Conversely, if you have a strong rule, you will be able to coin lots of new words with it.

[See Lüdeling, Evert & Heid 2000[1], Evert & Lüdeling 2001[2] 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):

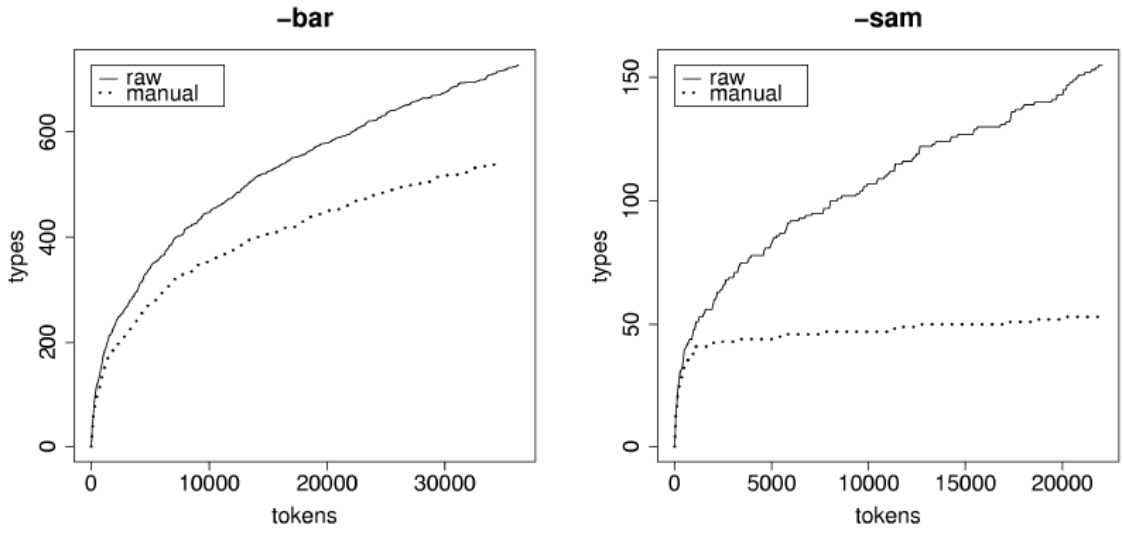


Figure 2: Raw and manually corrected vocabulary growth curves for *–sam* and *–bar*.

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

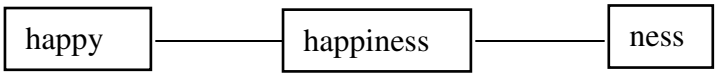
4 Flash forward: basic idea of Hay (2003)

[In Class 5 we'll see Hay's evidence and an implementation]

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

[2] 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.

(13) Representation of morphologically complex words
For Hay, any reasonably frequent complex word is represented as a whole, and any reasonably transparent complex word is connected to its subparts.

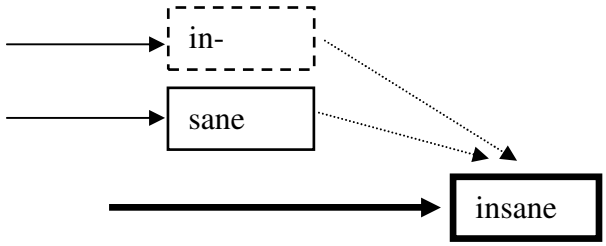


Therefore, most complex words can be accessed in two ways: directly or through their subparts.

This is a common assumption among many psycholinguists; what Hay does is point out that *relative* strength of the whole vs. the subparts is predicted to be important (not just frequency of the whole word). And, she provides evidence from English to back up this prediction.

Boundary signals should also be important: Hay assumes a **fast phonological preprocessor** stage in word recognition, in which possible boundaries are identified through phonological information alone—this information about possible chunks is then passed on to the lexicon.

(14) Dual-route model of lexical access
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—a model could allow this to affect the speed of the decomposed route.

5 Affix ordering

(15) Hay 2003, ch. 8
"[A]n affix which can be easily parsed out should not occur inside an affix which can not." (p. 161) E.g. English **helpfulism*: *pf* sequence is **boundary signal**—a sequency that can't occur within a monomorpheme in English—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 $\wp$	average = .002	average = .030

(16) Prefixes vs. suffixes (ch. 8.10)

Recall from last time the cross-linguistic claim that—where and to the extent that there is an asymmetry—prefixes are less phonologically cohering than suffixes: (prefix (stem suffix))—e.g., syllabification in Dutch (Booij; similar in German, I think):

(prefix)(stem-suffix)	(on).-(aar.d-ig)	'unkind'
(word clitic) (word)	(koch.te.n- 't).(boek)	'bought the book'

To see if the difference really can follow from lexical access, we need an implemented model, but we can imagine...

TIME ---->

heard so far	prefix	prefix stem	
getting activated	prefix	prefix, stem, whole_word	upshot: prefix can get activated before whole word does

heard so far	stem	stem suffix	
getting activated	stem	stem, suffix, whole_word	upshot: suffix doesn't get heard until whole word is heard. But! stem also could get activated earlier in suffixed words—and wouldn't that favor the decomposed parse?

(17) Prefixes vs. suffixes in bracketing paradoxes (ch. 8.10)

Classic examples like *un-grammatical-ity*:

- Morphologically, *un* has to attach first, because it attaches to adjectives (like *grammatical*), not nouns (like *grammaticality*): [[un [grammatical]$_A$]$_A$ ity]$_N$
- But, in derivational theories, we think *–ity* is an earlier-level affix than *un–*.

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

Consider the timecourse of lexical retrieval (for the hearer):

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

So, if prefix+root is a word, you can recognize it (*decongest*) and then hear the suffix. The idea would be that this gives an advantage to ((*prefix stem*) *suffix*), even if the suffix is, in general, a more "inner"/"early" one.

But there is no such advantage for (*prefix* (*stem suffix*)), because the stem+suffix (*careful*) can't get recognized until relatively late...

- Anything in this reasoning that we need to spell out further?

(18) Affix ordering (Hay's experiments 7a & 7b)
Hypothesis: decomposed suffixed words should be less able to be further suffixed with *–al* than directly-accessed suffixed words.

- Why do we expect this?

Subjects' 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*).

6 Affixation vs. compounding—slippery slope or tidy terraces?

Is there a clean grammatical distinction between affixation and compounding, or are they merely extremes of a distributionally-defined continuum? (Preview: I have no answer to offer!)

English example (from Plag): *error-free* vs. *errorless*

Numbers from BNC (full corpus, no cut-offs)—**token frequency is messed up**, because the numbers are occurrences per million, rounded to the nearest integer; thus, for most word's it's zero.

	-free	-less
type frequency	450	947
token frequency	8	874
token frequency, reestimated by adding 1 to each word's frequency	458	1821
number of words with morpheme that have token freq. <0.5 per million (as proxy for hapaxes)	443	839
Baayen's $\wp$ **rough (over)estimate**	443/458 = .9	839/1821 = .46

So *–free* does indeed look, distributionally, as though it's more productive than *–less*. But is that the cause of its phonological behavior (e.g., stress)? Or are both distribution and phonological behavior the results of differing morphological status?

7 Influence of whole lexicon on individual items

More work from Baroni (2002), this time on English.

(19) Inspiration: Rissanen's Minimum Description Length principle
You could describe a lexicon/corpus by listing it:

kind
unkind
kindly
kind
fair
unfair
clear
unclear
unkind
unfair
fair
kind
unclear (69 characters [to do this properly, we put everything in 0s and 1s, consider most efficient code, etc.])

Or you could take out some morphemes

1. un	2	
2. kind	1,2	
3. fair	kindly	
4. clear	2	
	3	
	1,3	
	4	
	1,4	
	1,2	
	1,3	
	3	
	1,4	(52 characters: it's a bit of a savings)

What if you take out *–ly* too?

1. un	2	
2. kind	1,2	
3. fair	2,5	
4. clear	2	
5. ly	3	
	1,3	
	3,5	
	4	
	1,4	
	1,2	
	1,3	
	3	
	1,4	(56 characters: less of a savings)

So, whether it's worth pulling out a morpheme and putting it in the lexicon depends on how long it is, how many items it occurs in, and what morphemes it occurs with (since pulling them out will cost something).

(20) Baroni's distribution-driven learner

- Guess one morpheme boundary to add somewhere in the lexicon (heuristics to help the choice be good: the "prefix" must have some minimum frequency, and the "stem" must have some minimum length).
- Calculate the new length of the lexicon + encoding
- Choose the best single guess—if it's better than not adding a morpheme boundary at all
- Repeat until no futher improvements.

(21) Results for English

(22) *prefixes postulated by DDPL*
ad- auto- co- com- con- cor- de- dis- ex- extra- juris- in- inter- man- mis- non- over- para- pre psycho- radio- re- sub- sup- super- sur- tele- un- under- (p. 17 of ms.)

(23) *prefixes missed by DDPL*
a- an- anti- arch- contra- counter- fore- hyper- il- im- ir- mal- mini- outpost- pro- pseudo- trans- ultra- (p. 18 of ms.)

Some of them weren't in the corpus, and others didn't occur often enough for the frequency criterion above to consider them. What's left is *anti-*, *contra-*, *pro-*, *trans-*.

(22) Validity of individual items' parses

Survey: 8 English speakers asked to rate morphological complexity of words like *detachment* (not complex, according to learner) vs. *de+fender* (complex, according to learner).

- The 8 subjects were highly correlated with each other
- There was a significant difference between ratings that subjects assigned to words predicted to be complex and ratings that subjects assigned to words predicted to be simple...
- ...even "when only semantically opaque words are considered"