Relative entropy, and naive discriminative learning

Harald Baayen

in collaboration with

Petar Milin, Peter Hendrix, Dusica Filipovic-Markovic, and Marco Marelli

San Diego, January 15–16, 2011

overview

- Milin, Filipovic-Durdevic & Moscoso del Prado (2009)
- Experiment 1: replication with primed self-paced reading
- Modeling with naive discriminative learning
- Experiment 2: relative entropy in syntax (lex. dec.)
- Experiment 3: relative entropy in syntax (eye-tracking)
- Relative entropy, random intercepts, and stem support

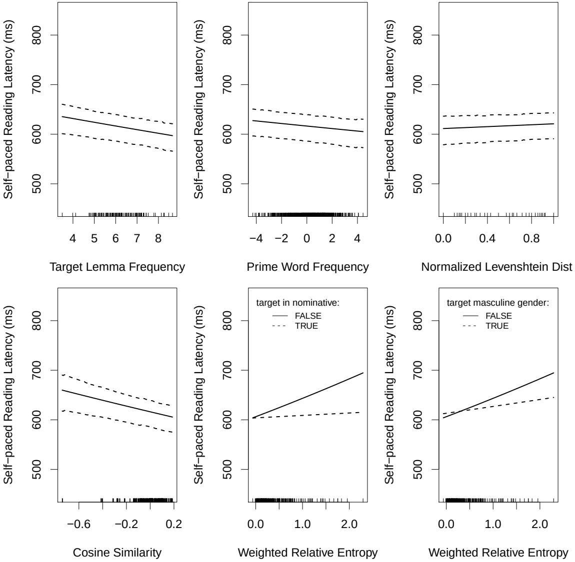
Milin et al. 2009

- $\{p\}$: the probability distribution of exponents of a given lemma
- $\{q\}$: the probability distribution of exponents across all lemmata in an inflectional class
- relative entropy $\mathrm{RE} = \sum_i p_i \log_2(p_i/q_i)$
- **greater relative entropy, longer lexical decision latencies**

Replication study using primed self-paced reading

- weighted relative entropy: $\sum_i \frac{p_i w_i}{\sum_i p_i w_i} log_2 \frac{p_i}{q_i}$
- weights $w_i = \frac{f(target_i)}{f(prime_i)}$
- **a greater WRE predicts longer latencies**
- but interactions with masculine gender and nominative case

Interactions with weighted relative entropy

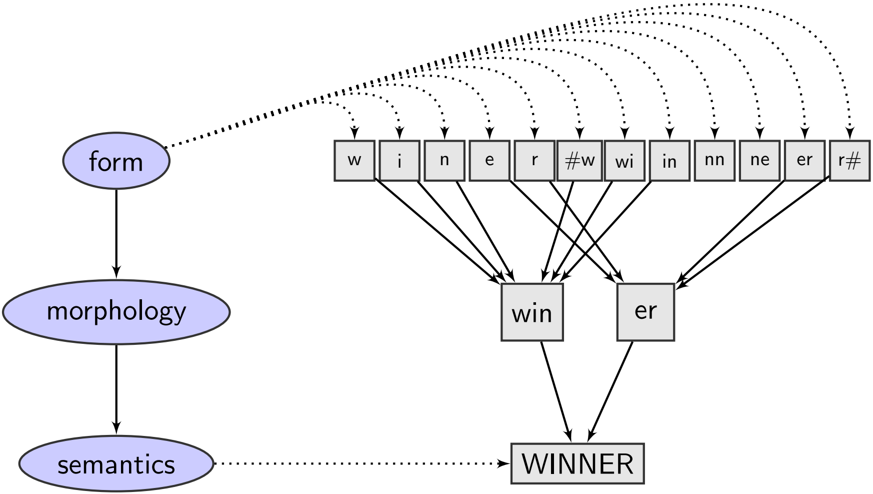


Modeling (weighted) relative entropy effects

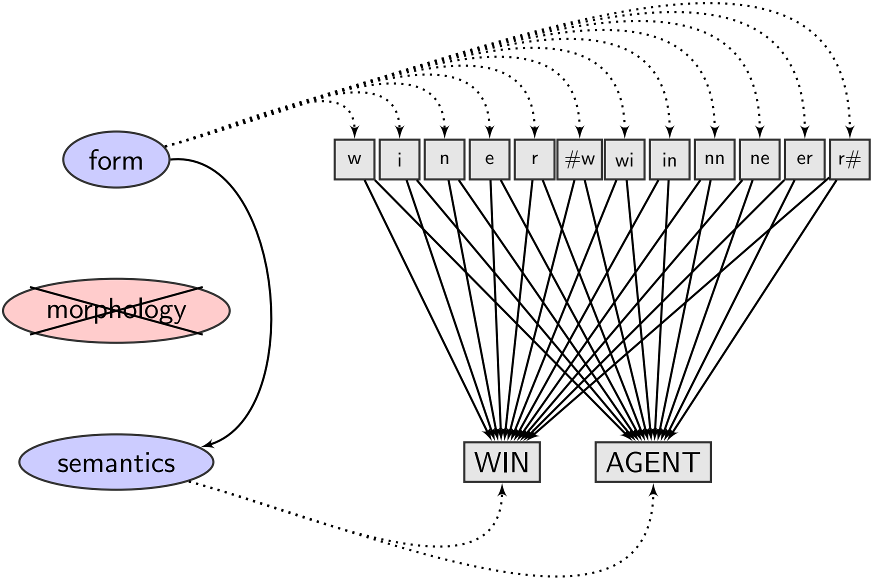
sources of inspiration

- recent work by Michael Ramscar on the Rescorla-Wagner equations in language acquisition
- old work by Fermin Moscoso del Prado Martin (PhD thesis, chapter 10)
- discussions with Jim Blevins

Models of morphological processing: the ‘standard’ model (Rastle, Davis)



Our approach: amorphous morphology



orthographic cues

- letters and letter pairs as cues for meanings
- legal scrabble words beginning with **qa**

orthographic cues

- letters and letter pairs as cues for meanings
- legal scrabble words beginning with **qa**
  - **qaid** (Muslim tribal chief)

orthographic cues

- letters and letter pairs as cues for meanings
- legal scrabble words beginning with **qa**
  - **qaid** (Muslim tribal chief)
  - **qanat** (gently sloping underground tunnel for irrigation)

orthographic cues

- letters and letter pairs as cues for meanings
- legal scrabble words beginning with **qa**
  - **qaid** (Muslim tribal chief)
  - **qanat** (gently sloping underground tunnel for irrigation)
  - **qat** (leaf of the shrub Catha edulis)

orthographic cues

- letters and letter pairs as cues for meanings
- legal scrabble words beginning with **qa**
  - **qaid** (Muslim tribal chief)
  - **qanat** (gently sloping underground tunnel for irrigation)
  - **qat** (leaf of the shrub Catha edulis)
- our model is based on a generalization of this idea

naive discriminative learning

- Links between orthography (cues) and semantics (outcomes) are established through **discriminative learning**
  - **Rescorla-Wagner equations** for discriminative learning (Rescorla & Wagner, 1972)
  - **Equilibrium equations** for the Rescorla-Wagner equations (Danks, 2003)
- The activation for a given meaning outcome is the sum of all associative links between the (active) input letters and letter pairs and that meaning

Rescorla-Wagner equations

$$V_i^{t+1} = V_i^t + \Delta V_i^t$$

with

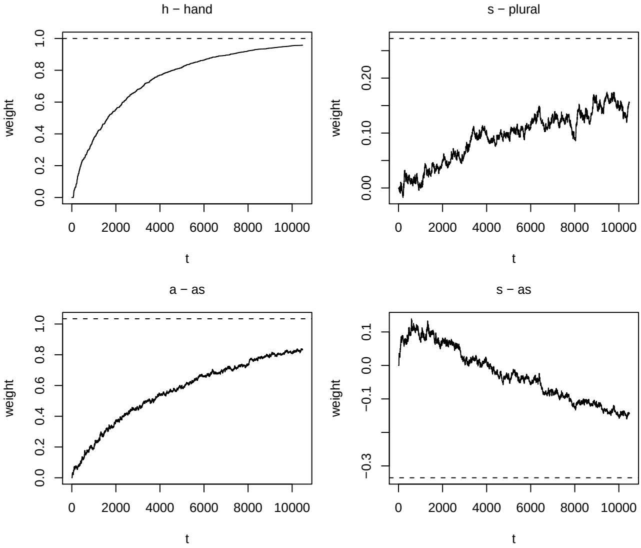
$$\Delta V_i^t = \begin{cases} 0 & \text{if } \textsc{absent}(C_i, t) \\ \alpha_i \beta_1 \left(\lambda - \sum_{\textsc{present}(C_j,\, t)} V_j \right) & \text{if } \textsc{present}(C_j, t) \;\&\; \textsc{present}(O, t) \\ \alpha_i \beta_2 \left(0 - \sum_{\textsc{present}(C_j,\, t)} V_j \right) & \text{if } \textsc{present}(C_j, t) \;\&\; \textsc{absent}(O, t) \end{cases}$$

- if a cue is **reliable**, it's connection strength will increase
- if a cue is unreliable, it's connection strength will decrease
- if many cues are relevant simultaneously, the contribution of a single cue from the set will be small

Example lexicon

Word	Frequency	Lexical Meaning	Number
hand	10	HAND	
hands	20	HAND	PLURAL
land	8	LAND	
lands	3	LAND	PLURAL
and	35	AND	
sad	18	SAD	
as	35	AS	
lad	102	LAD	
lads	54	LAD	PLURAL
lass	134	LASS	

The Rescorla-Wagner equations applied



a shortcut straight to the adult stable state

- **equilibrium equations (Danks)** when the system is in a stable state, the connection weights to a given meaning can be estimated by solving a set of linear equations

$$\begin{pmatrix} \Pr(C_0|C_0) & \Pr(C_1|C_0) & \dots & \Pr(C_n|C_0) \\ \Pr(C_0|C_1) & \Pr(C_1|C_1) & \dots & \Pr(C_n|C_1) \\ \dots & \dots & \dots & \dots \\ \Pr(C_0|C_n) & \Pr(C_1|C_n) & \dots & \Pr(C_n|C_n) \end{pmatrix} \begin{pmatrix} V_0 \\ V_1 \\ \dots \\ V_n \end{pmatrix} = \begin{pmatrix} \Pr(O|C_0) \\ \Pr(O|C_1) \\ \dots \\ \Pr(O|C_n) \end{pmatrix}.$$

  V_i: association strength of i-th cue C_i to outcome O

- **the association strengths V_j optimize the conditional outcomes given the conditional co-occurrence probabilities characterizing the input space**

from weights to meaning activations

- the activation a_i of meaning i is the sum of its incoming connection strengths

$$a_i = \sum_j V_{ji}$$

- the greater the meaning activation, the shorter the response latencies
  - simplest case:
    $\mathrm{RTsim}_i \propto -a_i$
  - a log transformation may be required to remove the right skew from the distribution of simulated RTs:
    $\mathrm{RTsim}_i \propto \log(1/a_i)$

the naive discriminative reader

- basic engine is parameter-free, and driven completely and only by the language input
- the model is computationally undemanding: building the weight matrix from a lexicon of 11 million phrases takes 10 minutes on my desktop
- implementation in R

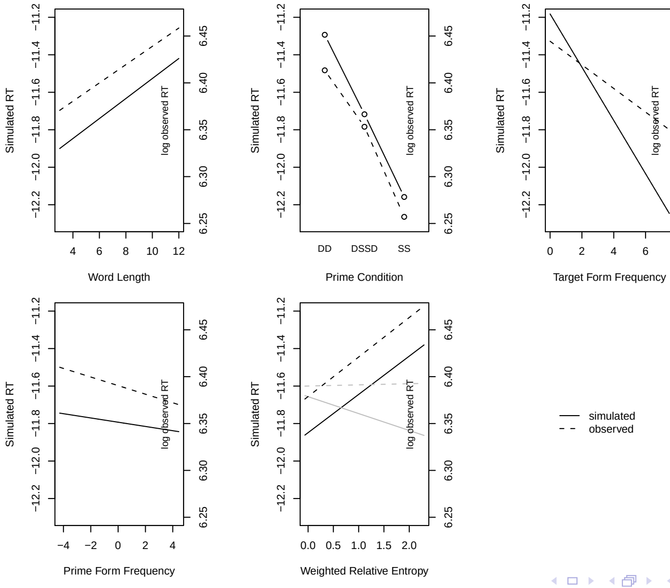
from weights to meaning activations

- for Serbian case-inflected nouns, sum over lexical meanings and grammatical meanings
- for priming, we use Ratcliff-McKoon's compound cue theory:

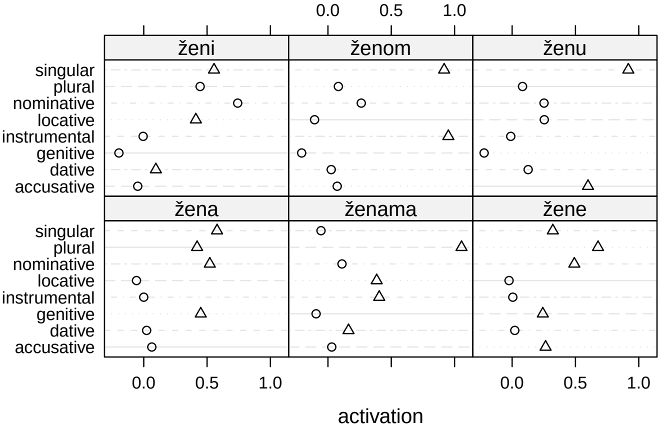
$$S = \sum_{i=1}^{10} (a_{Pi}^{w} \cdot a_{Ti}^{1-w}) \quad (0 \leq w \leq 0.5) \tag{1}$$

- this introduces a free parameter for the prime duration
- we also use one free parameter to model the time required to plan and execute a second fixation for longer words

Observed and simulated latencies (r = 0.24)



Activation of case meanings



Summary Experiment 1

- relative entropy effects persist in sentential reading
- they are modified, but not destroyed by priming
- the interaction with masculine gender follows from the distributional properties of the lexical input
- the interaction with nominative case remains unaccounted for (functions and meanings?)
- frequency effects for complex words and paradigmatic effects can arise without representations for complex words or representational structures for paradigms

Experiment 2: Relative entropy in syntax

phrase	phrasal frequency	phrasal probability	preposition	prepositional frequency
on a plant	28608	0.279	*on*	177908042
in a plant	52579	0.513	*in*	253850053
under a plant	7346	0.072	*under*	10746880
above a plant	0	0.000	*above*	2517797
through a plant	0	0.000	*through*	3632886
behind a plant	760	0.007	*behind*	3979162
into a plant	13289	0.130	*into*	25279478

40 spatial prepositions

prepositional relative entropy

training data

- the model is trained on 11,172,554 two and three-word phrases from the British National Corpus, comprising 26,441,155 word tokens
- phrases have as last word one of 24710 monomorphemic words, or any bimorphemic compounds, derived and inflected words containing one of the 24710 monomorphemic words

constructions sampled

PREPOSITION + ARTICLE + NOUN	*about a ballet*
PREPOSITION + POSSESSIVE PRON. + NOUN	*about her actions*
PREPOSITION + X + NOUN	*about actual costs*
PREPOSITION + NOUN	*about achievements*
X'S + NOUN	*protege's abilities*
ARTICLE + NOUN	*a box*
ARTICLE + X + NOUN	*the abdominal appendages*
POSSESSIVE PRONOUN + NOUN	*their abbots*
ARTICLE + X'S + NOUN	*the accountant's bill*
PRONOUN + AUXILIARY + VERB	*they are arrested*
PRONOUN + VERB	*he achieves*
AUXILIARY + VERB	*is abandoning*
ARTICLE + ADJECTIVE	*the acute*

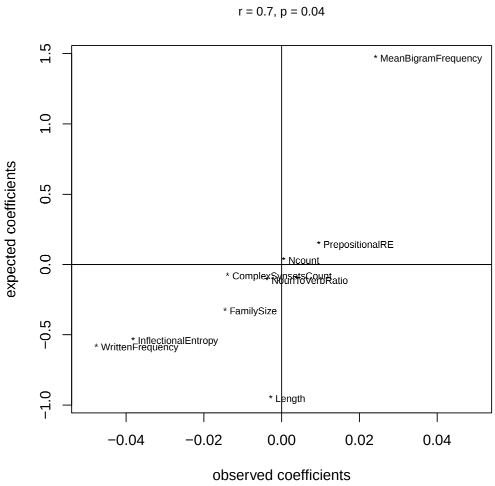
processing of monomorphemic words

- stimuli: 1289 monomorphemic nouns
- lexical decision latencies from the English Lexicon Project
- simulated lexical decision latencies

- predictors
  - Family Size
  - Inflectional Entropy
  - Written Frequency
  - Number of Morphologically Complex Synonyms
  - Neighborhood Density
  - Mean Bigram Frequency
  - Noun-Verb Ratio
  - Length
  - **Prepositional Relative Entropy**

results

correlation for the observed and simulated response latencies:
$r = 0.55$, $t(1287) = 23.83$, $p < 0.001$



Summary Experiment 2

- lexical paradigmatic effects (family size, inflectional entropy) modeled successfully without representations for inflections and derivations
- the phrasal paradigmatic effect is also modelled correctly, without representations for phrases
- the paradigmatic distributional properties of a word can affect single-noun reading

Other results obtained

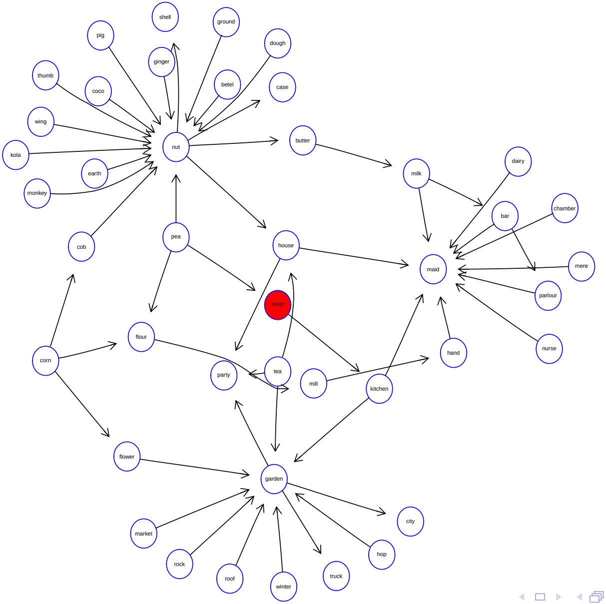
- phrasal frequency effects
- phonaestheme effects
- corn-corner effects (pseudoderived words)
- family size effects, whole-word frequency effects, and base frequency effects for complex words
- the interaction between first-constituent frequency and whole-word frequeny in compound words (Kuperman et al., 2009)
- interaction of regularity by tense in English

intermezzo: strong connectivity

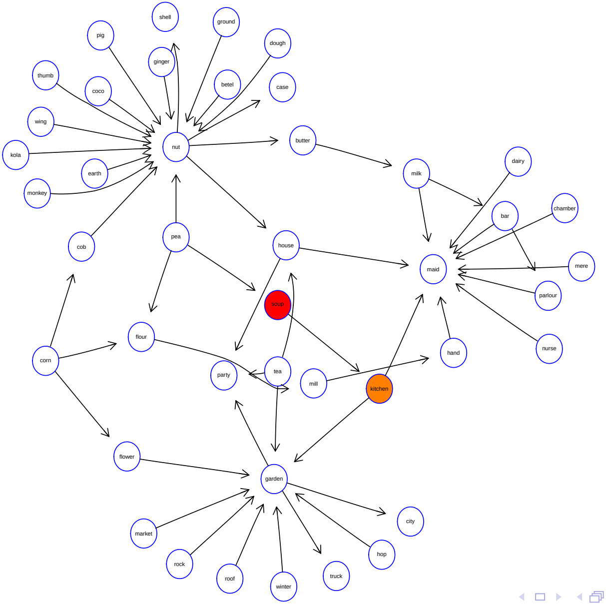
- **mediated priming** (Balota & Lorch, 1986)
  - **cat** → cab → **taxi**
  - **lion** → tiger → **stripes**

- **priming chains for compounds?**
  - **tea** trolley → trolley **bus**
  - **tea** trolley → trolley bus → bus **stop**

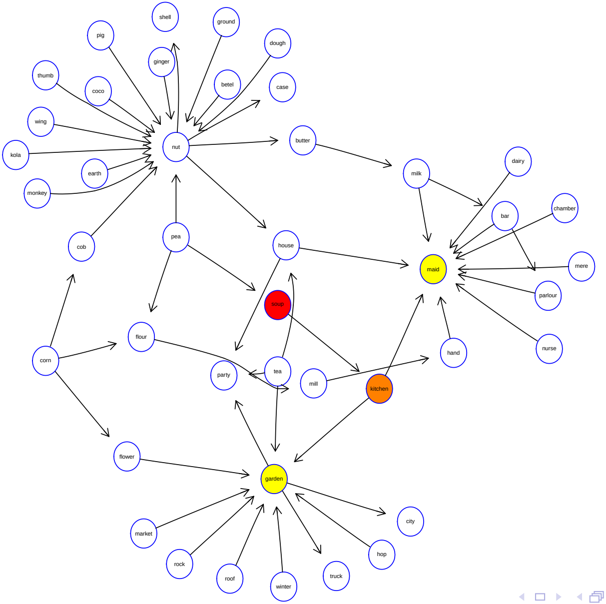
spreading activation: weak connectivity



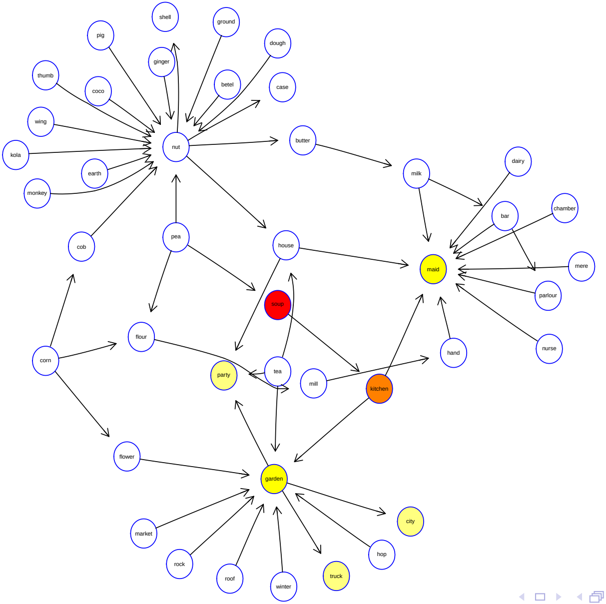
spreading activation: weak connectivity



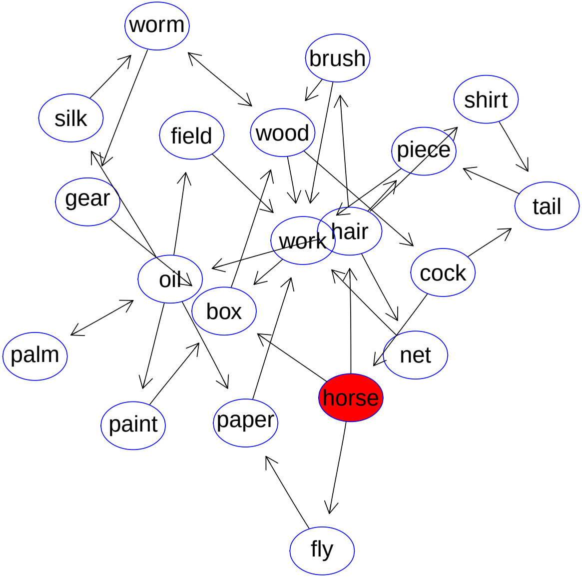
spreading activation: weak connectivity



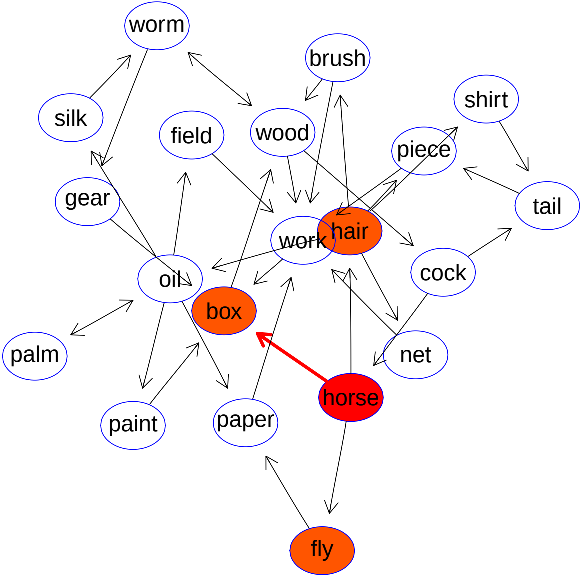
spreading activation: weak connectivity



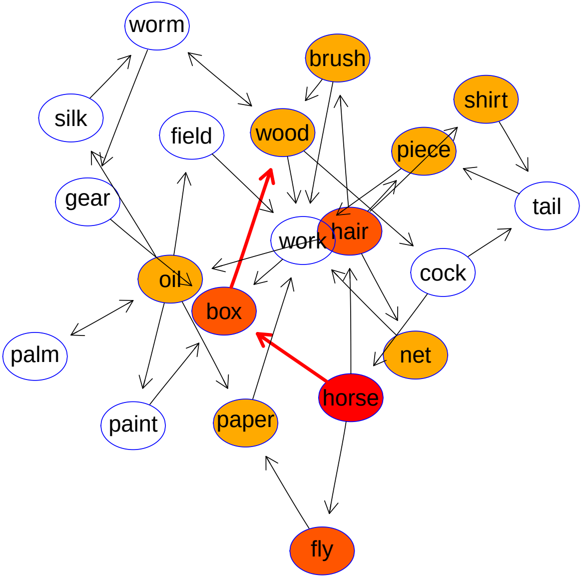
spreading activation: strong connectivity



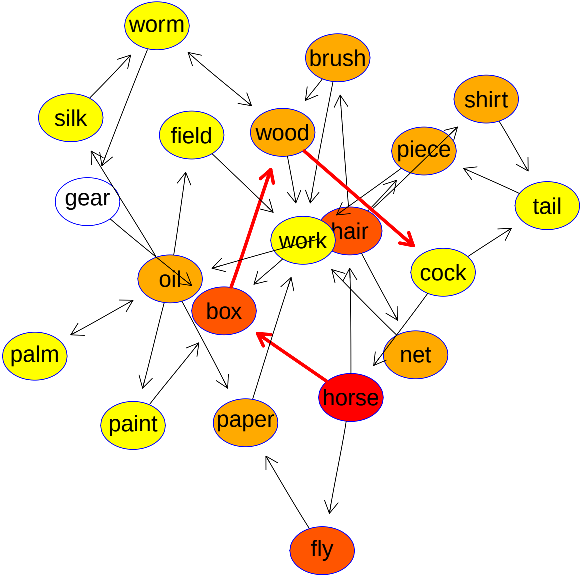
spreading activation: strong connectivity



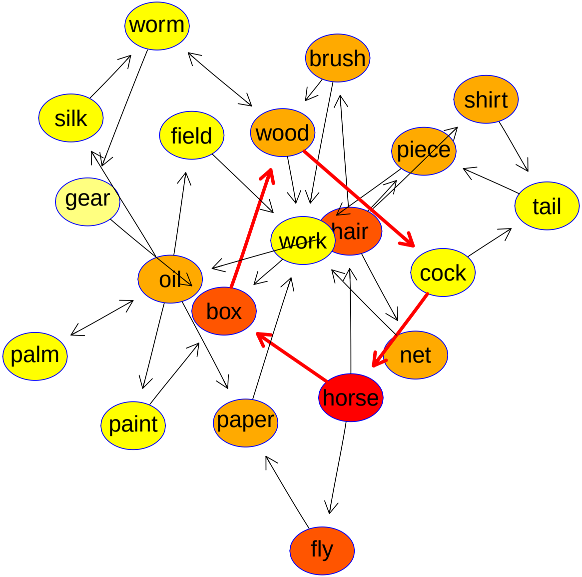
spreading activation: strong connectivity



spreading activation: strong connectivity



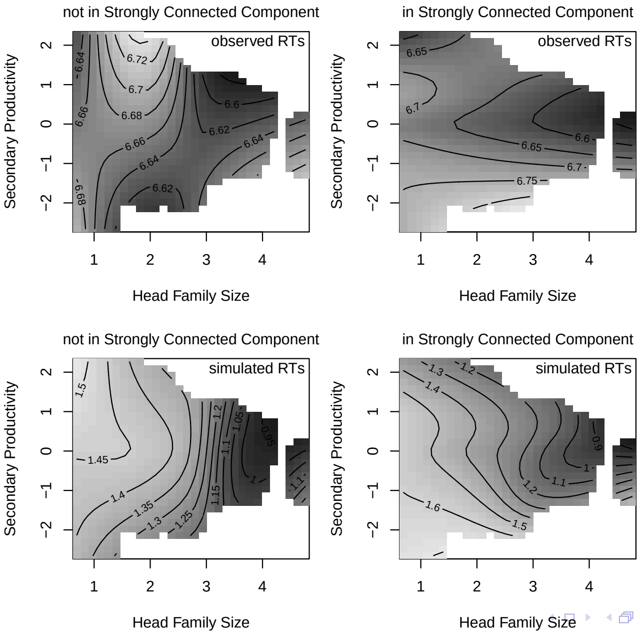
spreading activation: strong connectivity



is strong connectivity advantageous?

- is strong connectivity advantageous?
    - possibly yes — more integrated learning
    - possibly no — might cause confusion secondary family size
- this kind of connectivity should be beyond what the naive discriminative reader can handle — but it isn't

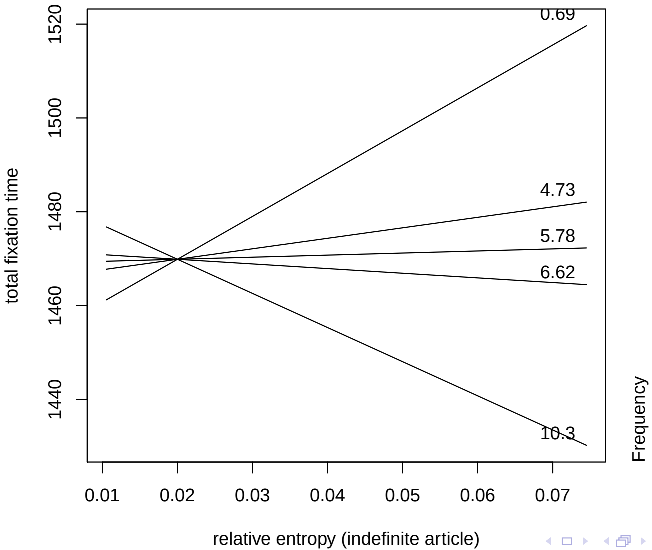
lexical connectivity



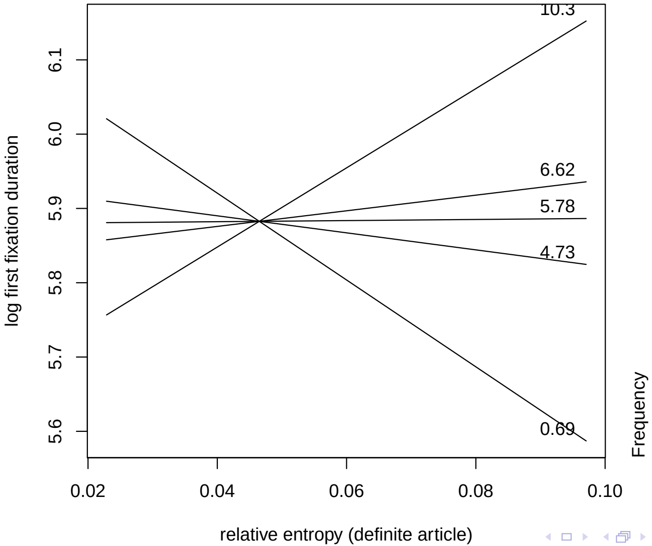
Experiment 3: More on relative entropy in syntax

- reading aloud combined with eye tracking
- first experiment: reading aloud single words
  (e.g., *table*)
- second experiment: reading aloud prepositional phrases
  (e.g., *on the* + *table*)

Experiment 3: single words, total fixation time



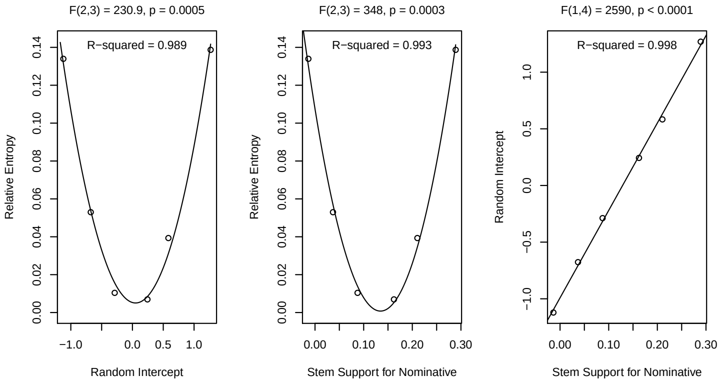
Experiment 3: phrases, total fixation time



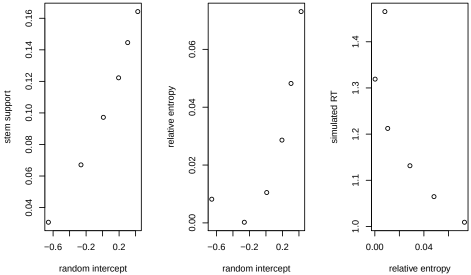
Naive discriminative and mixed-effects classifiers

Word Form	Frequency	Case	Lemma	Relative Entropy	Ranef	Stem Support Nominative	Stem Support Genitive	Exponent Support
AQEa	10	nom	A	0.134	-1.121	-0.014	0.260	0.353
AQEi	20	gen	A	0.134	-1.121	-0.014	0.260	0.740
AQEu	30	acc	A	0.134	-1.121	-0.014	0.260	0.595
AQEa	40	acc	A	0.134	-1.121	-0.014	0.260	0.127
ABCa	15	nom	B	0.053	-0.676	0.037	0.260	0.353
ABCi	22	gen	B	0.053	-0.676	0.037	0.260	0.740
ABCu	28	acc	B	0.053	-0.676	0.037	0.260	0.595
ABCa	35	acc	B	0.053	-0.676	0.037	0.260	0.127
APQa	20	nom	C	0.010	-0.288	0.087	0.260	0.353
APQi	24	gen	C	0.010	-0.288	0.087	0.260	0.740
APQu	26	acc	C	0.010	-0.288	0.087	0.260	0.595
APQa	30	acc	C	0.010	-0.288	0.087	0.260	0.127
ZPEa	30	nom	D	0.007	0.243	0.162	0.260	0.353
ZPEi	26	gen	D	0.007	0.243	0.162	0.260	0.740
ZPEu	24	acc	D	0.007	0.243	0.162	0.260	0.595
ZPEa	25	acc	D	0.007	0.243	0.162	0.260	0.127
EPBa	35	nom	E	0.039	0.583	0.210	0.260	0.353
EPBi	28	gen	E	0.039	0.583	0.210	0.260	0.740
EPBu	22	acc	E	0.039	0.583	0.210	0.260	0.595
EPBa	20	acc	E	0.039	0.583	0.210	0.260	0.127
DPBa	40	nom	F	0.139	1.269	0.289	0.260	0.353
DPBi	30	gen	F	0.139	1.269	0.289	0.260	0.740
DPBu	20	acc	F	0.139	1.269	0.289	0.260	0.595
DPBa	10	acc	F	0.139	1.269	0.289	0.260	0.127

stem support, random intercepts, and unsigned relative entropy

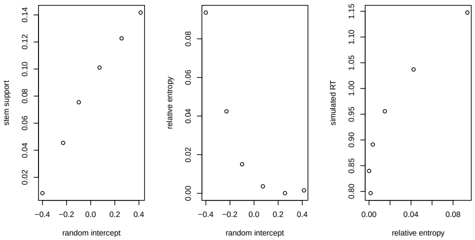


the main trend depends on the balance




```
c(10,20,30,40)*20,
c(15,24,32,40)*10,
c(20,28,33,40)*3,
c(25,32,35,40)*2,
c(30,34,37,40)*1,
c(35,37,38,40)*1
```

trend depends on position prototype




```
c(10,20,30,40)*1,
c(15,24,32,40)*1,
c(20,28,33,40)*2,
c(25,32,35,40)*3,
c(30,34,37,40)*10,
c(35,37,38,40)*20
```

trend depends on position prototype

- in a complex system, the same measure can have slopes with opposite signs depending on the distributional properties of the language input
- this may help explain the changes in sign of RE in the eye-tracking+naming study
- **our distributional measures provide partial and potentially distorting views on the complex structure arising from simple principles of learning**

Discussion

- Our model shows morphological effects in the absence of morphological representations, including paradigmatic effects
- This is consistent with a-morphous views on morphology (e.g.: Anderson, 1992; Blevins, 2003)
- The model is a classifier (for the dative alternation, it outperforms mixed models)
  - relative entropies are functionally equivalent to unsigned random intercepts in a mixed-effects model
  - relative entropies capture the total association strengths from stems to grammatical meanings

Discussion

- Our model is similar in spirit to the reading part of the triangle model (Seidenberg & Gonnermann, 2000)
- Both models map orthography onto semantics without morphological representations
- Our computational engine, however, is much simpler than that of the triangle model: we do not assume hidden layers or use back-propagation to estimate connection weights.
- Furthermore, our model is more radically a-morphous in that there is no hidden layer that can covertly represent morphology.

Discussion

- Our model is also similar in spirit to the Bayesian Reader (Norris, 2006)
- Both models map forms onto 'central' representations without intercession by morphemes
- Our computational engine, however, is much simpler than that of the Bayesian reader: the complexity of the Bayesian reader is quadratic in the number of orthographic 'units', whereas our model is linear in the number of elementary meanings

Summary

- Discriminative learning provides a good fit to a wide range of experimental data
- The model is trained on realistic input, it is as sparse as possible in its number of representations, and it is computationally efficient
- The model does not make an a priori distinction between phrasal learning and morphological learning, and therefore can straightforwardly handle gradient phenomena at the interface of morphology and syntax (cf. construction morphology, Booij 2010)