Bruce Hayes[1]
University of California, Los Angeles

Talk at Cornell University
March 30, 2018

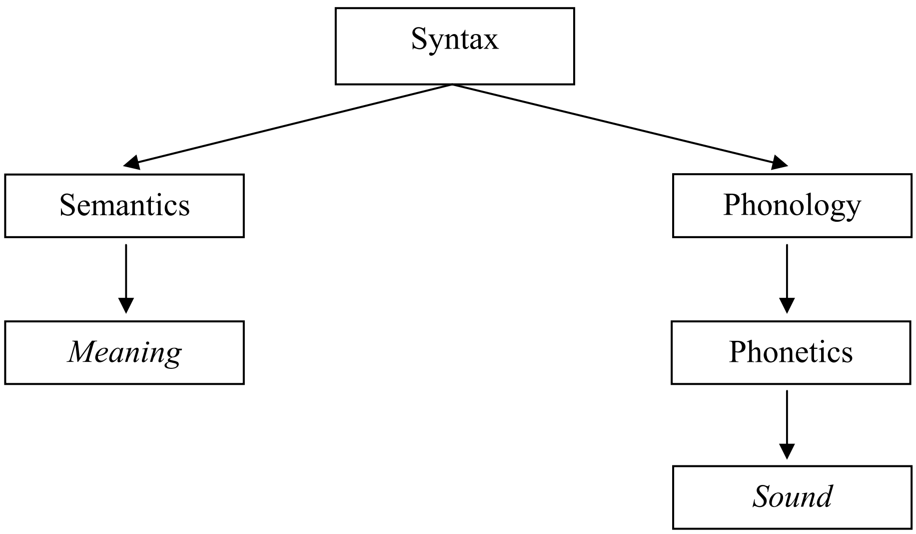
Phonological markedness effects in syntax: subtle but ubiquitous

I. CONTEXT

1. Research question

- How do the domains of human linguistic knowledge (“components”) interact in the creation of sentences?

2. The classical feed-forward model (e.g. Chomsky 1965)



- Key prediction: the construction of sentences is blind to any phonological consequences of word-concatenation.

3. Challenges to the feed-forward model arrive from multiple directions

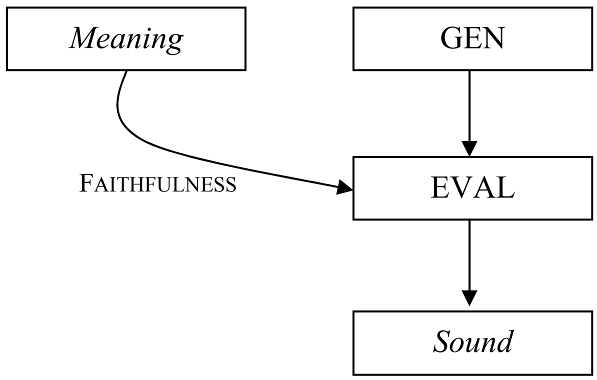
- Recent work:
  - Shih and Zuraw (2017): Tagalog speakers use the twin syntactic constructions Adj. $\left\{ \begin{matrix} \textit{ŋ} \\ \textit{na} \end{matrix} \right\}$ Noun vs. Noun $\left\{ \begin{matrix} \textit{ŋ} \\ \textit{na} \end{matrix} \right\}$ Adj. in ways that statistically avoid violating these constraints:

    — *[+nasal][+nasal]
    — *HIATUS
    — *NC̥

[1] This talk gives the results of a collaboration with **Canaan Breiss**, a 2nd-year graduate student in the UCLA Linguistics Department.

- Much other work, e.g. Inkelas and Zec (1990), Zuraw (2015), Shih et al. (2015), Shih (2012, 2017), Anttila (2016), Ryan (in press)

4. A pure-parallelist alternative to (2)



- … where candidates are deeply structured objects with semantic, syntactic, phonological, morphological, and phonetic structure.
- This is just one conception of a parallelist architecture; for established research programs see e.g. Jackendoff (2002, 2010), Bresnan et al. (2015).

5. Goal

- Earlier work employs scalpel-like precision on specific domains like Tagalog adjective-noun word order.
- We seek to learn how general these effects are from an **across-the-board, brute force** approach:
  - Look at whole sentences — *all* word concatenation.
  - Since we are indiscriminate in our choice of sequences, we can look at a wide variety of phonological constraints.

II. METHOD

6. A precedent: Martin (2011) on *GEMINATE in English compounds

- Key finding: compounds that violate *GEMINATE, such as *bookkeeper* [ˈbʊkkipɚ], are *statistically underrepresented* relative to those that don't.
- He uses a **Monte Carlo** method (Good 2005), summarized below.

7. Step 1: Form a list of *bigrams* (consecutive two-word sequences)

- Example: if the text is Jane Austen's novel *Emma* (first sentence given below) …

> Emma Woodhouse, handsome, clever, and rich, with a comfortable home and happy disposition, seemed to unite some of the best blessings of existence.

- … then the bigram sequence begins:

  [Emma Woodhouse], [Woodhouse, handsome], [handsome, clever], …

- Look up words in our augmented version of the CMU Pronouncing Dictionary:

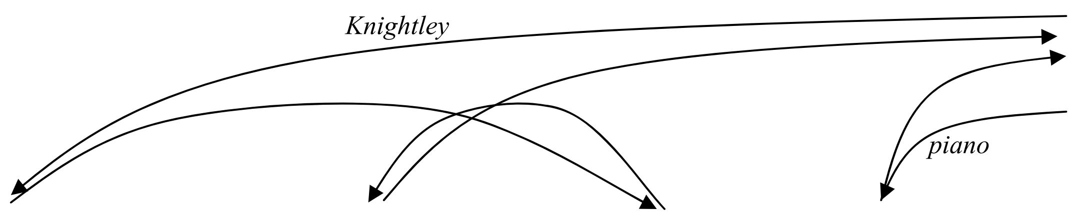
  [ɛmə, wʊdhaʊs][wʊdhaʊs, hændsəm][hænsəm, klɛvɚ] …

8. Step 2: Count how many times the constraints are violated in the bigrams

- In the bigrams of *Emma*, *GEMINATE is violated 590 times.
  - ➢ Example: *success supposes* [səkˈsɛs səˈpoʊzəz],

9. Step 3: Randomly shuffle the bigrams of the corpus

- For each bigram in the list, replace its Word2 with the Word2 of a randomly-selected bigram.



  [Emma *Woodhouse*], [Woodhouse, *handsome*], [handsome, *clever*], [clever, *and*], …

  [Emma *Knightley*], [Woodhouse, *clever*], [handsome, *Woodhouse*], [clever, *piano*], …

- Purpose: estimate the number of violations that might occur under the null hypothesis that word concatenation is phonologically free.

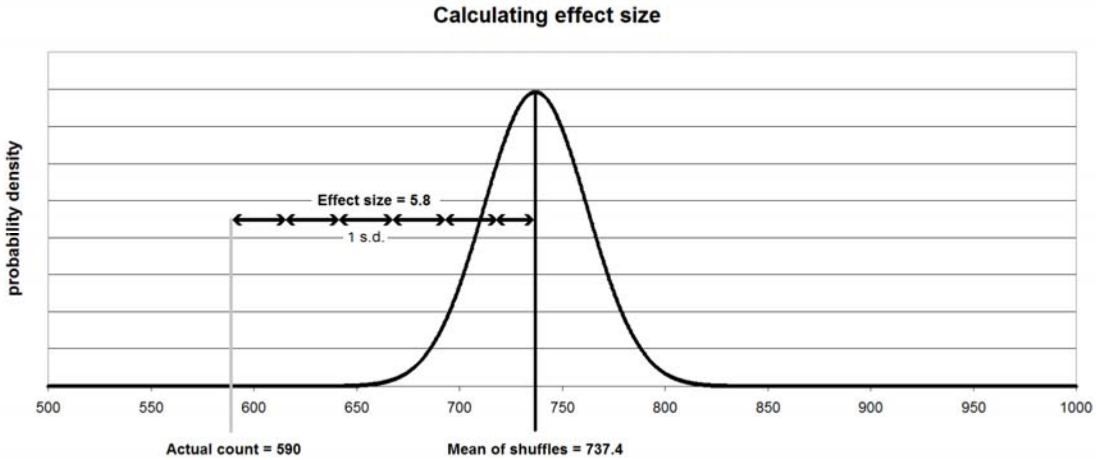
10. Step 4: Shuffle repeatedly, obtaining means and standard deviations of violation counts

- Shuffling *Emma* 1000 times (our standard), we find that *GEMINATE is violated an average of 737.4 times. Standard deviation is 25.2.

11. Step 5: Assessing underrepresentation in the real count

- We use a standard metric (Coe 2002): **effect size**
  - ➢ = number of standard deviations that the observed value falls below the mean of the shuffles

12. Computing effect size for *GEMINATE in *Emma*



- Mean violations of *GEMINATE in the shuffles = 737.4
- Standard deviation = 25.2
- Number of violations of *GEMINATE in corpus = 590
- Effect size = (737.2 − 590) / 25.2 = **5.8**

13. Interpreting effect size

- Rule of thumb: an effect size of magnitude 2.4 corresponds to a *p*-value of 0.01.
- 5.8 is a very substantial effect, suggesting that Jane Austen may have been avoiding geminates when she composed *Emma*.

III. REFINEMENTS

14. Refinement 1: taking into account phonological phrasing

- Across languages, phrasal phonology tends to be blocked at large prosodic breaks.
- So we might get cleaner results if we only consider bigrams falling within the same large prosodic phrase.
- To operationalize, we select the bigrams that are not separated in text by punctuation:

[Emma Woodhouse],	[Woodhouse, handsome],	[handsome, clever],	[clever, and], …
↓	↓	↓	↓
[Emma Woodhouse]	∅	∅	∅

- And the complement set can serve as a control case.

[Emma Woodhouse],	[Woodhouse, handsome],	[handsome, clever],	[clever, and], …
↓	↓	↓	↓
∅	[Woodhouse, handsome],	[handsome, clever],	[clever, and], …

- We expect such violations to be more freely tolerated.

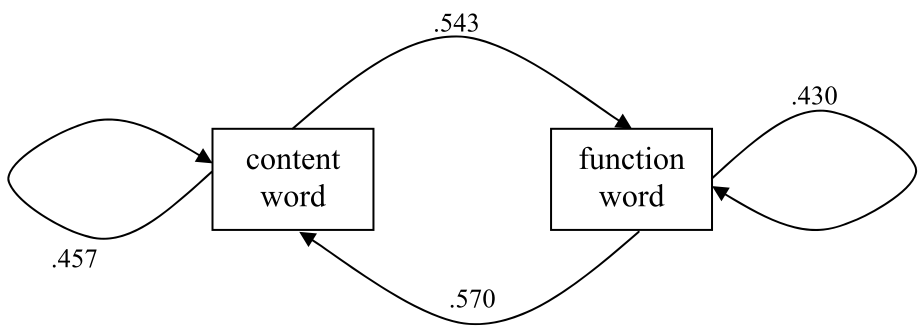
15. Refinement 2: dealing with frequent bigrams

- Frequent bigrams have three origins:
  - Common syntactic patterns, e.g. *subject pronoun* + *copula* (*I am*).
  - Frequent phrases (*a great deal*, *very good*, *a few minutes*)
  - Items frequent in a particular text: e.g. *Mr. Knightley* in *Emma.*
- If our interest is in studying *productive word concatenation*, it would be worthwhile to remove these phrases, which are likely to be lexicalized.
- To do this: discard all but the **hapax** bigrams (unique in corpus).

16. Refinement 3: controlling for syntax

- Even sticking just to hapaxes, it remains a fact that syntax has major effects on word order (!).
- We would like to control for this statistically.
- Ideally, we would implement model (4) but we need a lot of help …
- So until this is feasible, we use …

17. The poor man's syntax: a grammar for *Emma*



- Function words tend to be followed by content words and vice versa.
- The simplest control procedure, adopted for this talk, is just to *throw away all but content + content bigrams*.[2]
  - Reduces the *Emma* bigram set like this:

[with a]	[a comfortable]	[comfortable home]	[home and]	[and happy]	[happy disposition]
↓	↓	↓	↓	↓	↓
∅	∅	[comfortable home]	∅	∅	[happy disposition]

[2] "Segregating" the shuffles into content-initial and function-initial populations has a similar effect; we report content-only shuffling because the effects it finds are weak than segregated shuffling.

IV. PHONOLOGICAL CONSTRAINTS EXAMINED

18. Constraints examined

a.	*CLASH	Stressed syllables flanking #: *táll trées*
b.	*IAMBIC CLASH	Iambic word before stress, like **seréne lúte*
c.	*C#C	Simple ban on CC cluster across word boundary
d.	*3+ CONSONANTS	*C#CC or *CC#C
e.	*3+ OBSTRUENTS	ditto for obstruents only
f.	*4+ CONSONANTS	
g.	*4+ OBSTRUENTS	
h.	*GEMINATE	
i.	*BAD SONORITY	Violate Syllable Contact Law (Vennemann 1988)
j.	*HIATUS	*V#V
k.	*NC̥	voiceless consonant after nasal (Pater 1999)
l.	*SIBILANT CLUSTER	Sibilants are: {s, ʃ, z, ʒ, tʃ, dʒ}

19. We also studied "virtues"

- We made up a group of configurations that struck us as phonologically completely ordinary, calling them "Virtues".

a.	VOWEL # ALVEOLAR STOP	
b.	r # t	
c.	V # r	
d.	NAS # VCED HOMORG. STOP	Obeys *NC̥, place and voice assimilation
e.	V#CV	
f.	GOOD SONORITY	opposite of BAD SONORITY
g.	UNSTRESSED # STRESS	part of the complement set of *CLASH

- We expect no underrepresentation for classically unmarked configurations, so they serve as a check on the method.

V. CORPORA EXAMINED

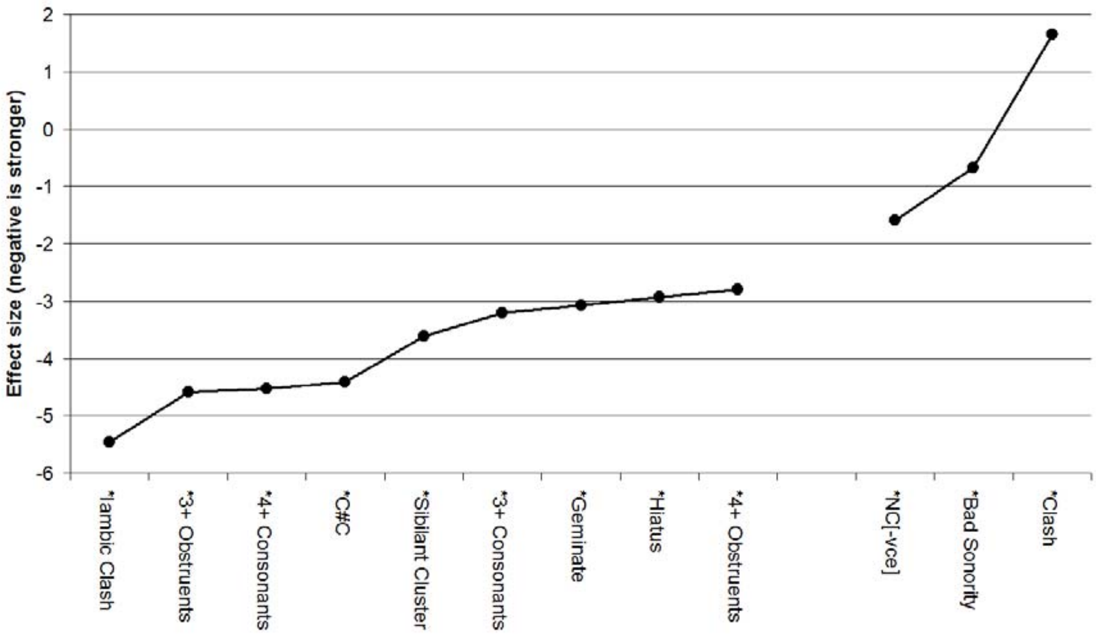
20. List

- Six novels by Jane Austen (722,000 words)
- Six novels by Mark Twain (568,000 words)
- Six novels by Nathaniel Hawthorne (592,000 words)
- Six non-fiction works by Charles Darwin (935,000 words)
- A mélange of conversations from six corpora of spoken English (~ 1,000,000 words)
  - 2016 Primary Debates corpus, Buckeye Corpus, Beatles corpus (Stanton 2016), Michigan Corpus of Academic English, British Academic Spoken English corpus, Human Communications Research Center Map Task corpus.

VI. RESULTS

21. Results I

- We plot effect size, averaged across all five corpora.
- Mostly, there is underrepresentation.
- Gap separates out those that were found to be significantly underrepresented (effect size magnitude > 2.4) from those that were not. xxx



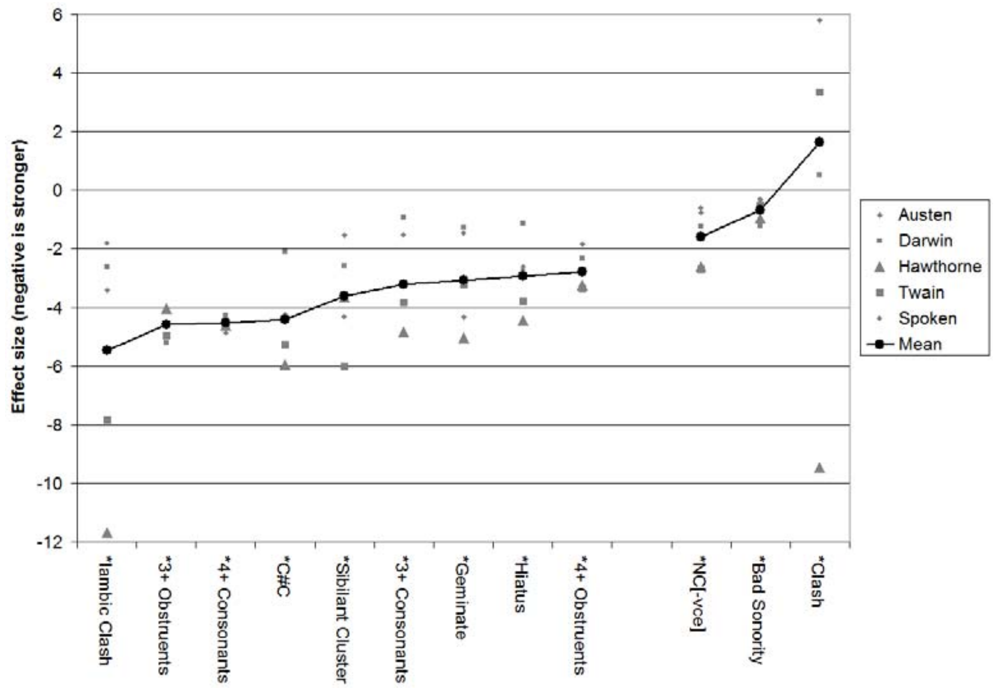
- These effects are small enough that they are likely to be hard to detect without statistical methods.[3]

22. Our sore thumb — *CLASH

- Unlike most of the others, this constraint did not show any underrepresentation.
- Yet:
  - It has an impeccable typological pedigree.
  - It is active in English phrasal phonology (Liberman and Prince 1977, Hayes 1984, etc. etc.)
- We discuss it further below.

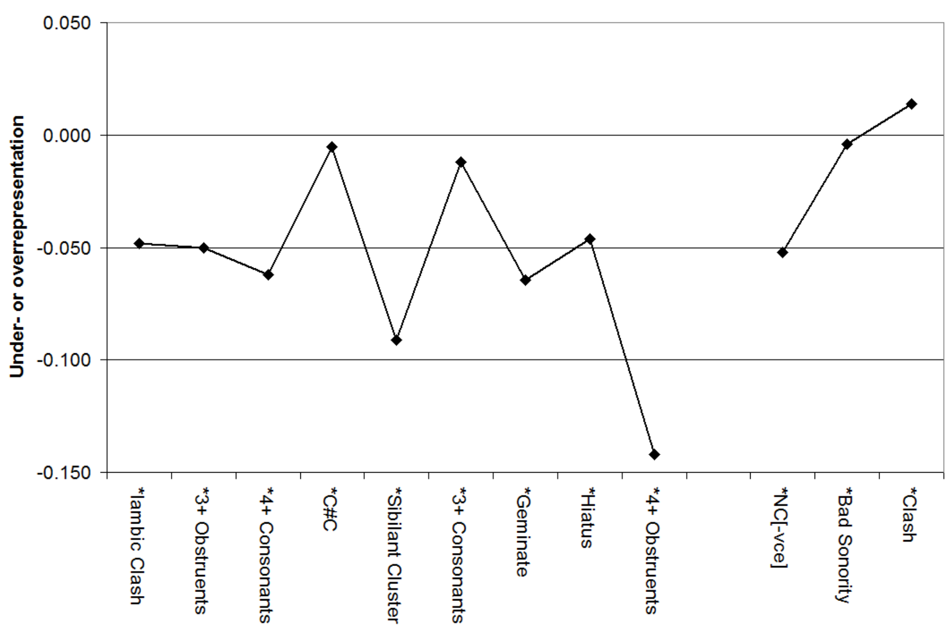
[3] See Bolinger (1965) for extensive anecdotal evidence supporting both *IAMBIC CLASH and (problematically) *CLASH in both word choice and syntactic choice.

23. Results II: breaking the results down by the five individual corpora



- Underepresentation is found for all cases except *CLASH.
- Note massive difference between corpora for *CLASH: spoken corpus/Austen are clash-seekers; Hawthorne is an avoider.

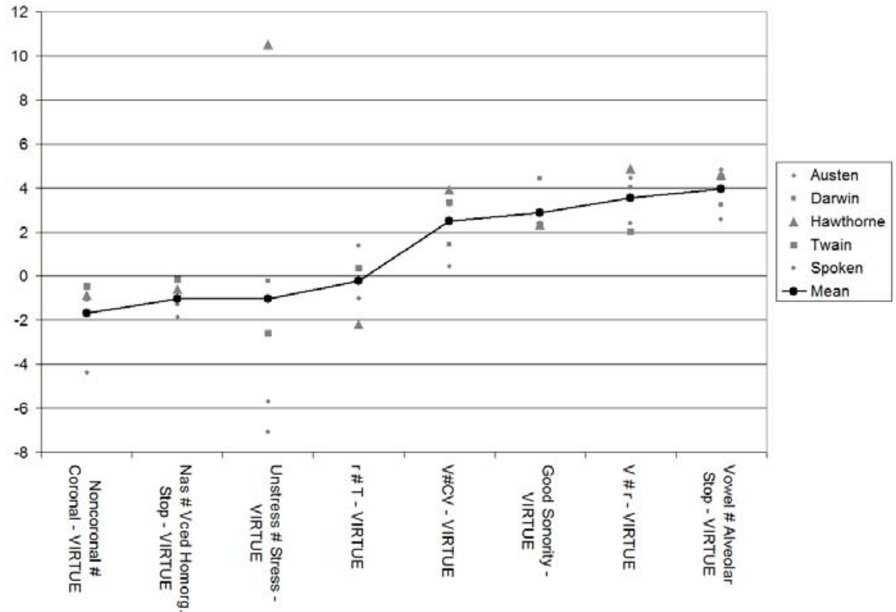
24. Results III: How big is the effect on a raw percentage basis?



- N.B. this is not a rigorous way to look at the data, just an aid to intuition.
  - Percentage size not closely tied to the effect size, since the standard deviation of the distribution also matters.

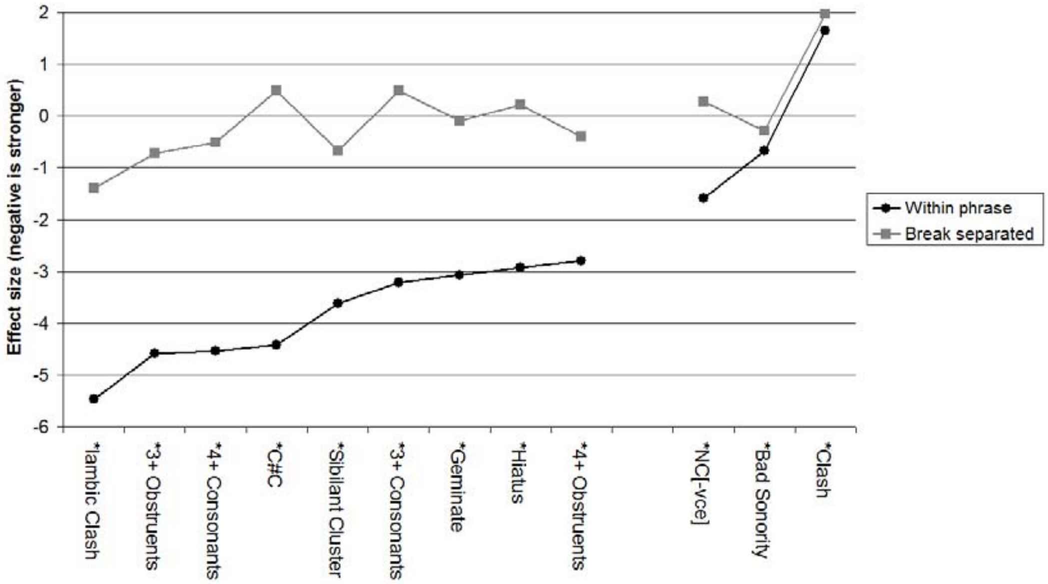
VII. CONTROL PROCEDURES

25. Our "virtues" (19) do not yield significant effect sizes



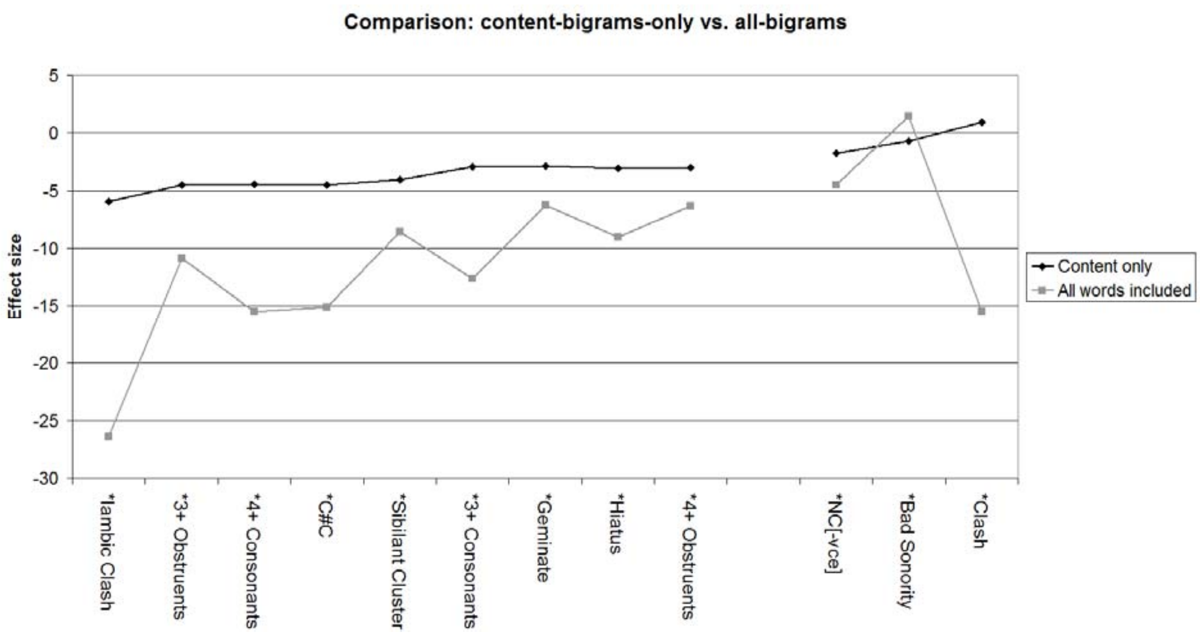
- None test out with significant average underrepresentation.
- Often, they instead involve overrepresentation, as expected.
- The aberrant Virtue UNSTRESSED # STRESS is partly the complement of *CLASH, so is in a sense the same mystery.

26. Per (14), significant effects are *not* found for bigrams formed across phonological breaks



- Basically, no effect — large breaks inhibit phonology.

27. What happens when you leave all the function words in?



- Key:
  - Black line = main results of (21), repeated but rescaled
  - Grey line = results obtained by including function words — far more dramatic
- We see that culling to just content words is indeed a conservative procedure.
- Note that *CLASH is strongly underrepresented when function words are included — suggesting that syntactically-driven content-function alternation likely plays a role in clash avoidance.

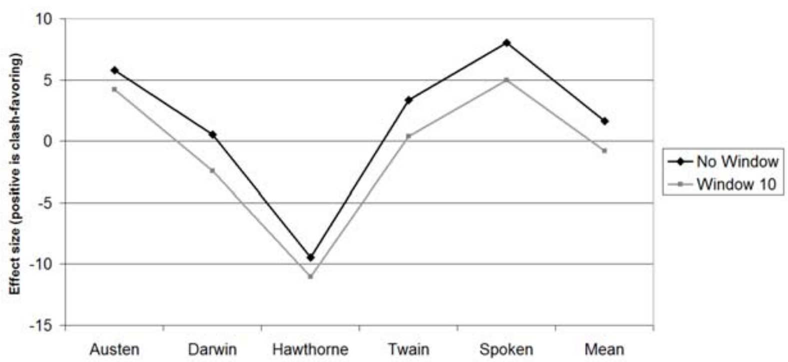
VIII. SOME THOUGHTS ON *CLASH

28. Burstiness?

- In texts, *CLASH violations have a **bursty** distribution (Altmann et al. 2009) — concentrated in clumps.
- Shuffling suppresses the bursts, eliminating sequences locally rich in *CLASH violations.
- Result: an artifact — text looks clash-seeking, but is merely bursty.

29. This explanation is incomplete

- We've tried **short-distance shuffling**, attempting to preserve burstiness.
- It lowers the overrepresentation a bit, but even so Austen and the spoken corpus still indicate clash-seeking.



IX. LOOKING AHEAD

30. Vowel harmony and language-specificity

- Could our results reflect general articulatory difficulty or actual phonology?
- Key test: assess whether these effects are language-specific.
- Empirical domain: vowel harmony
- So far: Finnish, Hungarian, Turkish each underrepresent harmony violations (i.e. of their own harmony pattern) across word boundaries; needs further checking.

X. CONCLUSIONS

31. What is the mechanism?

- Our favorite explanation: Martin (2011)
- He describes and models a mechanism of "leakage" in phonological acquisition,
  - Learners slightly misjudge the domain of word-bounded phonological constraints,
  - … causing them to "leak" into higher-domain versions, such as phrasal.

32. Summarizing

- Earlier work on phonological effects in syntax detected effects of specific constraints in specific syntactic constructions (per (3) and (6) above).
- Our "brute force" approach suggests that (modulo the *CLASH mystery) such effects are pervasive, found across the board and in essentially any text.
  - Per talk title: subtle but ubiquitous.
- Our results offer encouragement to the pursuit of parallelist (non-feed-forward) models of grammatical organization.

REFERENCES

Altmann, Eduardo, Janet B. Pierrehumbert and Adilson E. Motter. (2009). Beyond word frequency: bursts, lulls, and scaling in the temporary distribution of words. *PLoS ONE* 4 (11):e7678.

Anttila, Arto. (2016). Phonological effects on syntactic variation. *Annual Review of Linguistics* 2:115-137.

Boersma, Paul and Joe Pater (2016) Convergence properties of a gradual learning algorithm for Harmonic Grammar. In John J. McCarthy and Joe Pater, eds., *Harmonic Grammar and Harmonic Serialism*. Sheffield: Equinox, pp. 389–434.

Bolinger, Dwight (1965) Pitch accent and sentence rhythm. In his *Forms of English: Accent, Morpheme, Order*. Cambridge, MA: Harvard University Press.

Bresnan, Joan, Ash Asudeh, Ida Toivonen, and Stephen Wechsler. 2015. *Lexical-Functional Syntax, 2nd edition.* Wiley-Blackwell.

Chomsky, Noam (1965) *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Coe, Robert (2002) It's the effect size, stupid: what effect size is and why it is important. Paper presented at the Annual Conference of the British Educational Research Association, University of Exeter, England, 12-14 September 2002. Available online.

Good, Phillip I. (2005). *Permutation, Parametric and Bootstrap Tests of Hypotheses*. Springer.

Hayes, Bruce (1984) The phonology of rhythm in English. *Linguistic Inquiry* 15:33-74.

Inkelas, Sharon and Draga Zec (1990) *The Phonology-Syntax Connection*. Chicago: University of Chicago Press.

Jackendoff, Ray. 2002. *Foundations of Language*. New York: Oxford University Press.

Jackendoff, Ray. 2010. *The Parallel Architecture and its place in cognitive science*. In B. Heine and H. Narrog (eds.), *The Oxford Handbook of Linguistic Analysis*, Chapter 23, pages 583–605, Oxford: Oxford University Press.

Liberman, Mark and Alan Prince. 1977. On stress and linguistic rhythm. *Linguistic Inquiry* 8: 249–336.

Martin, Andrew (2011) Grammars leak: Modeling how phonotactic generalizations interact within the grammar. *Language* 87:751-770

Pater, Joe. (1999). Austronesian Nasal Substitution and other NÇ̥ Effects. In René Kager, Harry van der Hulst, and Wim Zonneveld, eds., *The Prosody Morphology Interface*. Cambridge University Press, pp. 310-343.

Ryan, Kevin (in press) Prosodic end-weight reflects phrasal stress. To appear in *Natural Language and Linguistic Theory*.

Shih, Stephanie S. (2012) Linguistic determinants of English personal name choice. Paper given at the Annual Meeting of the Linguistic Society of America, Portland, OR, 7 January 2012.

Shih, Stephanie S. (2017) "Phonological influences in syntactic choice." In Vera Gribanova and Stephanie S. Shih (eds). *The morphosyntax-phonology connection: locality and directionality at the interface.* Oxford University Press. 223–252.

Shih, Stephanie S., Jason Grafmiller, Richard Futrell, and Joan Bresnan (2015) Rhythm's role in predicting genitive alternation choice in spoken English. In Vogel, R and R. van de Vijver (eds). *Rhythm in Cognition and Grammar: A Germanic Perspective*. Berlin, Germany: De Gruyter Mouton. 207–234.

Shih, Stephanie S. and Kie Ross Zuraw (2017) Phonological conditions on variable adjective-noun word order in Tagalog. *Phonological Analysis.*

Stanton, Juliet. 2016. Learnability shapes typology: the case of the midpoint pathology. *Language* 92: 753–791.

Vennemann, Theo. 1988. *Preference laws for syllable structure and the explanation of sound change: With special reference to German, Germanic, Italian, and Latin.* Berlin: Mouton de Gruyter.

Zuraw, Kie (2015) Allomorphs of French *de* in coordination: a reproducible study. *Linguistics Vanguard* 1(1): 57–68.