Heinrich Heine Universität Düsseldorf

Farewell to relative frequency?

A closer look at frequency effects on the acoustic duration of English derivatives

Simon David Stein
Ingo Plag

DFG Deutsche Forschungsgemeinschaft
FOR2373

What is relative frequency?

The more frequently a linguistic unit occurs in a language, the more likely it is to be acoustically reduced.

frequency measures
e.g. word frequency
e.g. base frequency

but: relative frequency
the frequency of the base relative to
the frequency of the word

$$relative\ frequency = \frac{base\ frequency}{word\ frequency}$$

see, e.g., Losiewicz 1995, Bybee 2000, Jurafsky et al. 2001, Bell et al. 2003, Jurafsky 2003, Gahl 2008, Pluymaekers et al. 2005a, 2005b, Hay 2001, 2003

What is relative frequency?

relative frequency

might tap into morphological segmentability

Hay 2001, 2003, examples partly from Hay 2007, figure adapted from Hay 2001: 1045, frequencies from COCA, Davies 2008

What is relative frequency?

relative frequency
might tap into morphological segmentability

- base is more frequent than derivative

Word	Frequency
boring	7483
unboring	4

Hay 2001, 2003, examples partly from Hay 2007, figure adapted from Hay 2001: 1045, frequencies from COCA, Davies 2008

What is relative frequency?

relative frequency
might tap into morphological segmentability

- base is more frequent than derivative
  - derivative perceived as more complex

Word	Frequency	Segmentability
boring	7483	high
unboring	4	

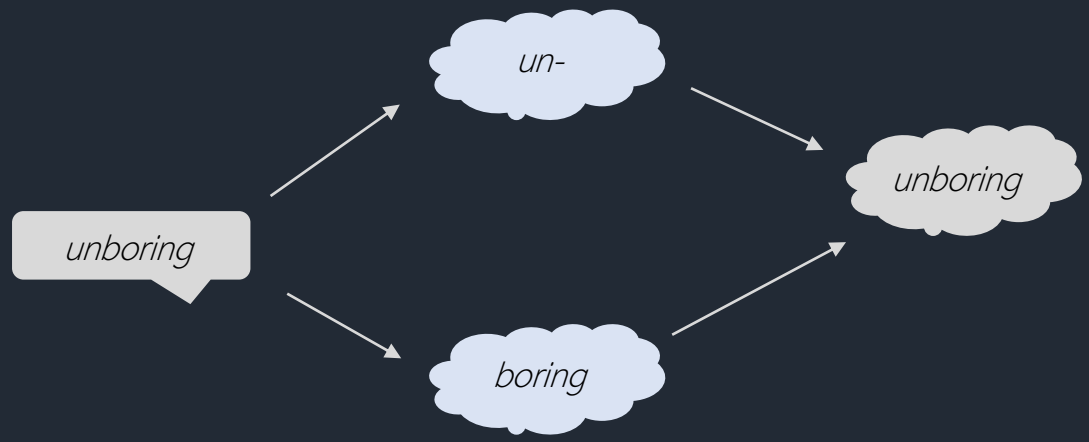
Hay 2001, 2003, examples partly from Hay 2007, figure adapted from Hay 2001: 1045, frequencies from COCA, Davies 2008

What is relative frequency?

relative frequency
might tap into morphological segmentability

- base is more frequent than derivative
  - derivative perceived as more complex
  - derivative processed compositionally

Word	Frequency	Segmentability
boring	7483	high
unboring	4	



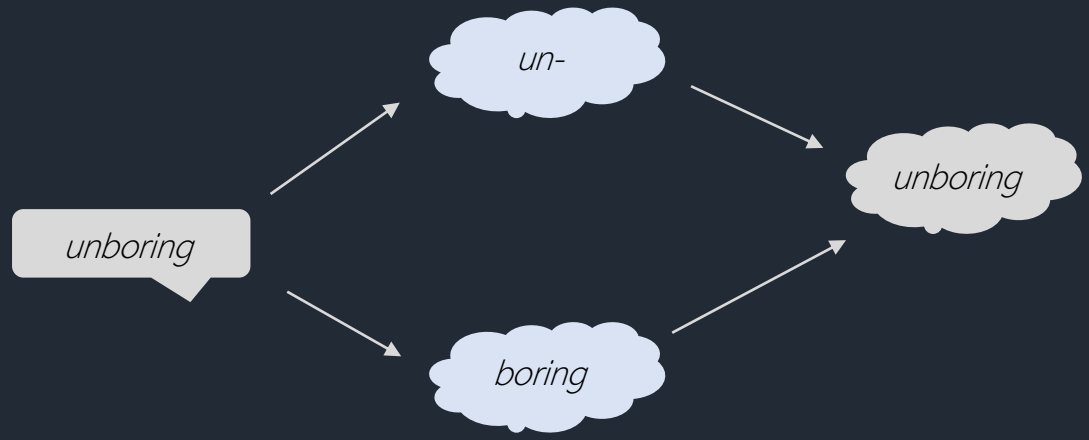
Hay 2001, 2003, examples partly from Hay 2007, figure adapted from Hay 2001: 1045, frequencies from COCA, Davies 2008

What is relative frequency?

relative frequency
might tap into morphological segmentability

- base is more frequent than derivative
  - derivative perceived as more complex
  - derivative processed compositionally
- derivative is more frequent than base
  - derivative perceived as more simplex

Word	Frequency	Segmentability
boring	7483	high
unboring	4	
sinkable	4	low
unsinkable	117	



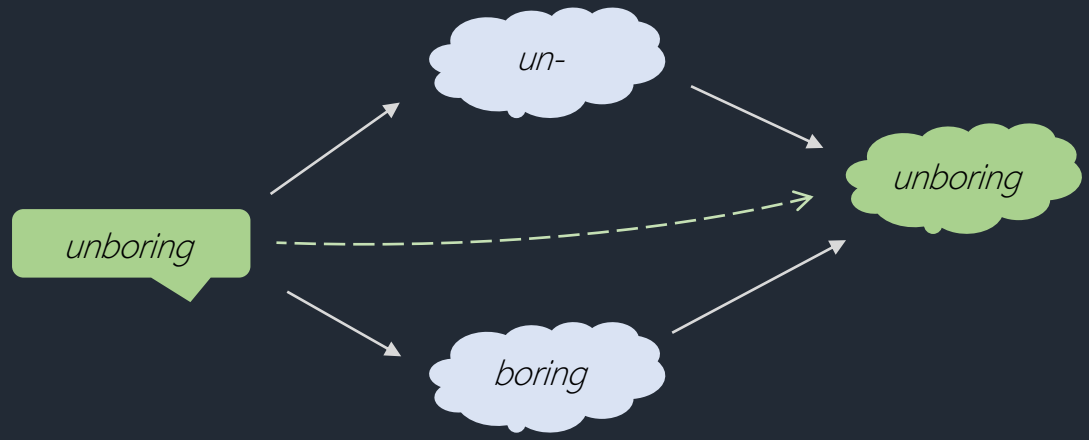
Hay 2001, 2003, examples partly from Hay 2007, figure adapted from Hay 2001: 1045, frequencies from COCA, Davies 2008

What is relative frequency?

relative frequency
might tap into morphological segmentability

- base is more frequent than derivative
  - derivative perceived as more complex
  - derivative processed compositionally
- derivative is more frequent than base
  - derivative perceived as more simplex
  - derivative processed as a whole

Word	Frequency	Segmentability
boring	7483	high
unboring	4	
sinkable	4	low
unsinkable	117	



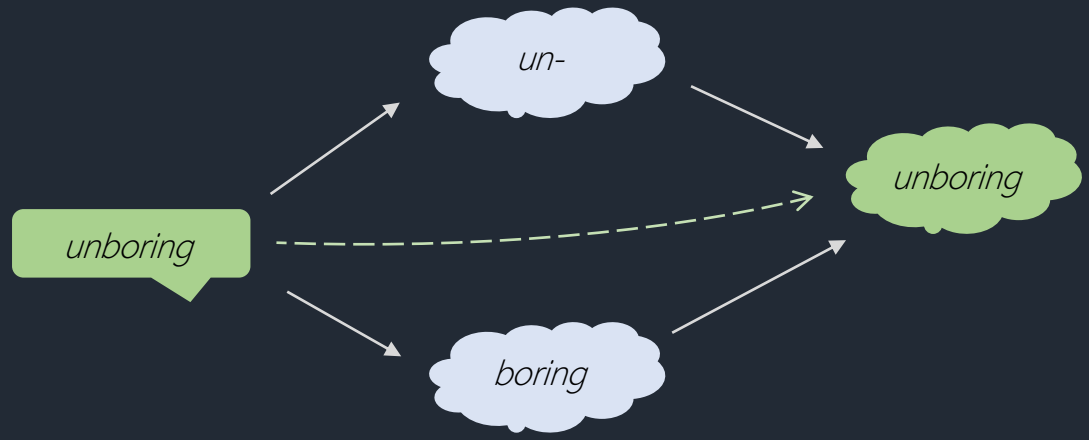
Hay 2001, 2003, examples partly from Hay 2007, figure adapted from Hay 2001: 1045, frequencies from COCA, Davies 2008

What is relative frequency?

relative frequency
might tap into morphological segmentability

- base is more frequent than derivative
  - derivative perceived as more complex
  - derivative processed compositionally
- derivative is more frequent than base
  - derivative perceived as more simplex
  - derivative processed as a whole
- the segmentability hypothesis: units processed compositionally should be protected against reduction, i.e. longer

Word	Frequency	Segmentability	Prediction
boring	7483	high	less reduction
unboring	4		
sinkable	4	low	more reduction
unsinkable	117		



Hay 2001, 2003, examples partly from Hay 2007, figure adapted from Hay 2001: 1045, frequencies from COCA, Davies 2008

Previous research on relative frequency

Higher relative frequency has been found to be associated with:

longer durations	no change in duration	shorter durations
Plag & Ben Hedia 2018	Pluymaekers et al. 2005b	Pluymaekers et al. 2005b
Zuraw et al. 2020	Plag & Ben Hedia 2018	Schuppler et al. 2012
Hay 2003	Zuraw et al. 2020	
Hay 2007		

- We need to test systematically for frequency effects across different affixes, measuring different durational domains.

Method

Data

	AUDIOBNC		QUAKEBOX		ONZE	
AFFIX	TOKENS	TYPES	TOKENS	TYPES	TOKENS	TYPES
-ness	363	130	156	39	121	41
-less	216	59				
pre-	123	71				
-ize	476	67				
-ation	3,979	373	492	94	1,040	186
dis-	689	170	179	58	251	68
un-	960	278	295	67	342	80
in-	342	72				
-able			199	50	285	61
-ity			594	46	447	79
-ment			398	46	705	73
re-			379	72	403	95

forced alignment, manual cleaning of results

Modeling

Responses
- word duration
- affix duration
- base duration

Covariates
- number of syllables
- biphone probability
- expected duration
- bigram frequency
- speech rate (corpus)

Predictors

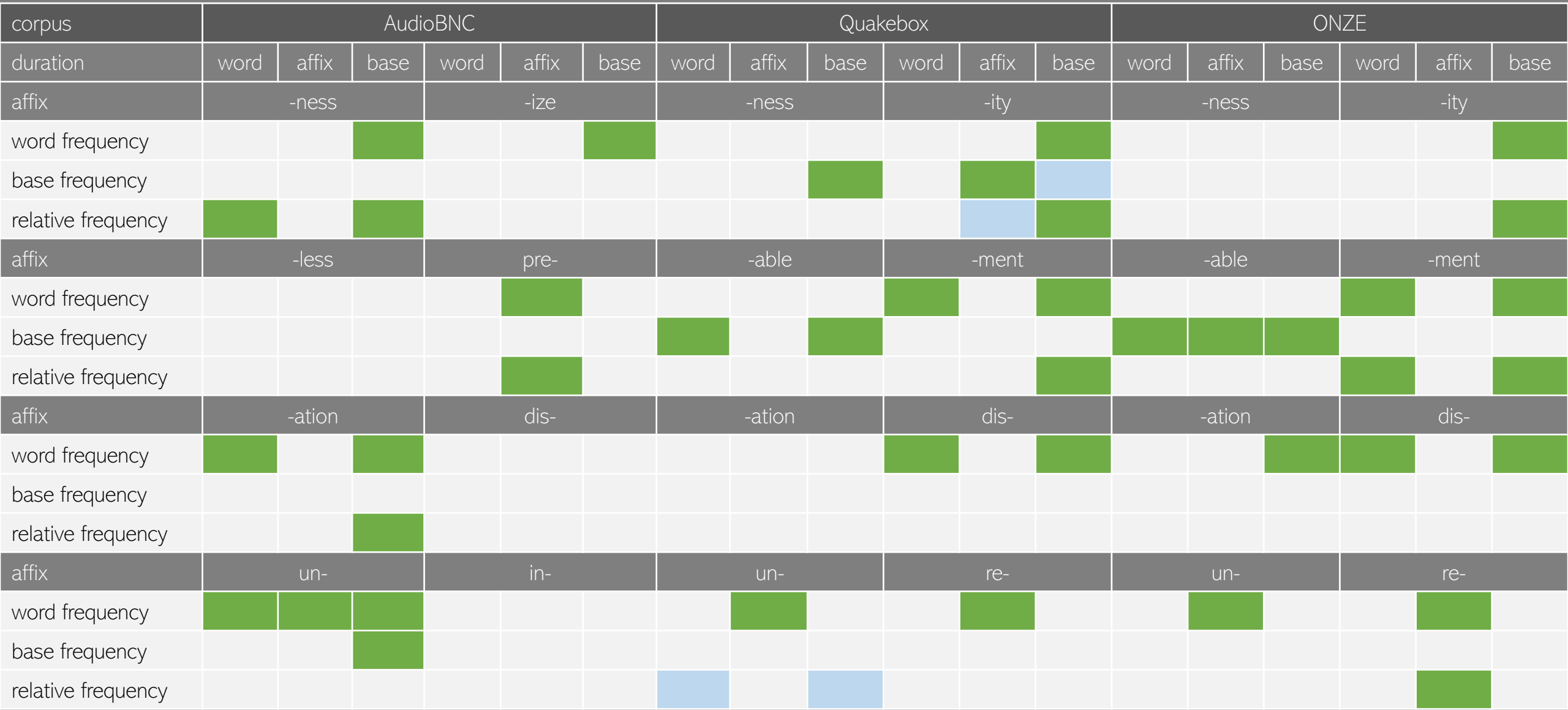
word frequency	higher → shorter durations
base frequency	higher → shorter durations
relative frequency	higher → longer durations

multiple linear regression

separate models for durations and frequencies:
72 models in each corpus study

Coleman et al. 2012, Boersma and Weenik 2014, R Core Team 2017, Davies 2008, Vitevitch and Luce 2004

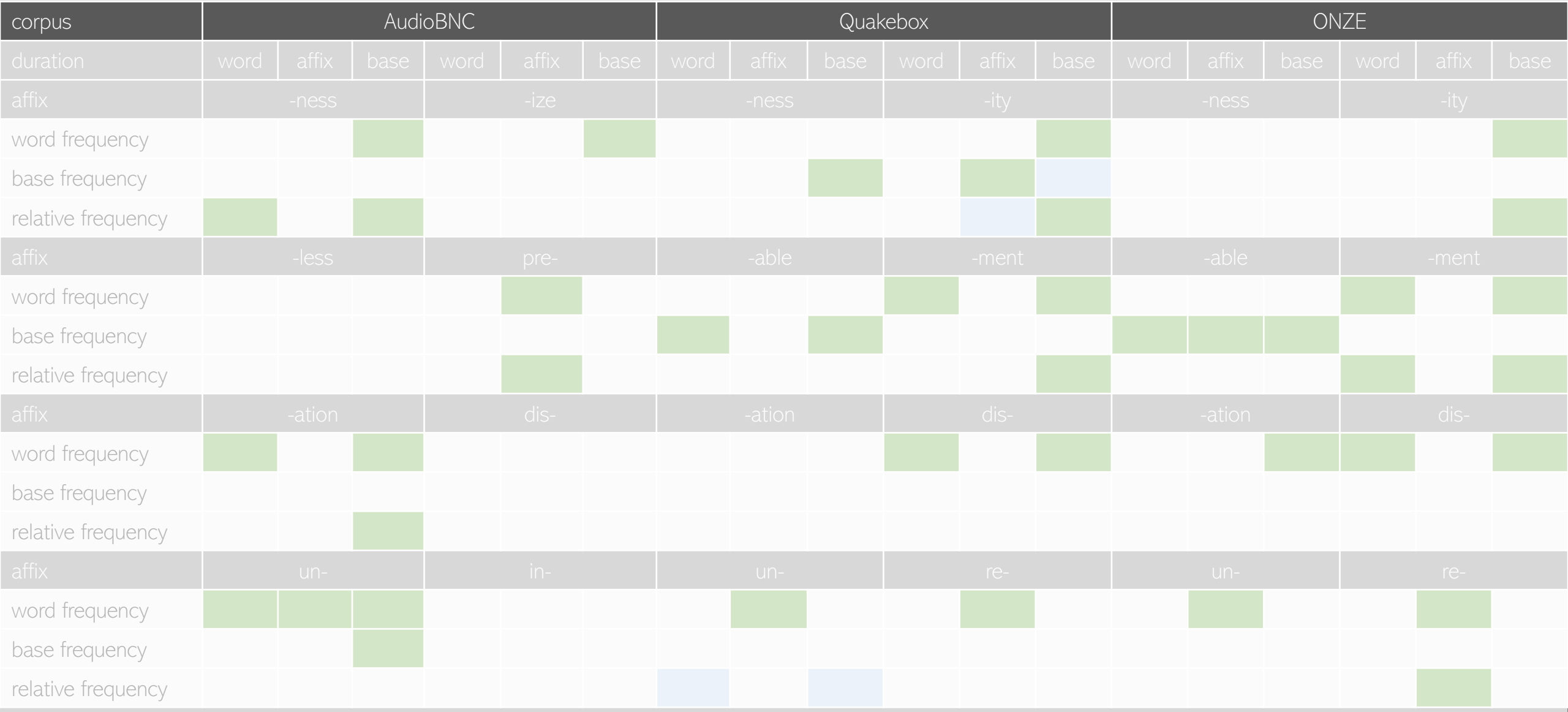
Overview of significant frequency effects



p < .001 expected direction

p < .001 unexpected direction

Overview of significant frequency effects



p < .001 expected direction

p < .001 unexpected direction

Overview of significant frequency effects

corpus	AudioBNC						Quakebox						ONZE					
duration	word	affix	base	word	affix	base	word	affix	base	word	affix	base	word	affix	base	word	affix	base
affix	-ness			-ize			-ness			-ity			-ness			-ity		
word frequency			■			■						■						■
base frequency									■		■							
relative frequency	■		■									■						■
affix	-less			pre-			-able			-ment			-able			-ment		
word frequency					■					■		■				■		■
base frequency							■		■				■	■	■			
relative frequency					■							■				■		■
affix	-ation			dis-			-ation			dis-			-ation			dis-		
word frequency	■		■							■		■			■	■		■
base frequency																		
relative frequency			■															
affix	un-			in-			un-			re-			un-			re-		
word frequency	■	■	■					■			■			■			■	
base frequency			■															
relative frequency							■		■								■	

■ $p < .001$ expected direction ■ $p < .001$ unexpected direction

Overview of significant frequency effects

corpus	AudioBNC						Quakebox						ONZE					
duration	word	affix	base	word	affix	base	word	affix	base	word	affix	base	word	affix	base	word	affix	base
affix	-ness			-ize			-ness			-ity			-ness			-ity		
word frequency			■			■						■						■
base frequency									■		■							
relative frequency	■		■									■						■
affix	-less			pre-			-able			-ment			-able			-ment		
word frequency					■					■		■				■		■
base frequency							■		■				■	■	■			
relative frequency					■							■				■		■
affix	-ation			dis-			-ation			dis-			-ation			dis-		
word frequency	■		■							■		■			■	■		■
base frequency																		
relative frequency			■															
affix	un-			in-			un-			re-			un-			re-		
word frequency	■	■	■					■			■			■			■	
base frequency			■															
relative frequency																	■	

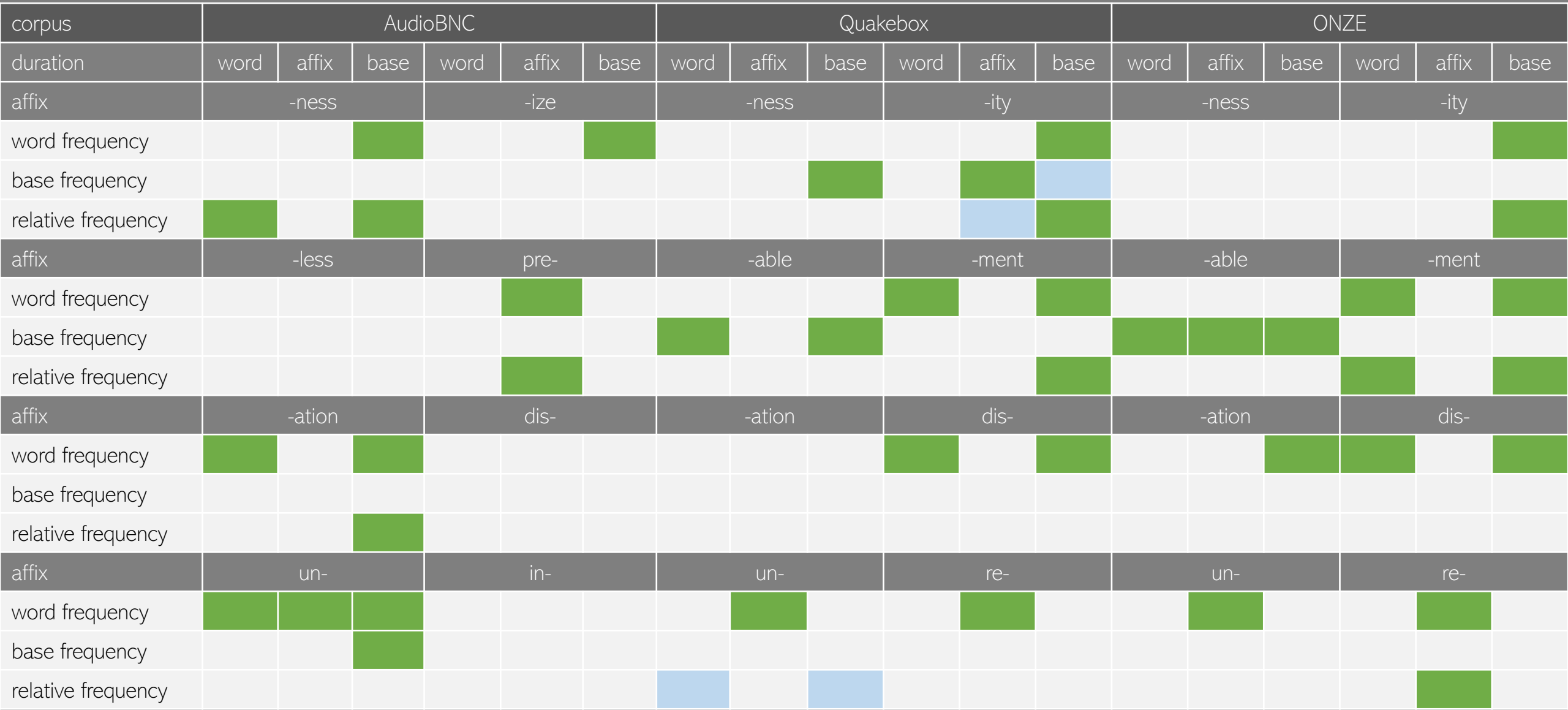
■ $p < .001$ expected direction ■ $p < .001$ unexpected direction

Overview of significant frequency effects

corpus	AudioBNC						Quakebox						ONZE					
duration	word	affix	base	word	affix	base	word	affix	base	word	affix	base	word	affix	base	word	affix	base
affix	-ness			-ize			-ness			-ity			-ness			-ity		
word frequency			■			■						■						■
base frequency									■		■							
relative frequency	■		■									■						■
affix	-less			pre-			-able			-ment			-able			-ment		
word frequency					■					■		■				■		■
base frequency							■		■				■	■	■			
relative frequency					■							■				■		■
affix	-ation			dis-			-ation			dis-			-ation			dis-		
word frequency	■		■							■		■			■	■		■
base frequency																		
relative frequency			■															
affix	un-			in-			un-			re-			un-			re-		
word frequency	■	■	■					■			■			■			■	
base frequency			■															
relative frequency																	■	

■ $p < .001$ expected direction ■ $p < .001$ unexpected direction

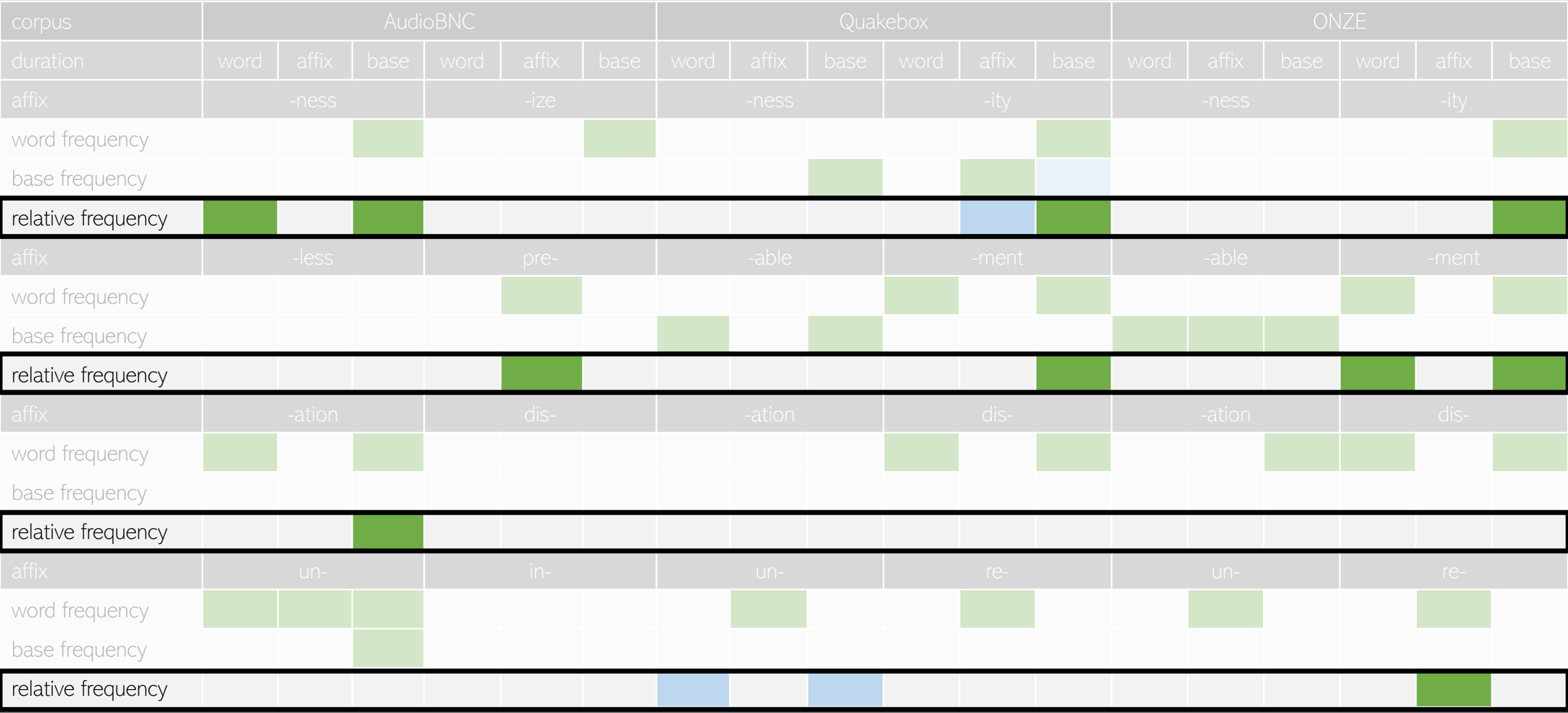
Overview of significant frequency effects



p < .001 expected direction

p < .001 unexpected direction

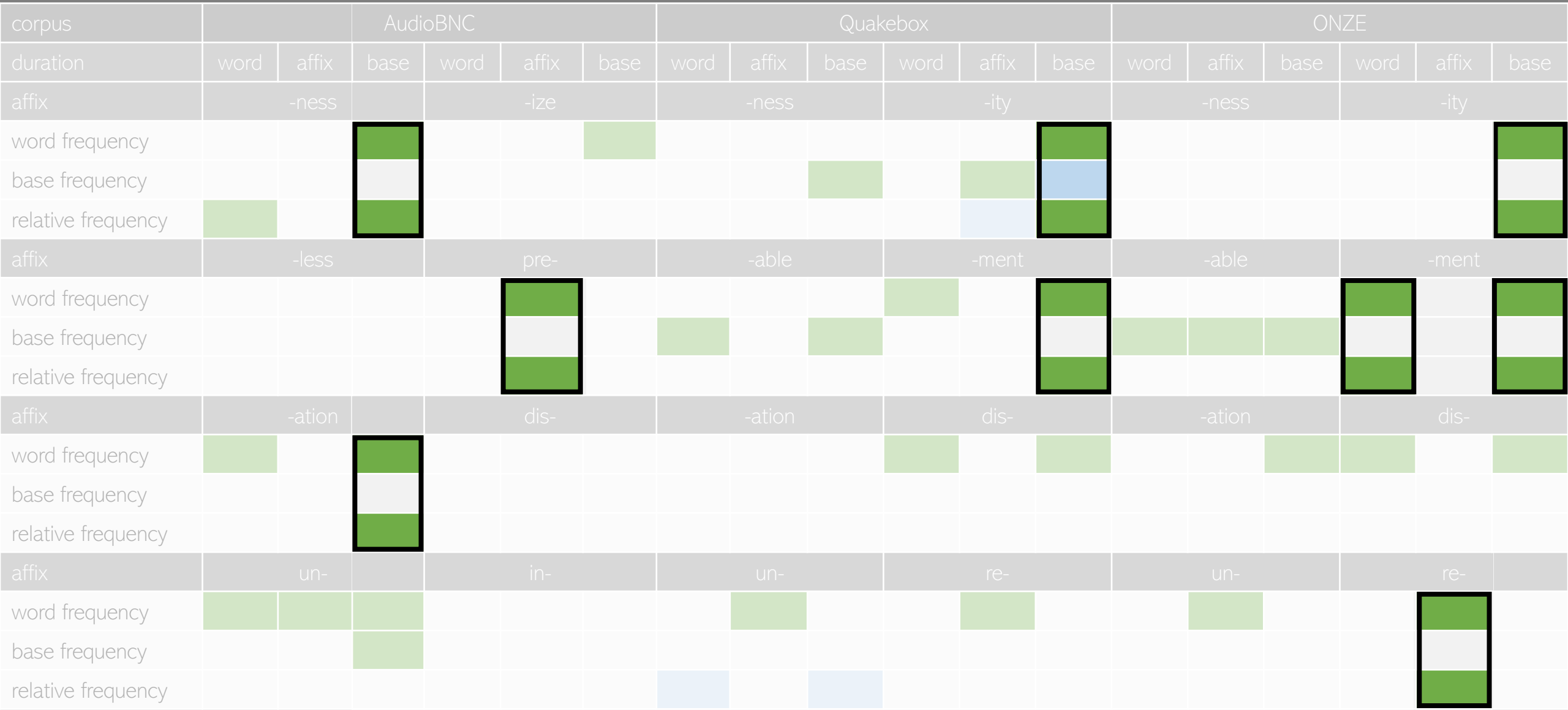
Overview of significant frequency effects



p < .001 expected direction

p < .001 unexpected direction

Overview of significant frequency effects



p < .001 expected direction

p < .001 unexpected direction

Additional analyses

- Solving the puzzle …
  - Can the varying emergence of frequency effects be explained by:
    - the domain of durational measurement (word, affix, base)?
    - the type of affix (prefix vs. suffix)?
    - the segmentation (automatic forced alignment vs. manual resegmentation)?
    - the size of the datasets (number of observations for a given affix)?
    - the type of prosodic affix integration (pword vs. clitic group vs. integrating)?
    - the semantic information load of the affix?
    - the conditional affix probability given the preceding unit?

Raffelsiefen 1999, 2007
Ben Hedia 2019

Summary

- Frequency effects are not completely arbitrary. If they emerge, higher word and base frequency are usually associated with reduction.
- The case is less clear for higher relative frequency (higher segmentability), but it is usually associated with less reduction, which is in line with the segmentability hypothesis.

- HOWEVER:
  - There are many null results: Relative frequency rarely has an effect at all.
    - Only some affixes are sensitive to it, independent of prosodic structure and affix informativity.
    - Effects emerge (yet) unpredictably for different affixes, durational domains, corpora.
  - Relative frequency effects almost always only appear together with word frequency.
    - Relative frequency is of course highly correlated with word frequency.
    - We might not measure anything beyond what word frequency already captures.
  - Relative frequency seems to be an unreliable measure for segmentability effects on duration.
    - We may need to explore other factors for the morphology-phonetics interaction and for processing in the mental lexicon.

Thank you for listening!

- Bell, Alan, Daniel Jurafsky, Eric Fosler-Lussier, Cynthia Girand, Michelle Gregory & Daniel Gildea. 2003. Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *The Journal of the Acoustical Society of America* 113.2: 1001–1024.
- Ben Hedia, Sonia. 2019. *Gemination and degemination in English affixation: Investigating the interplay between morphology, phonology and phonetics* (Studies in Laboratory Phonology 8). Berlin: Language Science Press.
- Boersma, Paul & David J. M. Weenik. 2001. *Praat: Doing phonetics by computer*. Version 5.4.04. http://www.praat.org/.
- Bybee, Joan L. 2000. The phonology of the lexicon: Evidence from lexical diffusion. In Michael Barlow & Suzanne Kemmer (eds.), *Usage-based models of language*, 65–85. Cambridge: Cambridge University Press.
- Coleman, John, Ladan Baghai-Ravary, John Pybus & Sergio Grau. 2012. *Audio BNC: The audio edition of the Spoken British National Corpus*. Phonetics Laboratory, University of Oxford. http://www.phon.ox.ac.uk/AudioBNC.
- Davies, Mark. 2008–. *The Corpus of Contemporary American English: 450 million words, 1990–present*. http://corpus.byu.edu/coca/.
- Gahl, Susanne. 2008. Thyme and time are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language* 84.3: 474–496.
- Guy, Gregory R. 1980. Variation in the group and the individual: The case of final stop deletion. In William Labov (ed.), *Locating language in time and space* (Quantitative Analyses of Linguistic Structure 1), 1–36. New York, London: Academic Press.
- Guy, Gregory R. 1991. Explanation in variable phonology: An exponential model of morphological constraints. *Language Variation and Change* 3.1: 1–22.
- Hay, Jennifer. 2001. Lexical frequency in morphology: Is everything relative? *Linguistics* 39.6: 041–1070.
- Hay, Jennifer. 2003. *Causes and consequences of word structure*. New York, London: Routledge.

cont.

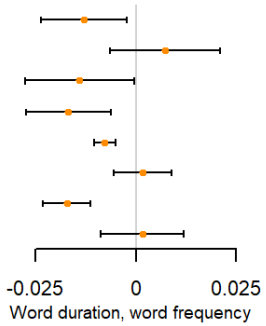
- Hay, Jennifer. 2007. The phonetics of un. In Judith Munat (ed.), *Lexical creativity, texts and contexts*, 39–57. Amsterdam, Philadelphia: John Benjamins.
- Hildebrandt, Kristine A. 2015. The prosodic word. In John R. Taylor (ed.), *The Oxford Handbook of the Word*. Oxford: Oxford University Press.
- Jurafsky, Daniel, Alan Bell, Michelle Gregory & William D. Raymond. 2001. Probabilistic relations between words: Evidence from reduction in lexical production. In Joan Bybee & Paul J. Hopper (eds.), *Frequency and the emergence of linguistic structure* (Typological Studies in Language 45), 229–254. Amsterdam: John Benjamins.
- Jurafsky, Daniel. 2003. Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In Rens Bod, Jennifer Hay & Stefanie Jannedy (eds.), *Probabilistic linguistics*, 39–95. Cambridge, London: MIT Press.
- Labov, William. 1989. The child as linguistic historian. *Language Variation and Change* 1.1: 85–97.
- Losiewicz, Beth L. 1995. Word frequency effects on the acoustic duration of morphemes. *The Journal of the Acoustical Society of America* 97.5: 3243.
- Plag, Ingo & Sonia Ben Hedia. 2018. The phonetics of newly derived words: Testing the effect of morphological segmentability on affix duration. In Sabine Arndt-Lappe, Angelika Braun, Claudine Moulin & Esme Winter-Froemel (eds.), *Expanding the lexicon: Linguistic innovation, morphological productivity, and ludicity*, 93–116. Berlin, New York: Mouton de Gruyter.
- Pluymaekers, Mark, Mirjam Ernestus & R. Harald Baayen. 2005a. Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica* 62.2–4: 146–159.
- Pluymaekers, Mark, Mirjam Ernestus & R. Harald Baayen. 2005b. Lexical frequency and acoustic reduction in spoken Dutch. *The Journal of the Acoustical Society of America* 118.4: 2561–2569.

cont.

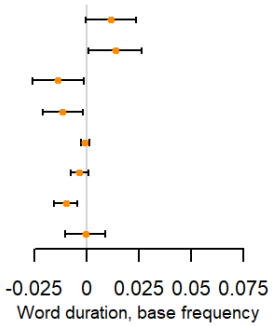
- R Core Team. 2020. *R: A language and environment for statistical computing*. Version 4.0.1. Vienna: R Foundation for Statistical Computing. http://www.R-project.org/.
- Raffelsiefen, Renate. 1999. Diagnostics for prosodic words revisited: The case of historically prefixed words in English. In Tracy A. Hall & Ursula Kleinhenz (eds.), *Studies of the phonological word*. 133–201. Amsterdam, Philadelphia: Benjamins.
- Raffelsiefen, Renate. 2007. Morphological word structure in English and Swedish: The evidence from prosody. In Geert Booij, Luca Ducceschi, Bernard Fradin, Ernesto Guevara, Angela Ralli & Sergio Scalise (eds.), *Online Proceedings of the Fifth Mediterranean Morphology Meeting* (MMM5), Fréjus, 15-18 September 2005, 209–268.
- Schuppler, Barbara, Wim A. van Dommelen, Jacques Koreman & Mirjam Ernestus. 2012. How linguistic and probabilistic properties of a word affect the realization of its final /t/: Studies at the phonemic and sub-phonemic level. *Journal of Phonetics* 40.4: 595–607.
- Vitevitch, Michael S. & Paul A. Luce. 2004. A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, and Computers* 36.3: 481–487.
- Zuraw, Kie, Isabelle Lin, Meng Yang & Sharon Peperkamp. 2020. Competition between whole-word and decomposed representations of English prefixed words. *Morphology*. 10.1007/s11525-020-09354-6.

AudioBNC

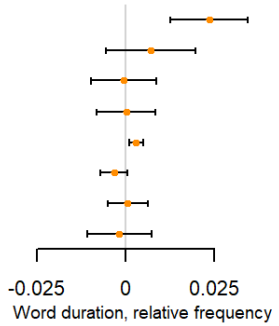
Affix	β	p
-ness	-0.013	*
-less	0.007	–
pre-	-0.014	–
-ize	-0.017	**
-ation	-0.008	***
dis-	0.002	–
un-	-0.017	***
in-	0.002	–



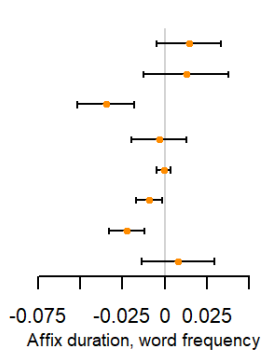
Affix	β	p
-ness	0.011	–
-less	0.014	–
pre-	-0.014	*
-ize	-0.012	*
-ation	-0.001	–
dis-	-0.003	–
un-	-0.01	***
in-	-0.001	–



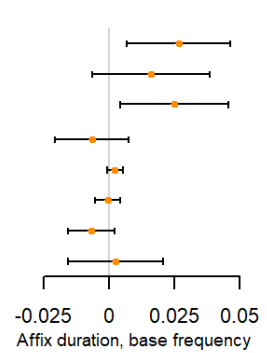
Affix	β	p
-ness	0.024	***
-less	0.007	–
pre-	-0.001	–
-ize	0	–
-ation	0.003	**
dis-	-0.003	–
un-	0.001	–
in-	-0.002	–



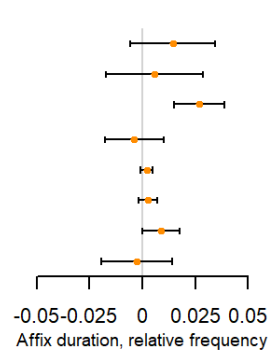
Affix	β	p
-ness	0.014	–
-less	0.013	–
pre-	-0.035	***
-ize	-0.003	–
-ation	-0.001	–
dis-	-0.009	–
un-	-0.022	***
in-	0.008	–



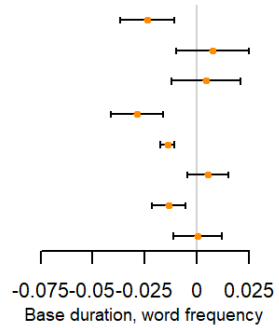
Affix	β	p
-ness	0.027	**
-less	0.016	–
pre-	0.025	*
-ize	-0.006	–
-ation	0.002	–
dis-	0	–
un-	-0.007	–
in-	0.002	–



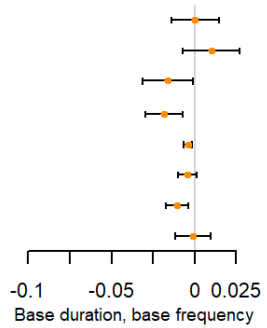
Affix	β	p
-ness	0.014	–
-less	0.006	–
pre-	0.027	***
-ize	-0.004	–
-ation	0.002	–
dis-	0.003	–
un-	0.009	–
in-	-0.003	–



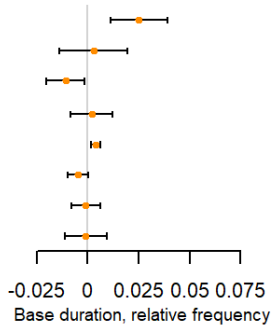
Affix	β	p
-ness	-0.024	***
-less	0.008	–
pre-	0.004	–
-ize	-0.029	***
-ation	-0.014	***
dis-	0.005	–
un-	-0.014	***
in-	0	–



Affix	β	p
-ness	0	*
-less	0.01	–
pre-	-0.016	–
-ize	-0.018	**
-ation	-0.004	**
dis-	-0.004	–
un-	-0.01	**
in-	-0.001	–

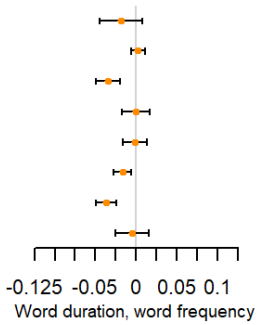


Affix	β	p
-ness	0.025	***
-less	0.003	–
pre-	-0.011	*
-ize	0.002	–
-ation	0.004	***
dis-	-0.005	–
un-	-0.001	–
in-	-0.001	–

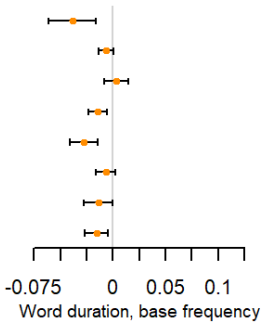


Quakebox

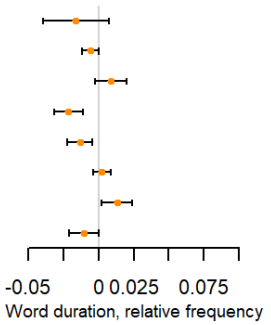
Affix	β	p
-ness	-0.019	–
-ation	0.002	–
dis-	-0.034	***
un-	0	–
-able	-0.002	–
-ity	-0.017	**
-ment	-0.037	***
re-	-0.005	–



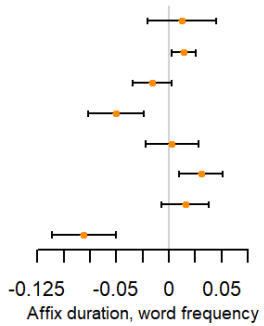
Affix	β	p
-ness	-0.038	**
-ation	-0.006	–
dis-	0.003	–
un-	-0.014	**
-able	-0.027	***
-ity	-0.007	–
-ment	-0.014	–
re-	-0.015	**



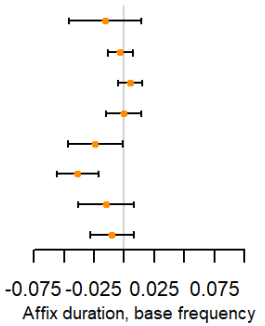
Affix	β	p
-ness	-0.016	–
-ation	-0.006	–
dis-	0.009	–
un-	-0.021	***
-able	-0.013	**
-ity	0.002	–
-ment	0.013	*
re-	-0.01	–



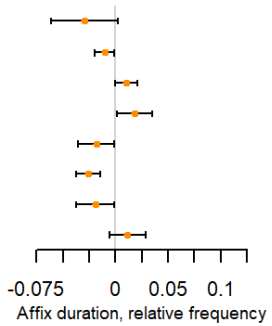
Affix	β	p
-ness	0.012	–
-ation	0.014	*
dis-	-0.016	–
un-	-0.05	***
-able	0.003	–
-ity	0.031	**
-ment	0.016	–
re-	-0.081	***



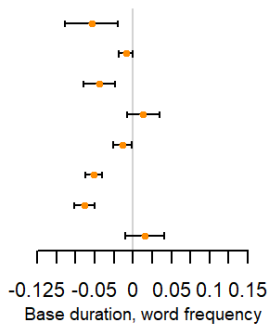
Affix	β	p
-ness	-0.016	–
-ation	-0.003	**
dis-	0.005	–
un-	0	–
-able	-0.024	*
-ity	-0.039	***
-ment	-0.015	–
re-	-0.01	–



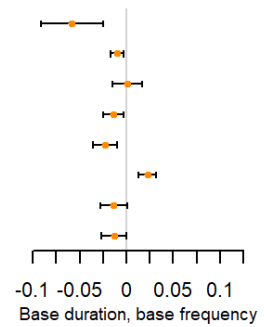
Affix	β	p
-ness	-0.03	–
-ation	-0.01	–
dis-	0.01	–
un-	0.018	*
-able	-0.018	*
-ity	-0.026	***
-ment	-0.019	–
re-	0.011	–



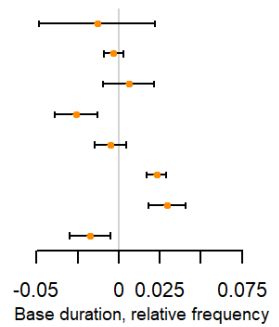
Affix	β	p
-ness	-0.054	**
-ation	-0.009	–
dis-	-0.043	***
un-	0.014	–
-able	-0.014	*
-ity	-0.051	***
-ment	-0.064	***
re-	0.015	–



Affix	β	p
-ness	-0.058	***
-ation	-0.01	–
dis-	0.001	–
un-	-0.014	*
-able	-0.023	***
-ity	0.023	***
-ment	-0.013	–
re-	-0.013	*

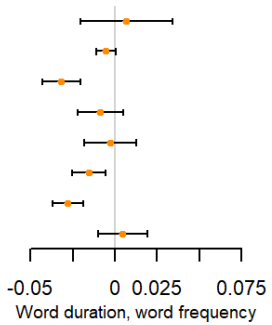


Affix	β	p
-ness	-0.013	–
-ation	-0.003	–
dis-	0.006	–
un-	-0.026	***
-able	-0.005	–
-ity	0.023	***
-ment	0.029	***
re-	-0.017	**

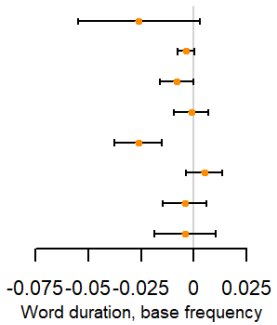


ONZE

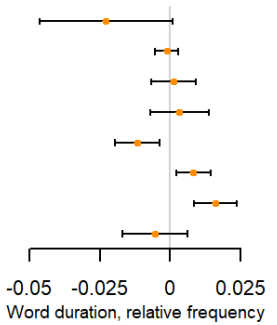
Affix	β	p
-ness	0.007	–
-ation	-0.005	–
dis-	-0.032	***
un-	-0.009	–
-able	-0.003	–
-ity	-0.015	**
-ment	-0.028	***
re-	0.005	–



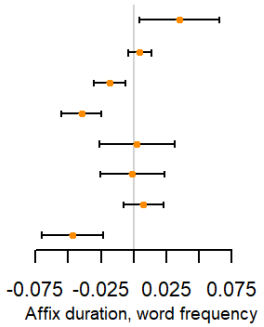
Affix	β	p
-ness	-0.026	–
-ation	-0.004	–
dis-	-0.008	*
un-	-0.001	–
-able	-0.026	***
-ity	0.005	–
-ment	-0.004	–
re-	-0.004	–



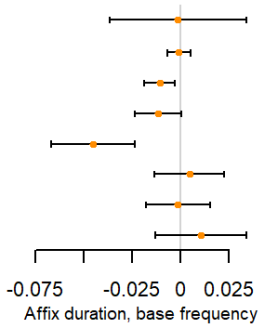
Affix	β	p
-ness	-0.023	–
-ation	-0.001	–
dis-	0.001	–
un-	0.003	–
-able	-0.012	**
-ity	0.008	*
-ment	0.016	***
re-	-0.005	–



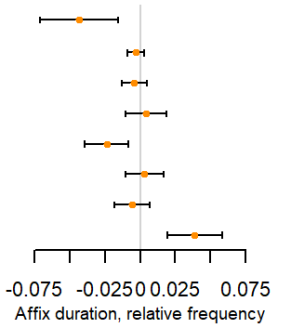
Affix	β	p
-ness	0.035	*
-ation	0.005	–
dis-	-0.018	**
un-	-0.04	***
-able	0.003	–
-ity	-0.001	–
-ment	0.008	–
re-	-0.047	***



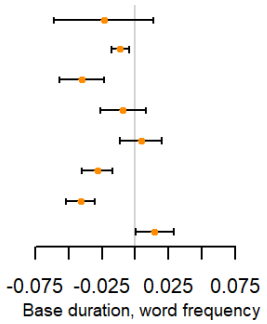
Affix	β	p
-ness	-0.001	–
-ation	-0.001	–
dis-	-0.011	**
un-	-0.012	–
-able	-0.045	***
-ity	0.005	–
-ment	-0.001	–
re-	0.011	–



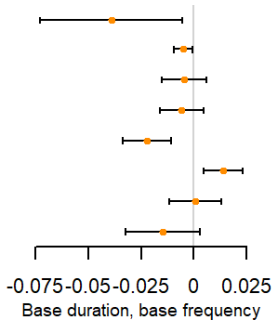
Affix	β	p
-ness	-0.043	**
-ation	-0.003	–
dis-	-0.004	–
un-	0.004	–
-able	-0.024	**
-ity	0.003	–
-ment	-0.006	–
re-	0.039	***



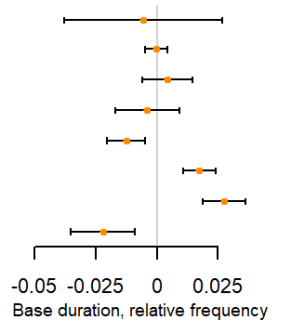
Affix	β	p
-ness	-0.024	–
-ation	-0.011	***
dis-	-0.04	***
un-	-0.009	–
-able	0.004	–
-ity	-0.028	***
-ment	-0.041	***
re-	0.014	*



Affix	β	p
-ness	-0.039	*
-ation	-0.005	*
dis-	-0.005	–
un-	-0.006	–
-able	-0.022	***
-ity	0.014	**
-ment	0.001	–
re-	-0.015	–



Affix	β	p
-ness	-0.005	–
-ation	0	–
dis-	0.004	–
un-	-0.004	–
-able	-0.012	**
-ity	0.017	***
-ment	0.028	***
re-	-0.022	**



Prosodic structure

The prosodic hierarchy

Some pword-diagnostics

- onset or coda conditions, LOI-violations, ambisyllabicity
- stress and relative prominence
- trisyllabic laxing, vowel reduction
- minimal word requirements
- compositionality, type of base

Morpho-prosodic alignment

- A morpheme cannot include multiple pwords,
  but a pword can include multiple morphemes.

Hildebrandt 2015, Raffelsiefen 1999, 2007

Prosodic structure

pword-forming

clitic group

integrating

following Raffelsiefen 1999

Prosodic structure

- Mixed-effects models with interactions on the response duration difference
  - relative frequency · type of morpheme
  - relative frequency · prosodic category
  - prosodic category · type of morpheme
  - speech rate
  - number of syllables
  - bigram frequency
  - sum of biphone probabilities
  - random intercepts for word type

The interaction between relative frequency and prosodic category is not significant in any of the three corpora.

Informativity

Measured in two ways for the AudioBNC:

Semantic information load score

5-point Likert scales coded for:

- clearness of semantic meaning
- type of base: free vs. bound root
- semantic transparency
- productivity

Affix-specific semantic segmentability hierarchy

H: The higher the semantic information load, the longer the duration.

Conditional affix probability C_{aff}

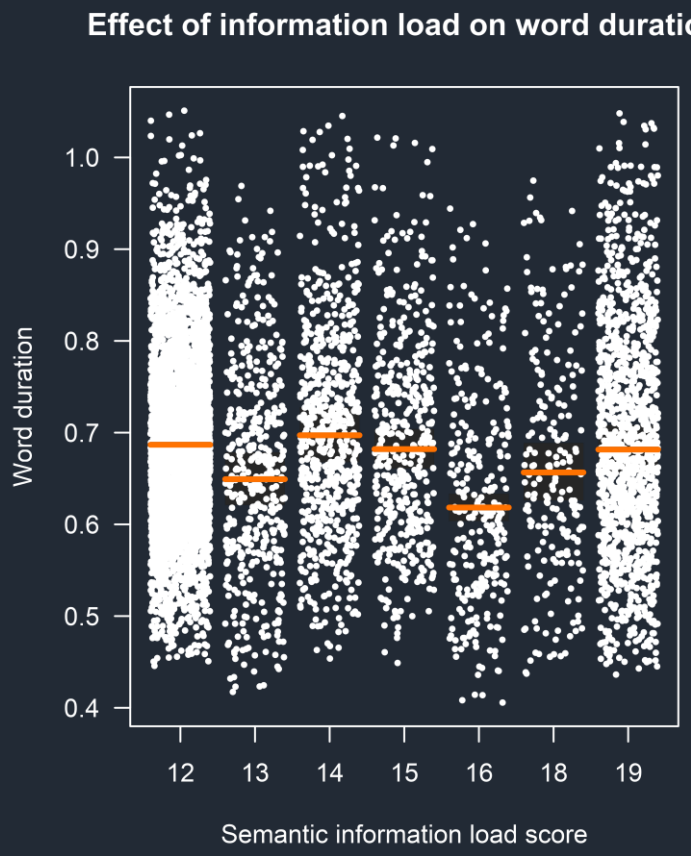
Affix probability given preceding word:

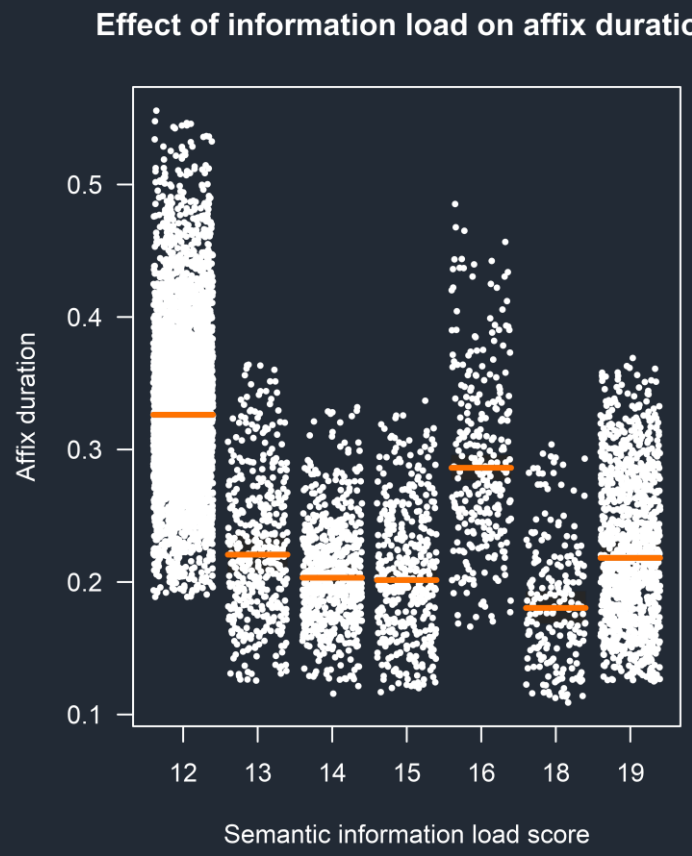
SUFFIX EXAMPLE		PREFIX EXAMPLE		
A	B	A	B	C
random	*ize*	*her*	*pre-*	…

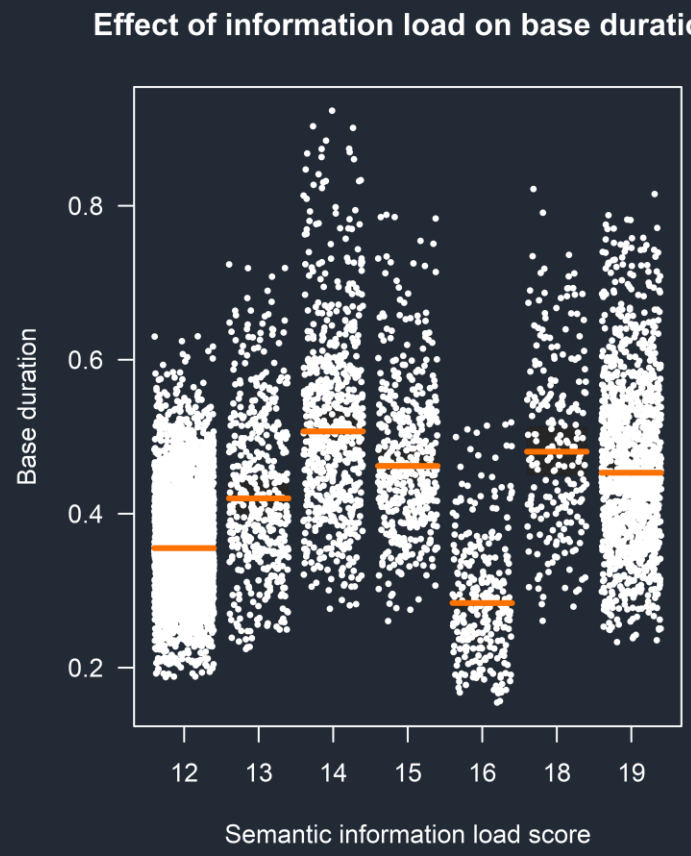
$$C_{aff} = \frac{Freq(AB)}{Freq(A)}$$

H: The higher the conditional affix probability, the shorter the duration.

Semantic information load







Conditional affix probability

duration	word	affix	base	word	affix	base	word	affix	base	word	affix	base
affix	-ness			-ize			-ation			dis-		
affix probability												
affix	-less			pre-			un-			in-		
affix probability												

p < .001 reduction