LSA.360 The Phonological Status of Morphologically Complex Words

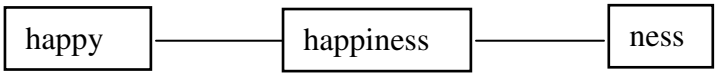
Class 4: A parsing model: evidence (Hay) and implementation (Baayen & colleagues)

Ideas about unithood that we've seen rely implicitly on models of lexical learning and access. Today we'll look at one in particular—first, the idea and evidence for it (from Hay 2003), then an implemented model (Baayen et al. 2000, Baayen & Schreuder 2000, adopted in Hay & Baayen 2002).

1 Hay 2003

(1) Representation of morphologically complex words

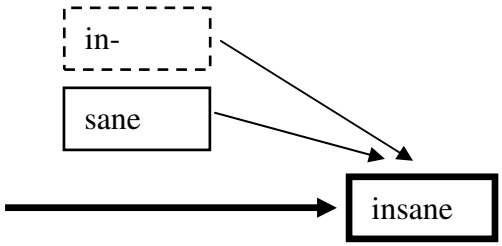
For Hay, any reasonably frequent complex word is represented as a whole, and any reasonably transparent complex word is connected to its subparts.



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

(2) 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 (although the Baayen model doesn't).

(3) 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 (but seldom test for) an effect of *relative* frequency, because direct and decomposed access are in competition.

(4) 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 (*fast phonological preprocessor, FPP*—Pierrehumbert 2001, 2002 and references therein) that, before any lexical access has occurred, attempts to segment the speech stream using only phonological cues. In Pierrehumbert's proposal, these hypothesized segmentations would be the places where (overlapping) new lexical searches are launched (instead of being launched, say, every 5 msec.).

Would be worth looking at languages like Dutch (and English) 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 pursues something along these lines for English.)

(5) 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 word (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.

(6) 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***. (Will it be pronounced *ilLEgible*, as normally, or *ILlegible*, with contrastive focus?)

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

- 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?

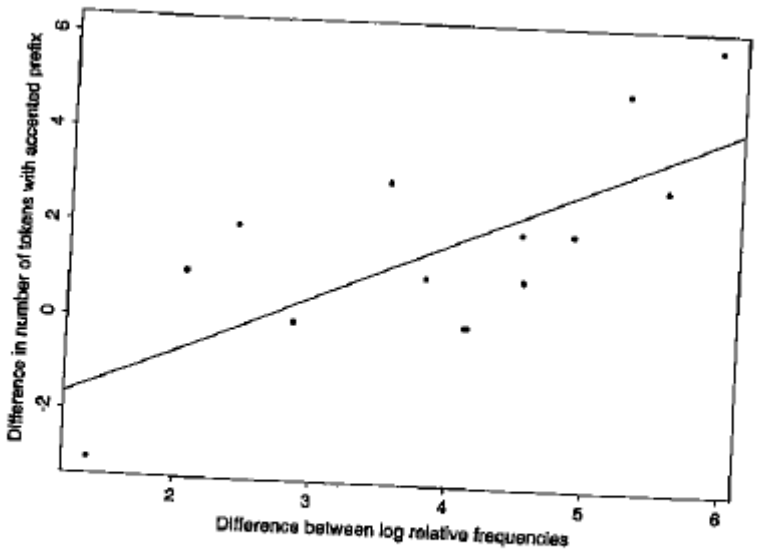


Figure 4.4: Difference in number of tokens attracting a pitch accent to the prefix, as a function of the difference between the log(base freq/derived freq) for each matched pair. (r^2=.53, p<.005)

(7) *t*-deletion in a reading task (experiment 6)
Hypothesis: 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 [-> possibly no lexical entry]: expect very much *t*)

*Chris dropped by very **briefly*** (control: expect no *t*)

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

Results: 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 in isolation should show less *t* than infrequent bases (given all the results of reduction/automaticity of frequent words). But, holding derived-word frequency roughly constant, words containing frequent bases show more *t*.

(8) 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 frequently it is accessed in complex words—but how frequently the base is accessed in complex words should depend in part on how frequent the base is.

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! You also need to build up item frequencies from scratch, mimicking lexical acquisition.

(9) Semantic drift (ch. 3, 5)
Hypothesis: Words that are accessed directly are less likely to be related in meaning to their bases (here, operationalized as less likely to have their base appear in their dictionary definition).

Prefixed words
Results (frequency): 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.

Results (phonotactics): 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
Results (frequency): 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.

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

(10) Aside: cf. Raffelsiefen 1999

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.

(11) 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: I 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, so I'm 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.

Would be interesting (because of worry above) to plot derived word's number of meanings against base's number of meanings, separately for word>base and base>word—do we see a difference in the tightness of the correlation, or just its slope?

(12) 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.

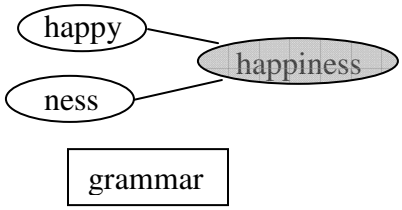
Suffixed words
Result: For both words with legal transitions and words with illegal transitions, 8% are more frequent than their bases—i.e., no relationship.

- Chicken or egg?[1] 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? [The first possibility could be investigated philologically.]

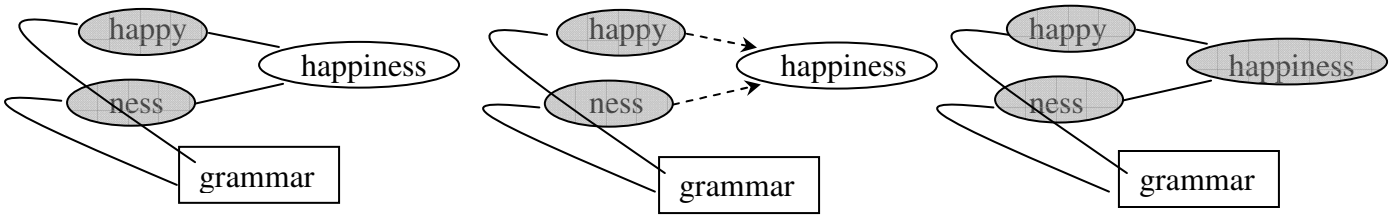
(13) Thoughts on effects of raw (whole-word) frequency
There could really be three routes of access in this model:
[shading represents activation; arrow indicates spreading of activation]

[1] From the multilingual saying "which came first, the chicken or the egg?", apparently having its roots in Aristotle, who didn't know about evolution.

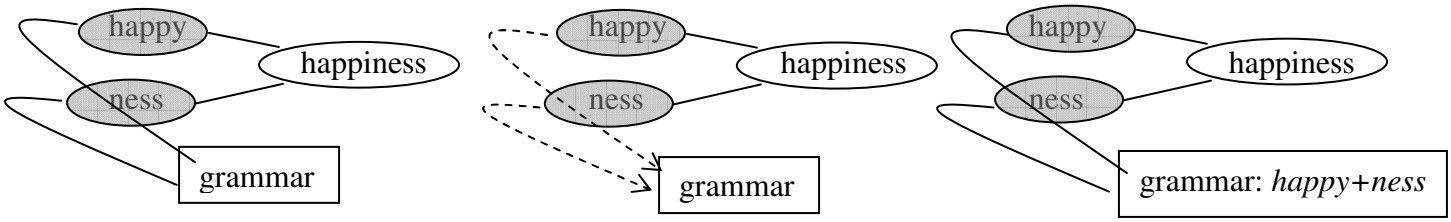
1. Direct: whole word reaches an activation threshold first [some subparts could reach threshold earlier, but crucially not all of them do]



2. Indirect: subparts reach activation threshold before the whole-word entry does; the parts then spread that activation to the whole-word entry, and it reaches threshold



3. Synthesized: the subparts reach threshold before the whole-word entry does, and those parts are then put together (somewhere in the grammar); this process is completed before the whole word can get activated.



- Direct route is favored by frequent whole word, infrequent parts.
- Indirect route is favored by frequent parts, whole word that is not too frequent but also not too infrequent (since the whole word does have to get activated), and strong connection between parts and whole (ideas on where those should come from?).
- Synthesized route is favored by frequent parts, infrequent whole, weak connection between parts and whole.

(14) 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 sequence 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

(15) 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?

(16) 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*): $[[\text{un } [\text{grammatical}]_A]_A \text{ ity}]_N$
- 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?

(17) 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*).

2 Implemented model

(18) Baayen et al.'s MATCHECK model

Given a lexicon (morphemes and morpheme combinations), predict which parse will win.

Dutch Example (BS p. 1281)

bestelauto	*be+stel+auto*	('delivery van')	possible and most likely to come to mind
	bes+tel+auto	('berry counting car')	possible but unlikely to come to mind
	bes+t+el+auto	--	real morphemes but illegal combination

MATCHECK models the timecourse of word recognition.

- When a lexical entry reaches an activation threshold (actually, when its share of the total activation, $p(w, t)$, reaches a threshold), it is copied into a memory buffer.
- When a set of lexical entries that span the target word (e.g. *be*, *stel*, *auto*) exists in the buffer, that parse of the target becomes available.

- First full parse to become available is assumed to have "won" (but we could also assign interpretations to other information, such as how long it took the first parse to become available, and which other parses become available when).

(B&S p. 1286)

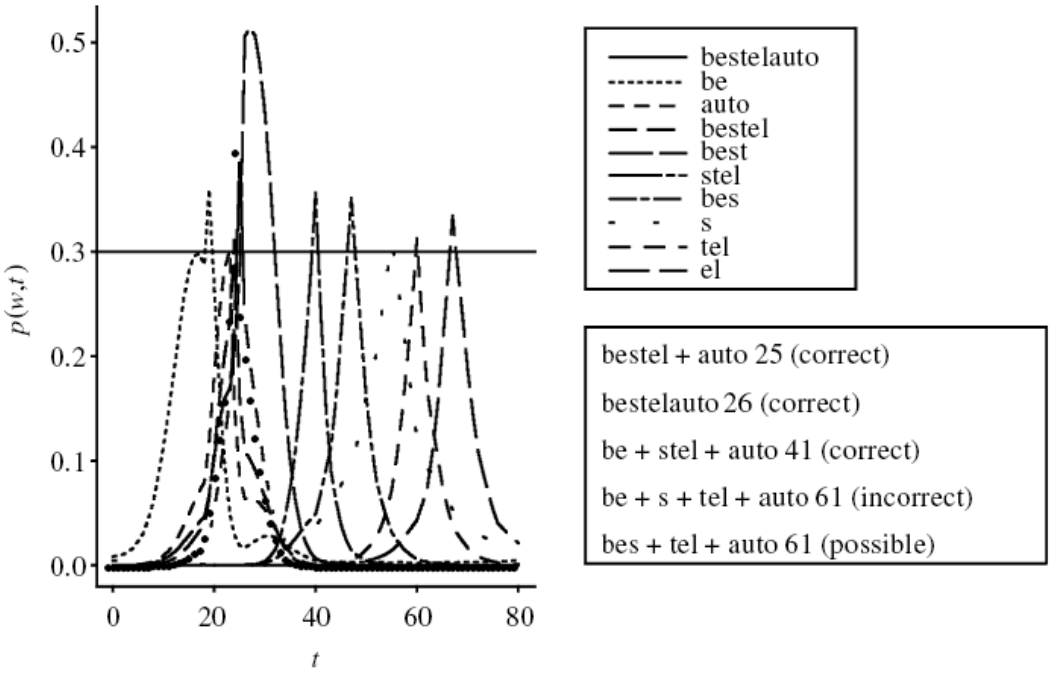


Figure 2. Probability of identification $p(w, t)$ for selected access representations as a function of time-step t, with activation threshold $\theta = 0.3$, for *bestelauto*, 'delivery van'. The time-steps at which full spannings become available are listed in the lower right-hand corner.

(19) What determines activation?

If an entry's weight is still being allowed to increase, the preceding activation is just multiplied by that node's decay rate:

$a(w, t) = a(w,t\text{-}1)/\delta_w$ (*a* stands for 'activation'; *w* identifies the entry; *t* is the current timestep; δ_w is the decay rate for that node)

If the entry's weight is determined to have peaked, the activation asymptotes out to the original (resting) activation:

$a(w, t) = |a(w,t\text{-}1)\text{-}a(w,0)|*\delta_w$

What determines when an entry starts decreasing its activation?

If an entry has not yet reached threshold, and it is either edge-aligned with the target[*] or similar enough[†] to the target, then it gets to keep increasing.

([*]or, edge-aligned with a substring of the target that can be formed by stripping off outer affixes that have reached threshold; e.g., *in* is edge-aligned with *uninformed* if *un* has already reached threshold)

([†]similarity is length of the target, if entry is superstring of target; otherwise, similarity is length of the target minus Levenshtein edit distance between entry and target; entry is "similar enough" if its similarity ≥ *t*)

What determines an entry's decay rate? Informally, it's a combination of its length, its resting activation, and how much of the target it matches.

$$\delta_w = \delta \frac{L(w)}{L(w) + \frac{\alpha}{L(w)} \log(a(w,0))} + \left(1 - \delta \frac{L(w)}{L(w) + \frac{\alpha}{L(w)} \log(a(w,0))}\right) \left(\frac{|L(w) - L(T)|}{\max(L(w), L(T))}\right)^{\zeta}$$

if $\zeta > 0$, and otherwise $\delta_w = \delta$

where L(*i*) is the length of *i*, *T* is the target word, and there are free parameters $\alpha > 0$ (spike parameter), $\zeta > 0$ (forest parameter), and $0<\delta<1$ (overall decay rate).

(20) Testing the model's predictions—example from the literature

Hay & Baayen (2002) find that, for a set of Dutch words in *-heid* Matcheck's parsing times are correlated with subjects' lexical-decision reaction times (from a previous study).

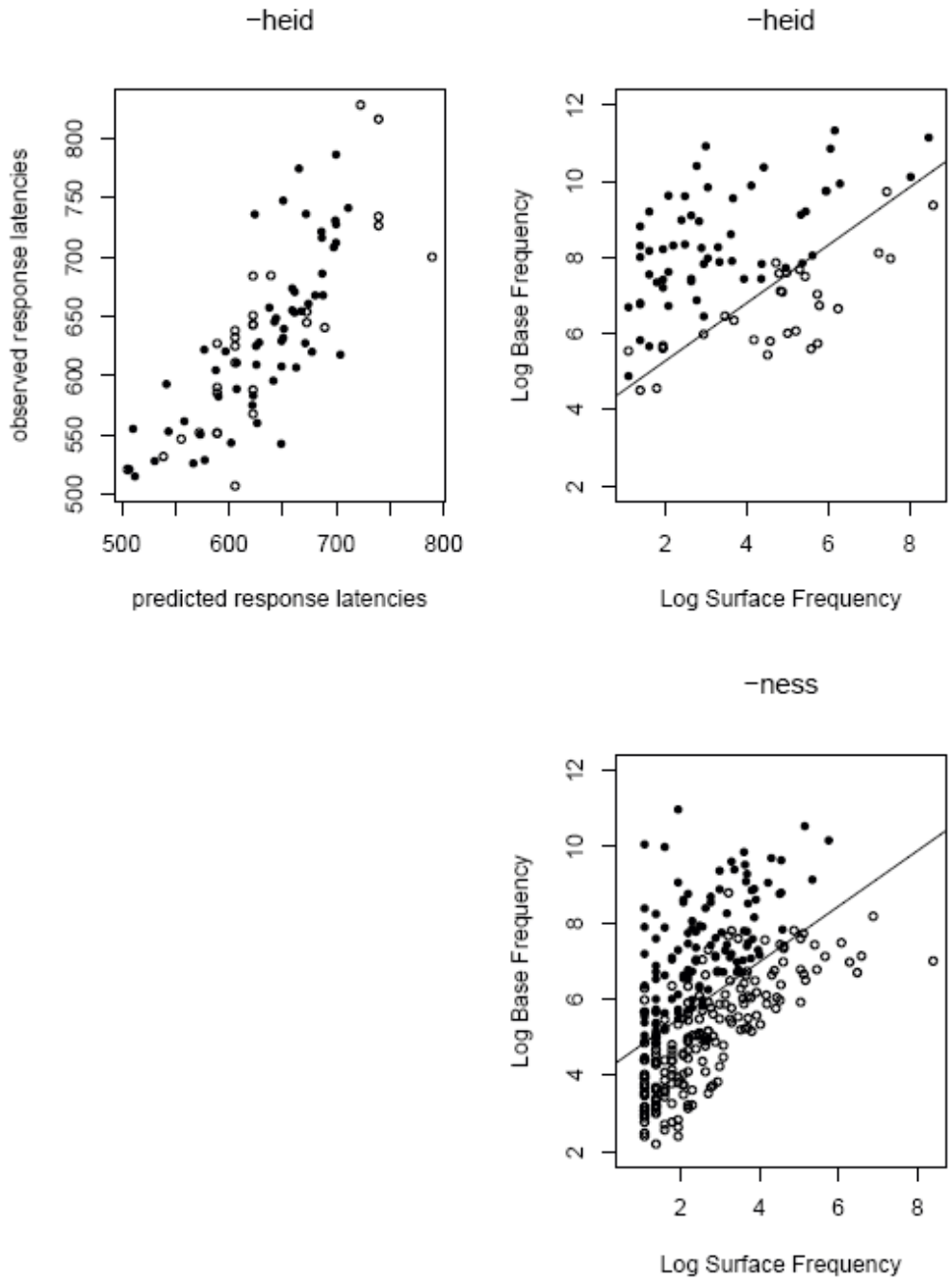


Figure 2: Left Panel: The correlation between the model times produced by MATCHECK and observed response latencies for the Dutch suffix *-heid*. Upper Right Panel: The relation between log derived frequency and log base frequency for *-heid*. Bottom Right Panel: The relation between log derived frequency and log base frequency for *-ness*. Solid points represent forms for which the parsing route is the first to produce a complete spanning in MATCHECK. Open points represent forms for which the derived form is the first to provide a complete spanning. The lines in the right panels optimally divide the words that are parsed from those that are recognized on the basis of their derived forms.

If we split up the words...

- For words where the whole-word parse is faster, only the time to the whole-word parse is significantly correlated with RT.

- For words where the decomposed parse is faster, the time to the whole-word parse and the morphological family size are sig. correlated with RT.

➢ Why does a large family speed RT in the decomposedly-parsed items?

Assume that, on top of the form-similarity relations implicit in Matcheck, there are explicit connections between words that contain the same morpheme, and they spread activation to each other. Thus, the stem gets activated faster when it occurs in lots of other words.

"Now consider the case in which the parsing route wins the race. The present experimental data on *-heid* suggest that in this case activation spreads into the morphological family. This makes sense, as initially the comprehension system knows only that it is dealing with a stem that has to be combined with some affix. By allowing the morphological family members to become co-activated, all and only the possibly relevant candidates are made more available for the processes which combine the stem with the derivational suffix and which compute the meaning of their combination.
"In fact, since the derived form is one of the family members of the base, it will be activated more quickly when the base has a large morphological family. This is because it is embedded in a larger network of morphologically related words, and the morphologically related words spread activation to each other. This may explain why log derived frequency remains such a strong predictor of the response latencies even for the words in the P set [i.e., the words where the decomposed parse wins]." (pp. 13-14 of ms. version)

➢ Why doesn't family size matter for whole-word-parsed items?

This part of the paper is fuzzy to me. I'll just quote:

"Consider what happens when the direct route is the first to provide a complete spanning of the target word, say, *snel-heid*, "quickness", i.e., "speed". Once the derived form has become available, the corresponding meaning is activated, apparently without activation spreading into the morphological family of its base, "snel". In other words, family members such as *ver-snel-en* ("to increase speed") and *snel-weg* ("freeway") are not co-activated when *snel-heid* has been recognized by means of the direct route." (p. 13 of ms. version)

It's hard to reconcile this with the idea that words containing the same stem are connected. Ideas?

➢ Why doesn't the time to the decomposed parse matter?

"We think that derived frequency and family size may conspire to mask an effect of the timestep itself at which the base word itself becomes available, leading to the absence of a measurable effect of base frequency and t_P [the time at which the decomposed parse becomes available]." (p. 14 of ms.)

References—for this handout and the last one

Baayen, Harald & Robert Schreuder (2000). Towards a psycholinguistic computational model for morphological parsing. *Transactions of the Royal SocietyLondon A* 358: 1281-1293.

Baayen, Harald; Robert Schreuder; and Richard Sproat (2000). Modeling morphological segmentation in a parallel dual route framework for visual word recognition. In Frank van Eynde & David Gibbon (eds.) *Lexicon Development for Speech and Language Processing.* Pp. 267-293.

Baroni, Marco (2000). Distributional Cues in Morpheme Discovery: A Computational Model and Empirical Evidence. UCLA Ph.D. dissertation.

Baroni, Marco (2003). Distribution-driven morpheme discovery: A computational/experimental study. In Geert Booij and Jaap van Marle (eds.), Yearbook of Morphology 2003, Dordrecht: Springer. 213-248.

Booij, Geert (1995). *The Phonology of Dutch.* Oxford: Clarendon.

Booij, Geert (2002). *The Morphology of Dutch.* Oxford: Oxford University Press.

Evert, Stefan & 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.

Hay, Jennifer & Harald Baayen (2002). Parsing and productivity. In Geert Booij & Jap van Marle (eds.) *Yearbook of Morphology 2001.* Kluwer Academic Publishers. Pp. 203-235.

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

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

Pierrehumbert, Janet (2001). Why phonological constraints are so coarse-grained. In J. McQueen and A. Cutler (eds.) SWAP special issue, *Language and Cognitive Processes* 16: 691-698.

Pierrehumbert, Janet (2002). Word-specific phonetics. *Laboratory Phonology VII*. Berlin: Mouton de Gruyter. Pp. 101-139.

Plag, Ingo (2004). Productivity. In Keith Brown (ed.) *Encyclopedia of Language and Linguistics*, Second Edition. Oxford: Elsevier.

Raffelsiefen, Renate (1999). Diagnostics for prosodic words revisited: the case of historically prefixed words in English. In T. Alan Hall & Ursula Kleinhenz (eds.) *Studies on the Phonological Word.* Amsterdam: Benjamins.

Raffelsiefen, Renate (2005). Paradigm Uniformity Effects versus Boundary Effects. In: L. J. Downing, T. A. Hall, and R. Raffelsiefen (eds.) *Paradigms in Phonological Theory.* Oxford: Oxford University Press. Pp. 211-262.

van Oostendorp, Marc (1994). Affixation and integrity of syllable structure in Dutch. In *Linguistics in the Netherlands 1994*. Amsterdam: John Benjamins.