On potential statistical universals of grammar in discourse: Evidence from Multi-CAST

Geoffrey Haig,[1] **Nils N. Schiborr,**[1] **Stefan Schnell**[1,2]

[1] *University of Bamberg,* [2] *Centre of Excellence for the Dynamics of Language*

DGfS2020, Hamburg

Workshop *Corpus-based typology:*

Spoken language from a cross-linguistic perspective

4–6th March 2020

Overview

1. Corpus-based typology with Multi-CAST
2. Cross-linguistic uniformity in the distribution of full (‘lexical’) expressions
3. Light Human Subjects
4. The subject/object person asymmetry
5. Conclusions

(1) Corpus-based typology with Multi-CAST

Traditional research in discourse & grammar

- roots in the functionalist tradition: Chafe, Givòn, Prince, Du Bois, among many others
- ‘grammar’ shaped and constrained by demands of successful communication, rather than an autonomous module
- explicitly cross-linguistic, empirical perspective
- remains hugely influential, e.g. in Cognitive Grammar, Grammaticalization

Corpus-based typology

- couples the functionalist tradition with digital corpora, and methodologies from corpus linguistics and variationist sociolinguistics (Schnell & Barth 2018)
- complements grammar-based, or “data-reduction” typology (Wälchli 2009)
- bottom-up, data-driven, probabilistic rather than categorial generalizations
- attends to variation, attends to context
- in Multi-CAST: focus on spoken language, monologic, indigenous content, sample breadth rather than corpus breadth

Multi-CAST

Multilingual Corpus of Annotated Spoken Texts

Multi-CAST, the *Multilingual Corpus of Annotated Spoken Texts*, is a collection of annotated texts from a typologically diverse section of languages.

- multiple levels of parallel annotations, time-aligned with audio recordings,
- including comparative morphosyntactic annotations for cross-corpus typological research
- chiefly monologic, natural narrative texts from twelve languages, encompassing roughly 21 500 clause units
- available in multiple file formats, including as EAF files for the linguistic annotation software ELAN, as XML and TSV files, and via the *multicastR* package for R
- freely accessible under a CC-BY 4.0 licence

Getting started with Multi-CAST

collection overview (!)	*PDF*	355 KB	v2.1	20/01/12	*archive*
research context	*PDF*	337 KB	v1.1	18/05/25	*archive*
full collection	*XML*	39 MB	2001	20/01/12	*archive*
	TSV	6.9 MB	2001	20/01/12	*archive*
full metadata	*TSV*	4 KB	2001	20/01/12	*archive*

Citing Multi-CAST

Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts.* (multicast.aspra.uni-bamberg.de/) (date accessed)

(2) The cross-linguistic uniformity in the use of full (“lexical”) expressions

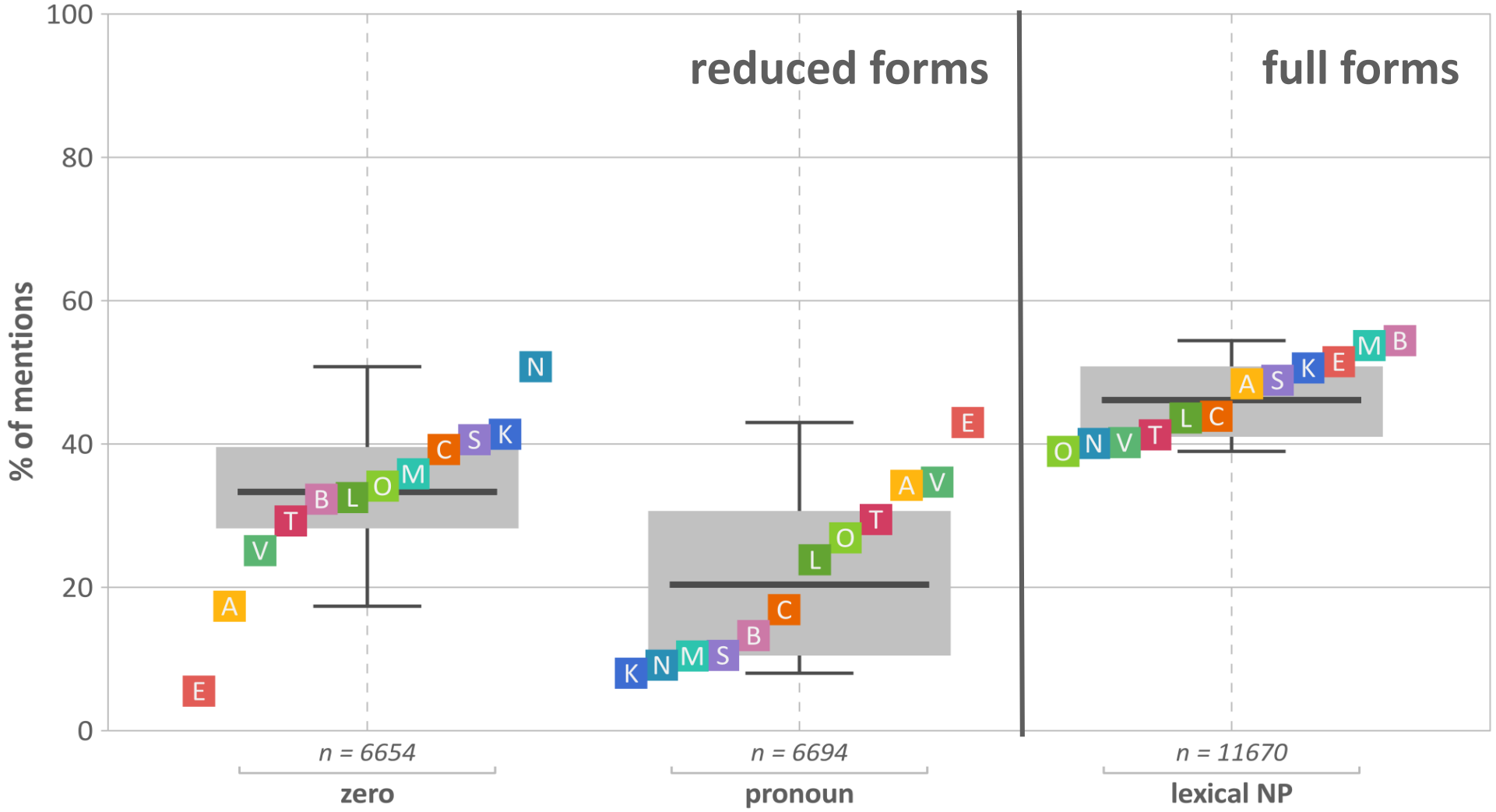
Lexical versus reduced (Kibrik 2011) forms of referring expressions

lexical ‘full’ expressions / lexical NPs:

A new syntax professor / Amanda / that woman / the supervisor ...

reduced / ‘light’ forms:

she / her / Ø


reduced forms
full forms
% of mentions
100
80
60
40
20
0
n = 6654
zero
n = 6694
pronoun
n = 11670
lexical NP
A Arta
C Cypriot Greek
E English
M Mandarin
N Nafsan
K Northern Kurdish
S Sanzhi Dargwa
B Tabasaran
T Teop
O Tondano
L Tulil
V Vera'a

Uniformity of lexical expressions: interim summary

- discourse is carried by a **relatively uniform bedrock of lexical expressions** (40–60%), regardless of language

- the locus of cross-linguistic variability is **the respective contributions of zero and pronouns among the reduced expressions** (Schiborr, in prep., Schnell & Barth 2018)

Explanations for uniformity of lexical expression

- lexical forms are used with similar rates across languages because their use is largely determined by **the same factors**

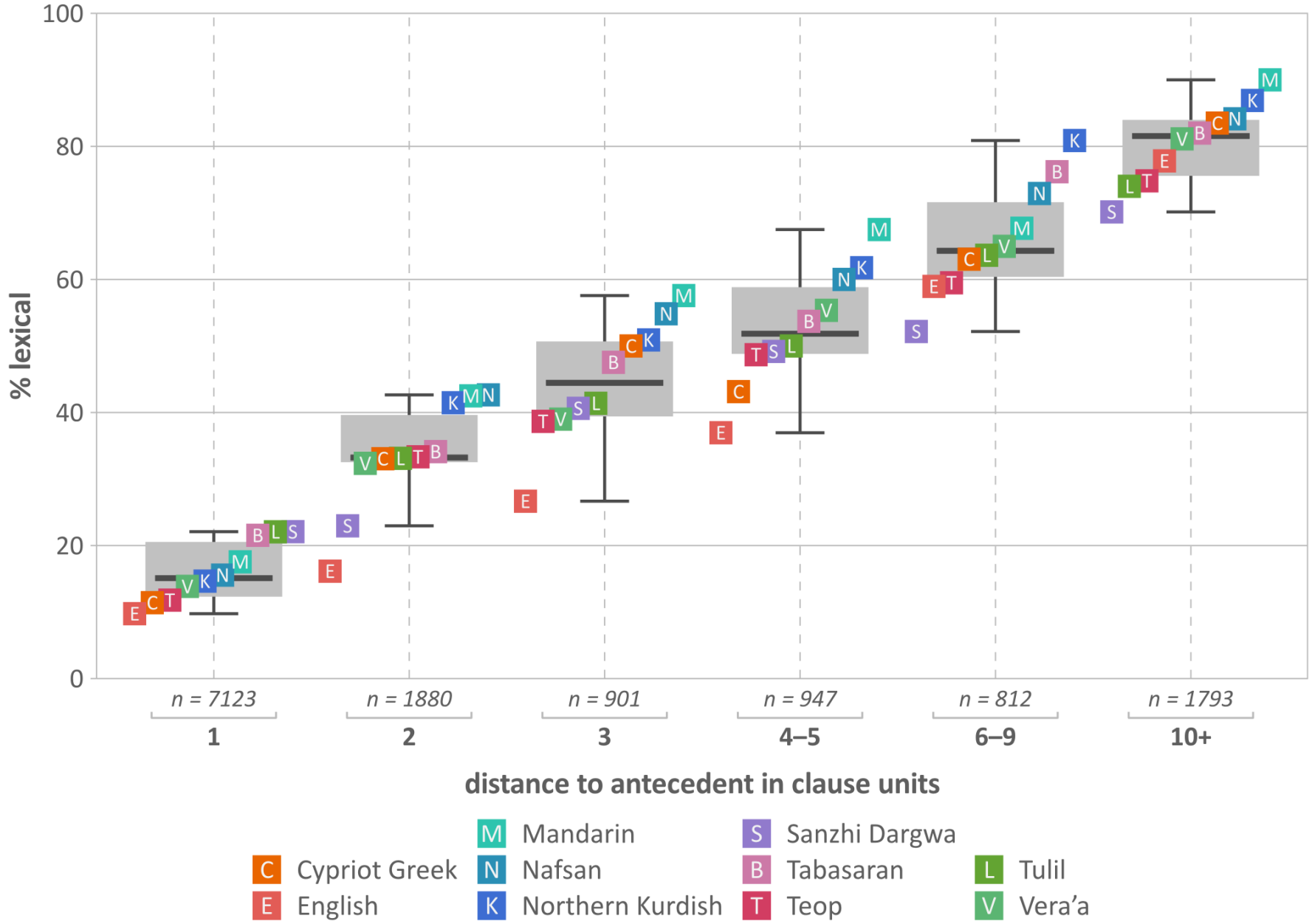
- most powerful factor: **anaphoric distance**

- the pronoun vs. zero choice is tempered by language-specific inherited historical accidents of morphosyntax, not treated in today's presentation, e.g.

  - presence of agreement morphology
  - informativity of pronouns (gender, number etc.)
  - differing effects with subjects and objects (Schnell & Barth 2018, resub.; Schwenter 2006, 2014)

  (**exception**: same-subject clause sequence contexts favour zero subjects across all languages; Torres Cacoulos & Travis 2019, Vollmer 2019, Schiborr, in prep.)

Anaphoric distance and lexical expression (Schiborr, in prep.)



Uniformity of lexical expressions: theoretical implications

- suggests a re-evaluation of the view that informativeness of discourse is language specific (i.e. that some languages are apparently ‘less explicit’, rely more on ‘pragmatic inference’, typologies of ‘pragmatic vs. syntactic’, ‘hot vs. cold’ languages; Stoll & Bickel 2009, Huang 2000)

- e.g. Mandarin: actually among the highest levels of lexical expressions in our sample (Vollmer 2019)

- little evidence for an across-the-board impact of ‘accessibility’ dictating zero vs. pronoun, and lexical vs. reduced (Schiborr, in prep.)

Uniform rates of lexical NPs: candidate universal

- in spontaneous unplanned discourse, between 40–60% of referring expressions are lexical NPs, regardless of language

(3) Light Human Subjects:
The skewed distribution of new referents in syntax

Light human subjects

- original observation by Du Bois 1987: **Avoid new/lexical A**
- **lexical** referential forms (with new referents) vs. **reduced** referential forms (pronouns, zero) are not evenly distributed across syntactic functions (Du Bois 1987, 2003, 2017)
- **transitive subjects** (A) apparently particularly favour reduced as opposed to lexical forms
- transitive (A) and intransitive subjects (S) apparently differ in this respect, with S clustering with P (objects)
- Du Bois' explanation for Avoid lexical A is related to information management in discourse, e.g. avoidance of more than one new referent per clause

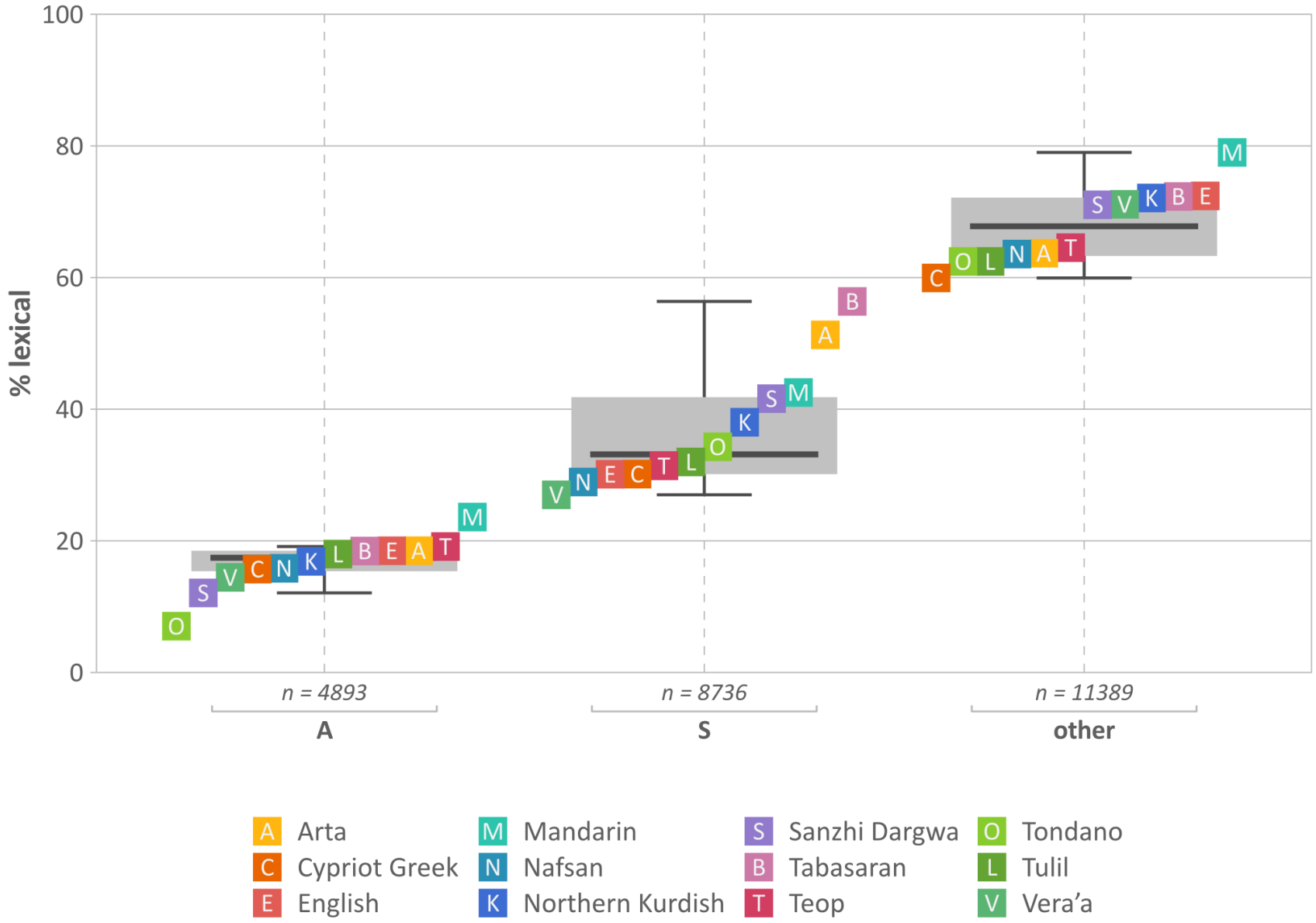
Light human subjects

- these claims have been questioned on empirical and conceptual grounds (e.g. Payne 1987, Kärkkäinen 1996, Haspelmath 2003, Everett 2009, Haig & Schnell 2016):

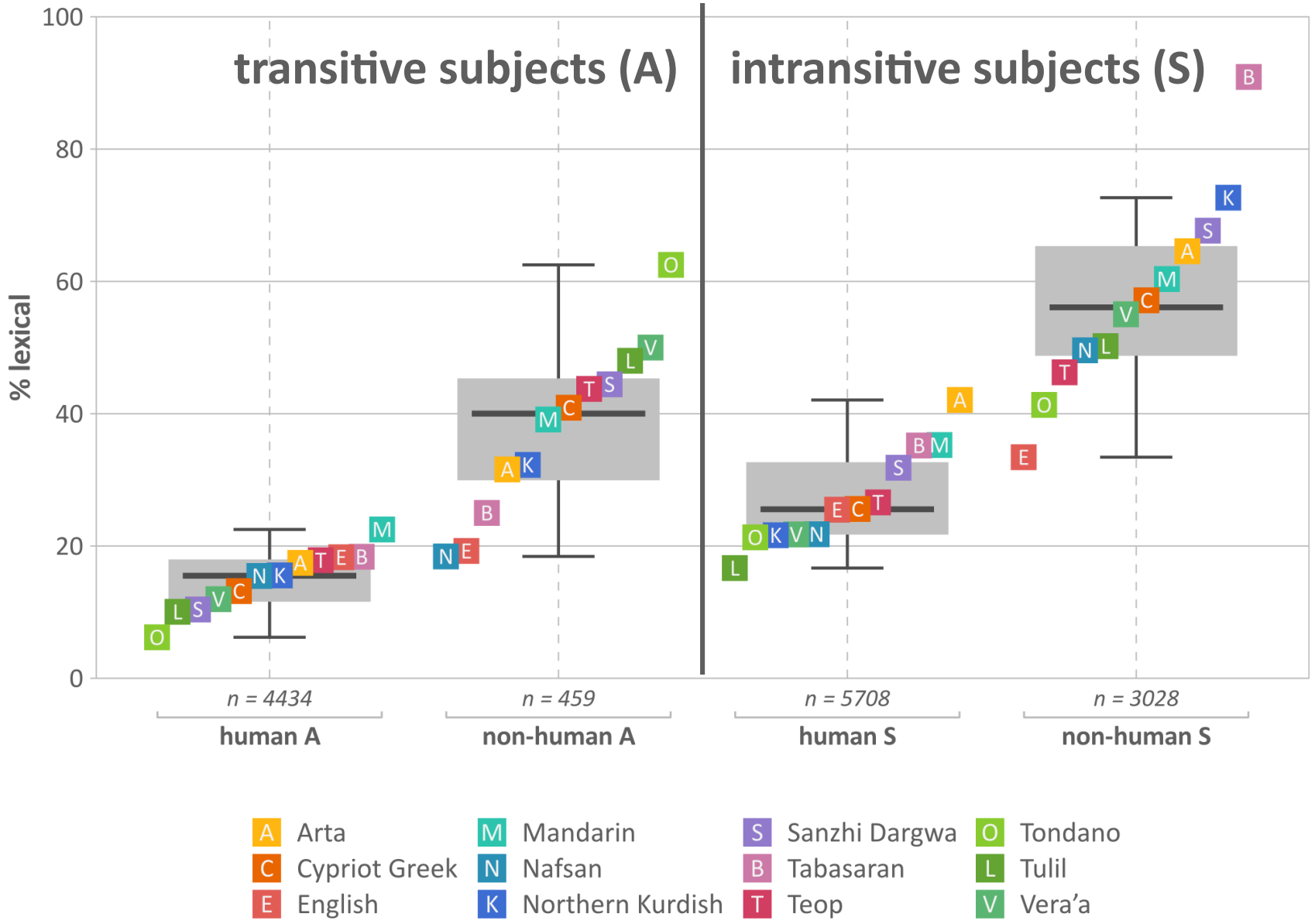
- no clustering of S and P; S and A closer than predicted
- role of information management overestimated; animacy accounts for most of the variation

- data from Multi-CAST …

Distribution of lexical arguments: A, S, and other



Human vs. non-human A and S

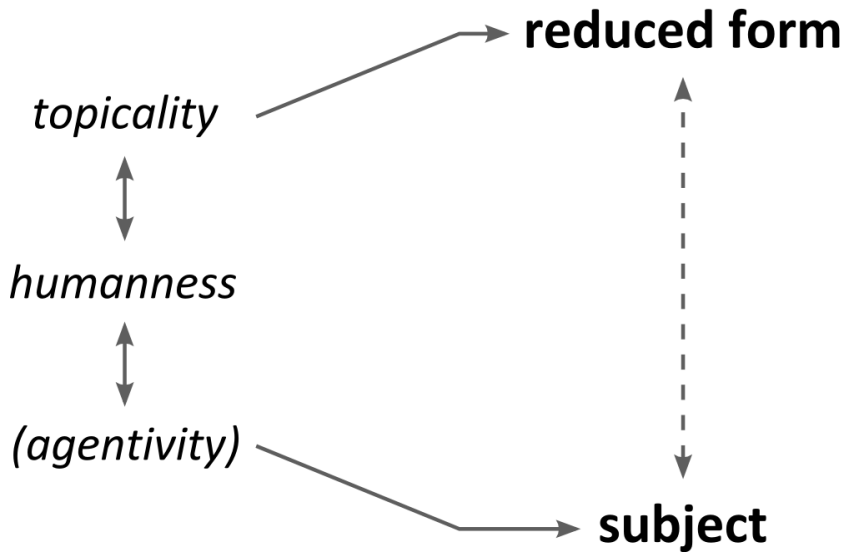


Light human subjects: interim summary

- the impact of transitivity (A vs. S) has been overrated
- the relevant generalization couples ‘humanness’ with ‘subject’ (S or A)
- not a question of ‘constraints on information management in discourse’, but a more general strategy reflecting cognitive prominence of human, topical entities

Explanations

- rather than a direct link of syntactic role and transitivity ('A') with information status, a more general concern with **human actors** drives the distribution:



- **the significant factor is the pragmatic and semantic prominence of human referents**

Light human subjects: candidate universal

- in spontaneous unplanned discourse, human subjects are generally (>75%) reduced

(4) The person asymmetry across subjects and objects

Subject/object asymmetry: main finding

- asymmetry between subjects and objects wrt. to various parameters regularly noted (e.g. Haig 2018; Schnell & Barth, resubm.; Dalrymple & Nikolaeva 2011)

- objects exhibit more complex patterns of pronominalization, with language-specific factor weightings (Schnell & Barth 2018)

- the single most robust difference appears to be robust regularities in the distribution of person values in transitive clauses

Subject/object asymmetry: main finding

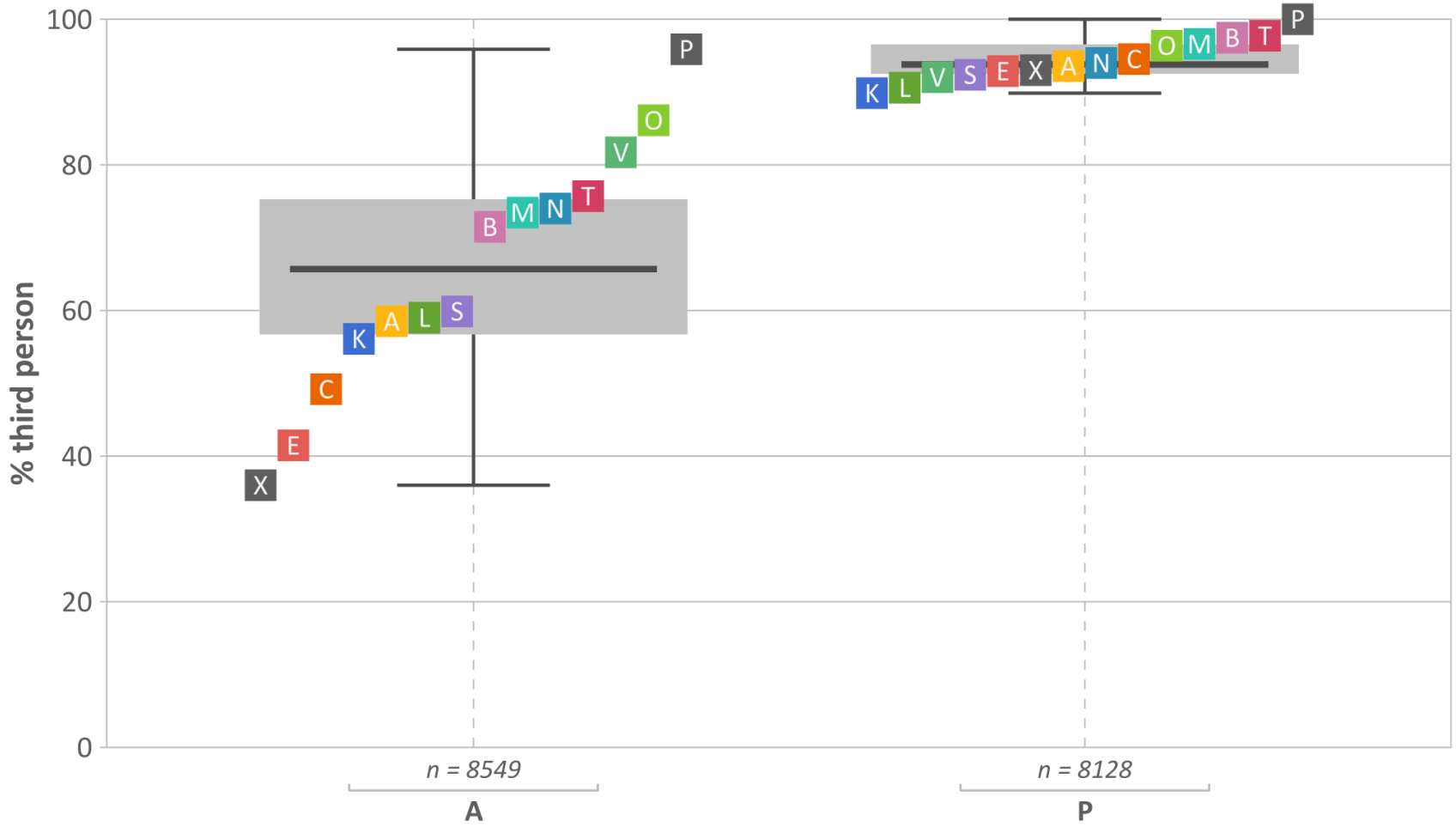
- the person value of transitive subjects is determined by **content and genre**:

  conversational → high levels of 1st/2nd person, low levels of 3rd person
  narratives → low 1st/2nd person, high 3rd person

- the person value of objects is **impervious to content and genre**:

  all genres: → overwhelmingly 3rd person

Person values, subjects vs. objects



The person asymmetry: candidate universal

- in spontaneous unplanned discourse, objects are overwhelmingly (> 90%) third person, regardless of content and genre

- the person values of transitive subjects, on the other hand, are dependent on content and genre

(5) Summary: candidate universals

Spontaneous unplanned discourse appears to comply with the following quantitative universals:

- **Light Human Subjects**
  the majority (> 75%) of human subjects are reduced in form (pronominal, zero)

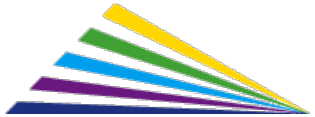
- **Uniform rates of lexical expression**
  between 40–60% of referring expressions are lexical NPs;
  the respective rates of pronoun and zero, on the other hand, are subject to cross-linguistic variability

- **The subject/object asymmetry in person values**
  at least 90% of all objects are third person;
  there is no comparable constant rate for subject person values

Thanks!

ARC CENTRE OF EXCELLENCE FOR
THE DYNAMICS OF LANGUAGE

University of Bamberg



DOBES

DFG Deutsche Forschungsgemeinschaft

Australian Government
Australian Research Council

VolkswagenStiftung

References

Dalrymple, Mary & **Nikolaeva**, Irina. 2011. *Objects and information structure*. Cambridge: Cambridge University Press.

Du Bois, John. 1987. The discourse basis of ergativity. *Language* 63(4), 805–855.

Du Bois, John. 2003. Discourse and grammar. In Tomasello, Michael (ed.), *The new psychology of language: Cognitive and functional approaches to language structure*, 47–88. Mahwah, NJ: Erlbaum.

Du Bois, John. 2017. Ergativity in discourse and grammar. In Coon, Jessica & Massam, Diane & Travis, Lisa D. (eds.), *The Oxford handbook of ergativity*, 23–57. Oxford: Oxford University Press.

Everett, Caleb. 2009. A reconsideration of the motivations for preferred argument structure. *Studies in Language* 33(1), 1–24.

Haig, Geoffrey. 2018. The grammaticalization of object pronouns: Why differential object indexing is an attractor state. *Linguistics* 56(4), 781–818. (DOI: 10.1515/ling-2018-0011)

Haig, Geoffrey & **Schnell**, Stefan. 2015. *Multi-CAST: The Multilingual Corpus of Annotated Spoken Texts*. (multicast.aspra.uni-bamberg.de)

Haig, Geoffrey & **Schnell**, Stefan. 2016. The discourse basis of ergativity revisited. *Language* 92(3), 591–618. (DOI: 10.1353/lan.2016.0049)

Haspelmath, Martin. 2003. *Ditransitive constructions in the world's languages*. Handout, March 2003, University of California, Berkeley.

Huang, Yan. 2000. *Anaphora: A cross-linguistic study*. Oxford: Oxford University Press.

Kärkkäinen, Elise. 1996. Preferred argument structure and subject role in American English conversational discourse. *Journal of Pragmatics* 25(5), 675–701.

Payne, Doris L. 1987. Information structuring in Papago narrative discourse. *Language* 63(4), 783–804.

Schiborr, Nils N. In preparation. *Lexical anaphora: A corpus-based typological study of referential choice*. PhD dissertation, University of Bamberg.

References

Schnell, Stefan & **Barth**, Danielle. 2018. Discourse motivations for pronominal and zero objects across genres in Vera'a. *Language Variation and Change* 30(1), 51–81. (DOI: 10.1017/S0954394518000054)

Schnell, Stefan & **Barth**, Danielle. Resubmitted. Towards subject--predicate agreement in Vera'a. Submitted to *Language Variation and Change*.

Schwenter, Scott. 2006. Null objects across South America. In Face, Timothy L. & Klee, Carol A. (eds), *Selected proceedings of the 8th Hispanic Linguistics Symposium*, 23–36. Somerville, MA: Cascadilla Proceedings Project.

Schwenter, Scott. 2014. Two kinds of object marking in Portuguese and Spanish. In Amaral, Patrícia & Carvalho, Ana M. (eds.), *Portuguese-Spanish interfaces: Diachrony, synchrony, and contact*, 237–260. Amsterdam: John Benjamins.

Stoll, Sabine & **Bickel**, Balthasar. 2009. How deep are differences in referential density? In Guo, Jiansheng & Lieven, Elena & Budwig, Nancy & Ervin-Tripp, Susan & Nakamura, Keiko & Özçaliskan, Seyda (eds.), *Crosslinguistic approaches to the psychology of language*, 543–555. London: Psychology Press.

Torres Cacoullos, Rena & **Travis**, Catherine E. 2019. Variationist typology: Shared probabilistic constraints across (non-)null subject languages. *Linguistics* 57(3), 653–692.

Vollmer, Maria C. 2019. *How radical is pro-drop in Mandarin? A quantitative corpus study on referential choice in Mandarin Chinese*. MA thesis, University of Bamberg.