Evaluating Lexical Frequency Measures for Sociolinguistic Variation

Ruaridh Purse & Meredith Tamminga

University of Pennsylvania

Penn

UKLVC 12

Lexical Frequency in Perception/Production

- Measures of lexical frequency are correlated with effects in perception…
  - Easier recognition in noise (Howes 1957, Luse *et al.* 1990, Savin 1963)
  - Faster lexical decision (Connine *et al.* 1990, Dupoux & Mehler 1990, Taft & Hambly 1986)
  - Rhyme monitoring (McQueen 1993), Word spotting (Freedman 1992), Cross-modal priming (Marslen-Wilson 1990) etc.

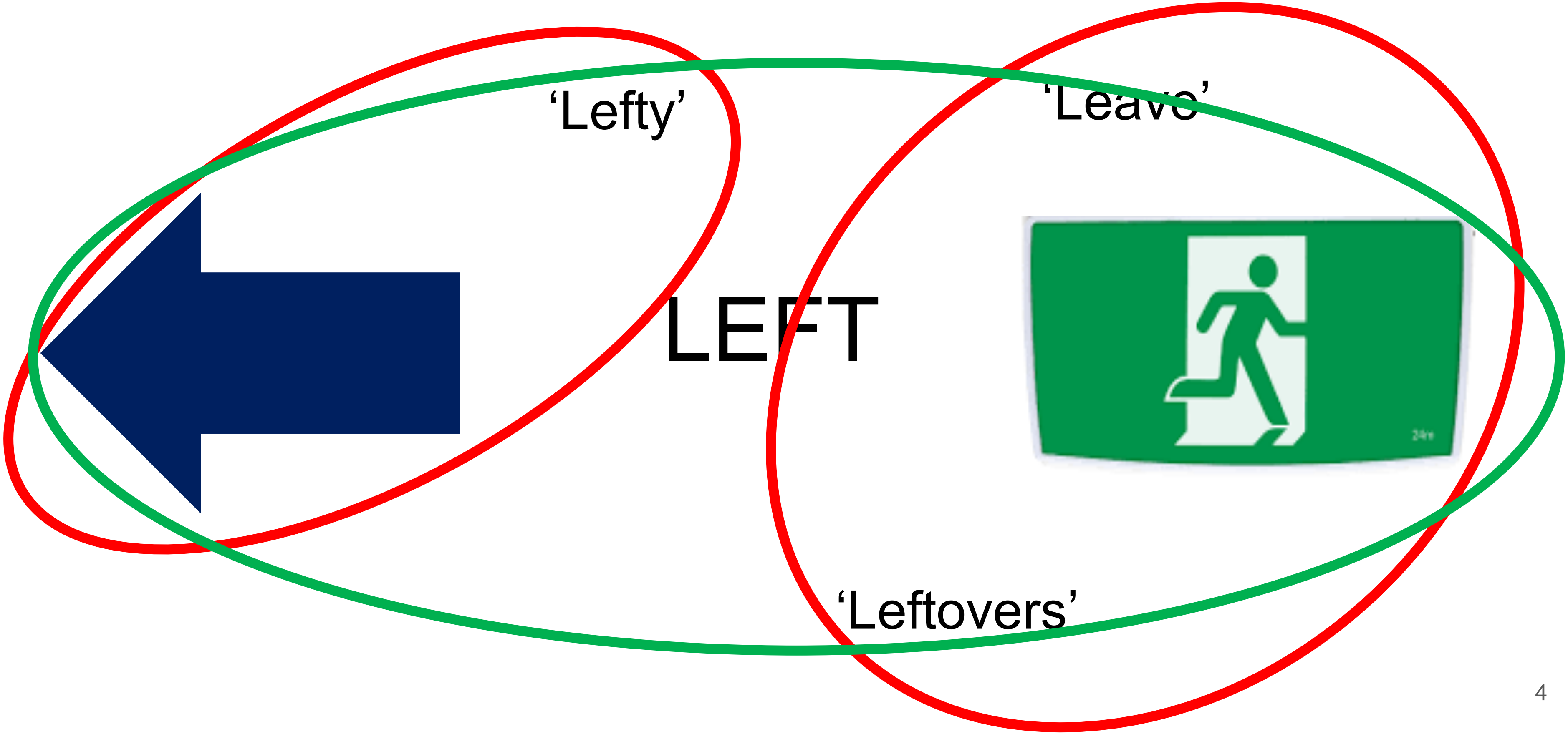
- … And production...
  - Faster picture naming (Oldfield & Wingfield 1965, Wingfield 1968)
  - Fewer speech errors (Dell 1990)
  - More 'lenition', and more advanced variants from changes in progress (Pierrehumbert 2002, Bybee 2002)

Why are frequent words produced differently?

- Speaker-oriented perspective
    - Articulatory routinisation (Bybee 2001)
    - Persistent leniting bias (Pierrehumbert 2002)
    - Accumulation of lenited exemplars

- Listener-oriented perspective
    - Frequency is correlated with predictability (Cohen Priva 2017, Bell *et al*. 2009)
    - High level of resting activation (Marslen-Wilson 1990, Tamminga *et al.* 2017)
    - Speaker can hypo- or hyper-articulate to attend to the listener's needs (Lindblom 1990)

But how do we count?



‘Lefty’

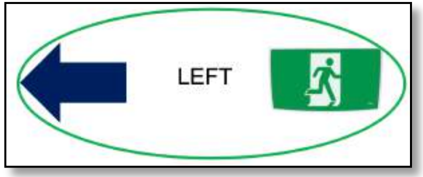
‘Leave’

LEFT

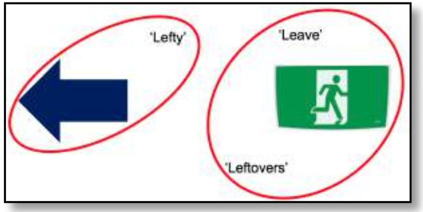
‘Leftovers’

Many measures of frequency

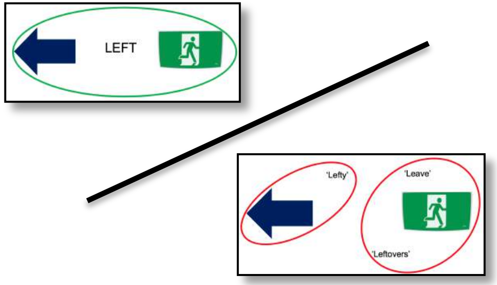
- Whole-word frequency
  - Every time a word appears, regardless of meaning
  - Standardly used; easy to automate (SUBTLEX count)
  - Some weird effects around homonyms



- Root frequency
  - Sum of all whole-word frequencies that share a root
  - Difficult to automate



- Conditional frequency
  - Probability of whole-word given the root
  - Whole-word frequency / Root frequency



Two morphophonological variables

TD

- Surface absence of underlying coronal stop in C_# context
  - e.g. *old* vs *ol’*
- More deletion in monomorphemes (*mist*) than complex forms (*missed*).

ING

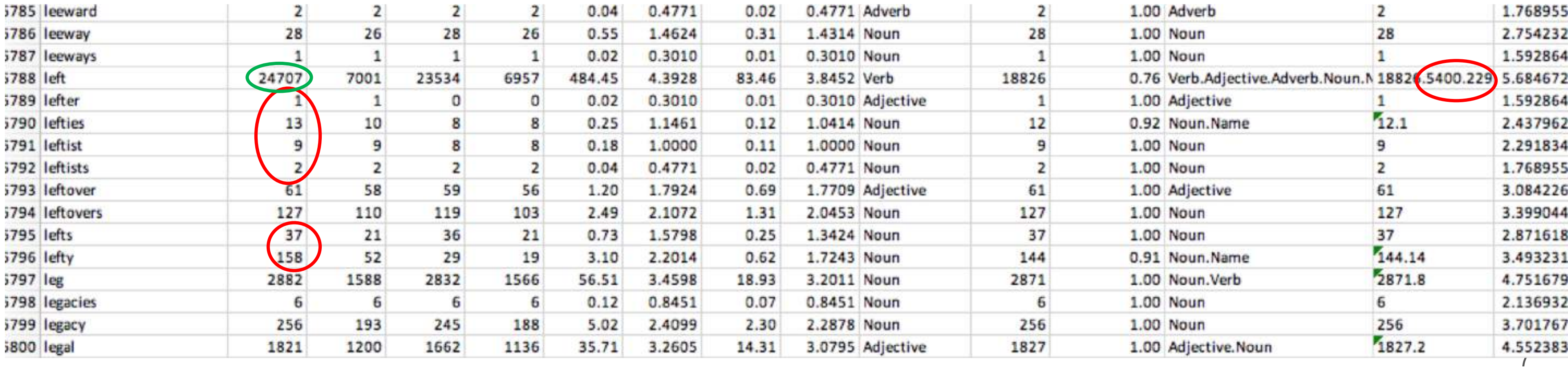
- Alternation between [ŋ] and [n] in word-final ING
  - e.g. *working* vs *workin’*
- More [n] in progressive forms (*I am working*) than gerundive forms (*Working is hard*).

➢ This study: Which frequency measure best accounts for variance?
How does lexical frequency interact with morphology?

Data & Methods

- Philadelphia Neighborhood Corpus
  - 118 white speakers
  - 11964 TD tokens, 5452 ING tokens

- Lexical frequency measures from SUBTLEX$_{US}$
  - Whole-word, root, conditional...



5785	leeward	2	2	2	2	0.04	0.4771	0.02	0.4771	Adverb	2	1.00	Adverb	2	1.768955
5786	leeway	28	26	28	26	0.55	1.4624	0.31	1.4314	Noun	28	1.00	Noun	28	2.754232
5787	leeways	1	1	1	1	0.02	0.3010	0.01	0.3010	Noun	1	1.00	Noun	1	1.592864
5788	left	24707	7001	23534	6957	484.45	4.3928	83.46	3.8452	Verb	18826	0.76	Verb.Adjective.Adverb.Noun.N	18826.5400.229	5.684672
5789	lefter	1	1	0	0	0.02	0.3010	0.01	0.3010	Adjective	1	1.00	Adjective	1	1.592864
5790	lefties	13	10	8	8	0.25	1.1461	0.12	1.0414	Noun	12	0.92	Noun.Name	12.1	2.437962
5791	leftist	9	9	8	8	0.18	1.0000	0.11	1.0000	Noun	9	1.00	Noun	9	2.291834
5792	leftists	2	2	2	2	0.04	0.4771	0.02	0.4771	Noun	2	1.00	Noun	2	1.768955
5793	leftover	61	58	59	56	1.20	1.7924	0.69	1.7709	Adjective	61	1.00	Adjective	61	3.084226
5794	leftovers	127	110	119	103	2.49	2.1072	1.31	2.0453	Noun	127	1.00	Noun	127	3.399044
5795	lefts	37	21	36	21	0.73	1.5798	0.25	1.3424	Noun	37	1.00	Noun	37	2.871618
5796	lefty	158	52	29	19	3.10	2.2014	0.62	1.7243	Noun	144	0.91	Noun.Name	144.14	3.493231
5797	leg	2882	1588	2832	1566	56.51	3.4598	18.93	3.2011	Noun	2871	1.00	Noun.Verb	2871.8	4.751679
5798	legacies	6	6	6	6	0.12	0.8451	0.07	0.8451	Noun	6	1.00	Noun	6	2.136932
5799	legacy	256	193	245	188	5.02	2.4099	2.30	2.2878	Noun	256	1.00	Noun	256	3.701767
5800	legal	1821	1200	1662	1136	35.71	3.2605	14.31	3.0795	Adjective	1827	1.00	Adjective.Noun	1827.2	4.552383

Data & Methods *continued*

- Whole-word and Root frequency log-transformed and centred

- Mixed effects logistic regression models
  - All combinations of lexical frequency measures
  - Controlled for grammatical class and speech rate

- ANOVAs to compare nested models
  - Optimal model minimizes AIC and BIC and significantly maximizes log-likelihood

Modeling TD

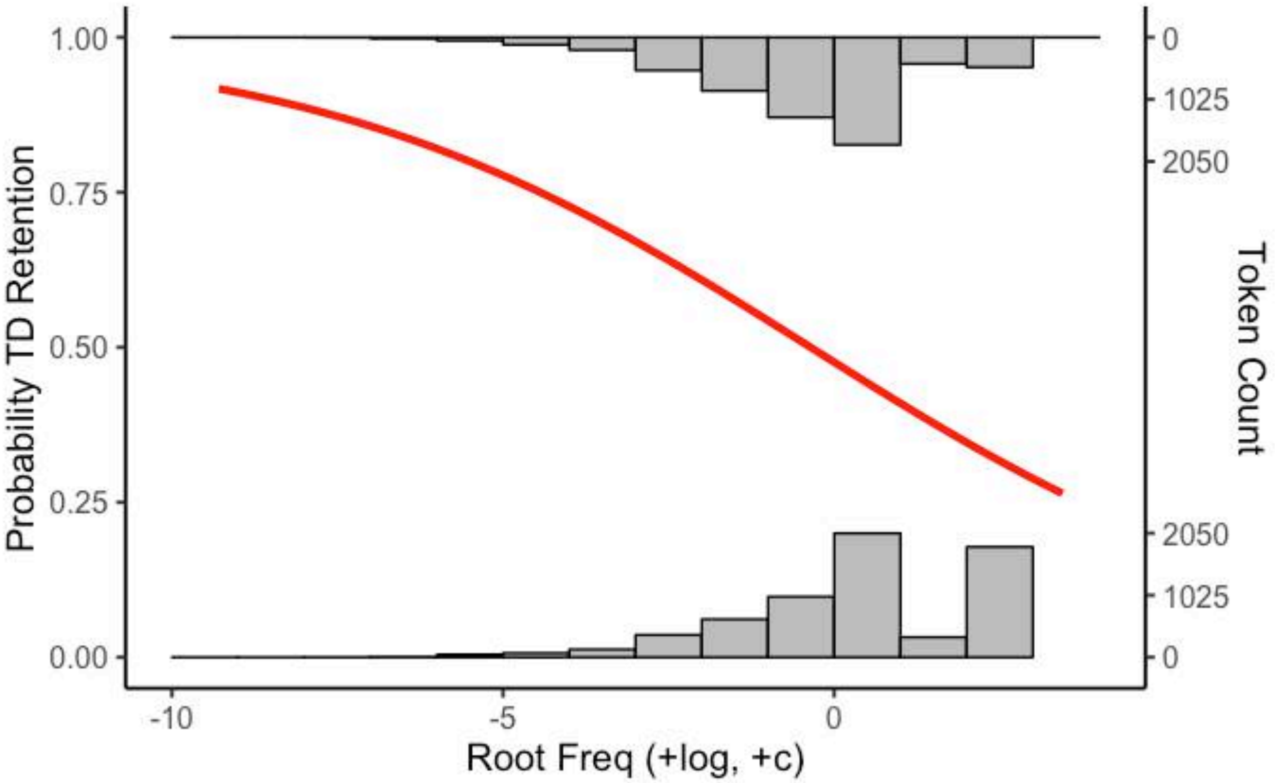
- Root frequency most effective measure
  - Root Freq improves a control model
  - Whatever the model has, Root Freq significantly improves it

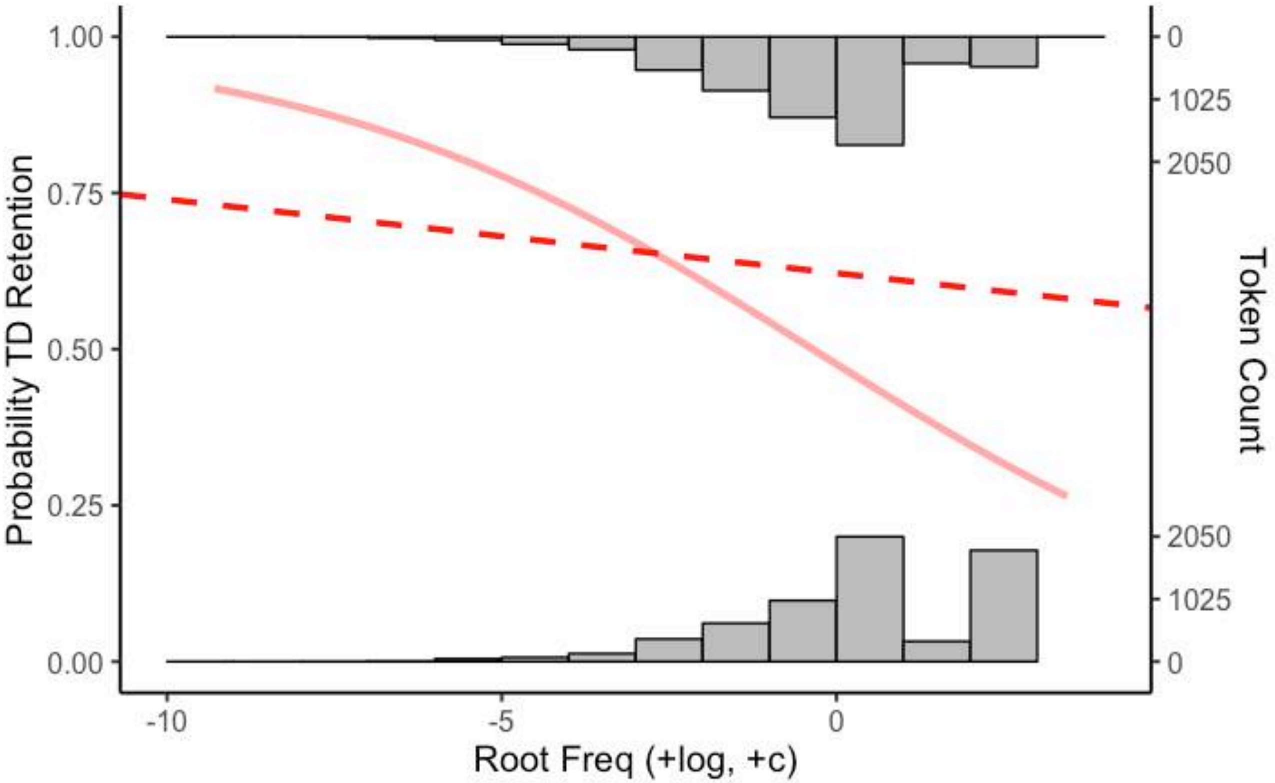
- Root Freq predicts TD outcomes
  - 64% retention at -2 RootFreq vs. 60% retention and 2 RootFreq

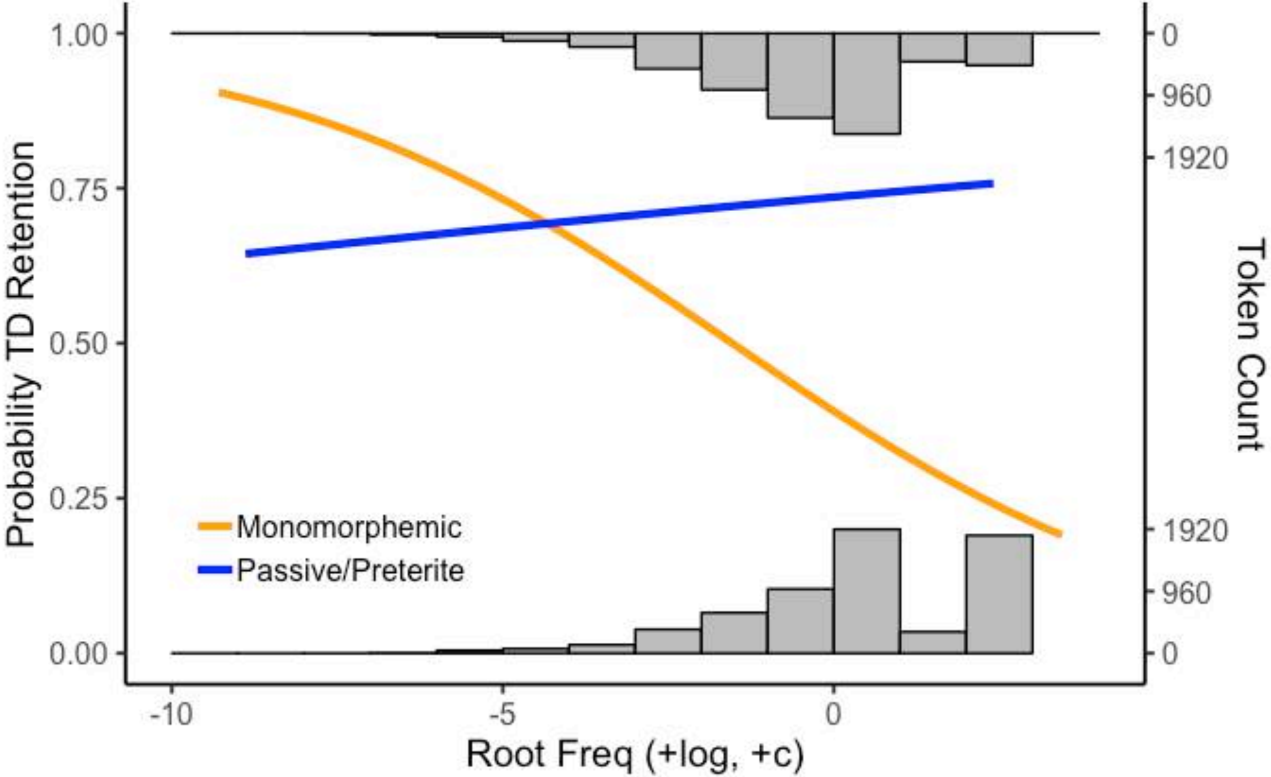
Start with...

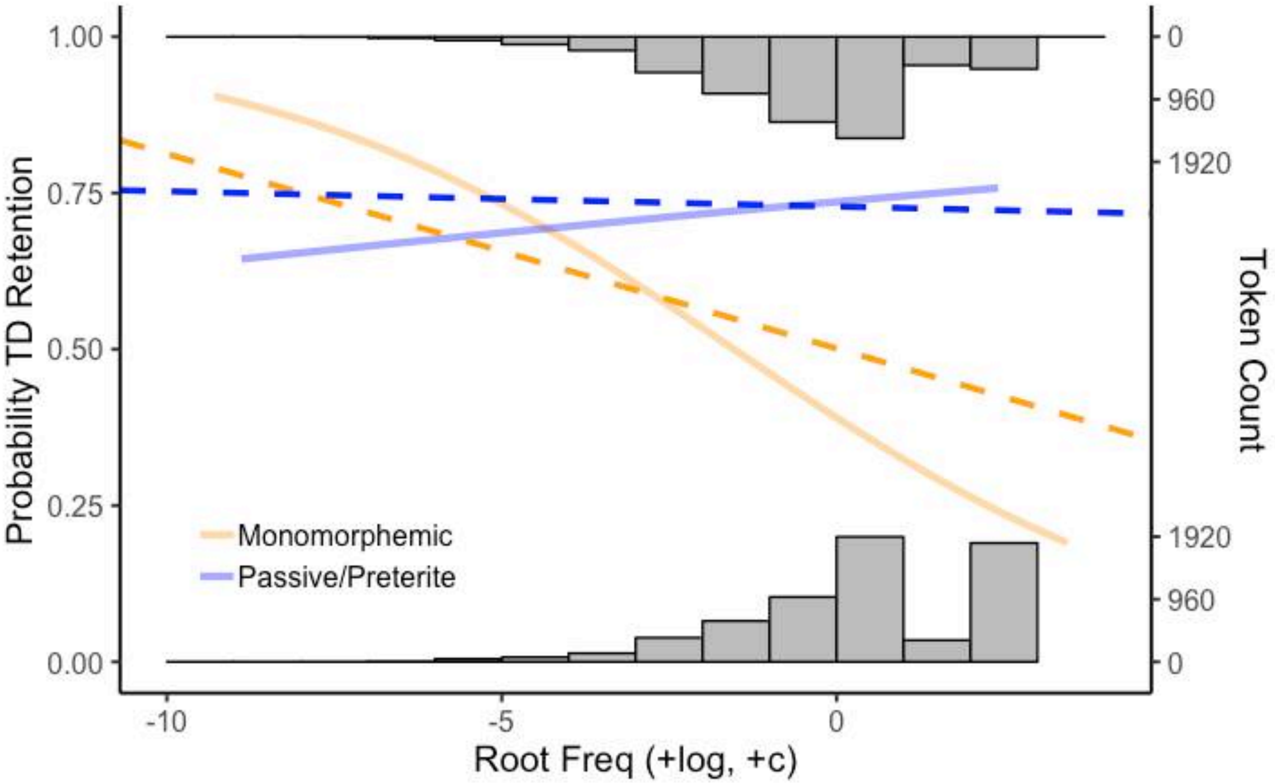
Add...

	Whole	**Root**	**Cond**
(control model)	.120	.008 **	.815
Whole		.022 *	.458
Root	.456		.982
Cond	.088	.007 **	
Whole + Root			.631
Whole + Cond		.026 *	
Root + Cond	.375		


Probability TD Retention
1.00
0.75
0.50
0.25
0.00
-10
-5
0
Root Freq (+log, +c)
Token Count
0
1025
2050
2050
1025
0


Probability TD Retention
1.00
0.75
0.50
0.25
0.00
Token Count
0
1025
2050
2050
1025
0
-10
-5
0
Root Freq (+log, +c)


Probability TD Retention
1.00
0.75
0.50
0.25
0.00
Token Count
0
960
1920
1920
960
0
Monomorphemic
Passive/Preterite
-10
-5
0
Root Freq (+log, +c)


Probability TD Retention
1.00
0.75
0.50
0.25
0.00
Token Count
0
960
1920
1920
960
0
-10
-5
0
Root Freq (+log, +c)
Monomorphemic
Passive/Preterite

Modeling ING

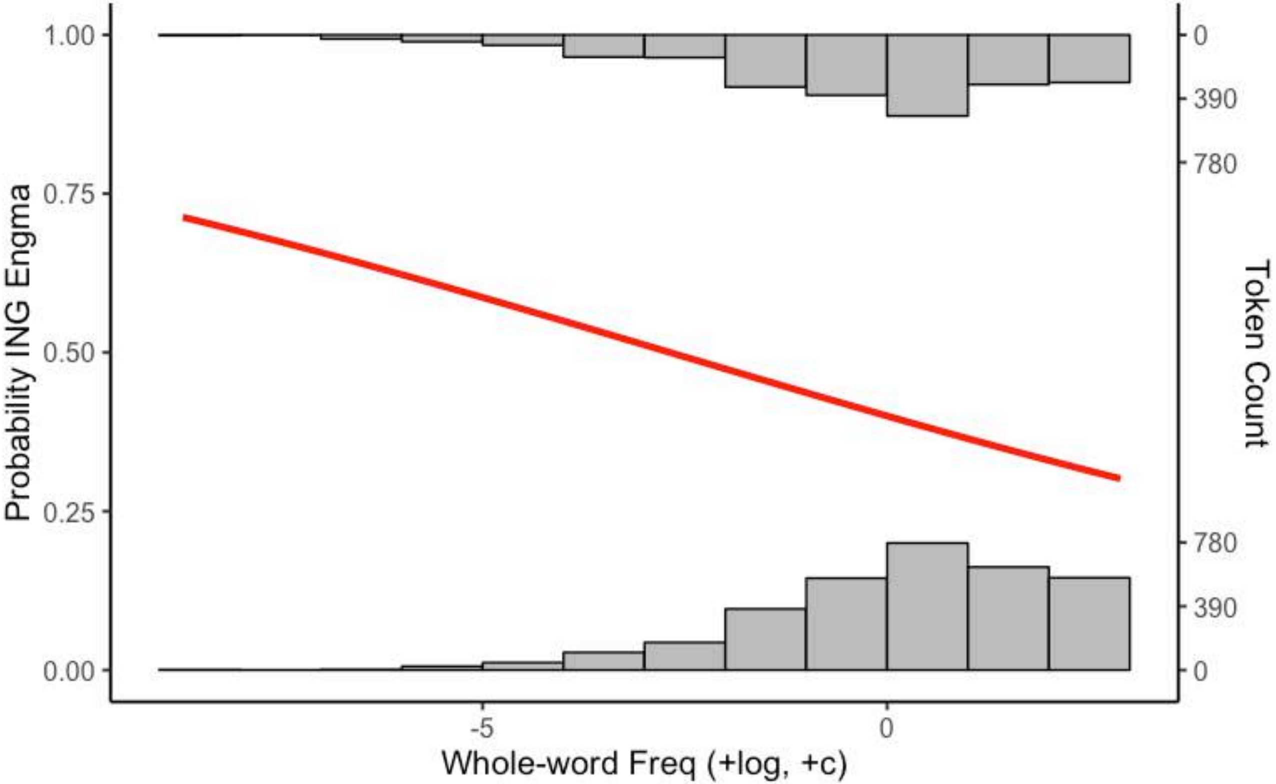
- Whole-word frequency most effective measure
  - Once whole-word is in the model, no other measure improves it
  - Whole-word frequency still improves all other models

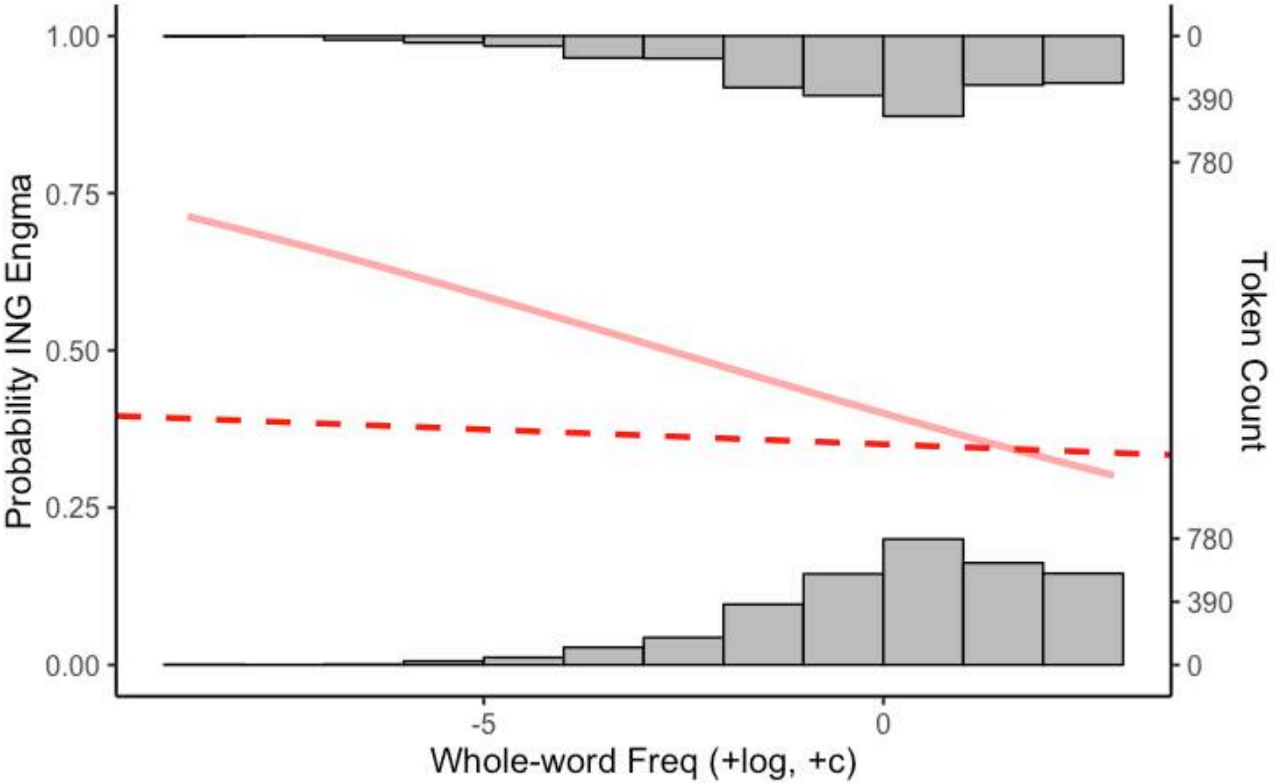
- Whole-word frequency predicts ING outcomes
  - 61% engma at -2 whole-word frequency vs. 60% engma for 2

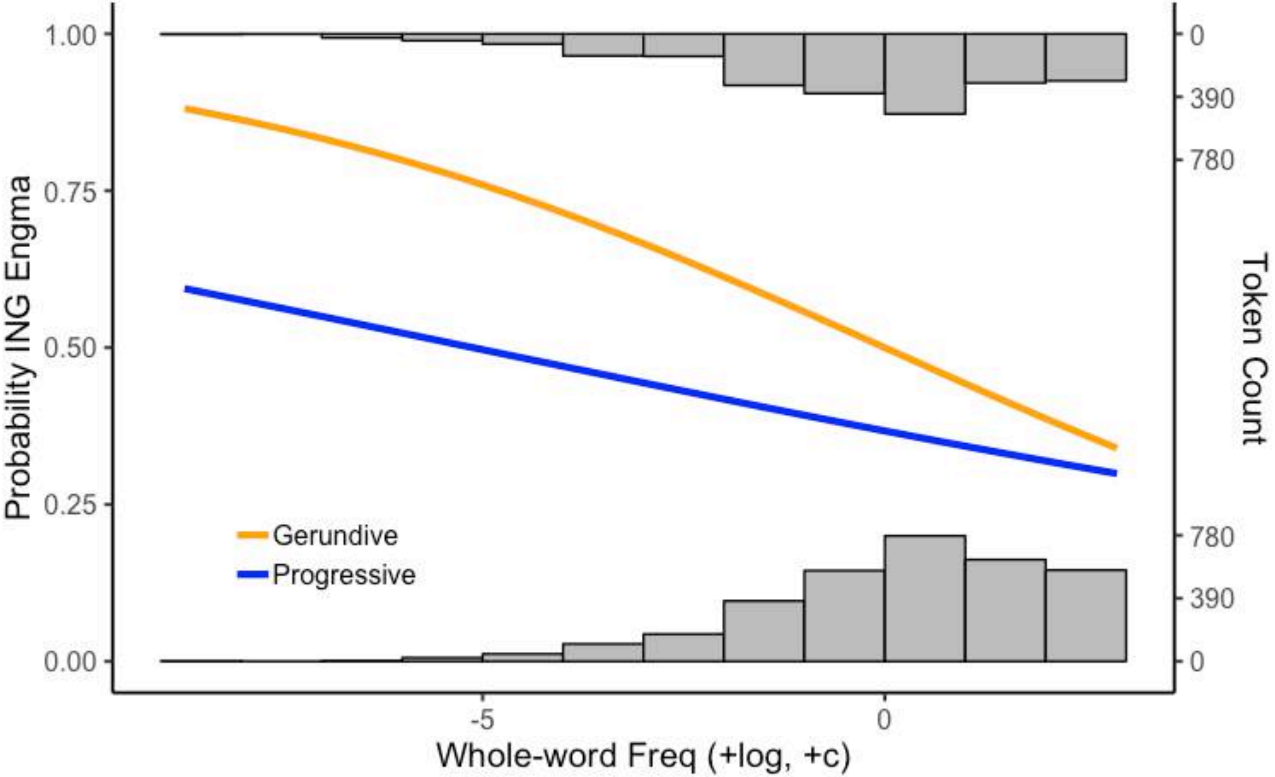
Start with...

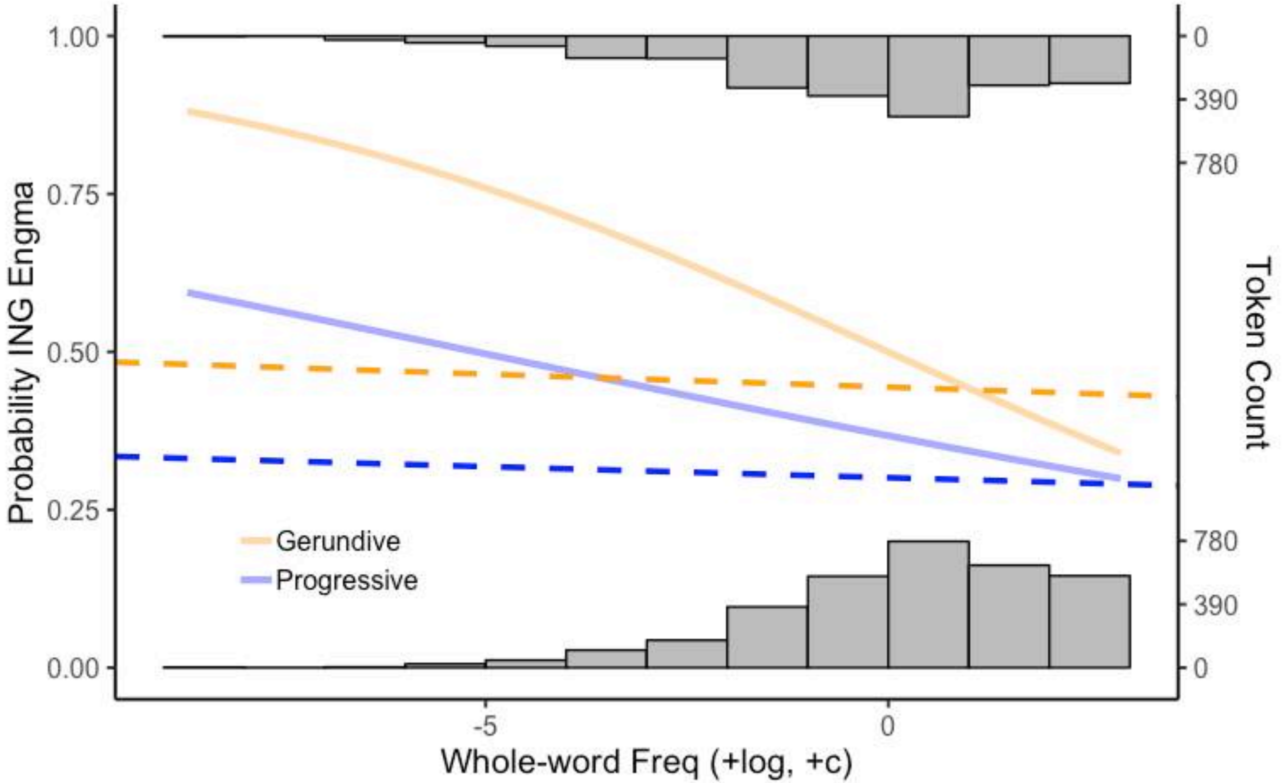
Add...

	Whole	**Root**	**Cond**
(control model)	<.001 ***	<.001 ***	.013*
Whole		1.000	1.000
Root	.022 *		.712
Cond	<.001 ***	<.001 ***	
Whole + Root			.070
Whole + Cond		.833	
Root + Cond	.004 **		


Probability ING Engma
1.00
0.75
0.50
0.25
0.00
Token Count
0
390
780
780
390
0
-5
0
Whole-word Freq (+log, +c)


Probability ING Engma
1.00
0.75
0.50
0.25
0.00
Token Count
0
390
780
780
390
0
-5
0
Whole-word Freq (+log, +c)


Probability ING Engma
1.00
0.75
0.50
0.25
0.00
Token Count
0
390
780
780
390
0
Gerundive
Progressive
-5
0
Whole-word Freq (+log, +c)



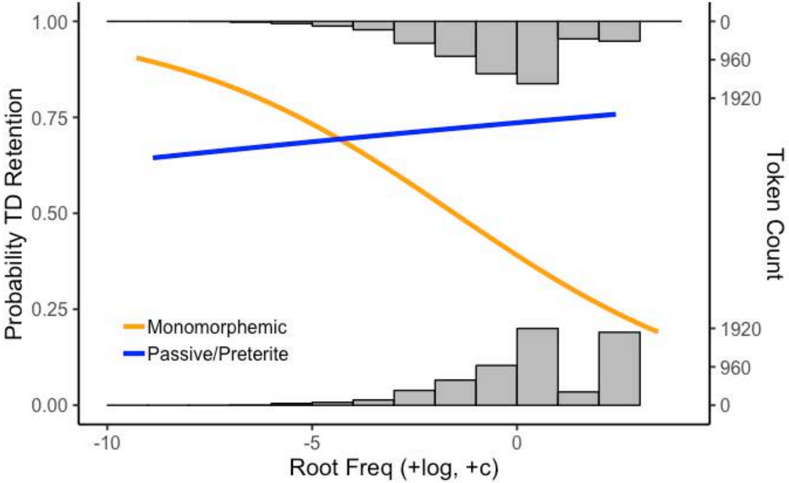
Probability ING Engma
1.00
0.75
0.50
0.25
0.00
Token Count
0
390
780
780
390
0
-5
0
Whole-word Freq (+log, +c)
Gerundive
Progressive

Discussion: ING

- Actual magnitude of the predicted effect is extremely small
  - Highly significant, but how much does it matter?
  - P-value on its own is not always informative (McShane *et al.* 2017, Nature Editorial 2019)

- Morphosyntactic categorisation is quite complicated (Tamminga, 2014)
  - Nuanced tests that require pragmatic context
  - Lots of ambiguity in the Gerundive category especially

Discussion: TD



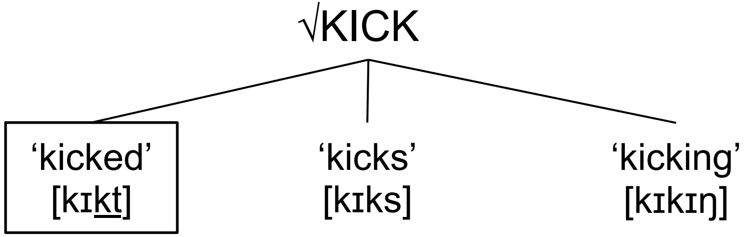
- Why is Root frequency the measure that matters?

- Why does it only matter for monomorphemes?

Discussion: TD

- For complex forms, the variable environment is formed with a suffix
- It does not reoccur in morphological relatives

- But for monomorphemes, the same variable environment appears across many related words

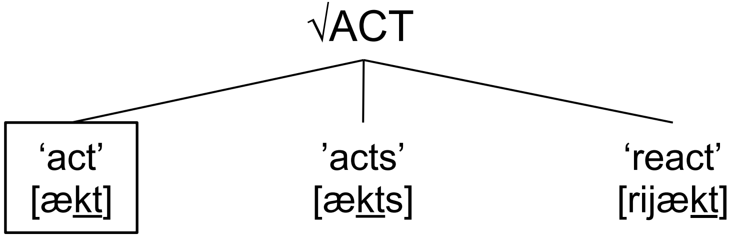


√KICK

‘kicked’ [kɪkt]

‘kicks’ [kɪks]

‘kicking’ [kɪkɪŋ]



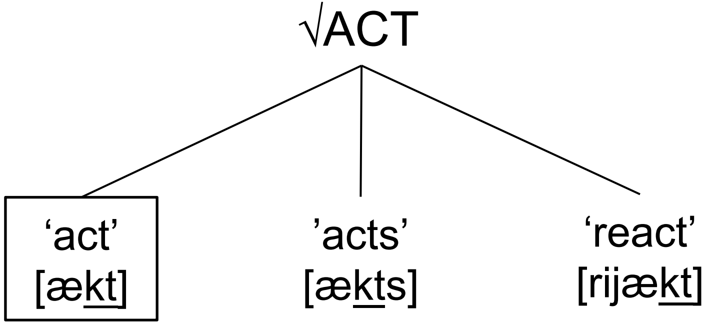
√ACT

‘act’ [ækt]

’acts’ [ækts]

‘react’ [rijækt]

Discussion: TD



√ACT

‘act’ [ækt]

’acts’ [ækts]

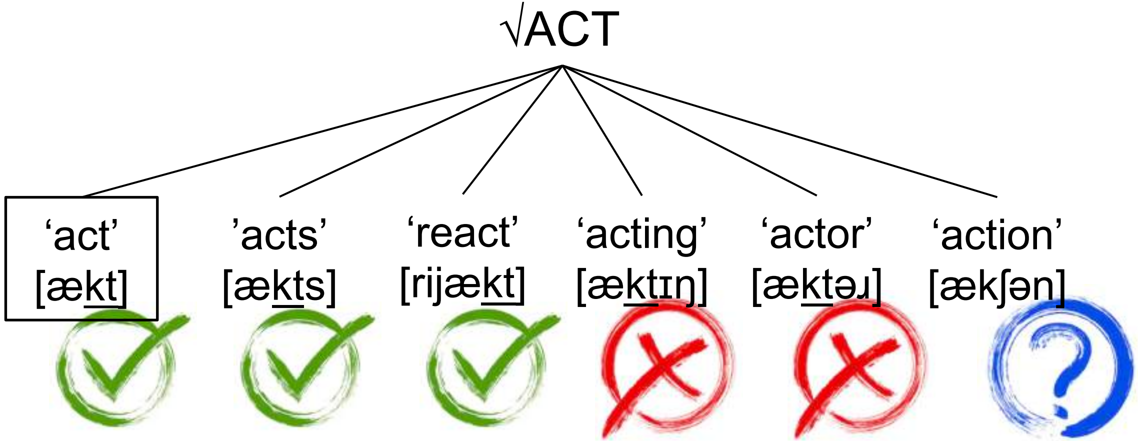
‘react’ [rijækt]

- Not all related words can feature deletion!

- Relative frequency of related words (Hay 2001) and frequency of contexts (Guy *et al.* 2008, Forrest 2017, Sloos 2019) both matter

- Next: proportions of related words *with* their contextual baggage
  - Accumulating exemplars: undeletable words contribute to increased retention
  - Increasing resting activation: all related words contribute to increased deletion

Discussion: TD



- Not all related words can feature deletion!

- Relative frequency of related words (Hay 2001) and frequency of contexts (Guy *et al.* 2008, Forrest 2017, Sloos 2019) both matter

- Next: proportions of related words *with* their contextual baggage
  - Accumulating exemplars: undeletable words contribute to increased retention
  - Increasing resting activation: all related words contribute to increased deletion

Interim Practical Recommendations

- Different measures of lexical frequency may capture different things
    - Predictability, resting activation, degree of articulatory routinisation, etc.
    - Look out for...
        - Interactions with other predictors
        - The magnitude of the effect

- Use a measure that is appropriate for your purposes
    - What are the (structural/social/phonetic) properties of your variable?
    - What is frequency a proxy for?