Incorporating phonological and syntactic factors into sentence probability distributions

Tim Hunter Canaan Breiss Bruce Hayes
Department of Linguistics, UCLA

Why *phonological* factors in *sentence* distributions?

Two rival views on the organization of grammar

1. Syntax determines which utterances are well-formed; phonology determines how they are realized (Chomsky 1965)
2. Different kinds of constraints (e.g. syntactic and phonological) jointly determine well-formedness of utterances (Sadock 1991, Jackendoff 1997, Bresnan 2000)

Three kinds of evidence **favor Option 2**.

Evidence I: Unacceptability for phonological reasons

Norwegian (Rice 2007): *SONORITYSEQUENCING

(1) Sykl opp bakken
bike up the.hill
"Bike up the hill!"

(2) * Sykl ned bakken
bike down the.hill
"Bike down the hill!"

This is not possible if phonology determines only how to realize what syntax gives it.

Evidence II: Choices among alternative constructions

Speakers whose grammar offers paired choices statistically prefer the phonologically unmarked one.

Example from Shih (2017): *CLASH

(3) the whéel of the cár ≫ the cár's whéel

See also Shih & Zuraw 2017, Antilla et al 2010, Shih & Graffmiller 2011, Shih et al 2015, Ryan 2018, Benor & Levy 2006, Gunkel & Ryan 2011

Evidence III: Statistical modeling in large corpora

Breiss and Hayes (submitted) model the frequency of word bigrams in spoken and written texts.

Multinomial logistic regression shows that authors/speakers prefer bigrams that obey, e.g.:

- *CLASH (adjacent stresses)
- *IAMBICCLASH (iambically-stressed before stress, e.g. *maróon swéater*)
- *CCC (triple cluster)
- *HIATUS (adjacent vowels)

relative to what would be expected from the two words' unigram frequencies.

Goal: an integrated model, deploying phonological and syntactic constraints in parallel

Much thinking is still needed about what the detailed architecture might be — let's try something concrete.

Intended advantages:

- Better testing of phonology: a full model will better evaluate our hypothesis that phonological markedness affects sentence probability.
- Better testing of syntax: a full model will let syntacticians control for phonological effects.

What qualities should such a model have?

- It should assign **probabilities** to sentences, depending on both syntactic and phonological principles.
- We should be able to fit its parameters to data for quantitative testing.

A tree-building model: Probabilistic left-corner grammars

Manning & Carpenter (1997) define probabilistic left-corner grammars (PLCGs):

- A tree determines a unique sequence of generative actions that construct it left-to-right.
- The probability of a tree is the product of the probabilities of the actions that construct it, each conditioned upon some surrounding syntactic context.

SHIFT: D → the
P(SHIFT: D → the | $\overline{S}$, $\overline{S}$)
P(SHIFT: D → the | $\overline{S}$, $\overline{S}$, ∅)

PROJECT: NP → D N
P(PROJECT: NP → D N | D, $\overline{S}$)
P(PROJECT: NP → D N | D, $\overline{S}$, the)

SHIFT: N → shoe
P(SHIFT: N → shoe | NP, $\overline{S}$)
P(SHIFT: N → shoe | NP, $\overline{S}$, the)

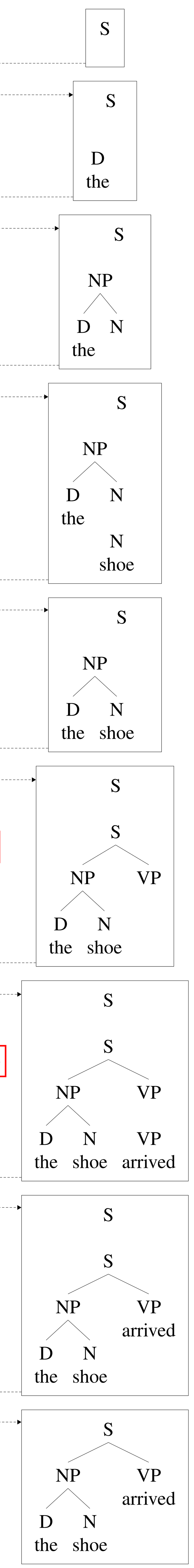
ATTACH
P(ATTACH | N, $\overline{N}$)
P(ATTACH | N, $\overline{N}$, shoe)

PROJECT: S → NP VP
P(PROJECT: S → NP VP | NP, $\overline{S}$)
P(PROJECT: S → NP VP | NP, $\overline{S}$, shoe)

SHIFT: VP → arrived
P(SHIFT: VP → arrived | $\overline{VP}$, $\overline{VP}$)
P(SHIFT: VP → arrived | $\overline{VP}$, $\overline{VP}$, shoe)

ATTACH
P(ATTACH | VP, $\overline{VP}$)
P(ATTACH | VP, $\overline{VP}$, arrived)

ATTACH
P(ATTACH | S, $\overline{S}$)
P(ATTACH | S, $\overline{S}$, arrived)



Our integrated probability model

We include the most recent word in the conditioning context, and use this to define the conditional probabilities of the PLCG's actions to be sensitive to not only

- the syntactic goodness of the hierarchical structure being built, but also
- the phonological goodness of any word junctures being created.

	Action	Context: top	Context: goal	Context: word	Property vector: Standard PLCG parameters	Property vector: *HIATUS
→	SHIFT: D → the	$\overline{S}$	$\overline{S}$	∅	1 0 0 0 0 0 0 0 ···	0
→	PROJECT: NP → D N	D	$\overline{S}$	the	0 1 0 0 0 0 0 0 ···	0
→	SHIFT: N → shoe	NP	$\overline{S}$	the	0 0 1 0 0 0 0 0 ···	0
→	ATTACH	N	$\overline{N}$	shoe	0 0 0 1 0 0 0 0 ···	0
→	PROJECT: S → NP VP	NP	$\overline{S}$	shoe	0 0 0 0 1 0 0 0 ···	0
→	SHIFT: VP → arrived	$\overline{VP}$	$\overline{VP}$	shoe	0 0 0 0 0 1 0 0 ···	1
	SHIFT: VP → arrived	$\overline{VP}$	$\overline{VP}$	toe	0 0 0 0 0 1 0 0 ···	1
	SHIFT: VP → arrived	$\overline{VP}$	$\overline{VP}$	boot	0 0 0 0 0 1 0 0 ···	0
	SHIFT: VP → arrived	$\overline{VP}$	$\overline{VP}$	sock	0 0 0 0 0 1 0 0 ···	0
→	ATTACH	VP	$\overline{VP}$	arrived	0 0 0 0 0 0 1 0 ···	0
→	ATTACH	S	$\overline{S}$	arrived	0 0 0 0 0 0 0 1 ···	0

$$harmony(action, context) = \sum_i \left(feature_i(action, context) \times weight_i\right)$$

$$\Pr(action \mid context) = \frac{\exp(harmony(action, context))}{\sum_a \exp(harmony(a, context))}$$

$$\Pr(\textsc{shift}: \mathrm{VP} \to \mathrm{arrived} \mid \overline{\mathrm{VP}}, \overline{\mathrm{VP}}, \mathrm{shoe}) = \frac{\exp(harmony(\textsc{shift}: \mathrm{VP} \to \mathrm{arrived}, (\overline{\mathrm{VP}}, \overline{\mathrm{VP}}, \mathrm{shoe})))}{\sum_a \exp(harmony(a, (\overline{\mathrm{VP}}, \overline{\mathrm{VP}}, \mathrm{shoe})))}$$

Results

We test by setting parameters to maximize likelihood of observed data, and comparing the fit of

- a phonologically-unaware baseline PLCG, and
- our phonologically-aware model.

The results below are based on sections wsj00–wsj04 of the Penn Treebank (9648 sentences), but results on other five-section samples are very similar.

Given this corpus, a straightforwardly-induced PLCG baseline model has 52173 parameters; each phonologically-aware variant adds one to this, for the weight of the tested phonological constraint.

	Log-likelihood	Delta log-likelihood	χ^2 test
■ PLCG baseline	−1562286		
■ PLCG + *IAMBICCLASH	−1562252	34	$p = 1.63 \times 10^{-16}$
■ PLCG + *HIATUS	−1562159	127	$p = 3.49 \times 10^{-57}$
■ PLCG + artificial ASCII code calculation	−1562284	2	$p = 0.0455$

We compare this with the results of adopting a unigram model for sentences, versus analogous phonologically-aware variants:

	Log-likelihood	Delta log-likelihood	χ^2 test
Unigram baseline	−1761162		
Unigram + *IAMBICCLASH	−1760874	288	$p = 2.78 \times 10^{-127}$
Unigram + *HIATUS	−1760767	395	$p = 8.06 \times 10^{-174}$
Unigram + artificial ASCII code calculation	−1761123	39	$p = 1.03 \times 10^{-18}$

Conclusion

- There is work for phonological constraints to do, even when we incorporate a more realistic model of sentence probabilities than unigrams.
- Some of the work that would be attributed to phonology in the context of the unigram sentence model (e.g. delta of 288 for *IAMBICCLASH) get taken over by the tree-based PLCG (delta down to 34).
- We now have a working system that we can scale up and extend to investigate other constraints.