university of groningen



# Predicting Grammatical Roles – Does Word Order Matter?

Statistics and Methodology

Anna Lobanova: a.lobanova@ai.rug.nl

# The Goals:

- To present a study where I looked at the correlation between grammatical roles and such properties of NPs as (1) animacy and (2) referential form;
- To show that contrary to previous research these features can help to predict the position of an NP in a sentence rather than its grammatical role;
- To discuss what went wrong statistically!

# Prototypicality (Aissen 2003):

- **Subjects** – Animate
- **Subjects** – Definite
- **Subjects** – Agent
- **Subjects** – Pronouns

**Objects** – Inanimate

**Objects** - Indefinite

**Objects** – Patients

**Objects** – full NPs

# A study on English and Swedish (Zeevat&Jäger 2002)

***Corpus Study***: disharmonic elements (e.g., pronominal human objects).

- English (the Wall Street Journal: ~250,000 NP's)
- Swedish (Conversations in Gothenburg)

***Expectations***: disharmonic combinations have lower frequencies than the frequency of the either element in the combination:

$p(\text{OBJ}|\text{HUM}) < p(\text{OBJ}|\text{NP})$

and

$p(\text{HUM}|\text{OBJ}) < p(\text{HUM}|\text{NP})$

# A study on English and Swedish (Zeevat&Jäger 2002)

***Corpus Study***: disharmonic elements (e.g., pronominal human objects).

- English (the Wall Street Journal: ~250,000 NP's)
- Swedish (Conversations in Gothenburg)

***Expectations***: disharmonic combinations have lower frequencies than the frequency of the either element in the combination:

$p(\text{OBJ}|\text{HUM}) = 42\% \quad < \quad p(\text{OBJ}|\text{NP}) = 75\%$

and

$p(\text{HUM}|\text{OBJ}) = 10\% \quad < \quad p(\text{HUM}|\text{NP}) = 13\%$

# A study on English and Swedish (Zeevat&Jäger 2002)

⇨ Given the features of an NP, it is possible to predict its grammatical role (abstract category).

# but

The word order in English and Swedish is relatively strict!

Are these results valid for free word order languages?

# Word Order in Russian:

*'Mary loves Kostja.'*

**SVO** - default

**Marija**(NOM) ljubit Kostju(ACC).

*Mary loves Kostja.*

or

**OVS**

Kostju(ACC) ljubit **Marija**(NOM).

*Mary loves Kostja.*

or ...

**SOV**, **OSV**, **VSO**, **VOS**

# What can influence the order of constituents?

- Information structure: given before new
- Thematic roles: agent before patient
- Animacy of NP's: animate NPs precede inanimate (Dahl 2000, Zeevat&Jäger 2002, Øvrelid 2004)
- Definiteness of NP's: definite NPs precede indefinite (Zeevat&Jäger 2002, Weber&Müller 2004)
- Length of NPs: shorter constituents precede longer c. (Heylen 2005)

⇨ Given properties of NPs, it is possible to predict their order in a sentence!

university of groningen



# Part I: Preliminary Study (Russian)

300 transitive sentences from a novel were divided according to the word order:

| | |
|---|---|
| 88% (or 265 clauses) | were **SVO**, |
| 6% (or 17 clauses) | were **OVS**, |
| 4% (or 11 clauses) | were OSV, |
| 1,5% (or 5 clauses) | were SOV, |
| 0,5% (2 clauses) | were VOS. |

# Part II: Main Study (Russian)

150 SVO & 150 OVS sentences from the novel;

150 SVO & 150 OVS sentences from the newspaper.

All nouns were annotated as:

- Animate/Inanimate Abstract/Inanimate Concrete;
- Pronoun/Proper Name/NP.

# Results 1.1. (Animacy)

| *SVO* (novel) | ***Subject*** | ***Object*** |
|---|---|---|
| *Animate* | 73% (109) | 28% (42) |
| *Inan Abst* | 21% (32) | 50% (75) |
| *Inan Conc* | 6% (9) | 22% (33) |

$\chi^2 = 60.723$

| *OVS* (novel) | ***Subject*** | ***Object*** |
|---|---|---|
| *Animate* | 50% (75) | 54% (81) |
| *Inan Abst* | 44% (66) | 39% (58) |
| *Inan Conc* | 6% (9) | 7% (11) |

$\chi^2 = 0.947$ p= 0.62

| *SVO* (news) | ***Subject*** | ***Object*** |
|---|---|---|
| *Animate* | 52% (78) | 15% (23) |
| *Inan Abst* | 43% (65) | 72%(108) |
| *Inan Conc* | 5% (7) | 13% (19) |

$\chi^2 = 46.177$

| *OVS* (news) | ***Subject*** | ***Object*** |
|---|---|---|
| *Animate* | 46% (69) | 33% (50) |
| *Inan Abst* | 47% (71) | 59% (89) |
| *Inan Conc* | 7% (10) | 7% (11) |

$\chi^2 = 5.106$ p= 0.077

# Results 1.2. (Animacy)

| | SVO | OVS |
|---|---|---|
| Novel | P(Obj \| Anim) = 28% < P(Obj \| NP) = 63% | **P(Obj \| Anim) = 52% > P(Obj \| NP) = 37%** |
| | | **P(Obj \| InanAbs) = 47% > P(Obj \| NP) = 37%** |
| | | |
| News | P(Obj \| Anim) = 23% < P(Obj \| NP) = 60% | P(Obj \| Anim) = 43% < P(Obj \| NP) = 50% |
| | | **P(Obj \| InanAbs) = 55% > P(Obj \| NP) = 50%** |

# Results 2.1. (Referential Form)

| ***SVO*** **(novel)** | ***Subject*** | ***Object*** |
|---|---|---|
| *Pronoun* | 19% (29) | 18% (27) |
| *Proper Name* | 37% (55) | 5% (7) |
| *NP* | 44% (66) | 77% (116) |

$X^2 = 50.969$

| ***OVS*** **(novel)** | ***Subject*** | ***Object*** |
|---|---|---|
| *Pronoun* | 2% (3) | 34% (51) |
| *Proper Name* | 33% (49) | 15% (22) |
| *NP* | 65% (98) | 38% (57) |

$X^2 = 62.671$

| ***SVO*** **(news)** | ***Subject*** | ***Object*** |
|---|---|---|
| *Pronoun* | 20% (30) | 5% (8) |
| *Proper Name* | 19% (28) | 5% (7) |
| *NP* | 61% (92) | 90% (135) |

$X^2 = 33.482$

| ***OVS*** **(news)** | ***Subject*** | ***Object*** |
|---|---|---|
| *Pronoun* | 0 | 18% (27) |
| *Proper Name* | 24% (36) | 7% (10) |
| *NP* | 76% (114) | 75% (113) |

$X^2 = 41.7$

# Results 2.2. (Referential Form)

| | SVO | OVS |
|---|---|---|
| Novel | P(Obj\|Pro) = 49% < P(Obj\|NP) = 63% | **P(Obj\|Pro) = 94% > P(Obj\|NP) = 37%** |
| | | P(Obj\|PN) = 30% < P(Obj\|NP) = 37% |
| | | |
| News | P(Obj\|Pro) = 22% < P(Obj\|NP) = 60% | **P(Obj\|Pro) = 100% > P(Obj\|NP) = 50%** |
| | | P(Obj\|PN) = 21% < P(Obj\|NP) = 50% |

# Conclusions:

- There is a correlation between animacy/referential form of NP and their grammatical function given a certain word order;
- Given animacy and referential form of NPs, it is possible to predict their position in a sentence;
- Consequently, in languages where the word order is fixed, predicting position will also “tell” the grammatical role of NPs.