Rethinking the arbitrariness of the sign

Giovanni Cassani CIMeC

08/01/2020

Tilburg University

In a nutshell

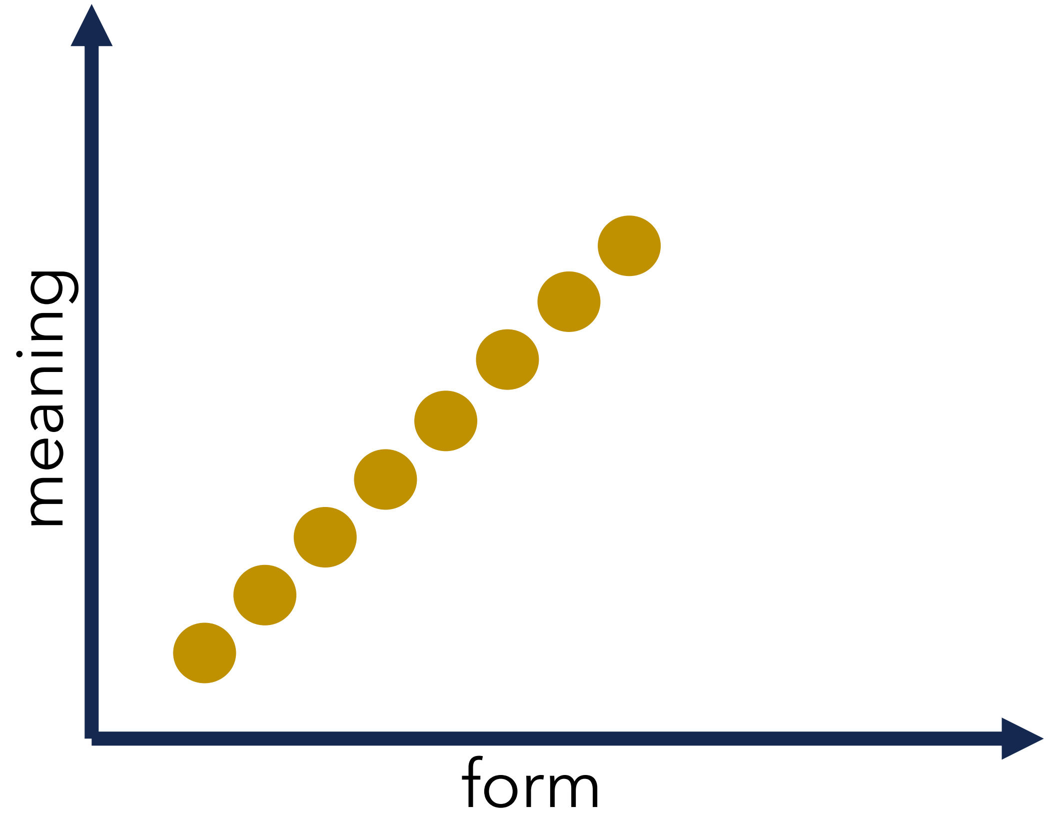
Starting point: if you know how a word sounds, you know nothing about what it means.

We find out about some stuff: sound symbolism, phonastaemes, iconicity, phonological bootstrapping

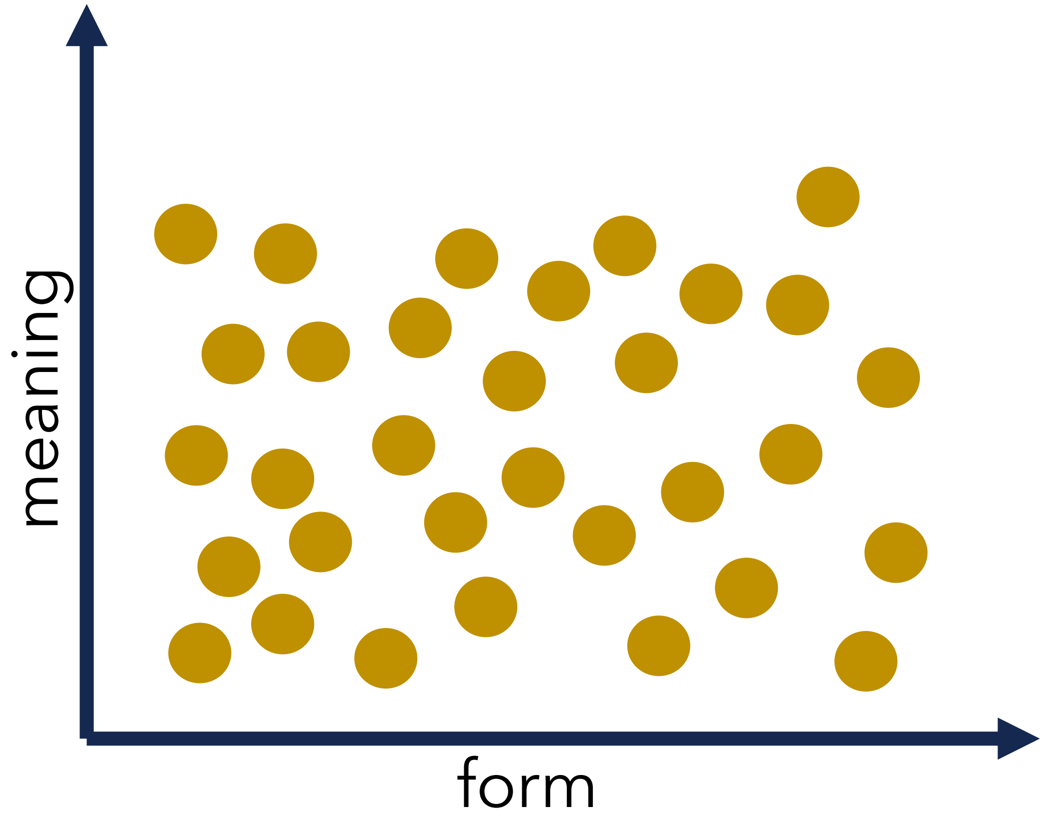
Now: some **systematicity in form-meaning mapping,** which affects acquisition and processing, maybe more.

By any other name?

A more efficient use of form



A more efficient use of form



Bad for learning, good for using

Arbitrariness helps processing by orthogonalizing words with similar contexts.

Arbitrariness makes learning a language hard by orthogonalizing words that refer to similar things or experiences.

But…

Languages are not entirely arbitrary: iconicity, onomatopoeias, phonaesthemes, statistical relationships between form and high-order grammatical functions, …

The bootstrapping hypothesis

Languages modulate the arbitrariness in the vocabulary according to when words are learnt: **early acquired words have more systematic form-meaning relations**.

This helps children recognize the referential nature of words and help them learn by exploiting similarities in the two domains.

Extended

All phenomena in which systematicity has been reported are somewhat peripheral.

What if it is actually possible to **guess the meaning of a word from its form only**? How pervasive is this process? What does it affect?

The start of this all

The semantics in non-words

If we can map form to meaning, then we should be able to infer the likely meaning of non-words.

How to test it?

Fitneva et al. 2009

Create non-words which phonologically resemble nouns and verbs, present them to children and ask them to pick a referent (either object or action).

Original claim: children can use form to infer grammatical function.
But action/entity is a primarily semantic distinction!

The approach

1. Extract word embeddings from corpora.
2. Create form vectors using word phonology.
3. Learn a mapping from form to meaning.
4. Apply this mapping function to non-words' form and get estimated semantic embeddings.
5. Classify the estimated embeddings as either action or entity.

An easy approach

Simple linear mappings:

$$F_w M_w = S$$
$$M = S_w F_w^{-1}$$

Use M to generate S_{nw} from F_{nw}

Classification strategies

Look at the semantic neighborhood of S_{nw} and count how many nouns, adjectives, verbs and so on occur in the vicinity.

Correlate S_{nw} with the embeddings of morpho-syntactic functions like plural, past or –ful.

Correlate S_{nw} with the embeddings of 20 early acquired words (10 nouns, 10 verbs)

Behavioral insights

Use LDA to classify S_{nw} on the basis of the three strategies.

Correlate LDA weights with the phonological typicality of the stimuli and with the proportion of children who picked the action referent.

Outcome

Classification accuracy is always significantly better than chance.

Correlations with phonological typicality and behavioral choices are significantly different from 0.

The source which brings the least information is the lexical category distribution over the semantic neighbors.

LDA

Source	Acc	p value	Errors	C	P	Wilks' λ	df	χ^2	p value
neighbors	15/16	< 0.001	sig	V	N	0.409	9	8.505	0.484
affixes	14/16	< 0.01	risp skik	N V	V N	0.281	12	10.156	0.602
anchors	16/16	< 0.001				0.335	13	8.194	0.831
neighbors + affixes	16/16	<0.001				0.045	14	21.661	0.086
neighbors + anchors	16/16	< 0.001				0.245	14	9.847	0.773
anchors + affixes	16/16	< 0.001				0.004	14	38.246	< 0.001
all	16/16	< 0.001				0.005	14	36.684	< 0.001

Correlations

Simulation 1	Simulation 2	corr 1	corr 2	Hotelling T	df	p
neighbors+ affixes	neighbors affixes	0.961	0.734	4.248	13	<0.001
		0.961	0.871	2.310	13	<0.05
neighbors+ anchors	neighbors anchors	0.861	0.734	1.314	13	0.106
		0.861	0.800	2.603	13	<0.05
anchors+ affixes	anchors affixes	0.981	0.800	5.650	13	<0.001
		0.981	0.871	4.071	13	<0.001
all	neighbors+ affixes	0.984	0.961	2.335	13	<0.05
	neighbors+ anchors	0.984	0.861	4.696	13	<0.001
	anchors+ affixes	0.984	0.981	0.380	13	0.355

Upshot

It is possible to **guess the meaning of a non-word based on its form** as far as the action/entity distinction is concerned.

The estimated meanings enter informative relations with the semantic neighborhood, the semantics of morpho-syntactic functions and early acquired words.

These relations reflect phonological typicality and children performance.

How easy is it to guess?

Beyond Whorf

The interest is in **statistical patterns which relate form to meaning**, not in finding out what a specific n-gram "means".

We don't dispute that a rose could be called by any other name. But once that thing is called *rose*, the degrees of freedom of the whole system decrease by more than one mapping becoming unavailable.

If it works for non-words, maybe…

It also works for words, and we **automatically** infer something about the meaning of a word from its form.

Maybe this process affects how early we learn words, how fast we process them, and so on…

The backbone

1. Get embeddings.
2. Get binary form vectors (1 if bigram present, 0 otherwise)
3. Learn a form-to-meaning mapping on all words but one.
4. Use the mapping to predict the meaning of the held out word.
5. **Compute cosine between estimated and observed embedding.**
6. Repeat for all words.

An array of hypotheses

Words for which *cosine(observed,estimated)* is high are predicted to be:

- acquired earlier
- processed faster
- more resistant to language change

Resources

- BLP: lexical decision RTs for visually presented words
- MALD: lexical decision RTs for aurally presented words
- AoA norms for 30K words (Kuperman et al 2012)
- Embeddings from Mandera et al 2017, already validated on psycholinguistics tasks, extracted from Subtlex+UkWaC.
- Morpholex
- CELEX

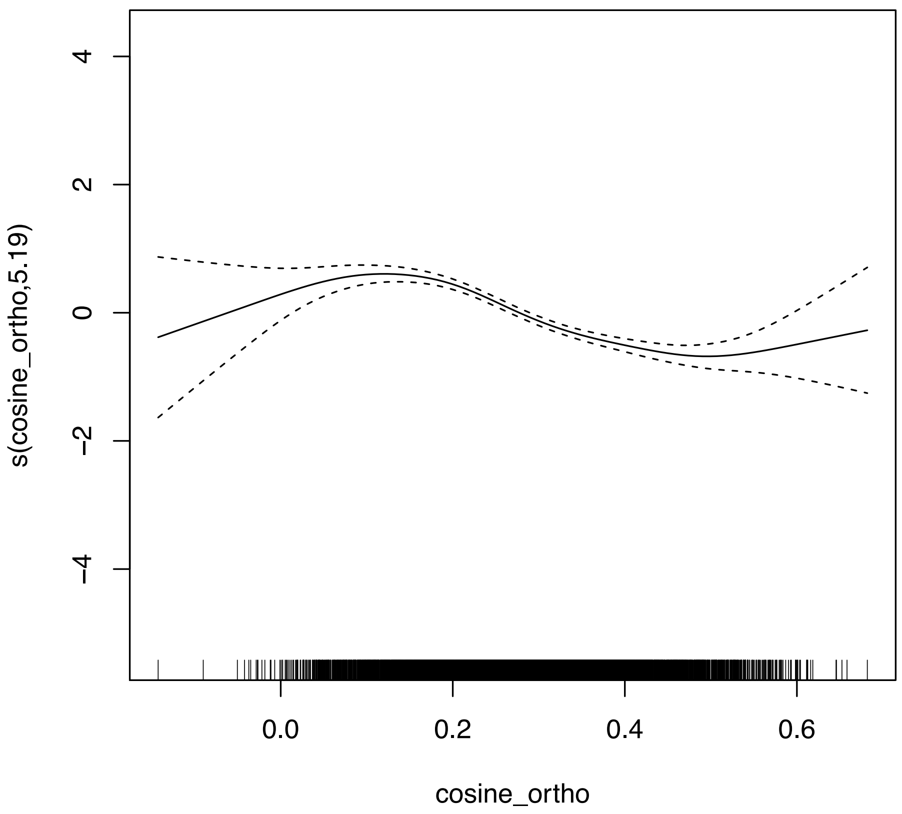
Make it hard for us

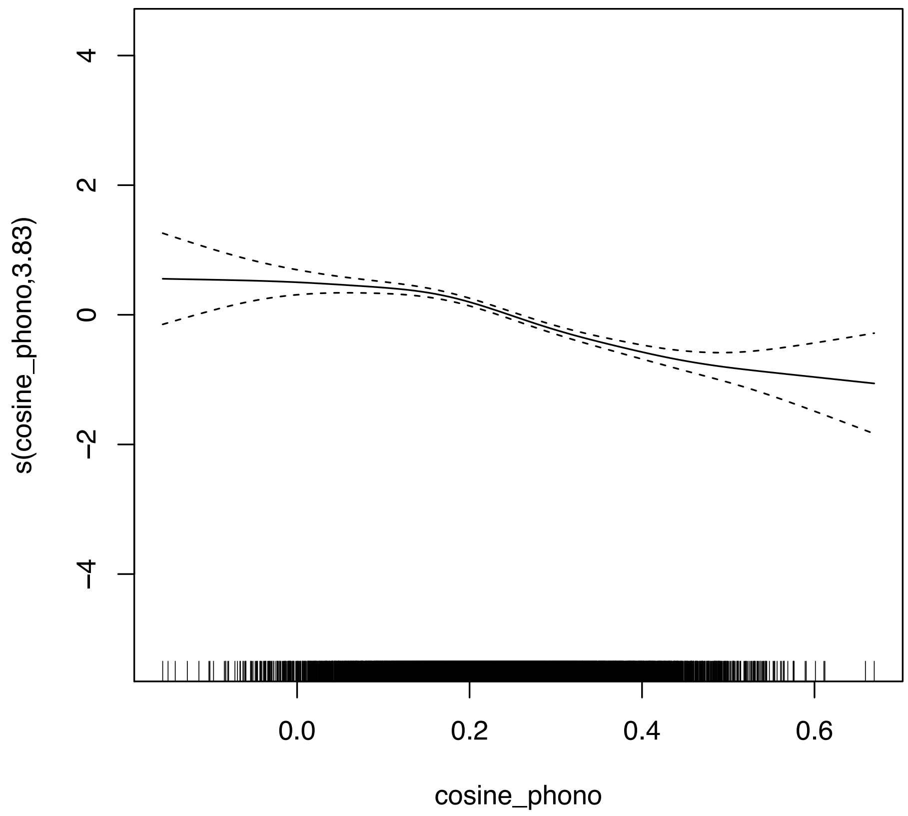
Focus on mono-morphemic words! Otherwise it could be just morpheme composition, which is very much systematic.

Predict AoA using GAMs

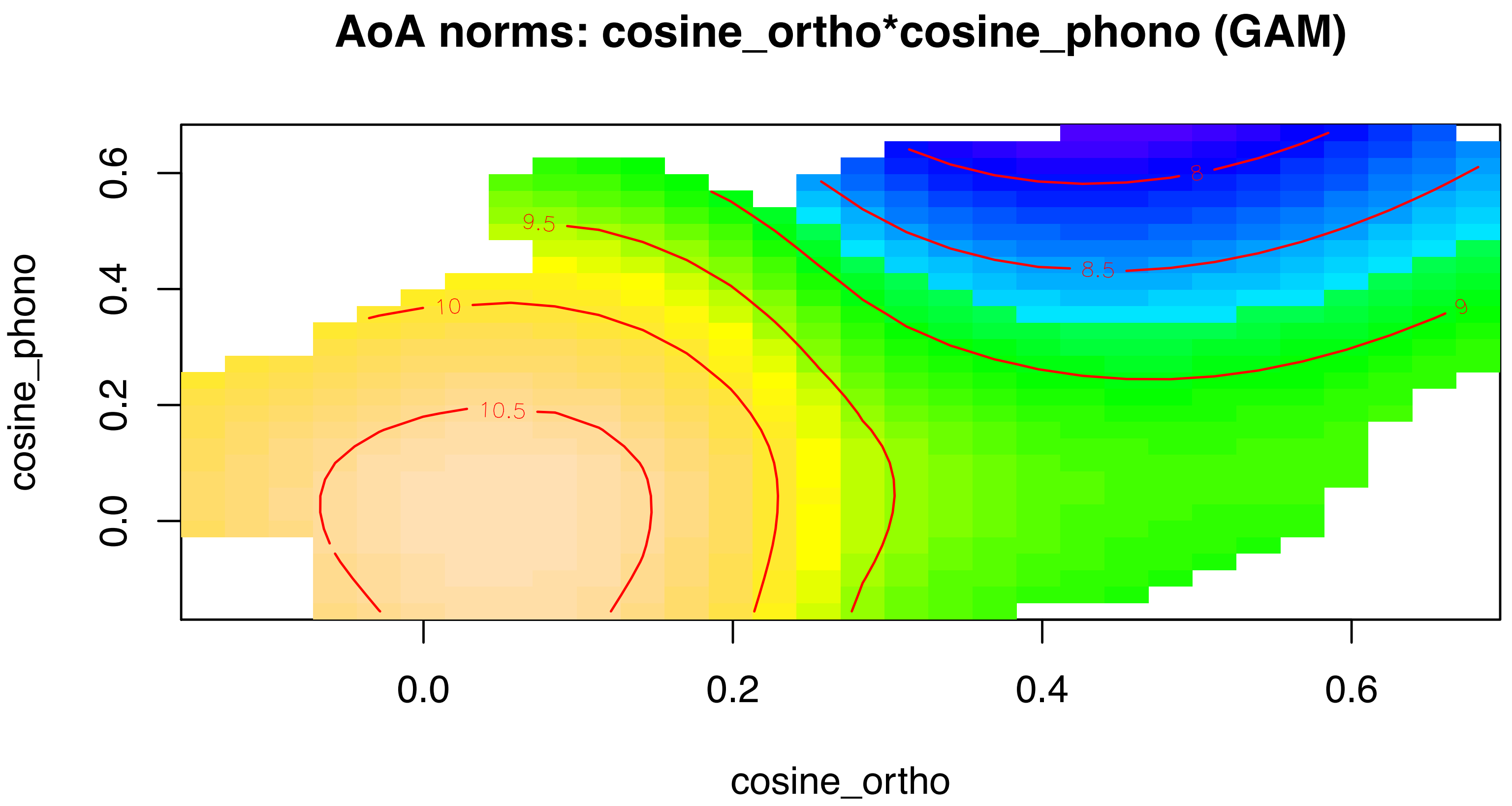
Baseline model: log(frequency), nletters, OLD20
Add cosine measures separately due to collinearity issues
Fit significant 2-way tensor products

AoA: orthographic v phonological form

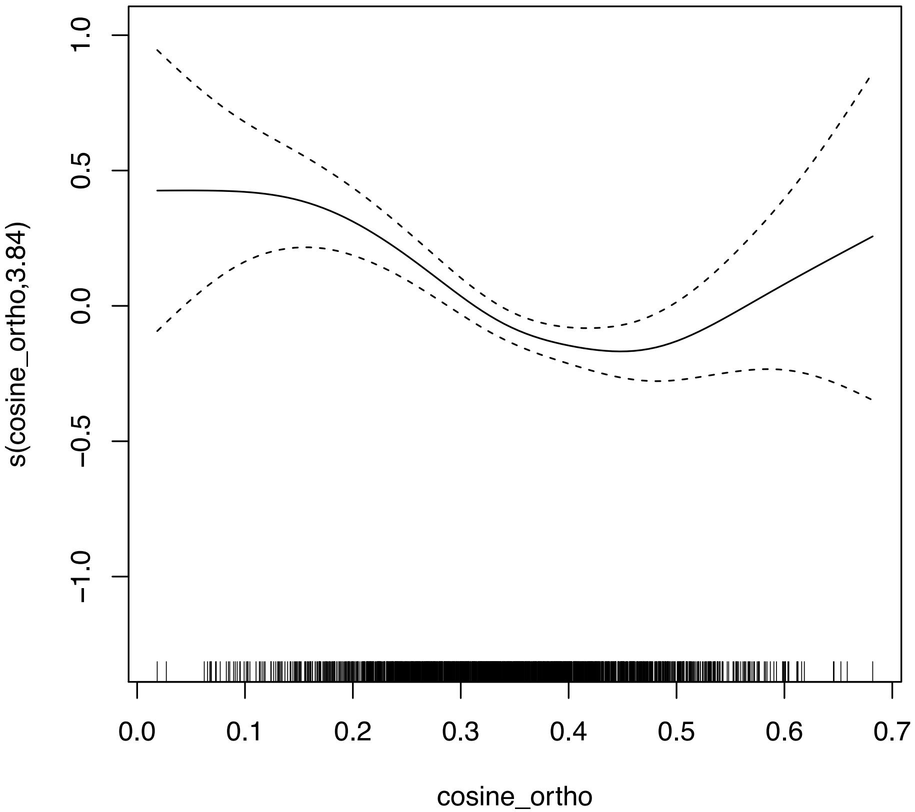


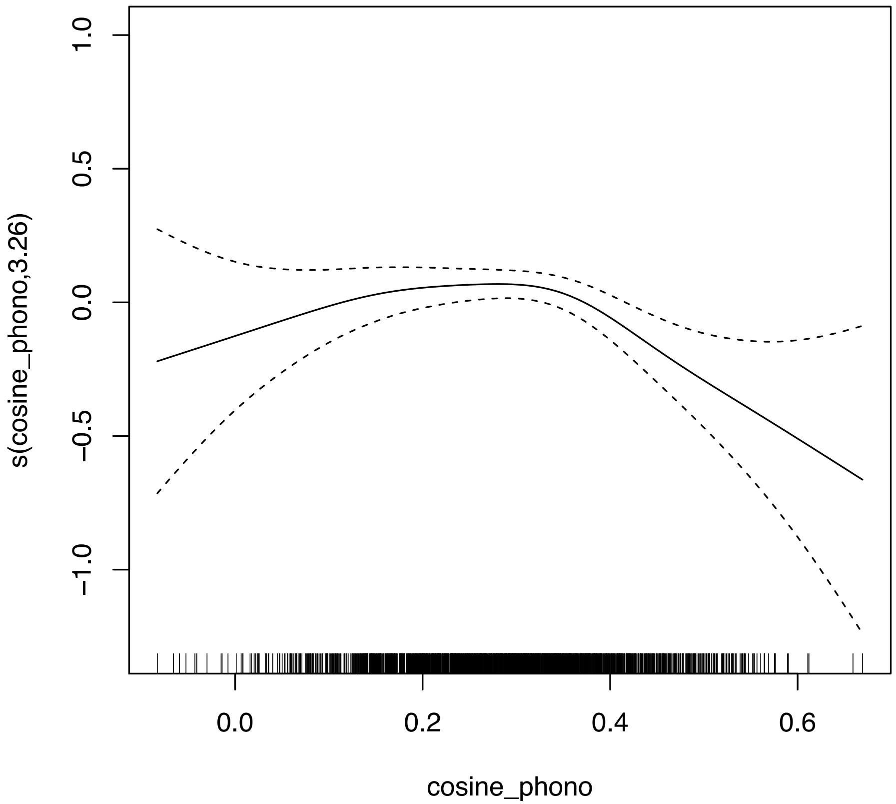


AoA: orthographic*phonological



Zoom on early acquired words



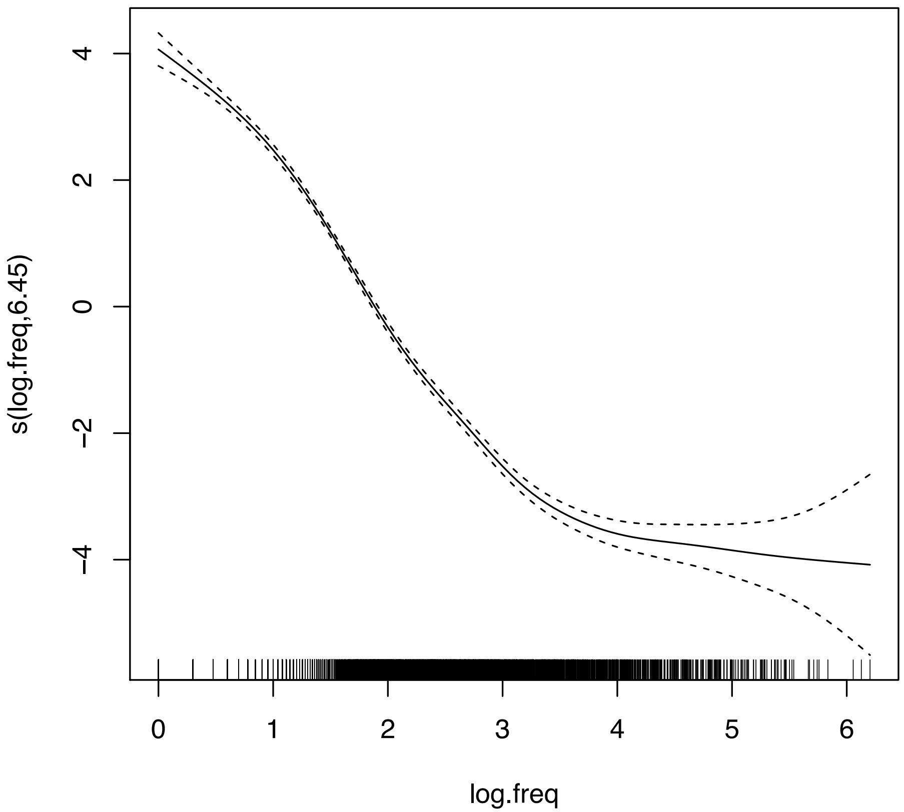


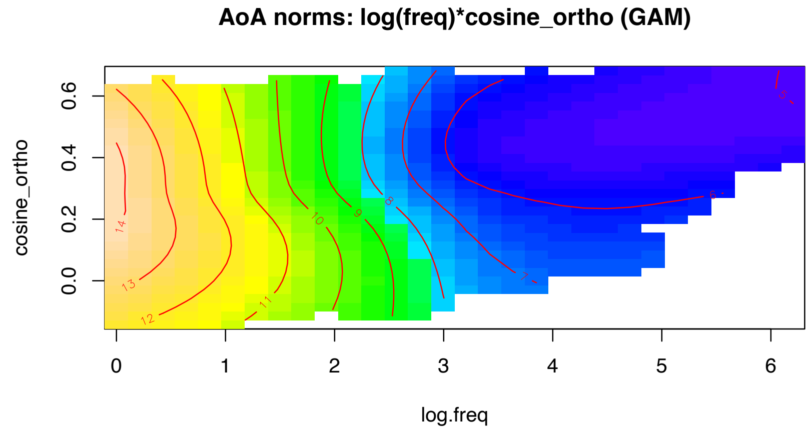
Upshots

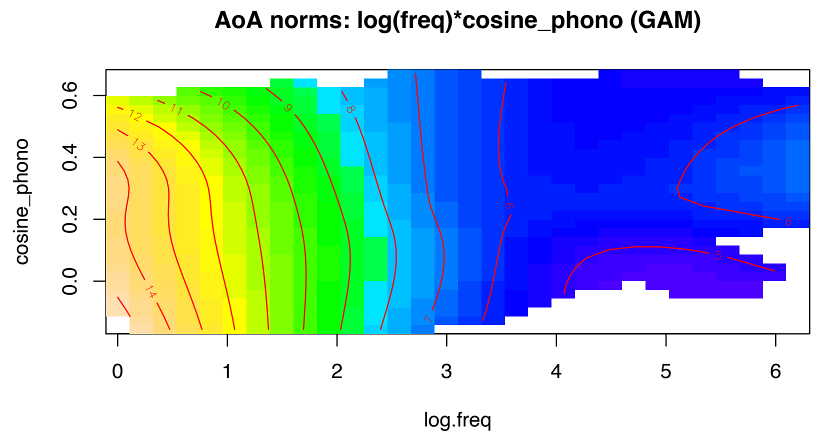
Cosine from orthography has a lower AIC, but its effect is sinusoidal. Cosine from phonology has a higher AIC but a consistent **facilitatory effect**.

The tensor product shows that for early acquired words, cosine from phonology dominates over cosine from orthography.

Frequency tho







What does it mean?

We don't know for sure.

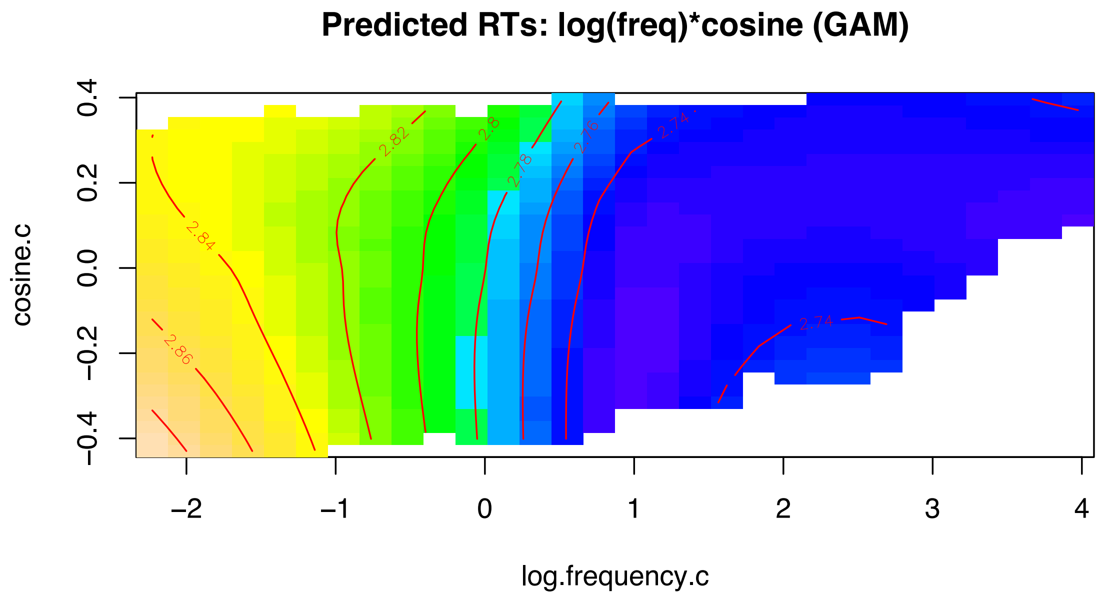
But the analysis suggests that it is indeed the case that words whose phonological form hints more to their meaning are learnt earlier, in line with the idea that **systematicy is beneficial to learning** and that it is found in natural languages (English at least).

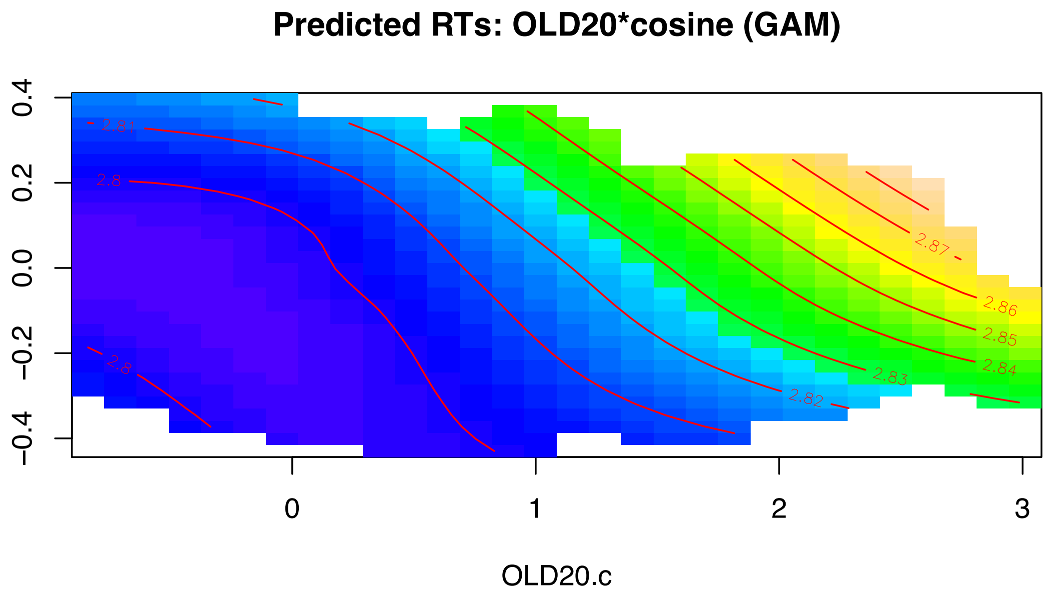
Predict LD RTs using LMEMs

Baseline model: log(frequency), nletters, OLD20
Add cosine measure
Test interaction between covariates and cosine

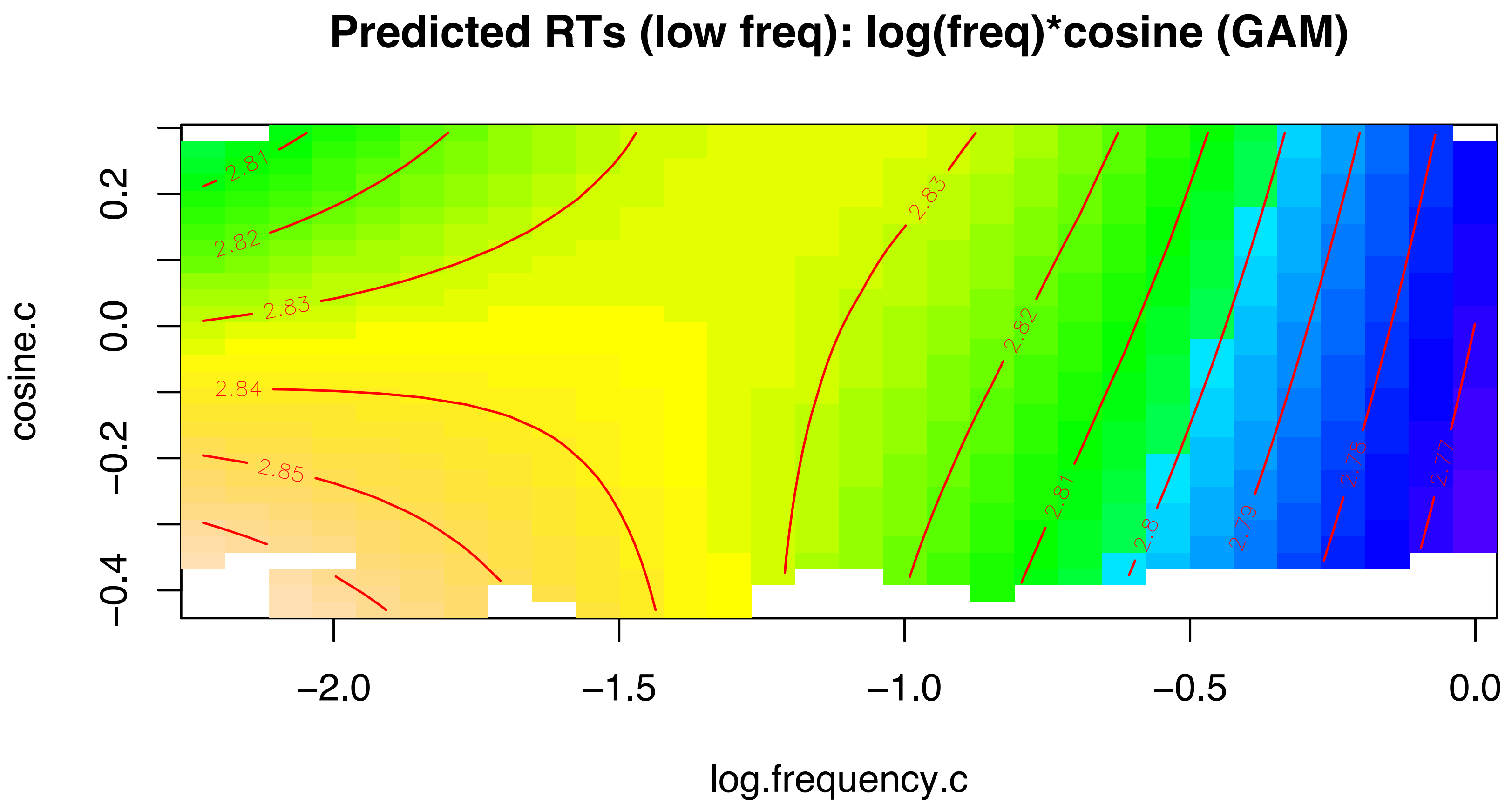
Cosine explains extra variance in RTs and enters a significant interaction with frequency and OLD20.

Visualizing effects using GAMs





Less entrenched



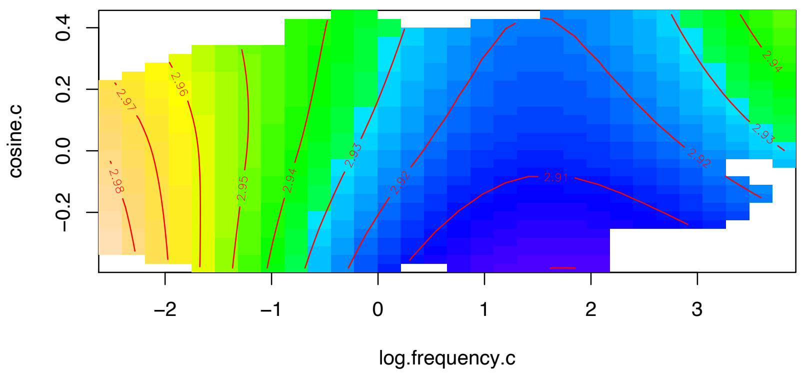
Same stuff for RTs on auditory LD

Baseline model: log(frequency), duration, Phonological Levensthein distance

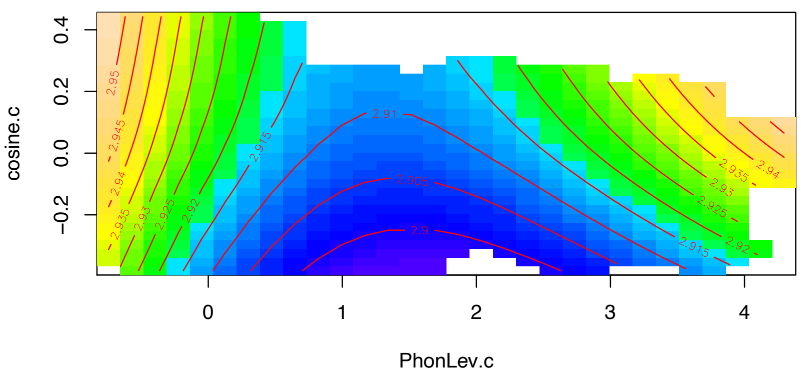
Cosine doesn't explain any extra variance, duration has the lion's share.

Weird stuff tho

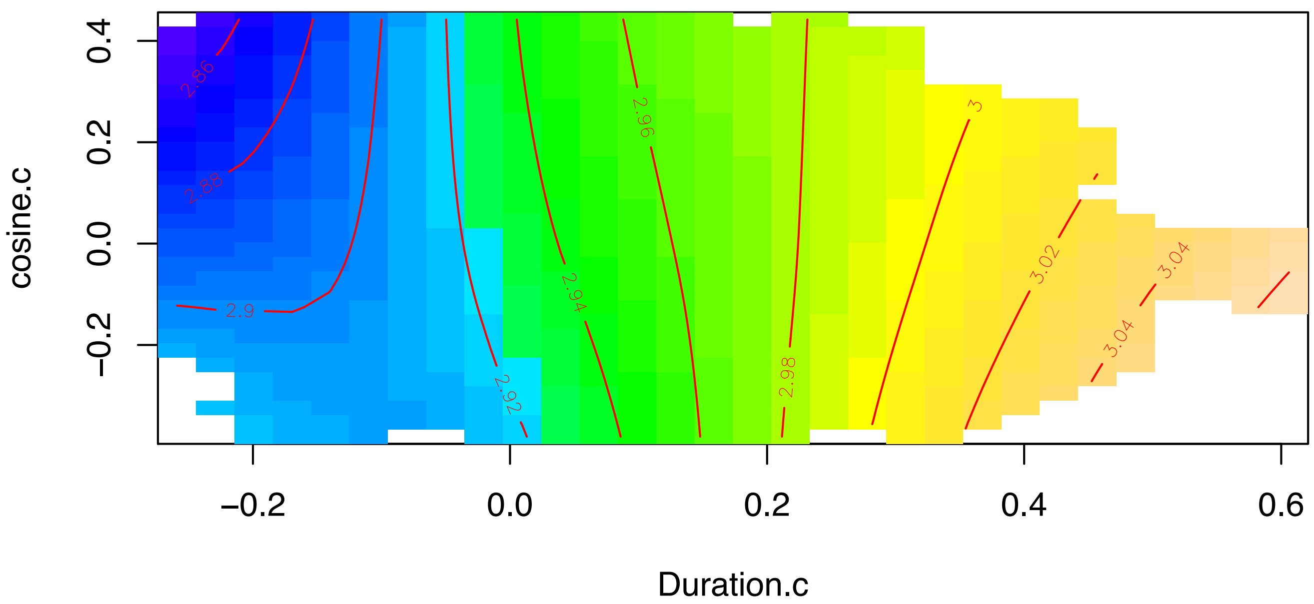
Predicted RTs: log(freq)*cosine (GAM)



Predicted RTs: PhonLev*cosine (GAM)



Predicted RTs: log(freq)*cosine (GAM)

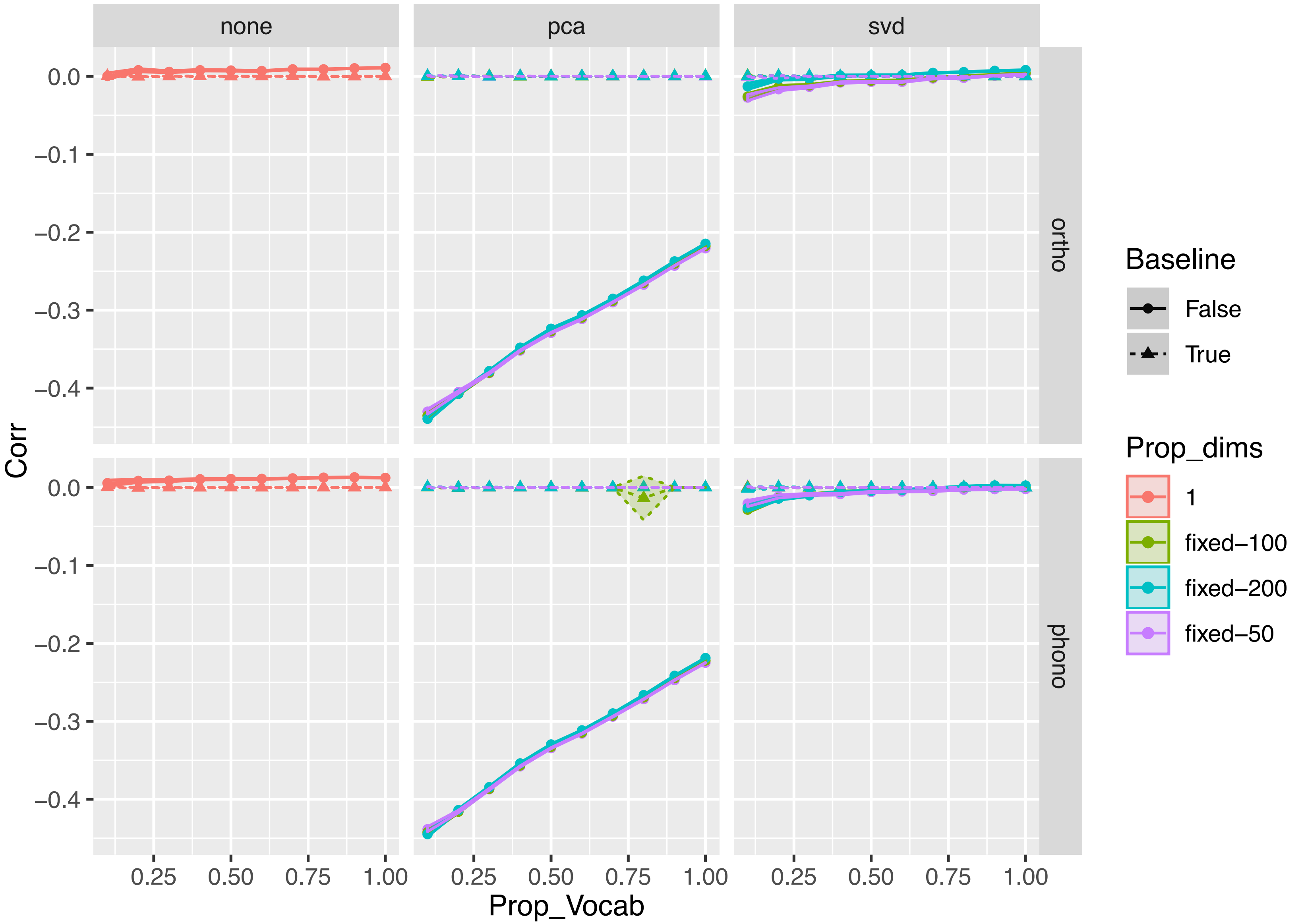


The uniqueness point

We need to run everything again estimating frequency, phonological levensthein, duration and cosine only up to the uniqueness point in words, since people make decisions then!

How similar?

RSA



none
pca
svd
ortho
phono
Corr
Prop_Vocab
0.0
−0.1
−0.2
−0.3
−0.4
0.25
0.50
0.75
1.00
Baseline
False
True
Prop_dims
1
fixed−100
fixed−200
fixed−50

Form and frequency

With binary vectors, correlation is statistically different from 0 but very low – no practical significance.

If form vectors are built putting word frequencies in cells rather than 1/0s that happens.

Robustness checks

I still have to run analysis with frequency sensitive form vectors reduced with PCA: with plain vectors, effects are consistent with binary vectors but messier and smaller.

But the RSA correlation is also similar to that of binary vectors so no big surprises there.

Next?

A number of tests

Priming at different SOAs (how automatic is this process?)
Use cosine(observed,predicted) to predict language change.
Include a covariate quantifying semantic neighborhood density.
Explore different ways to represent form.
Explore different mapping functions.
…

More interesting stuff

Disentangle the contribution of semantics inferred **from form v. from context** and their possible interaction with development and vocabulary size.

Check whether inferred semantics can help **counter referential uncertainty** in cross-situational mapping.