



Ghent University
New Ways of Analyzing Syntactic Variation
19.05.2016

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).

First through word order: Siewierska 1993, Branigan & Feleki 1999, Prat-Sala et al. 2000, Prat-Sala & Branigan 2000, Bornkessel et al. 2005, Grimshaw 1990, Haider 1993, Scheepers et al. 2000, etc.

First through subject choice: Bock & Warren 1985, Prat-Sala 1997, Ferreira 1994, Aissen 1999, Bresnan et al. 2001, Van Nice & Dietrich 2003, etc.

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:

animate $>$ _{likelihood to appear in spec,TopP} 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.).

highly identifiable in CG $>$ _{likelihood to appear in spec,TopP} 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**.

Relation to constituent structure

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.

- Topicalizing the lower argument of a canonical tr. active V



- Topicalizing the lower role of a passive V



Active and passive do not have identical extensions, but a significant overlap. They are only interchangeable to this extent.

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



EO-verbs may show exceptional syntactic properties:

- linearization
- passivization, extraction, binding, etc.

Belletti & Rizzi 1988, Pesetzky 1995, Haspelmath 2001, Reinhart 2002, Bayer 2004, Landau 2010, Verhoeven 2014, Temme & Verhoeven 2015, 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

(at least for accusative verbs)

Chinese
Turkish
Yucatec Maya
Korean

Typological difference in lexicon

languages having
a subclass of EO verbs with exceptional syntactic properties

yes

no

(at least for accusative verbs)

German

Greek

Islandic

Italian

Chinese

Turkish

Yucatec Maya

Korean

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 *mízhù* y 'x attracts y'

y *bèi* x *mízhù* 'y is attracted by x'



(b) Greek mediopassive

x *endīaféri* y 'x interests y'

y *endīaférete ja* x 'y is interested in x'



(c) German reflexive, stative passive

x *ärgert* y 'x annoys y'

y *ärgert sich über* x 'y is annoyed by x'



BASIS: EXPERIENCER-SUBJECT

(a) Turkish causativization

y x *sevin-di* 'y is happy about x'

x y *sevin-dir-di* 'x makes y happy'



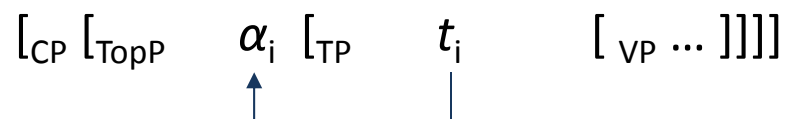
s. Nichols et al. 2004

typology of detransitivizing vs. transitivizing languages,

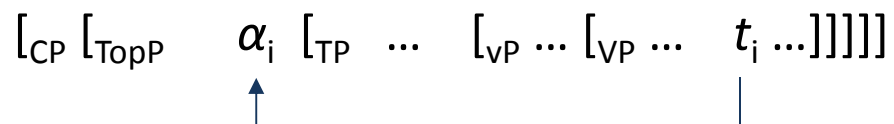
Intransitive morphological bases

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



- Causative

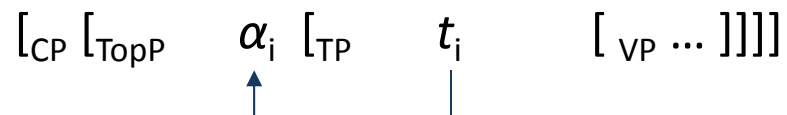


the intransitive structure should occur more frequently based on derivational cost AND nominative-first. While in German/Greek the lower-first linearization must be contextually licensed, in languages like Turkish, it is the actor-first linearization that must be contextually licensed.

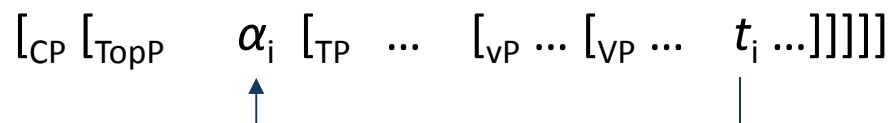
Intransitive morphological bases

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



- Causative



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?

	Non-canonical EO properties	Basis
Chinese	–	<i>tr</i>
Turkish	–	<i>intr</i>
Greek	+	<i>tr</i>
German	+	<i>tr</i>

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 >

Jackendoff 1987, Grimshaw 1990, Van Valin & LaPolla (1998:127), Primus 1999, Bresnan 2001, etc.

animacy

animate > inanimate

Silverstein 1976, Siewierska 1993, Dahl & Fraurud 1996; Comrie 1981, Branigan & Feleki 1999, Prat-Sala & Branigan 2000, Prat-Sala et al. 2000, etc.

referentiality

pronoun > definite DP > indefinite DP

Givón (ed.) 1983, 1994, Gundel et al. 1993, Fraurud 1990, Bickel 2008, Bickel et al. 2015, etc.

Verb classes

Canonical transitive verbs

διαλύω 'damage', καταστρέφω 'destroy', δηλητηριάζω 'poison', προειδοποιώ 'warn', εμποδίζω 'hinder', προστατεύω 'protect', βελτιώνω 'improve', σώζω 'rescue', etc.

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

Experiencer-Object verbs

ενδιαφέρω 'interest', στενοχωρώ 'sadden', ενθουσιάζω 'inspire', ενοχλώ 'annoy', απογοητεύω 'disappoint', τρομάζω 'frighten', εντυπωσιάζω 'impress', etc.

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** <_{animacy} **undergoer**)

(1) Greek: **actor=inanimate**, **undergoer=animate**

<i>Τον</i>	<i>Σαντ</i>	<i>τον</i>	<i>συγκινούσε</i>	<i>ακόμα</i>
the.ACC	PN	the.ACC	touched	even
<i>έντονα</i>	<i>η</i>	<i>τέχνη</i>	<i>της</i>	<i>κηροπλαστικής,</i>
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 <_{animacy} **undergoer**)

(2) Greek: **actor=animate**, **undergoer=animate**

<i>Πάντως</i>	<i>με</i>	<i>εξέπληξε ο</i>	<i>Μπάγεβιτς, ...</i>
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**

Shíchéng-de shàonán-xiǎohuǒmen jiù bèi huábǎn ... 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**

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 transitivity 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

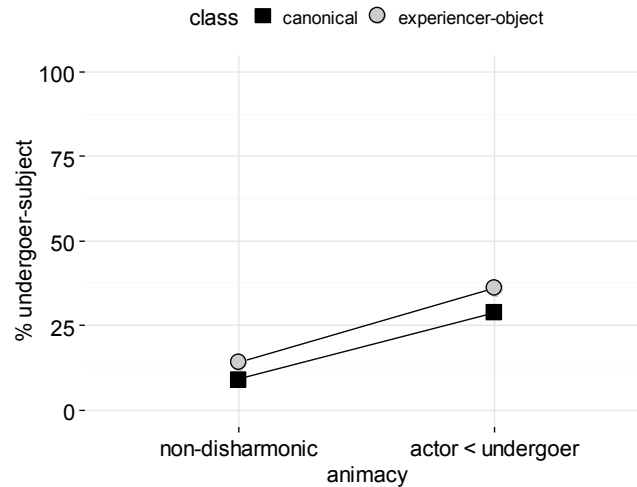
$n = 1187$

glmer:

verb, p n.s.

anim, $p < .001$

$v^{\wedge}a$, p n.s.



German

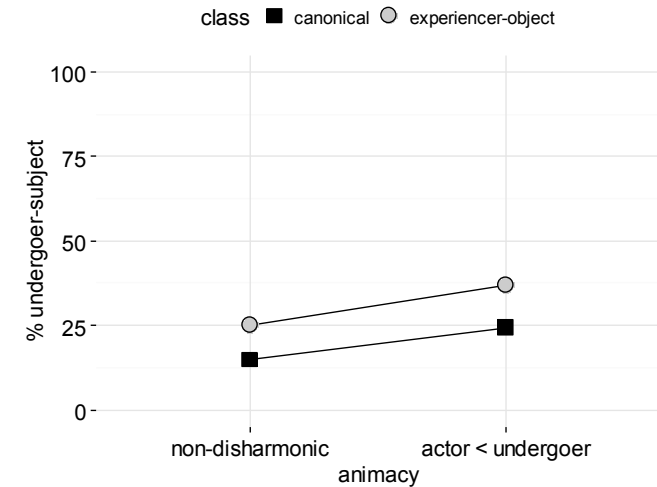
$n = 1805$

glmer:

verb, $p < .001$

anim, $p < .01$

$v^{\wedge}a$, p n.s.



Turkish

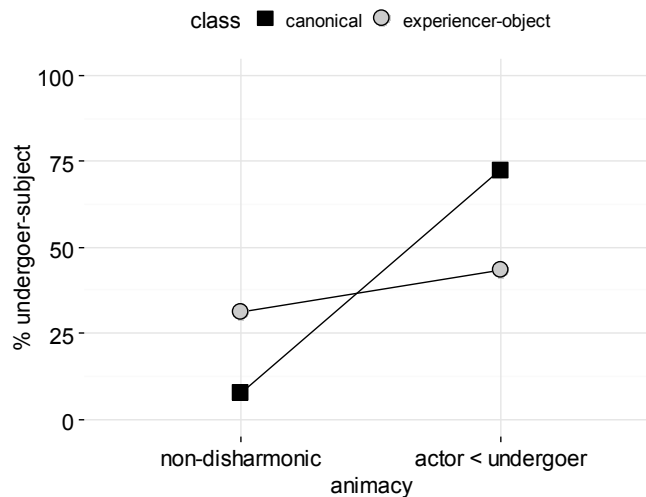
$n = 1054$

glmer:

verb, $p < .001$

anim, $p < .001$

$v^{\wedge}a$, $p < .001$



Chinese

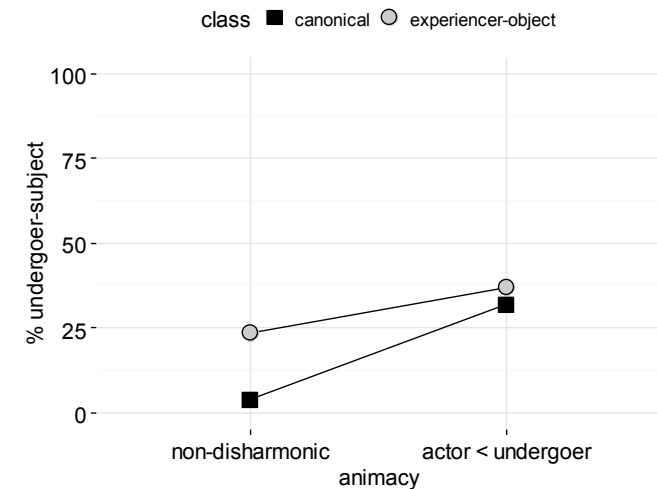
$n = 1391$

glmer:

verb, $p < .01$

anim, $p < .001$

$v^{\wedge}a$, $p < .001$



Complementary role of order?

RECALL

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.*

lang	class	SO	OS	%
		<i>n</i>	<i>n</i>	
Chinese	Canon	351	0	0
	Exp-O	220	0	0
German	Canon	138	9	6.1
	Exp-O	146	35	19.3
Greek	Canon	213	10	4.5
	Exp-O	158	12	7.1
Turkish	Canon	221	3	1.3
	Exp-O	210	2	0.9

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

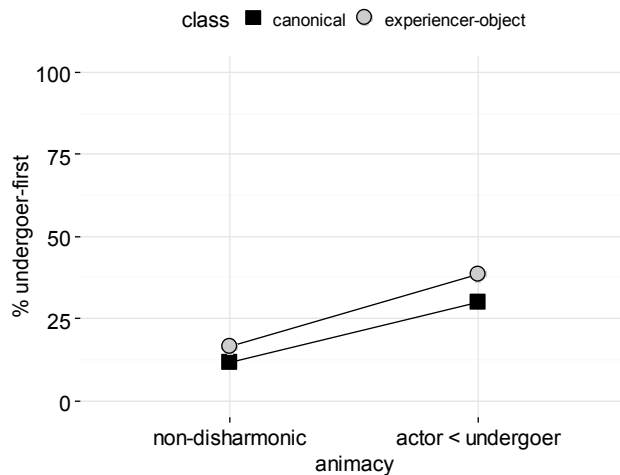
$n = 1187$

glmer:

verb, p n.s.

anim, $p < .001$

v^a , p n.s.



German

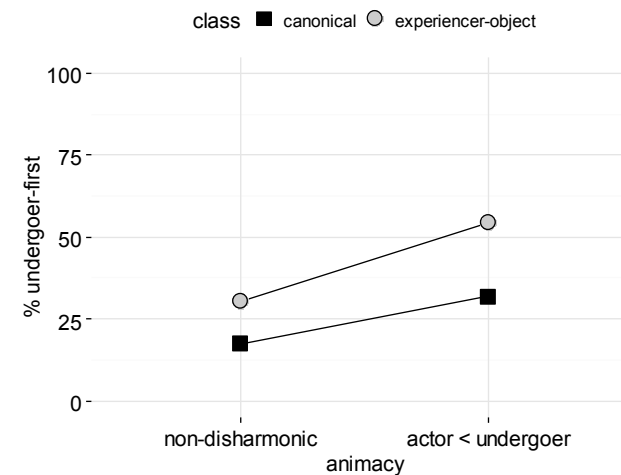
$n = 1805$

glmer:

verb, $p < .001$

anim, $p < .01$

v^a , $p < .01$



Turkish

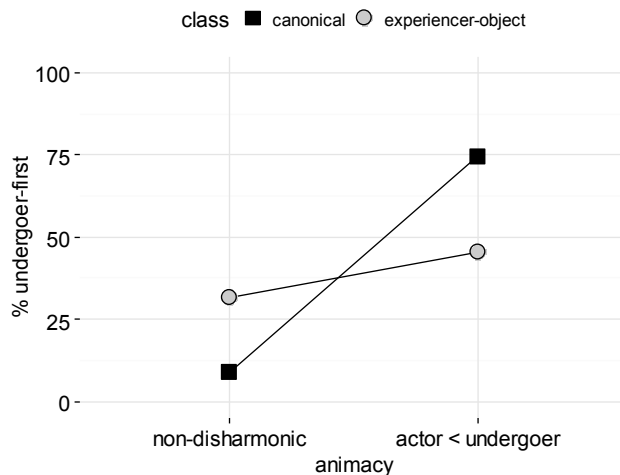
$n = 1054$

glmer:

verb, $p < .001$

anim, $p < .001$

v^a , $p < .001$



Chinese

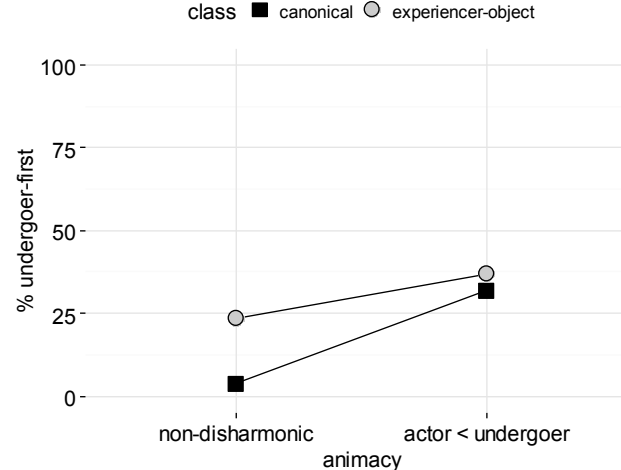
$n = 1391$

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.



Ghent University
New Ways of Analyzing Syntactic Variation
19.05.2016

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.