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 πολιτίκον)

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.)

\[ l_e = - \log_2 Pr_{\pi}(e) \]

\[ l'_e = - \log_2 \left( \frac{Pr_{\pi}(e)/R_e}{\sum_e Pr_{\pi}(e)/R_e} \right) \]

- 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)

\[ H = - \sum_e \text{Pr}_\pi(w_e) \log_2 \text{Pr}_\pi(w_e) \]

\[ I_R = I_W - H \]

- Derivational vs Inflectional entropy
  (Baayen et al., 2006)
## Inflected Nouns in Serbian

<table>
<thead>
<tr>
<th>Inflected variant</th>
<th>Frequency</th>
<th>Relative frequency</th>
<th>Exponent</th>
<th>Frequency</th>
<th>Relative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>planin-a</td>
<td>169</td>
<td>0.31</td>
<td>-a</td>
<td>18715</td>
<td>0.26</td>
</tr>
<tr>
<td>planin-u</td>
<td>48</td>
<td>0.09</td>
<td>-u</td>
<td>9918</td>
<td>0.14</td>
</tr>
<tr>
<td>planin-e</td>
<td>191</td>
<td>0.35</td>
<td>-e</td>
<td>27803</td>
<td>0.39</td>
</tr>
<tr>
<td>planin-i</td>
<td>88</td>
<td>0.16</td>
<td>-i</td>
<td>7072</td>
<td>0.10</td>
</tr>
<tr>
<td>planin-om</td>
<td>30</td>
<td>0.05</td>
<td>-om</td>
<td>4265</td>
<td>0.06</td>
</tr>
<tr>
<td>planin-ama</td>
<td>26</td>
<td>0.05</td>
<td>-ama</td>
<td>4409</td>
<td>0.06</td>
</tr>
</tbody>
</table>
Nominal Classes and Paradigms

knjiga (book)

snaga (power)

pučina (open-sea)

Pr

feminine class exponents
Nominal Classes and Paradigms

 knjiga (book)

 snaga (power)

 pučina (open–sea)
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

\[ f(\text{target}_e) \]
\[ f(\text{prime}_e) \]
DYNAMICS OF THE CLASSES AND PARADIGMS
<table>
<thead>
<tr>
<th>Target</th>
<th>Inflected variant</th>
<th>Frequency $F(w_e)_a$</th>
<th>Prime</th>
<th>Frequency $F(w_e)_b$</th>
<th>Weight $\omega_e$</th>
<th>Exponent</th>
<th>Exponent Frequency $F(e)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>planin-(a)</td>
<td>struj-(a)</td>
<td>169</td>
<td></td>
<td>40</td>
<td>4.23</td>
<td>-(a)</td>
<td>18715</td>
</tr>
<tr>
<td>planin-(u)</td>
<td>struj-(u)</td>
<td>48</td>
<td></td>
<td>23</td>
<td>2.09</td>
<td>-(u)</td>
<td>9918</td>
</tr>
<tr>
<td>planin-(e)</td>
<td>struj-(e)</td>
<td>191</td>
<td></td>
<td>65</td>
<td>2.94</td>
<td>-(e)</td>
<td>27803</td>
</tr>
<tr>
<td>planin-(i)</td>
<td>struj-(i)</td>
<td>88</td>
<td></td>
<td>8</td>
<td>11.0</td>
<td>-(i)</td>
<td>7072</td>
</tr>
<tr>
<td>planin-(om)</td>
<td>struj-(om)</td>
<td>30</td>
<td></td>
<td>9</td>
<td>3.33</td>
<td>-(om)</td>
<td>4265</td>
</tr>
<tr>
<td>planin-(ama)</td>
<td>struj-(ama)</td>
<td>26</td>
<td></td>
<td>17</td>
<td>1.53</td>
<td>-(ama)</td>
<td>4409</td>
</tr>
</tbody>
</table>
DYNAMICS OF THE CLASSES AND PARADIGMS

INFORMATION-THEORETIC PERSPECTIVE

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

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

\[
\Delta V_{i}^{t} = \begin{cases} 
0 & \text{if ABSENT}(C_{i}, t) \\
\alpha_{i}\beta_{1} \left( \lambda - \sum_{\text{PRESENT}(C_{i}, t)} V_{i} \right) & \text{if PRESENT}(C_{i}, t) \& \text{PRESENT}(O, t) \\
\alpha_{i}\beta_{2} \left( 0 - \sum_{\text{PRESENT}(C_{i}, t)} V_{i} \right) & \text{if PRESENT}(C_{i}, t) \& \text{ABSENT}(O, t) 
\end{cases}
\]

- 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

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
<th>Lexical Meaning</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>hand</td>
<td>10</td>
<td>HAND</td>
<td></td>
</tr>
<tr>
<td>hands</td>
<td>20</td>
<td>HAND</td>
<td></td>
</tr>
<tr>
<td>land</td>
<td>8</td>
<td>LAND</td>
<td></td>
</tr>
<tr>
<td>lands</td>
<td>3</td>
<td>LAND</td>
<td></td>
</tr>
<tr>
<td>and</td>
<td>35</td>
<td>AND</td>
<td></td>
</tr>
<tr>
<td>sad</td>
<td>18</td>
<td>SAD</td>
<td></td>
</tr>
<tr>
<td>as</td>
<td>35</td>
<td>AS</td>
<td></td>
</tr>
<tr>
<td>lad</td>
<td>102</td>
<td>LAD</td>
<td></td>
</tr>
<tr>
<td>lads</td>
<td>54</td>
<td>LAD</td>
<td>PLURAL</td>
</tr>
<tr>
<td>lass</td>
<td>134</td>
<td>LASS</td>
<td></td>
</tr>
</tbody>
</table>
THE RESCORLA-WAGNER EQUATIONS APPLIED

s – plural

s – as

a – as
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

\[
\begin{pmatrix}
\Pr(C_0|C_0) & \Pr(C_1|C_0) & \ldots & \Pr(C_n|C_0) \\
\Pr(C_0|C_1) & \Pr(C_1|C_1) & \ldots & \Pr(C_n|C_1) \\
\vdots & \vdots & \ddots & \vdots \\
\Pr(C_0|C_n) & \Pr(C_1|C_n) & \ldots & \Pr(C_n|C_n)
\end{pmatrix}
\begin{pmatrix}
V_0 \\
V_1 \\
\vdots \\
V_n
\end{pmatrix}
= \begin{pmatrix}
\Pr(O|C_0) \\
\Pr(O|C_1) \\
\vdots \\
\Pr(O|C_n)
\end{pmatrix}
\]

*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: $RT_{sim_i} \propto -a_i$
  - to remove the right skew: $RT_{sim_i} \propto \log(1/a_i)$
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

<table>
<thead>
<tr>
<th>Target</th>
<th>Inflected variant</th>
<th>Target Frequency</th>
<th>Prime Frequency</th>
<th>Target Frequency</th>
<th>Weight</th>
<th>Exponent</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>planin-a</td>
<td>169 struj-a</td>
<td>40</td>
<td>4.23</td>
<td>-a</td>
<td>18715</td>
<td></td>
<td></td>
</tr>
<tr>
<td>planin-u</td>
<td>48 struj-u</td>
<td>23</td>
<td>2.09</td>
<td>-u</td>
<td>9918</td>
<td></td>
<td></td>
</tr>
<tr>
<td>planin-e</td>
<td>191 struj-e</td>
<td>65</td>
<td>2.94</td>
<td>-e</td>
<td>27803</td>
<td></td>
<td></td>
</tr>
<tr>
<td>planin-i</td>
<td>88 struj-i</td>
<td>8</td>
<td>11.0</td>
<td>-i</td>
<td>7072</td>
<td></td>
<td></td>
</tr>
<tr>
<td>planin-om</td>
<td>30 struj-om</td>
<td>9</td>
<td>3.33</td>
<td>-om</td>
<td>4265</td>
<td></td>
<td></td>
</tr>
<tr>
<td>planin-ama</td>
<td>26 struj-ama</td>
<td>17</td>
<td>1.53</td>
<td>-ama</td>
<td>4409</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
EXPECTED AND OBSERVED COEFFICIENTS

Word Length

Target Form Frequency

Weighted Relative Entropy
EXPECTED AND OBSERVED COEFFICIENTS

**Graphs:**
- **Word Length**
- **Target Form Frequency**
- **Weighted Relative Entropy**
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

<table>
<thead>
<tr>
<th>Phrase Preposition</th>
<th>Frequency</th>
<th>Rel. freq.</th>
<th>Preposition</th>
<th>Frequency</th>
<th>Rel. freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>on a plant</strong></td>
<td>28608</td>
<td>0.279</td>
<td>on</td>
<td>177908042</td>
<td>0.372</td>
</tr>
<tr>
<td><strong>in a plant</strong></td>
<td>52579</td>
<td>0.513</td>
<td>in</td>
<td>253850053</td>
<td>0.531</td>
</tr>
<tr>
<td><strong>under a plant</strong></td>
<td>7346</td>
<td>0.072</td>
<td>under</td>
<td>10746880</td>
<td>0.022</td>
</tr>
<tr>
<td><strong>above a plant</strong></td>
<td>0</td>
<td>0.000</td>
<td>above</td>
<td>2517797</td>
<td>0.005</td>
</tr>
<tr>
<td><strong>through a plant</strong></td>
<td>0</td>
<td>0.000</td>
<td>through</td>
<td>3632886</td>
<td>0.008</td>
</tr>
<tr>
<td><strong>behind a plant</strong></td>
<td>760</td>
<td>0.007</td>
<td>behind</td>
<td>3979162</td>
<td>0.008</td>
</tr>
<tr>
<td><strong>into a plant</strong></td>
<td>13289</td>
<td>0.130</td>
<td>into</td>
<td>25279478</td>
<td>0.053</td>
</tr>
</tbody>
</table>
EXPECTED AND OBSERVED COEFFICIENTS

$r = 0.87, p < 0.0001$
 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 realised raises its paradigmatic structures
THE MEANING OF RELATIVE ENTROPY

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!