From nominal case in Serbian to prepositional phrases in English

Petar Milin

Department of Psychology, University of Novi Sad

Laboratory for Experimental Psychology, University of Belgrade

General background

- There exists huge diversity of how biological system cope with the environment
- Aristotle: human is ZOON POLITIKON (*ζωον πολίτίκον*)

We could add: ZOON PLIROFORIKON (*ζωον πληροφορίκον*)

General background

- Language is our sixth sense – extremely powerful input-output channel
- Language is complex adaptive system (CAS) The "Five Graces Group" (2009): Beckner, Ellis, Blythe, Holland, Bybee, Ke, Christiansen, Larsen-Freeman, Croft, and Schoenemann
- Information theory provides formal characterisations of parts of such a system

Historical overview

Information theory and lexical processing

- Amount of information

  (Kostić, 1991, 1995; Kostić et al., 2003 etc.)

$$
\begin{aligned}
I_e &= -\log_2 \Pr_\pi(e) \\
I'_e &= -\log_2 \left(\frac{\Pr_\pi(e)/R_e}{\sum_e \Pr_\pi(e)/R_e} \right)
\end{aligned}
$$

- Family size

  (Schreuder & Baayen, 1997)

- Singular/Plural dominance

  (Baayen et al., 1997)

Historical overview

Information theory and lexical processing

- Entropy

  (Moscoso del Prado Martín et al., 2004)

$$H = -\sum_{e} \Pr_{\pi}(w_e) \log_2 \Pr_{\pi}(w_e)$$

$$I_R = I_w - H$$

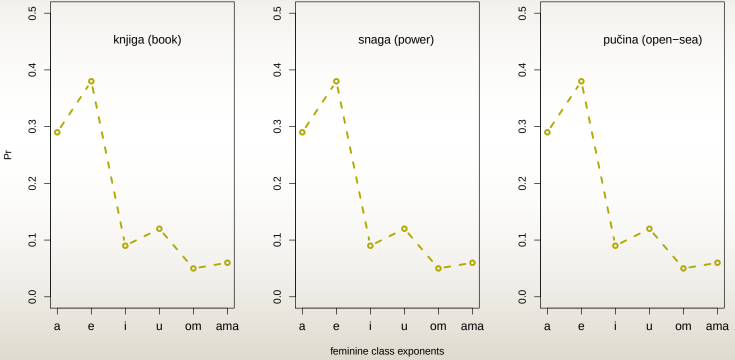
- Derivational vs Inflectional entropy

  (Baayen et al., 2006)

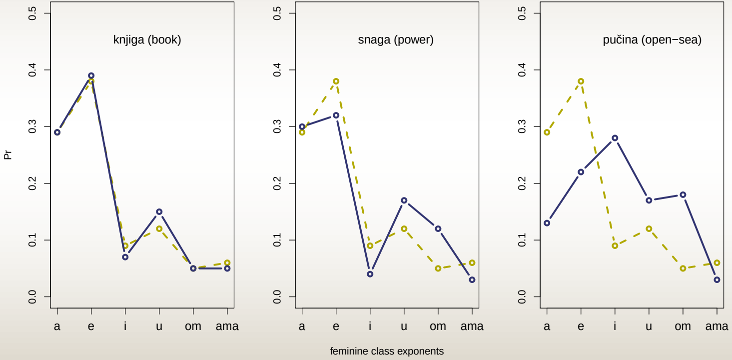
Inflected nouns in Serbian

	Inflected variant			Exponent	
	Frequency	Relative frequency		Frequency	Relative frequency
	$F(w_e)$	$Pr_\pi(w_e)$		$F(e)$	$Pr_\pi(e)$
planin-*a*	169	0.31	**-a**	18715	0.26
planin-*u*	48	0.09	**-u**	9918	0.14
planin-*e*	191	0.35	**-e**	27803	0.39
planin-*i*	88	0.16	**-i**	7072	0.10
planin-*om*	30	0.05	**-om**	4265	0.06
planin-*ama*	26	0.05	**-ama**	4409	0.06

Nominal classes and paradigms

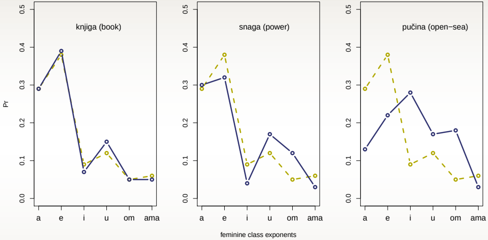


Nominal classes and paradigms



Nominal classes and paradigms

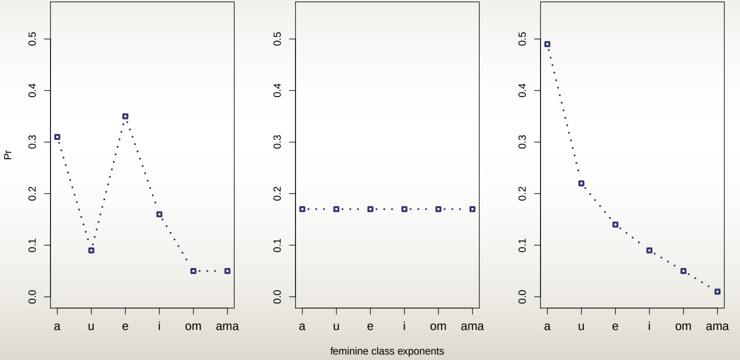
Information-theoretic perspective



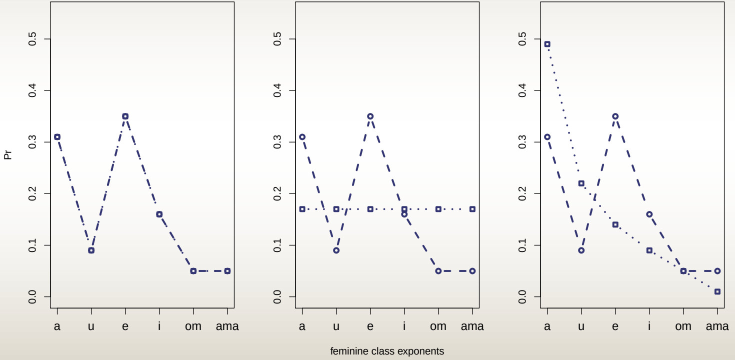
$$D(P||Q) = \sum_e \Pr_\pi(w_e) \log_2 \frac{\Pr_\pi(w_e)}{\Pr_\pi(e)}$$

(Milin, Filipović Đurđević, & Moscoso del Prado Martin, 2009)

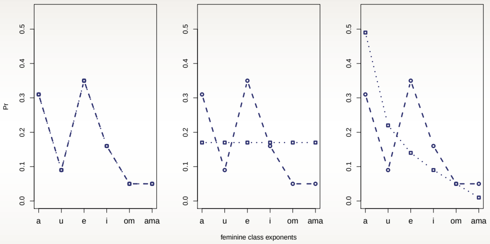
DYNAMICS OF THE CLASSES AND PARADIGMS



Dynamics of the classes and paradigms

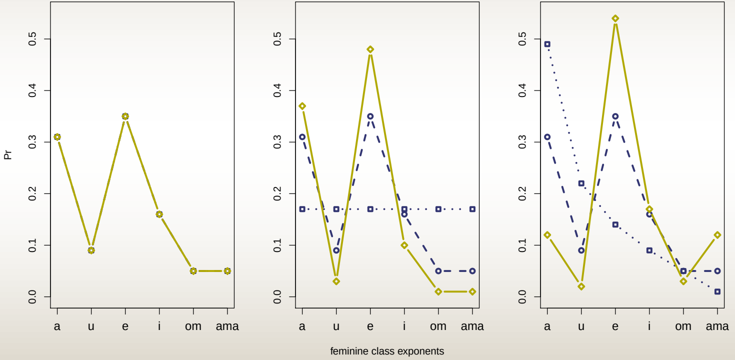


Dynamics of the classes and paradigms



$$\frac{f(target_e)}{f(prime_e)}$$

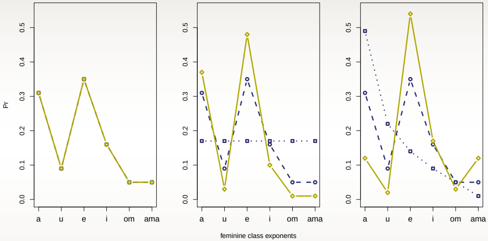
DYNAMICS OF THE CLASSES AND PARADIGMS



Target	Inflected variant Frequency	Prime	Frequency	Weight	Exponent	Frequency
	$F(w_e)_a$		$F(w_e)_b$	ω_e		$F(e)$
planin-*a*	169	struj-*a*	40	4.23	**-a**	18715
planin-*u*	48	struj-*u*	23	2.09	**-u**	9918
planin-*e*	191	struj-*e*	65	2.94	**-e**	27803
planin-*i*	88	struj-*i*	8	11.0	**-i**	7072
planin-*om*	30	struj-*om*	9	3.33	**-om**	4265
planin-*ama*	26	struj-*ama*	17	1.53	**-ama**	4409

Dynamics of the classes and paradigms

Information-theoretic perspective



$$D(P||Q;W) = \sum_e \frac{\Pr_\pi(w_e)\omega_e}{\sum_e \Pr_\pi(w_e)\omega_e} log_2 \frac{\Pr_\pi(w_e)}{\Pr_\pi(e)}; \qquad \omega_e = \frac{f(target_e)}{f(prime_e)}$$

(Baayen, Milin, Filipović Đurđević, Hendrix, & Marelli, 2011)

LIGHTER SHADE OF PALE

- Do we (really want to) believe that we are doing on-line entropy measuring while we listen/speak/read/write?
- Information-theoretic measures must take proper epistemological positioning in our way of thinking about language
- Levels of analysis (Marr, 1982):
  - **computational:** what does the system do, and why
  - **algorithmic (representational):** how does the system do, how it uses information
  - **implementational:** physical (biological) realisation

Language as a Complex Adaptive System

- COMPUTATIONALLY

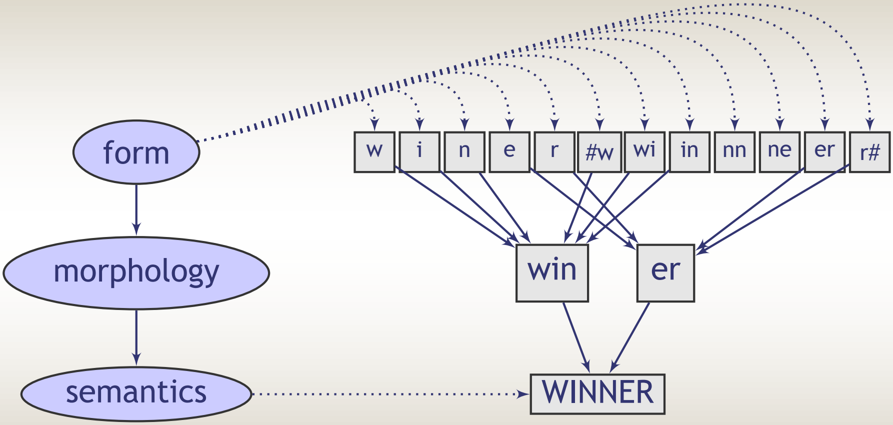
  Information theory is essential for understanding language as CAS

  It characterises what the system is doing

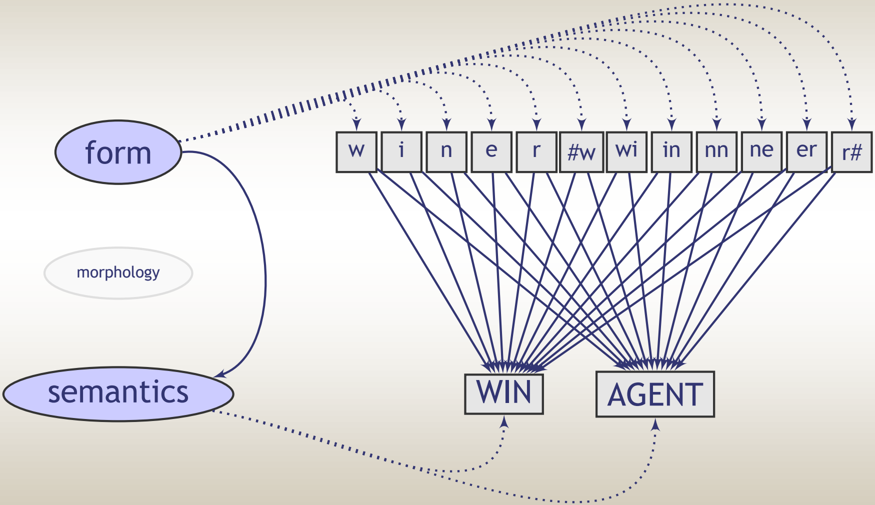
- ALGORITHMICALLY

  A simple model based on learning principles can give us insights into how language as CAS makes these dynamics

PROCESSING MORPHOLOGY: STANDARD MODEL



Processing morphology: Amorphous model



NAIVE DISCRIMINATIVE LEARNING PRINCIPLES

- Links between orthography (cues) and semantics (outcomes) are established through discriminative learning
  - Rescorla-Wagner discriminative learning equations

    (Rescorla & Wagner, 1972)
  - Equilibrium equations

    (Danks, 2003)
- The activation for a given outcome is the sum of all association weights between the relevant input cues and that outcome
  - **cues:** letters and letter combinations
  - **outcomes:** meanings

Rescorla-Wagner equations

Recursive discriminative learning

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

with

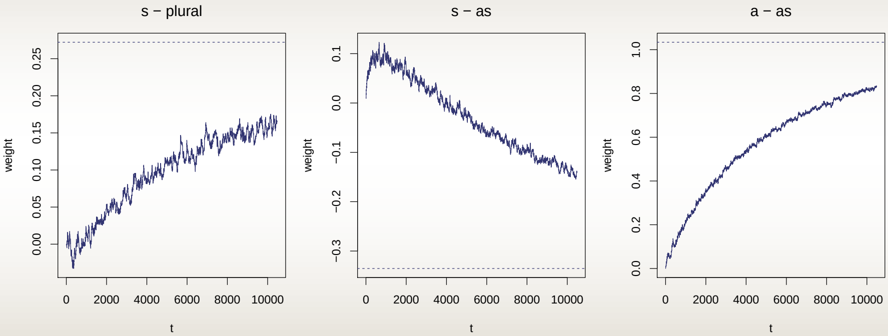
$$\Delta V_i^t = \begin{cases} 0 & \text{if } \textsc{absent}(C_i, t) \\ \alpha_i \beta_1 \left(\lambda - \sum_{\textsc{present}(C_i,\, t)} V_i \right) & \text{if } \textsc{present}(C_i, t) \ \&\ \textsc{present}(O, t) \\ \alpha_i \beta_2 \left(0 - \sum_{\textsc{present}(C_i,\, t)} V_i \right) & \text{if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

- connection strength increases if cue is informative
- it decreases if cue is not discriminative
- the larger the set of cues, the smaller the individual connections

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



Danks equilibrium equations

Stable state

- If the system is in the stable state, 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

- V_i optimises the conditional outcomes given the conditional co-occurrence probabilities of 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
  - **the simplest case:** $\mathrm{RTsim}_i \propto -a_i$
  - **to remove the right skew:** $\mathrm{RTsim}_i \propto \log(1/a_i)$

The Naive Discriminative Learning

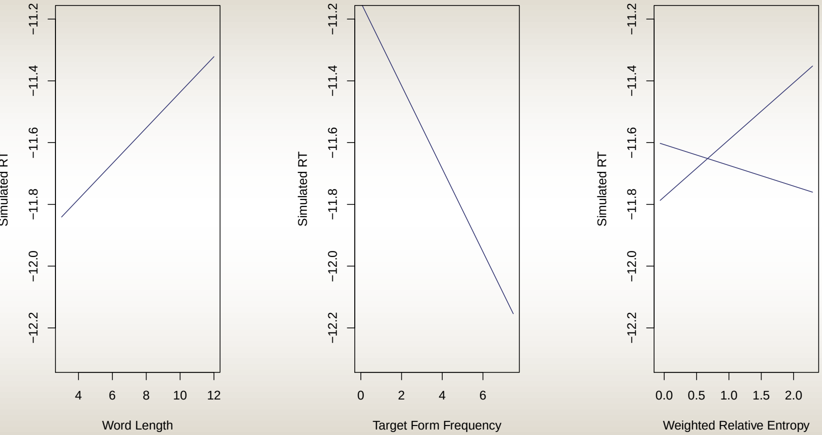
- Basic engine is parameter-free, and driven completely and only by the language input
- The model is computationally undemanding: building the weight matrix from a lexicon of 11 million phrases takes about 10 minutes
- Full implementation in R (`ndl` package on CRAN)

Serbian nominal case paradigms

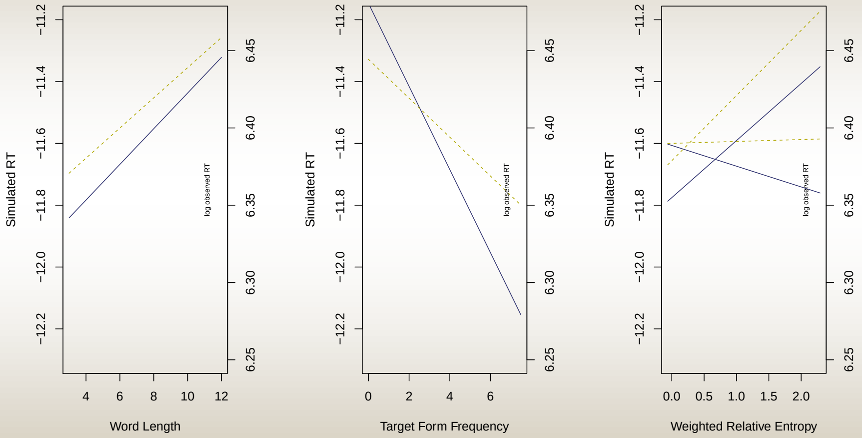
Training set: 270 nouns in 3240 inflected forms

	Inflected variant				Exponent	
Target	Frequency	Prime	Frequency	Weight		Frequency
	$F(w_e)_a$		$F(w_e)_b$	ω_e		$F(e)$
planin-*a*	169	struj-*a*	40	4.23	**-a**	18715
planin-*u*	48	struj-*u*	23	2.09	**-u**	9918
planin-*e*	191	struj-*e*	65	2.94	**-e**	27803
planin-*i*	88	struj-*i*	8	11.0	**-i**	7072
planin-*om*	30	struj-*om*	9	3.33	**-om**	4265
planin-*ama*	26	struj-*ama*	17	1.53	**-ama**	4409

Expected and observed coefficients



Expected and observed coefficients



Summary of results on Serbian data

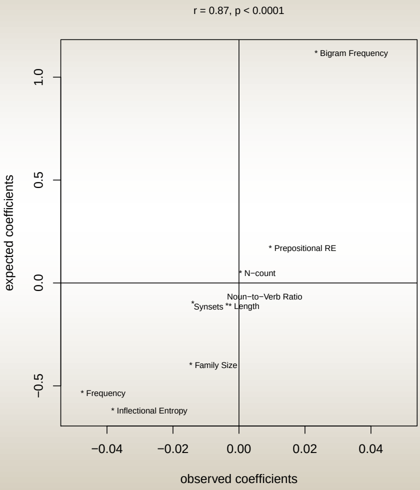
- Relative entropy effects persist in sentential reading
- They are modified, but not destroyed by the prime
- The interaction with masculine gender follows from the distributional properties of the lexical input
- The interaction with nominative case remains unaccounted; it could be caused by syntactic functions and meanings (cf., Kostić, 2003)
- Paradigmatic effects can arise without representations for complex words or representational structures for paradigms

English prepositional phrase paradigms

Training set: 11,172,554 two and three-word phrases from the British National Corpus, comprising 26,441,155 word tokens

	Phrase Frequency	Rel. freq.		Preposition Frequency	Rel. freq.
	$F(p_p)$	$\Pr_\pi(p_p)$		$F(p)$	$\Pr_\pi(p)$
on a plant	28608	0.279	**on**	177908042	0.372
in a plant	52579	0.513	**in**	253850053	0.531
under a plant	7346	0.072	**under**	10746880	0.022
above a plant	0	0.000	**above**	2517797	0.005
through a plant	0	0.000	**through**	3632886	0.008
behind a plant	760	0.007	**behind**	3979162	0.008
into a plant	13289	0.130	**into**	25279478	0.053

Expected and observed coefficients



Summary of results on English data

- Phrasal paradigmatic effect is modelled correctly, and without representations for phrases
- Again, we observed prototype and exemplar interplay, as expressed by the prepositional relative entropy, without explicit linkage between the two

- This confirms that syntactic context is relevant for word processing
- Crucially, word's syntactic realisation raises its paradigmatic structures

THE MEANING OF RELATIVE ENTROPY

Q What connections in our model carry information about Relative Entropy?

- Inflectional exponents or prepositions are not at all discriminative
- They are present (active) in many words

- Contrariwise, base cues are those that give support for the particular realisation of inflected variants or phrases
- They carry functional load which we measure as Relative Entropy

The meaning of Relative Entropy

- From the cognitive perspective:
  - words are part of our mental representations
  - they denote what denotee does in reality
  - this seems to be encoded in our personal experience
  - and, more importantly, in our sixth-sense — language

- From the linguistic perspective:
  - this puts some challenge to the notion of compositionality
  - part of knowledge about paradigms are present in the base

Concluding remarks

- Language as an COMPLEX ADAPTIVE SYSTEM has very rich dynamics, but optimality constraints
- Information theory is a fruitful tool that helps us understanding what are these constraints and why they emerge
- Relative Entropy does a beautiful job in revealing nature of WORDS and theirs PARADIGMS and CLASSES
- It even gives us insights into dynamics of words' paradigmatics

Concluding remarks

- Naive Discriminative Learning machinery is a simple model which does calculus of connectivity
- In Marrian spirit, it can be seen just one possible algorithmic realisation of Bybee's computational Network Model
- It is probably way to simple, but does not require hard statistics on the hidden layer
- It is useful for detailed linguistic and psychological analysis
- Please, help us make it better! ☺

http://cran.opensourceresources.org/web/packages/ndl/index.html

COLLABORATORS

R. Harald Baayen, University of Alberta

Antti Arppe, University of Helsinki

Marco Marelli, University of Milano-Bicocca

Peter Hendrix, University of Alberta

Thank you!

Department of Psychology
Faculty of Philosophy
University of Novi Sad

Laboratory for Experimental Psychology
Faculty of Philosophy
University of Belgrade