Radboud Universiteit

Using Universal Dependencies for testing hypotheses about communicative efficiency

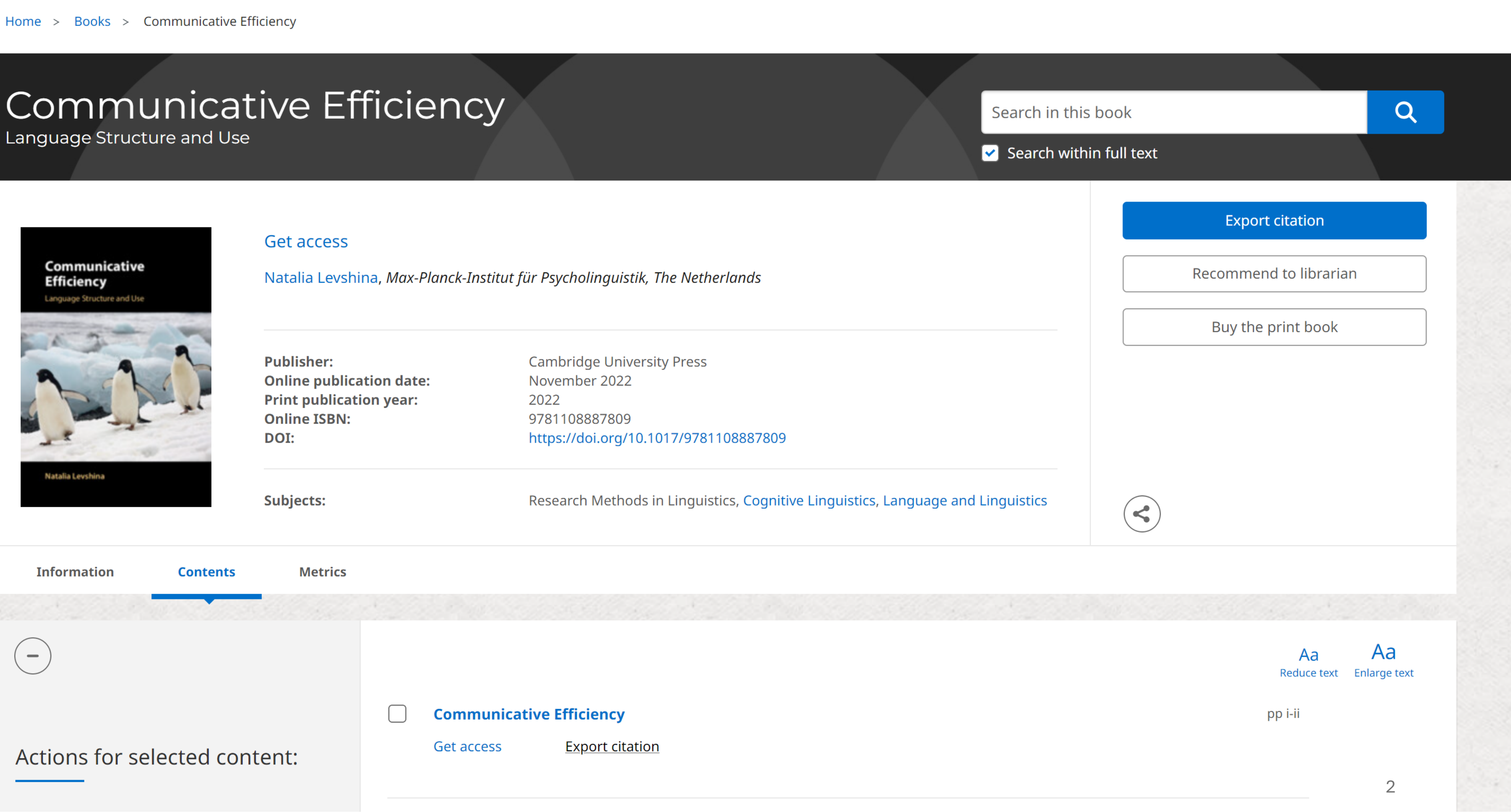
Natalia Levshina

Centre for Language Studies

Radboud University

Joint Workshop MWE-UD

Torino, May 25 2024



Home > Books > Communicative Efficiency

Communicative Efficiency

Language Structure and Use

Search in this book 🔍

☑ Search within full text

Get access

Natalia Levshina, *Max-Planck-Institut für Psycholinguistik, The Netherlands*

Publisher: Cambridge University Press
Online publication date: November 2022
Print publication year: 2022
Online ISBN: 9781108887809
DOI: https://doi.org/10.1017/9781108887809

Subjects: Research Methods in Linguistics, Cognitive Linguistics, Language and Linguistics

Export citation

Recommend to librarian

Buy the print book

Information Contents Metrics

Aa Reduce text Aa Enlarge text

Actions for selected content:

What is efficiency?

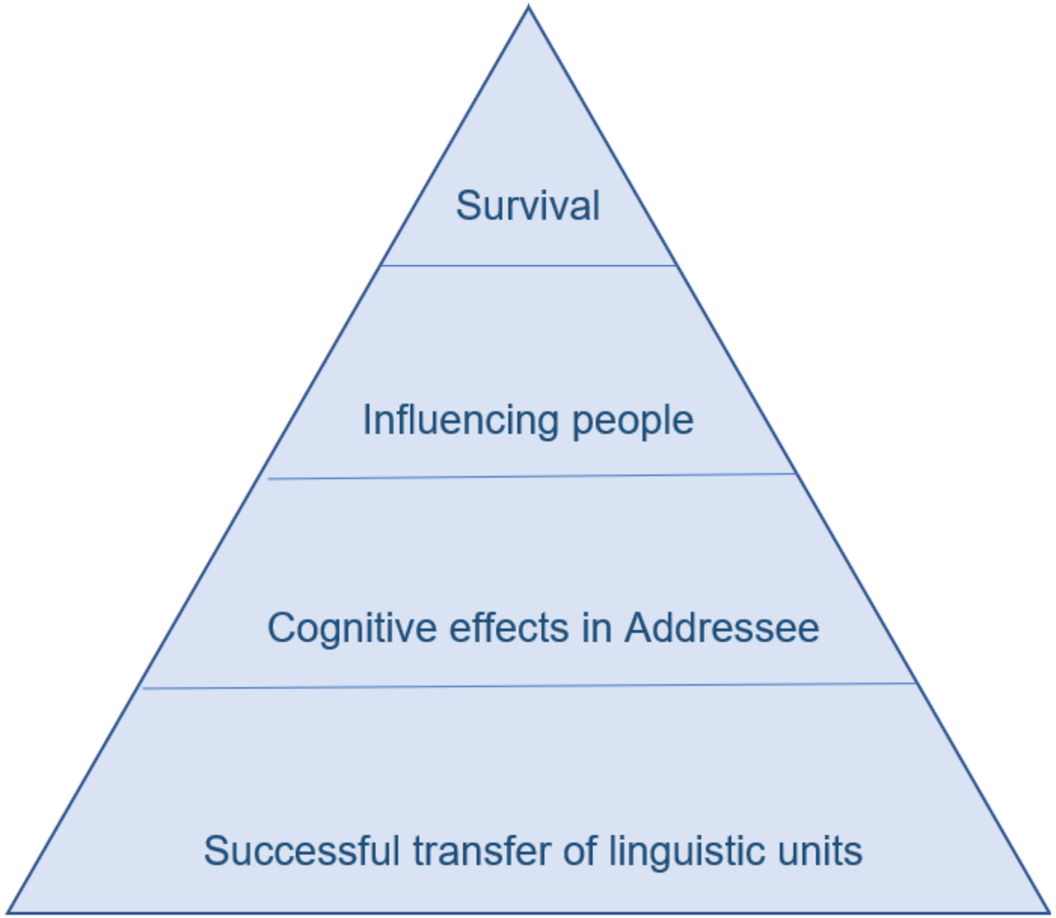
- Efficiency means minimization of a cost-to-benefit ratio. Being efficient means not spending more effort than necessary in order to achieve something.

- Living organisms try to save effort:
  - Penguins waddle because it conserves energy in comparison with walking.
  - Professional runners position their heels in such a way as to lower metabolic energy consumption.

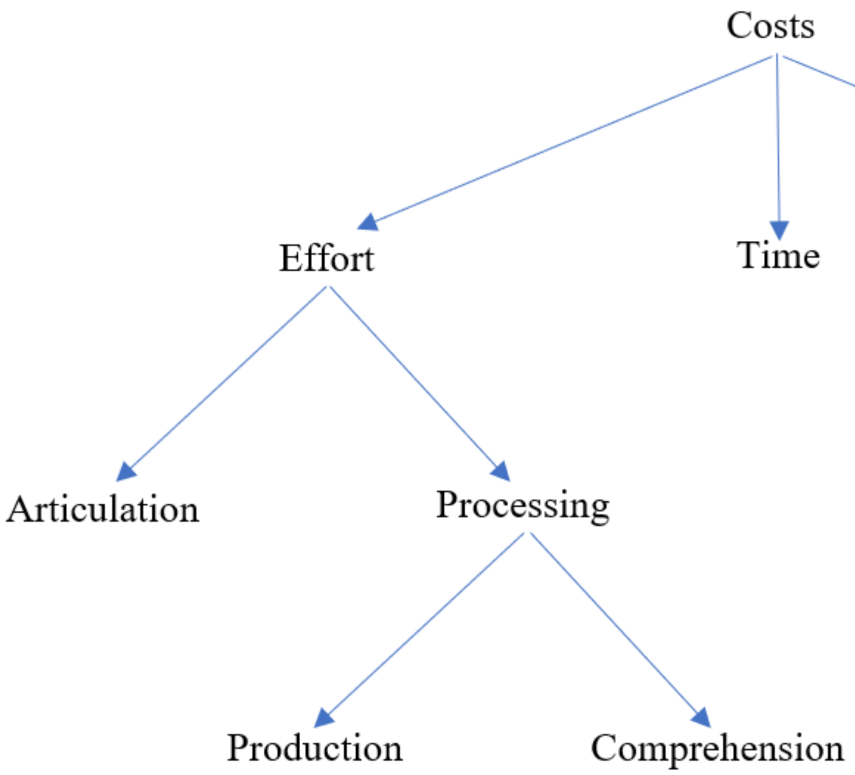
Benefits of linguistic communication

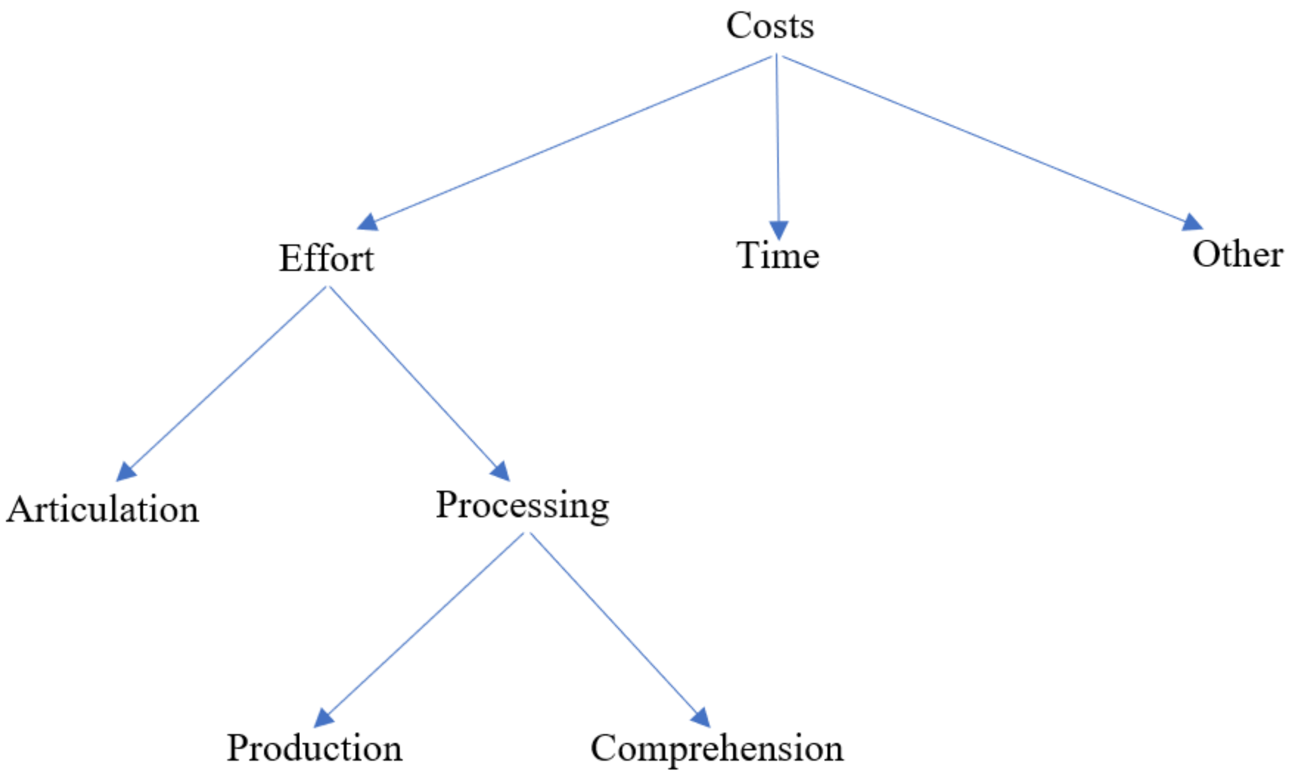
Jakobson's functions of language

- *referential:* describes a situation, object or mental state.
- *poetic*: "the message for its own sake"
- *emotive:* give information about the speaker's internal state
- *conative:* engages the Addressee directly
- *phatic*: language for the sake of interaction
- *metalinguistic*: the use of language


Survival
Influencing people
Cognitive effects in Addressee
Successful transfer of linguistic units

Costs of linguistic communication





Principles of efficient communication

- **Positive correlation between benefits and costs**
  - Don't spend effort and time on useless information
  - Extra costs should be justified by extra benefits

- **Negative correlation between accessibility and costs**
  - Spend less effort and time on more accessible (predictable, known, stereotypical, etc.) information
  - Spend more effort and time on less accessible information

- **Maximization of accessibility**
  - Minimize surprisal
  - Produce more accessible information first

Cross-linguistic evidence: illustrations

- Negative correlation between accessibility and costs:
  - More formally marked grammatical categories are less frequent. E.g., SG *book* vs. PL *book**s***.
  - Differential object marking when low P (ObjectRole|Feature). E.g., Spanish *Veo **a** la actriz* 'I see the actress'.
  - Causatives that express less frequent causation meanings are expressed by longer forms. E.g., *Harry Potter **caused** the cup **to rise***.
- Maximization of accessibility:
  - Subject-first preference
  - Dependency length minimization
  - Avoidance of crossing dependencies

Greenberg 1963, 1966, Hawkins 2004, Ferrer-i-Cancho 2006, Liu 2008,
Futrell et al. 2015, Haspelmath 2021, Yadav et al. 2021, Levshina 2022 and many others

Principles of efficient communication

- **Positive correlation between benefits and costs**
  - Don't spend effort and time on useless information
  - Extra costs should be justified by extra benefits

- **Negative correlation between accessibility and costs**
  - Spend less effort and time on more accessible (predictable, known, stereotypical, etc.) information
  - Spend more effort and time on less accessible information

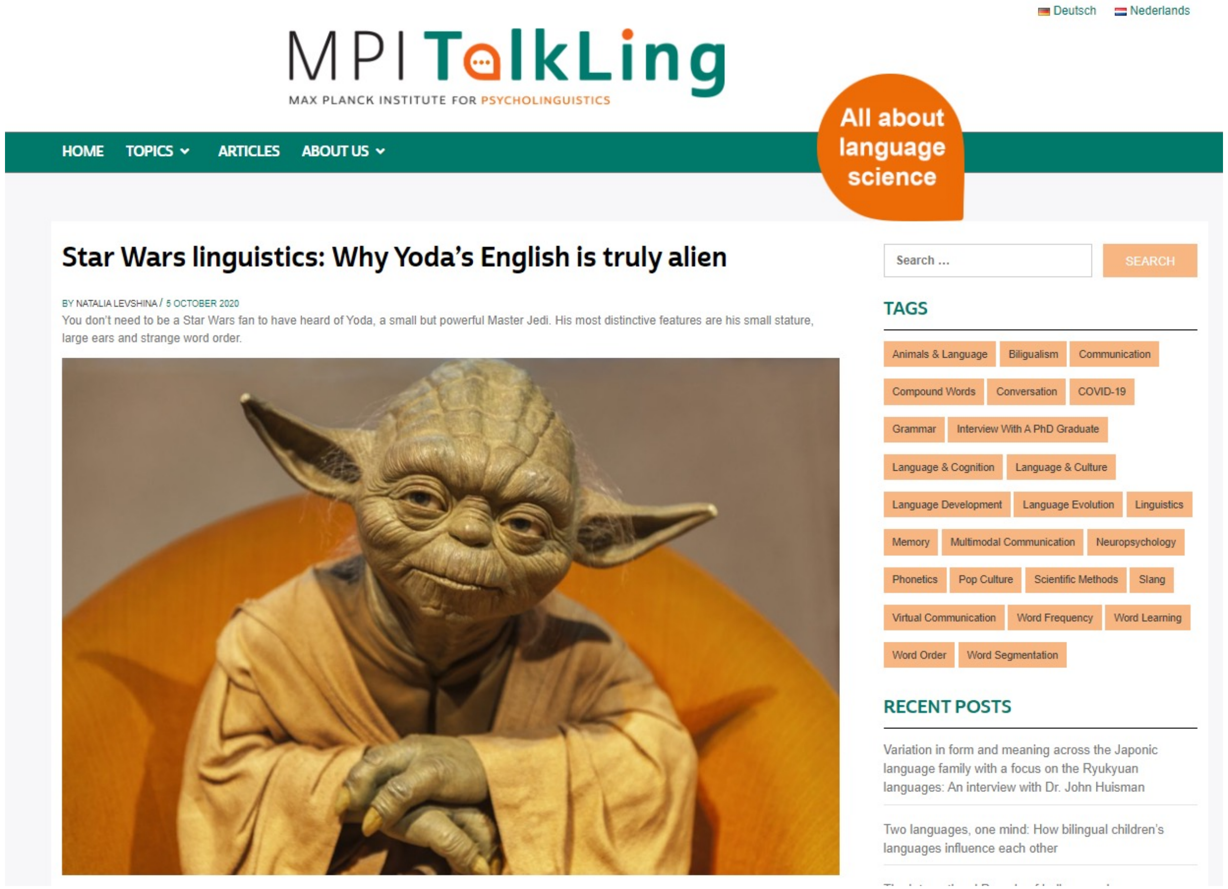
- **Maximization of accessibility**
  - Minimize surprisal
  - Produce more accessible information first

Example of an exception: Yodish

- *Hard to see, the dark side is.*
- *Friends you have there.*
- *Help you it will.*

The costs of processing Yodish are high, but there are extra benefits! (See first principle)

Levshina 2019 *SyntaxFest*



Star Wars linguistics: Why Yoda's English is truly alien

BY NATALIA LEVSHINA / 5 OCTOBER 2020

You don't need to be a Star Wars fan to have heard of Yoda, a small but powerful Master Jedi. His most distinctive features are his small stature, large ears and strange word order.

Variation in form and meaning across the Japonic language family with a focus on the Ryukyuan languages: An interview with Dr. John Huisman

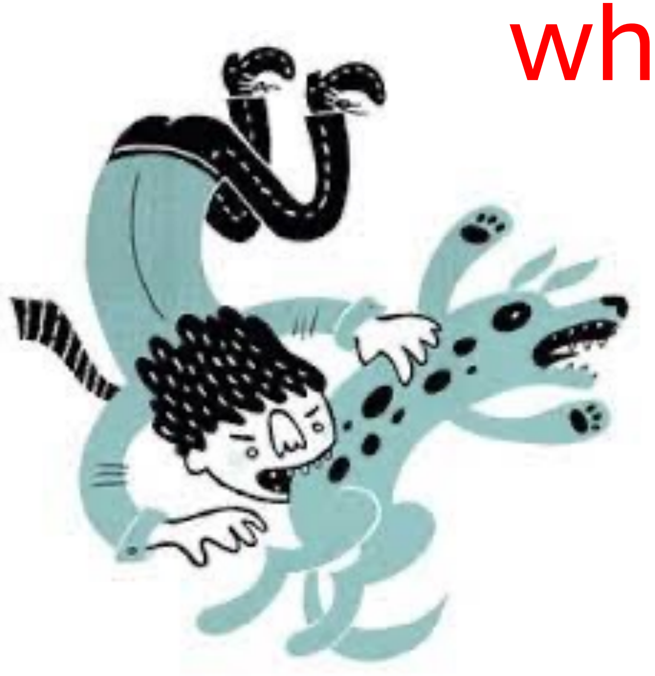
Two languages, one mind: How bilingual children's languages influence each other

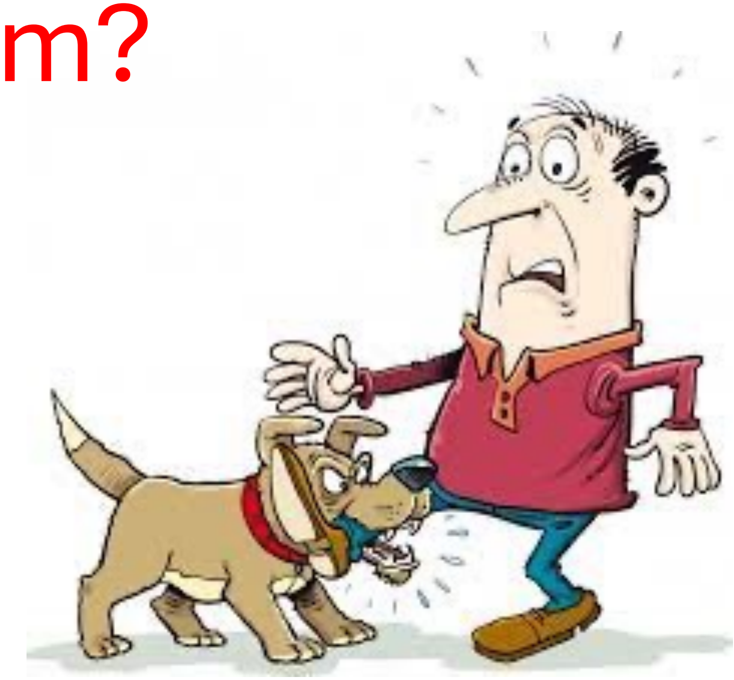
https://www.mpi-talkling.mpi.nl/?p=63&lang=en

A case study: Cues to A and P

(aka Subject and Object in many languages, as well as in UD)

Who did what to whom?





(man, dog, bite)

Images from https://dogdailynews.wordpress.com/2009/09/28/man-bites-dog/, https://www.seagull-tandem.eu/portfolio/b1-dog-bites-man/

Cues to A and P roles

- Case and agreement (German, Latin, Russian, Spanish)
- Rigid word order of core arguments (English, Mandarin Chinese)
- Semantics
  - categorical restrictions: Jakaltek (Mayan) and Halkomelem (Salishan) strictly exclude inanimate Subjects in transitive clauses (Aissen 2003)
  - probabilistic constraints: inanimate arguments are more likely to be Objects than Subjects
  - probabilistic constraints: encyclopaedic knowledge of typical frames and scenarios (Kurumada & Jaeger 2015)
- POS, person, information status… (Levshina 2021 *Ling Van*)

Corpora annotated with Universal Dependencies

+

Communicative Efficiency Theory

Hypothesis 1

- If language users and structures are efficient, we can expect a negative correlation between

  a) the rigidity of subject and object order in a transitive clause

  and

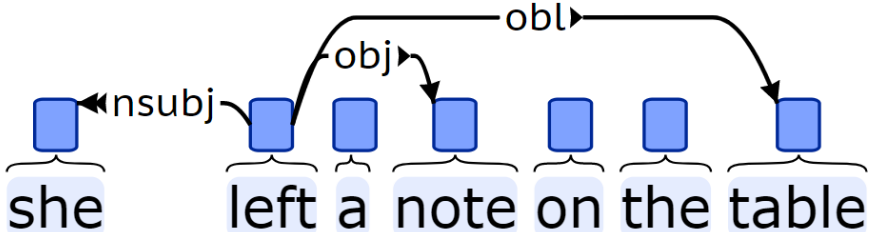
  b) the use of disambiguating case marking

- Why? The principle of negative correlation between accessibility and costs: if the word order is rigid enough to make the roles accessible, then we don't need to waste time and effort on case markers.

An online news dataset

- 30 online news corpora, 1M sentences each, from the Leipzig Corpora Collection (Goldhahn et al. 2012)

- Annotated with UDPipe (R package udpipe by Wijffels, Straka & Straková 2018)

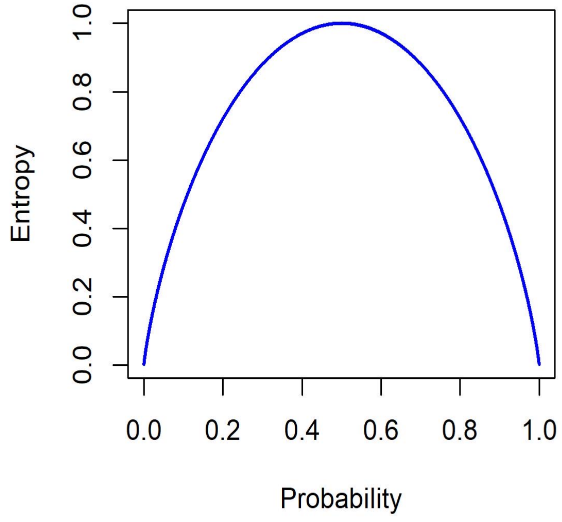


Subject - Object order rigidity

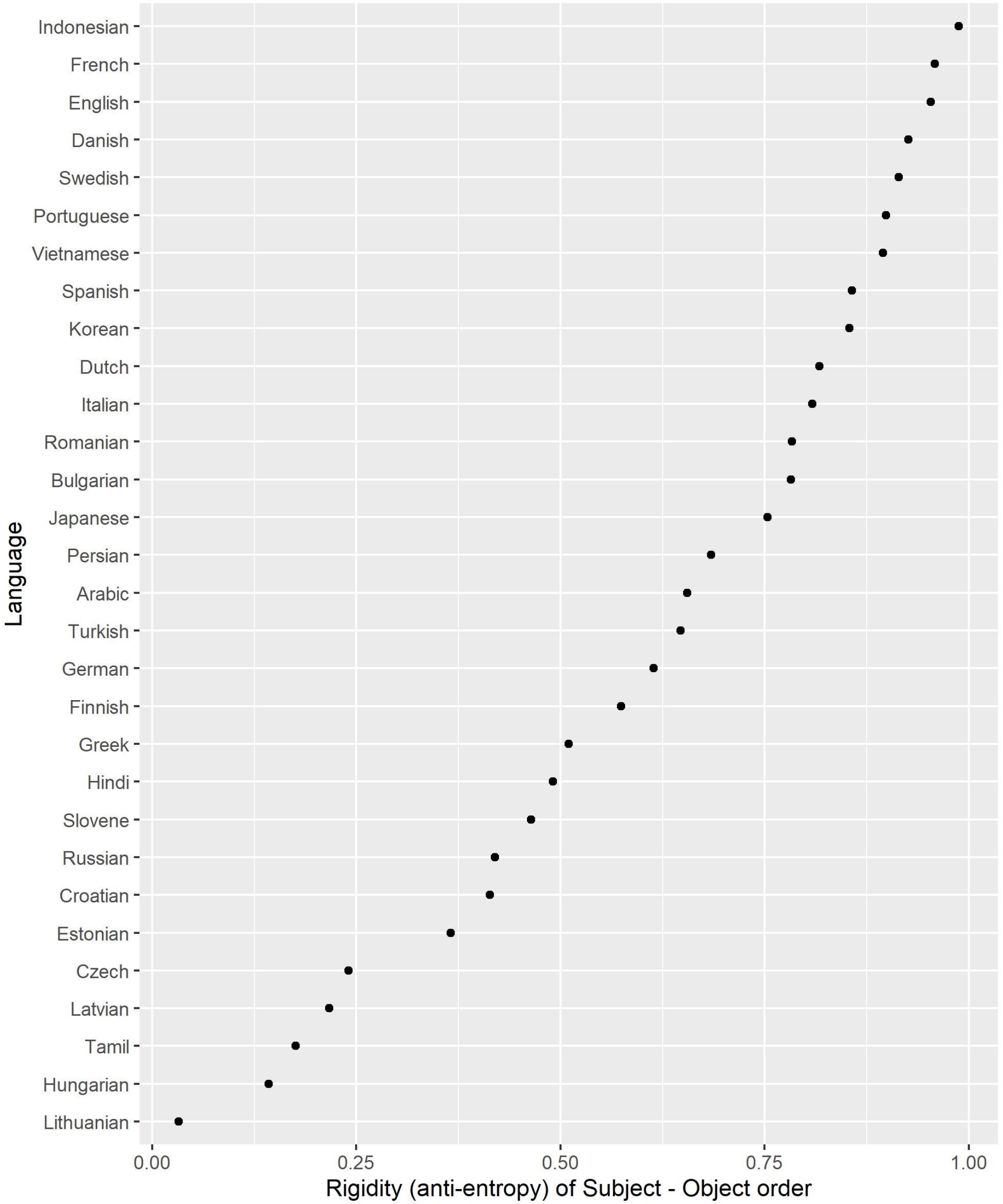
- Proportions of nsubj + obj and obj + nsubj (only common nouns) in a transitive clause

- The higher entropy H, the greater the variability

$$H(X) = -\sum_{i=1}^{2} P(x_i)\, log_2\, P(x_i)$$



- Rigidity is measured as 1 – H



Levshina 2021 *Front Psych*

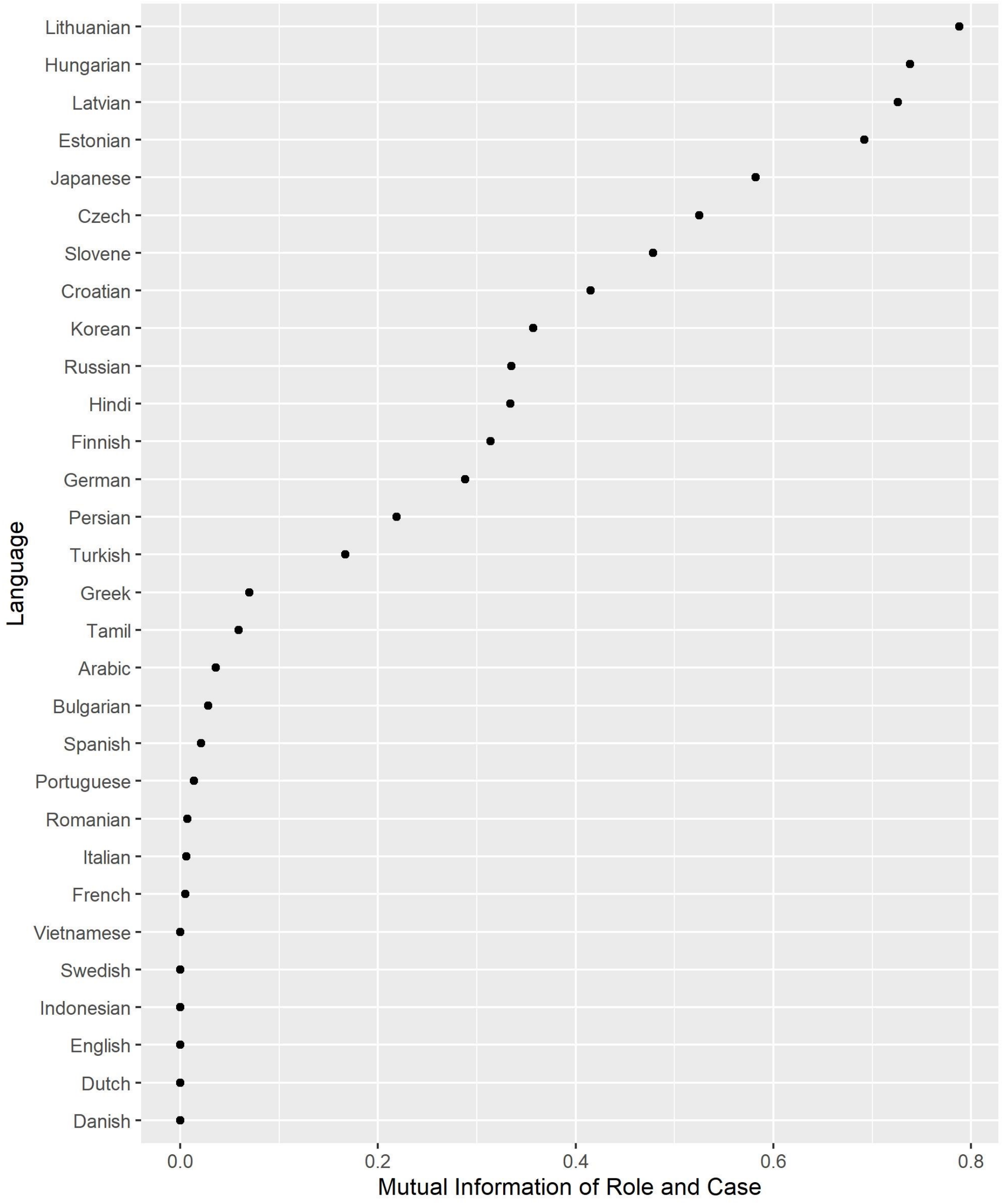
The role of nominal case in A and P disambugation

- Mutual Information of case forms and Subject/Object roles (only nominals)
- Example: Spanish

Case	Subject	Object
Zero marking	126,736	569,252
Preposition *a*	0	55,422

- No case differences: MI = 0
- Languages with morphological marking: Smaller samples of Subjects and Objects were analyzed manually, then the results were extrapolated, and MI were computed.

Levshina 2021 *Front Psych*



Levshina 2021 *Front Psych*

How to test typological hypotheses correctly?

- Method 1: Sampling one language per Genus/Family and geographic Area

- Method 2: Mixed-Models regression with Genus/Family and Area as random effects

- Method 3 (SOTA): Phylogenetic regression with genealogical trees and geographic distances as random effects (variance and covariance matrices)

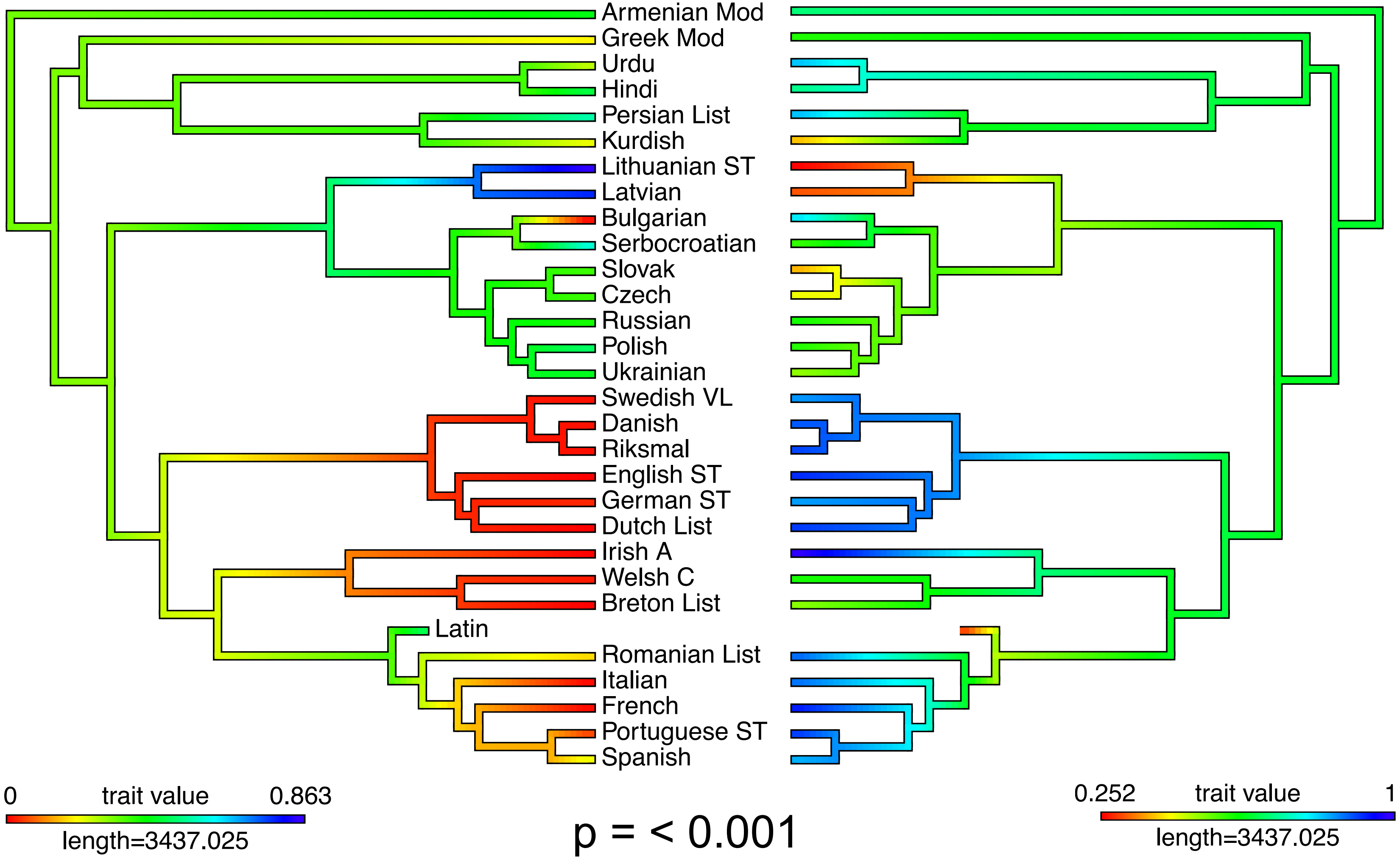
Hypothesis 1: Results

Sampling Method	Data	Effect size	l-95% CI	u-95% CI	Interpretation
Sampling from every genus 1K times	Ranked data	r = -0.67	*-0.67*	-0.66	Confirmed
Genera as random intercepts	Original data (beta)	β = -3.58	-5.09	-2.03	Confirmed
	Ranked data (Gaussian)	β = -0.81	-1.04	-0.58	
Genealogical trees and geographic distances as random effects	Original data (beta)	β = -4.05	-5.47	-2.52	Confirmed
	Ranked data (Gaussian)	β = -0.83	-0.99	-0.65	

Indo-European languages: CIEP+ corpus

Case

Rigid WO



p = < 0.001

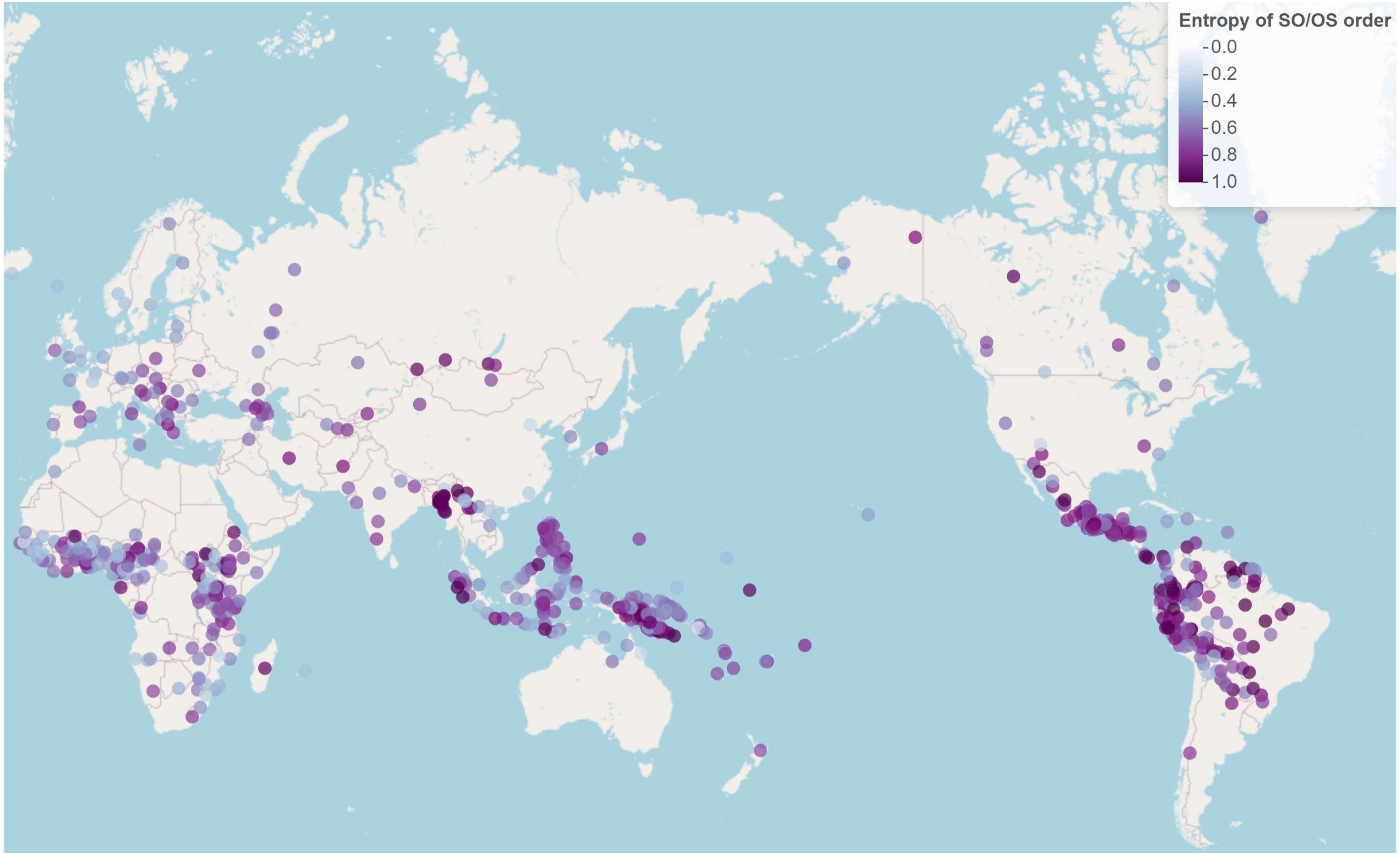
Larger dataset

- Robert Östling's (2015) multilingual alignment of New Testament translations
    - Order of nominal Subject, nominal Object and Verb
    - Sum frequency of all possible orders > 10
    - 954 unique ISO-639-3 codes

- Case marking: Yes or No (reference grammars and typological databases like WALS and Grambank).

- 689 languages in total

Entropy of Subject and Object order in NT



Phylogenetic regression

- Entropy ~ Case

- Bayesian Beta regression

- Weak generic priors

- Case=Yes vs. Case=No: β = 0.33, 95% CI 0.10 to 0.55.

- The hypothesis is confirmed again!

Hypothesis 1: Summary

- Regardless of the statistical method, typological data or dataset, the correlation between case and rigid word order remains robust.

- Languages are efficient in that regard.

Hypothesis 2

- Similar to Hypothesis 1, but instead of case marking, we test verb agreement.

- If a language has rigid word order, is it less likely to use verb agreement for disambiguation.

Ongoing project

- Althea Löfgren (PhD candidate, Paris Nanterre)
- Disambiguating effect of Verb agreement in the same sample of languages.
- Samples of 100 clauses with nominal nsubj and obj and verbal main clauses, retrieved from SUD corpora.
- Manually annotated: in how many clauses does the verb form help to disambigate between subject and object?
    - The dog chases the cat. NO
    - The dog chases the cats. YES (Number information)
    - Disambiguation index: proportion of clauses in which the verb form actually allows to tell who did what to whom.

Preliminary results

- Phylogenetic beta regression
- A negative correlation between disambiguation index and rigid order:
  - β = -2.32, 95% CI -4.98 to 0.14, but posterior P(β < 0) = 0.968.
- Note that subject agreement is extremely common (Siewierska 2013), but there is no consensus about its functional and discourse origins.
  - Different proposals, e.g., Givón 1976, Ariel 2000, Schell 2018.
- Next steps:
  - We need more languages with object agreement.
  - We should use conversational data to have representative frequencies of different persons as A and P. Our data: only 3rd person.

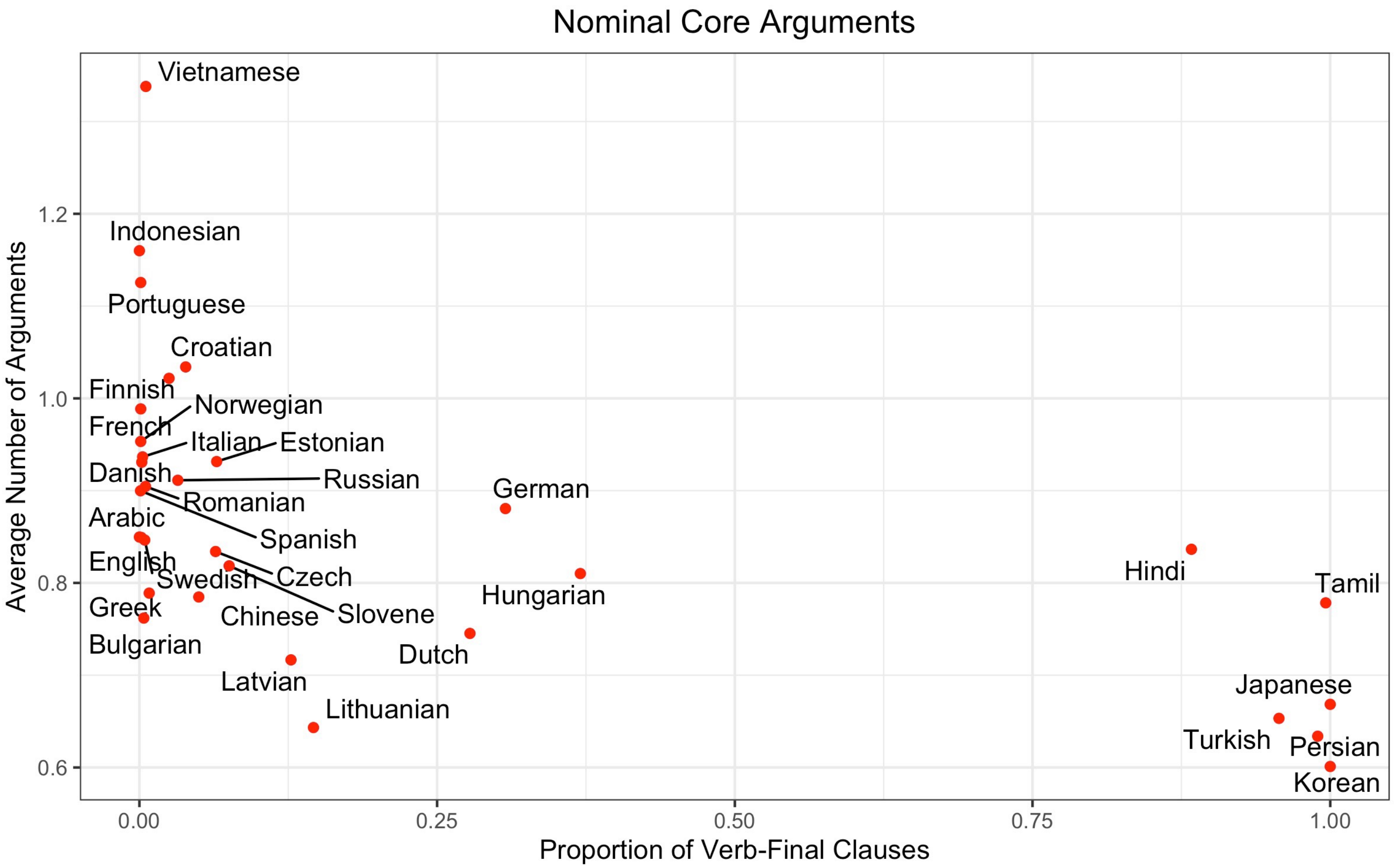
Hypothesis 3

- When the verb comes late, the processing costs required for keeping longer dependencies in mind are higher (cf. Ueno & Polinsky 2009).
- So it is more efficient to use fewer arguments in verb-final languages: either drop them arguments or use intransitive constructions.
- This is a way of maximizing accessibility.
- We can expect a negative correlation between the following variables:
  - relative frequency of verb-final clauses
  - average number of overt core arguments in a main clause

Data

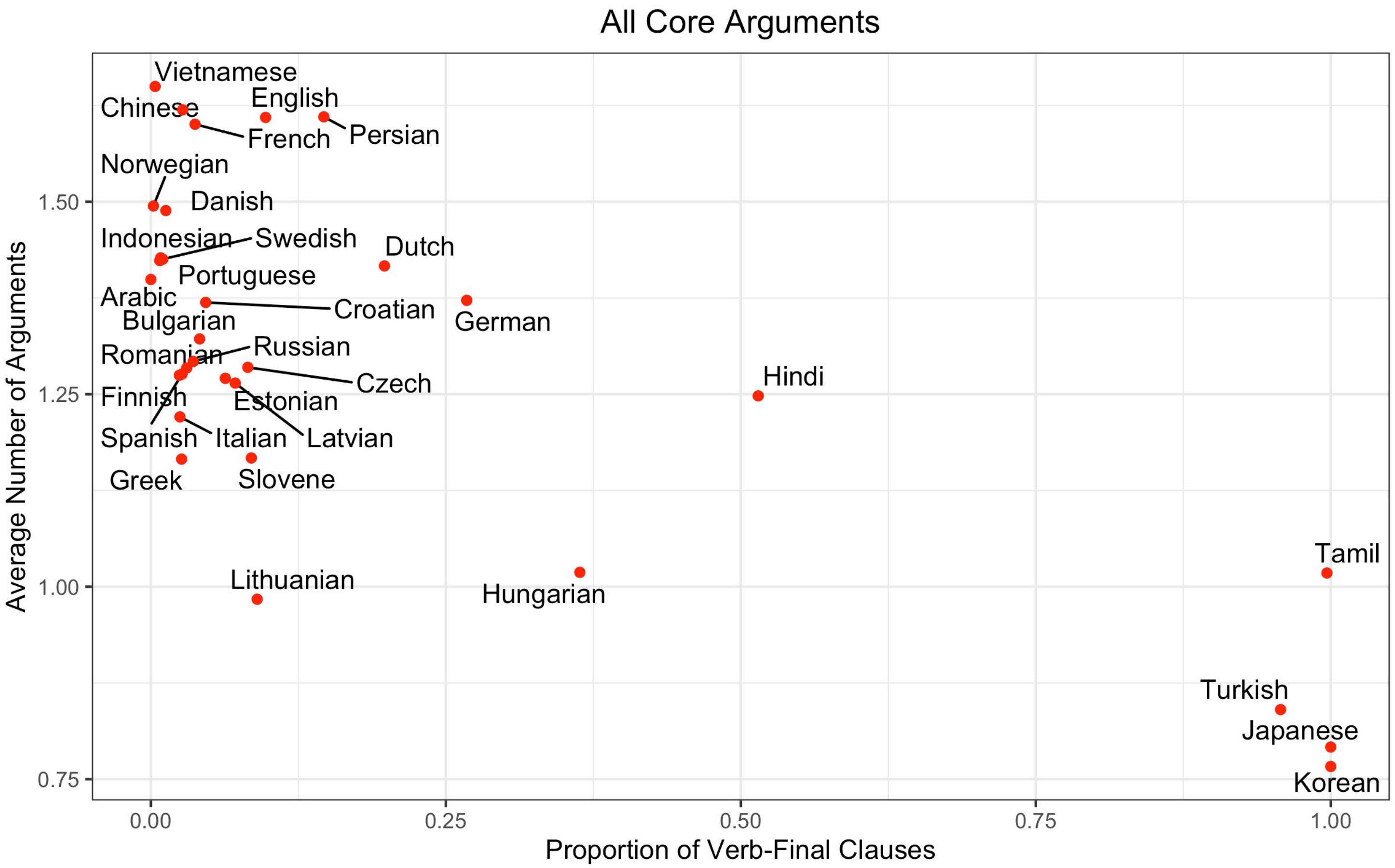
- 32 online news corpora from the Leipzig corpora collection. Important to control for register!
- Two approaches:
  - Nominal core arguments only
  - Any core arguments (nominal, pronominal, clausal complements)
- Variables:
  - Relative frequencies of verb-final clauses wrt. all verbal main clauses
  - The average number of core arguments per clause (nsubj and obj only, or also csubj, obj, xcomp, ccomp).

Nominal core arguments



Phylogenetic LMM β = -0.28, 95% CI [-0.44, -0.13]
Bayesian R^2 = 0.83 95% CI [0.47, 0.99]

All core arguments



Phylogenetic LMM β = -0.59, 95% CI [-0.80, -0.38]
Bayesian R^2 0.84, 95% CI [0.65, 0.99]

Conclusions and new questions

- We find support for the predictions based on Communicative Efficiency Theory:
  - Rigid word order → less disambiguating case marking
  - Rigid word order → less disambiguating agreement marking (only 3rd person core arguments!)
  - More verb-final clauses → fewer core arguments (is it due to pro-drop or use of intransitive strategies? Another ongoing project...)
- But we shouldn't forget that there is also counterevidence:
  - Levshina (2021) finds a positive correlation between case marking and MI of lexemes and roles → redundancy!
- Communicative efficiency is only part of the big picture...

Many thanks! Vielen Dank! Dank U wel! Spasibo!

natalia.levshina@ru.nl