

The Bayesian Basis for Linguistic Expectations in Language Processing

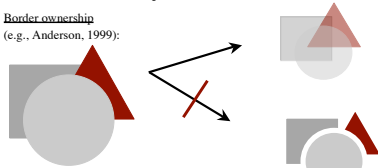
Evelina Fedorenko¹, Edward Gibson¹ & Maria Babyonyshev² (¹MIT, ²Yale University)

Background & Motivation

Background

“Non-accidentalness” assumptions in vision

Border ownership
(e.g., Anderson, 1999):



Generic View Principle

(e.g., Nakayama & Shimojo, 1992):



The Size Principle (Laplace, 1812; Hoffman, 1998; Tenenbaum, 2000)

The “number game”:

GUESS THE RULE WHICH I AM USING IN PICKING THE NUMBERS (1-100).

INPUT: 20, 70

SOME POSSIBLE HYPOTHESES:

- H1 - even numbers
- H2 - multiples of 5
- H3 - multiples of 10

the narrowest hypothesis

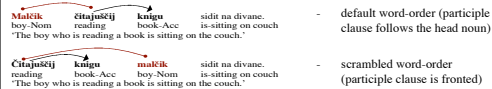
The Size Principle: Narrower hypotheses receive greater likelihood, and exponentially more so as n increases.

$$p(X|h) = \begin{cases} 1/|size(h)|^n & \text{if } x_1, \dots, x_n \in h \\ 0 & \text{if any } x_i \notin h \end{cases}$$

Linguistic Background

Participle constructions in Russian:

- (1) Transitive participles **assign case** to their objects.
- (2) Participles **agree** with the head noun in gender, number and case.
- (3) For a transitive participle, the embedded object must immediately follow the participle.



	[CASE]	[AGR]	[verb]	[subject]
Participle	[emb object]	[matrix object]		
Уважавший	скрипаčku	пианистку	разсердил	дирижер
Respecting fem-acc	violinist-fem-acc	pianist-fem-acc	angered	conductor-nom

The conductor angered the pianist who respected the violinist.

Motivating Intuition

IF (1) the sentences are reversible (both the embedded and the matrix objects are plausible subjects/objects of the participle), and (2) the embedded object happens to match the participle in gender, number and case, THEN there is a strong tendency to interpret the embedded object as the matrix object, which would result in an **ungrammatical** interpretation in Russian.

Experiment

Method

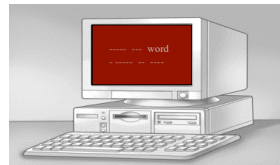
- Participants - 40 subjects (5 removed from the analyses on the basis of low accuracy performance)
- Design - 2 x 2
- Factors:
 - **Case Match** between the Participle and the Embedded Object (NP1) (+CaseMatch, -CaseMatch)
 - **Gender Match** between the Participle and the Embedded Object (NP1) (+GenMatch, -GenMatch)

Materials

a. +CaseMatch/+GenderMatch	Participle	EmbObj-NP1	MtrObj-NP2	Verb	Subj...
Уважавший	скрипаčku	пианистку	разсердил	дирижер...	angered conductor...
Respecting	violinist (acc/fem)	pianist (acc/fem)	angered conductor...		
The conductor angered the pianist who respected the violinist...					
b. +CaseMatch/-GenderMatch	Уважавший	скрипача (acc/masc)	пианисту	разсердил	дирижер...
Respecting	violinist (acc/masc)	pianist (acc/fem)	angered conductor...		
c. -CaseMatch/+GenderMatch	Позвоновивший	скрипаčke	пианисту	разсердил	дирижер...
Calling	violinist (dative/fem)	pianist (acc/fem)	angered conductor...		
d. -CaseMatch/-GenderMatch	Позвоновивший	скрипаču	пианисту	разсердил	дирижер...
Calling	violinist (dative/masc)	pianist (acc/fem)	angered conductor...		

EmbObj (NP1) - always satisfied the subcategorization requirement of the participle;
- satisfied 0, 1 or 2 agreement requirement(s) of the participle.

Procedure



-self-paced word-by-word reading with a moving-window display;

- the critical regions were defined as
 - (1) the Embedded Object (NP1);
 - (2) the Matrix Object (NP2)
 (for the reanalysis effect);
- (3) the Verb (for possible spill-over effects).

Predictions

Morphological constraints on the form of the upcoming nouns:

- *The agreement requirement of the participle:* case, gender, and number [for the matrix object].
- *The subcategorization requirement of the participle:* case [for the embedded object].
- (1) The agreement requirement is **narrower** than the subcategorization requirement.

A **Bayesian account** predicts that when two syntactic predictions (which differ in their specificity levels) are pending, the parser will attempt to satisfy the narrowest prediction first, overriding other possible constraints of the language (e.g. structural constraints).

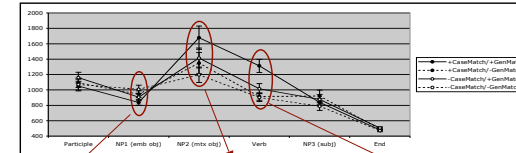
Specifically, it predicts that a noun consistent with both the agreement and the subcategorization requirements of the participle will be initially **preferentially interpreted as satisfying the agreement requirement**, despite the resulting ungrammaticlicity.

Reading time predictions:

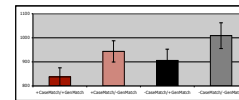
- [Bayesian account prediction]
 - NP1 - facilitation of processing when the noun can satisfy the narrowest requirement of the participle (i.e. the agreement prediction).
- [any sentence processing account that assumes incrementality]
 - NP2 - the reversal of this effect when it is discovered that NP1 is not the matrix object and reanalysis (i.e. reinterpreting NP1 as the embedded object of the participle) is required.
 - Verb - similar pattern due to possible spill-over effects from NP2.

Results & Conclusions

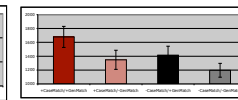
Reading Times



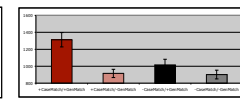
Reading Times - NP 1



Reading Times - NP 2



Reading Times - Verb



-Main effect of **CASE**:
+CaseMatch conditions faster
(F(1,33)=4.94; $p < .05$; F(2,128)=9.01; $p < .01$).

-Main effect of **CASE**:
+CaseMatch conditions slower
(F(1,33)=10.81; $p < .005$; F(2,128)=6.98; $p < .02$).

-Main effect of **CASE**:
+CaseMatch conditions slower
(F(1,33)=13.46; $p < .001$; F(2,128)=7.13; $p < .02$).

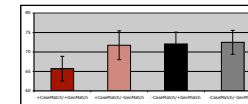
-Main effect of **GENDER**:
+GenMatch conditions faster
(F(1,33)=4.94; $p < .05$; F(2,128)=9.01; $p < .01$).

-Main effect of **GENDER**:
+GenMatch conditions slower
(F(1,33)=32.63; $p < .0001$; F(2,128)=19.06; $p < .0005$).

-Main effect of **GENDER**:
+GenMatch conditions slower
(F(1,33)=15.70; $p < .0005$; F(2,128)=16.81; $p < .0005$).

-**INTERACTION**:
(F(1,33)=8.82; $p < .01$; F(2,128)=7.46; $p < .02$).

Comprehension Performance



No significant effects or interactions
($F_s < 1.9$, $p_s > .18$).
Numerically, the pattern is consistent with the RT data.

Conclusions

The results were consistent with the Bayesian prediction.

NP1:

- **+CaseMatch/+GenMatch** condition (where the noun matched both of the agreement requirements of the participle - morphological case and gender) was the fastest.
- **+CaseMatch/-GenMatch** condition (where the noun did not match either of the agreement requirements of the participle) was the slowest.
- **-CaseMatch/+GenMatch** and **-CaseMatch/-GenMatch** conditions (where the noun matched only one of the agreement requirements) were in the middle.

NP2: the reverse pattern

- **+CaseMatch/+GenMatch** condition was the slowest.
- **-CaseMatch/-GenMatch** condition was the fastest.
- **+CaseMatch/-GenMatch** and **-CaseMatch/+GenMatch** conditions were in the middle.

The recovery from the initial misinterpretation is most difficult for the **+CaseMatch/+GenMatch** condition (continuing into the Verb region), because in that condition the initial interpretation of NP1 as the matrix object is the strongest due to the fact that NP1 *fully* satisfies the agreement requirement of the participle, compared to the **+CaseMatch/-GenMatch** and **-CaseMatch/+GenMatch** conditions, where NP1 only *partially* satisfies the agreement requirement of the participle.

To the best of our knowledge, no theory of sentence processing predicts the observed pattern of results.