Syntactic variation and uniformity across languages:
A crosslinguistic corpus study on linearization devices

Elisabeth Verhoeven
elisabeth.verhoeven@cms.hu-berlin.de
Introduction

FREQUENCIES IN DISCOURSE AND CONSTITUENT STRUCTURE

Background
Semantic and pragmatic asymmetries have an impact on the frequency of linearizations in discourse (e.g. choice of word order, of subject in passive or causative alternations).

Animate-first (see Siewierska 1993, ff.)

Given-first (see Clark & Haviland 1977, ff.), or - in corpus studies - Definite-first (see e.g. Weber & Müller 2004, Bresnan & Hay 2008).


Question of theoretical relevance
What is the relation of these preferences to properties of constituent structure?
Discourse asymmetries

• Animates are highly activated in memory and as such are very likely to be in the focus-of-attention (Bock and Warren 1985; Tomlin 1995). At the utterance level, this property means that:

\[
\text{animate} >_{\text{likelihood to appear in spec,TopP}} \text{inanimate}
\]

• Given information is part of the common ground and as such is more likely to be the topic of the utterance (Chafe 1976, ff.).

\[
\text{highly identifiable in CG} >_{\text{likelihood to appear in spec,TopP}} \text{less identifiable in CG}
\]

We assume that the correlations between animacy/referentiality and linearization options in discourse are the result of the likelihood of animates/identifiable information etc., to be the topic of the utterance.
Assume a configuration in which the lower role is more prominent than the higher role on a discourse prominence scale (animacy, referentiality or similar), such that it is more likely to be the topic of the utterance.

- Topicalizing the lower argument of a canonical tr. active V
  \[
  \left[\begin{array}{l}
  \text{CP} \left[ \begin{array}{l}
  \text{TopP} \\
  \alpha_i \\
  \text{TP} \\
  \vdots
  \end{array} \right] \\
  \text{[VP} \text{ ... } t_i \text{ ...}]\\
  \end{array} \right]
  \]

- Topicalizing the lower role of a passive V
  \[
  \left[\begin{array}{l}
  \text{CP} \left[ \begin{array}{l}
  \text{TopP} \\
  \alpha_i \\
  \text{TP} \\
  \vdots
  \end{array} \right] \\
  \text{[VP} \text{ ... ]}
  \end{array} \right]
  \]

We know from several empirical studies that in languages that have both options, the passive option occurs more frequently under several triggers (see e.g. Van Nice & Dietrich 2003 for animacy, Skopeteas & Fanselow 2009 for givenness, etc.). It is not clear why this is so (derivational cost or robust nominative-first preference?).
Relation to constituent structure

However, this prediction is not motivated for verb structures in which the non-nominative argument is higher in the constituent structure, i.e., for morphologically downgraded experiencers with non-canonical subject properties. In view of derivational costs, there is no reason for selecting a non-active voice in this case.

• Topicalizing the lower role of an EO verb with a quirky experiencer

\[
[\text{CP} \left[ \text{TopP} \; \alpha_i \; \left[ \text{TP} \; \left[ t_i \; \left[ \text{VP} \; \ldots \; ] \right] \right] \right] \right]
\]

EO-verbs may show exceptional syntactic properties:
- linearization
- passivization, extraction, binding, etc.


This is a contrast in the verbal lexicon and does not appear in all languages.
Typological difference in lexicon

languages having
a subclass of EO verbs with exceptional syntactic properties

yes

German
Greek
Islandic
Italian

no

Chinese
Turkish
Yucatec Maya
Korean

(at least for accusative verbs)

see Verhoeven 2010, 2014, Temme & Verhoeven 2015
Typological difference in lexicon

languages having a subclass of EO verbs with exceptional syntactic properties

<table>
<thead>
<tr>
<th>yes</th>
<th>no</th>
</tr>
</thead>
<tbody>
<tr>
<td>German</td>
<td>Chinese</td>
</tr>
<tr>
<td>Greek</td>
<td>Turkish</td>
</tr>
<tr>
<td>Icelandic</td>
<td>Yucatec Maya</td>
</tr>
<tr>
<td>Italian</td>
<td>Korean</td>
</tr>
</tbody>
</table>

(at least for accusative verbs)

see Verhoeven 2010, 2014, Temme & Verhoeven 2015
Typological difference in lexicon

languages having
a subclass of EO verbs with exceptional syntactic properties

yes  no
(at least for accusative verbs)

German  Chinese
Greek  Turkish
Islandic  Yucatec Maya

Prediction 1: Constituent structure and discourse preferences
The choice of non-active voice is not motivated for languages that have morphologically downgraded experiencers with subject properties (left side). – under the assumption that the surface order will reflect the fact that experiencers are higher in the syntactic structure.
Stems differ across languages

**Basis: Experiencer-Object**

(a) Chinese periphrastic passive

\[ x \text{ mízhù } y \text{ ‘}x\text{ attracts }y\text{’} \]

\[ y \text{ bèi } x \text{ mízhù } ‘y\text{ is attracted by }x’ \]

(b) Greek mediopassive

\[ x \text{ endiaféri } y \text{ ‘}x\text{ interests }y\text{’} \]

\[ y \text{ endiaférete ja } x \text{ ‘}y\text{ is interested in }x’ \]

(c) German reflexive, stative passive

\[ x \text{ ärgert } y \text{ ‘}x\text{ annoys }y\text{’} \]

\[ y \text{ ärgert sich über } x \text{ ‘}y\text{ is annoyed by }x’ \]

**Basis: Experiencer-Subject**

(a) Turkish causativization

\[ y \text{ x sevin-di } ‘y\text{ is happy about }x’ \]

\[ x \text{ y sevin-dir-di } ‘x\text{ makes }y\text{ happy’} \]

---

s. Nichols et al. 2004
typology of detransitivizing vs. transitivizing languages,
Intransitive morphological bases

Assume a configuration in which the lower role $a$ is more prominent than the higher role on a discourse prominence scale (animacy, referentiality or similar), such that it is more likely to be the topic of the utterance. In a language with an intransitive basis and the lower role $a$ as the subject (Turkish):

- **Intransitive basis**

  \[
  \text{[CP} \left[ \text{TopP} \alpha_i \text{[TP } t_i \text{[VP } \ldots \text{]]]} \right] \]

- **Causative**

  \[
  \text{[CP} \left[ \text{TopP} \alpha_i \text{[TP } \ldots \text{[vP } \ldots \text{[VP } \ldots \text{[VP } t_i \ldots\text{]]]} \right] \]

the intransitive structure should occur more frequently based on derivational cost AND nominative-first. While in German/Greek the lower-first linearization must be contextually licensed, in languages like Turkish, it is the actor-first linearization that must be contextually licensed.
Intransitive morphological bases

Assume a configuration in which the lower role a is more prominent than the higher role on a discourse prominence scale (animacy, referentiality or similar), such that it is more likely to be the topic of the utterance. In a language with an intransitive basis and the lower role a as the subject (Turkish):

- Intransitive basis
  \[
  \begin{array}{c}
  \text{CP} \quad (\text{TopP} \quad \alpha_i \quad [\text{TP} \quad t_i \quad [\text{VP} \ldots ]])
  \end{array}
  \]

- Causative
  \[
  \begin{array}{c}
  \text{CP} \quad (\text{TopP} \quad \alpha_i \quad [\text{TP} \quad \ldots \quad [\text{VP} \quad \ldots \quad [\text{VP} \quad \ldots \quad t_i \ldots ]]])
  \end{array}
  \]

Prediction 2: Morphological basis and discourse preferences
In transitivizing languages (Turkish), there is a bias for the intransitive structure, and the actor-first linearization must be contextually licensed. In detransitivizing languages (German, Greek), there is a bias for the transitive structure and the undergoer-first linearization must be contextually licensed.
Expectations from structure

Questions

Do the expected effects of referentiality and animacy differ dependent on (a) the type of the morphological psych alternation and (b) syntactic differences wrt exceptional/quirky experiencers?

In particular for (a):
- is there a frequency advantage of the base form (Haspelmath et al. 2014)?
- or are the effects of referentiality, animacy, and verb class similar across the psych alternation types?

In particular for (b):
- do languages with exceptional/quirky experiencers not (or to a lesser extent) use non-active forms (following comparable results in Verhoeven 2015, Lamers and de Hoop 2016)?
- or are the effects of referentiality, animacy, and verb class similar across the psych-verb types?
**Empirical study**

**Research Question:**
Do the differences in constituent structure and morphological type have an impact on discourse preferences?

<table>
<thead>
<tr>
<th>Non-canonical EO properties</th>
<th>Basis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese</td>
<td>–</td>
</tr>
<tr>
<td>Turkish</td>
<td>–</td>
</tr>
<tr>
<td>Greek</td>
<td>+</td>
</tr>
<tr>
<td>German</td>
<td>+</td>
</tr>
</tbody>
</table>

**Caveat:** The typological distinctions between languages relate to other types of data that are informative for scopal asymmetries. The idea is not to „validate“ these insights through corpus frequencies, but to figure out, whether they are *mapped* on discourse frequencies or not.
**Corpora**

**Chinese**  
*CCL Corpus, Beijing University*; 264 444 436 Modern Chinese characters, 84 127 123 Old Chinese characters;

**German**  
*W-öffentlich of COSMAS database, IDS*, 2.291.520.000 word forms;

**Greek**  
*Hellenic National Corpus (HNC), ILSP*, 47.000.000 word forms;

**Turkish**  
*TS Corpus, Taner Sezer, Mersin University*, 491.000.000 word forms;

**extracted**  
10 verbs for every verb class (two verb classes)  
250 tokens per verb (randomized)  
total 5000 sentences per language

**valid**  
declarative main clauses (with two arguments: sbj, obj, either lex. or pron.)
Fixed factors

**SEMANTIC AND PRAGMATIC FACTORS**

thematic roles
agent > causer > experiencer > stimulus > patient > ....

animacy
animate > inanimate
Prat-Sala & Branigan 2000, Prat-Sala et al. 2000, etc.

referentiality
pronoun > definite DP > indefinite DP
Verb classes

Canonical transitive verbs


(particular subclass of canonical verbs with include animacy configurations similar to EO verbs)

Experiencer-Object verbs

**Thematic role**

**ANNOTATING ACTOR AND UNDERGOER ARGUMENTS**

canonical transitive verbs

**ACTOR = agent; UNDERGOER = patient**

(1) *Τελικά, η Ευρώπη κατέστρεψε*  
finally the.NOM Europe.NOM destroyed  
την κυρία Θάτσερ.  
the.ACC Mrs.ACC Thatcher  
‘Finally, Europe destroyed Mrs Thatcher.’

experiencer object verbs

**ACTOR = stimulus; UNDERGOER = experiencer**

(2) *Το θέμα ενδιαφέρει την Ελλάδα ως ζήτημα αρχής: ...*  
the.NOM subject interests the.ACC Greece as matter principle  
‘The subject interests Greece as a matter of principle: ...’
Animacy

**ANNOTATING THE ANIMACY OF THE ARGUMENTS**

Animacy scale: animate > inanimate

**Disharmonic configuration (actor <\text{animacy} undergoer)**

1. **Greek:** actor=inanimate, undergoer=animate

   Τον Σαντ τον συγκινούσε ακόμα
   the.ACC PN the.ACC touched even
   έντονα η τέχνη της κηροπλαστικής, ....
   intensively the.NOM art.NOM the.GEN plastic.surgery.GEN
   ‘De Sade was affected even intensively by the art of plastic surgery, …’

**Other configuration (actor NOT <\text{animacy} undergoer)**

2. **Greek:** actor=animate, undergoer=animate

   Πάντως με εξέπληξε ο Μπάγεβιτς, ...
   anyway me.ACC surprised the.NOM PN
   ‘Anyway, I was surprised by Bajević, …’
**Animacy**

**ANNOTATING THE ANIMACY OF THE ARGUMENTS**

*Animacy scale: animate > inanimate*

**Disharmonic configuration (actor <\_\_\_animacy\_\_\_ undergoer)**

(1) Chinese: actor=inanimate, undergoer=animate

\[
\text{Shíchéng-de shàonán-xiǎohuǒmen jiù bèi huábān ... mízhù-le}
\]

Shicheng-ATTR young.fellows already BEI skateboard charm-PFV

‘the boys of Shicheng were already fascinated by skateboard ....’

**Other configuration (actor NOT <\_\_\_animacy\_\_\_ undergoer)**

(2) Turkish: actor=animate, undergoer=animate

\[
\text{Oğuz Ferit'e sinirlen-ir ...}
\]

O. F.-DAT upset-AOR:3.SG

‘Oğuz gets angry with Ferit ...’
Results

PREDICTIONS FROM CONSTITUENT STRUCTURE

Prediction 1: Constituent structure and discourse preferences
The choice of non-active voice is not motivated for languages that have morphologically downgraded experiencers with subject properties (German/Greek) – under the assumption that the surface order will reflect the fact that experiencers are higher in the syntactic structure.

Prediction 2: Morphological basis and discourse preferences
In transitivizing languages (Turkish), there is a bias for the intransitive structure, and the actor-first linearization must be contextually licensed. In detransitivizing languages (German, Greek), there is a bias for the transitive structure and the undergoer-first linearization must be contextually licensed.
Animacy and subject choice

Greek
\[ n = 1187 \]
\[ \text{glmer:} \]
verb, \( p \text{ n.s.} \)
anim, \( p < .001 \)
\( v^a, p \text{ n.s.} \)

Turkish
\[ n = 1054 \]
\[ \text{glmer:} \]
verb, \( p < .001 \)
anim, \( p < .01 \)
\( v^a, p \text{ n.s.} \)

German
\[ n = 1805 \]
\[ \text{glmer:} \]
verb, \( p < .001 \)
anim, \( p < .01 \)
\( v^a, p \text{ n.s.} \)

Chinese
\[ n = 1391 \]
\[ \text{glmer:} \]
verb, \( p < .01 \)
anim, \( p < .001 \)
\( v^a, p < .001 \)
Complementary role of order?

**Prediction 1: Constituent structure and discourse preferences**
The choice of non-active voice is not motivated for German/Greek – *under the assumption that the surface order will reflect the fact that experiencers are higher in the syntactic structure.*

**Observations:**

- **BETWEEN LANGUAGES:** Dominant surface order is *nominative-accusative* - also for German/Greek (which rejects the assumption of P1 and explains the frequency of non-active voice in G/G).

- **BETWEEN VERB CLASSES:**
  Experiencer-objects are more likely to occur first than patients in German/Greek.
Undergoer-first (non-active+OVS)

Greek
\(n = 1187\)

**glmer:**
verb, \(p \text{ n.s.}\)
anim, \(p < .001\)
\(v^a, p \text{ n.s.}\)

Turkish
\(n = 1054\)

**glmer:**
verb, \(p < .001\)
anim, \(p < .01\)
\(v^a, p < .01\)

German
\(n = 1805\)

**glmer:**
verb, \(p < .001\)
anim, \(p < .01\)
\(v^a, p < .01\)

Chinese
\(n = 1391\)

**glmer:**
verb, \(p < .01\)
anim, \(p < .001\)
\(v^a, p < .001\)
Conclusions

CROSS-LINGUISTIC PROPERTIES OF PROMINENCE SCALES

Typology
We did not find evidence for the typological prediction based on the differences in the morphological basis.

Cross-linguistic
Choice of subject is prominence-related and can be explained by effects of animacy (and referentiality). All languages independent of different morphosyntactic structures (transitivizing/detransitivizing; canonicity of EOs) display very similar effects.

Languages without syntactically prominent experiencers (Turkish/Chinese) show an interaction effect on the impact of the Verb Class (such that EO verbs may occur in non-active voice without contextual trigger).

This result is not visible for languages with syntactically prominent experiencers (German/Greek). For German, we found evidence that this is partly explained by the fact that Accusative-first order is used instead.
Syntactic variation and uniformity across languages:
A crosslinguistic corpus study on linearization devices

Special thanks are due to
Birgit Jänen, Julian A. Rott, Youjin Li, Yungang Zhang, Jiangling Zhang, Secil Sen, Nadire Biskin, and Dimitra Karmi
for their contribution to data annotation.