

USING STATISTICAL CLASSIFICATION TO DISCOVER CROSS-LINGUISTIC SEMANTIC PROTOTYPES: THE CAUSATION DOMAIN

2023 LSA 97TH ANNUAL MEETING
DENVER, CO, JANUARY 5-8, 2023



Jürgen Bohnemeyer,¹ Erika Bellingham,¹ Andrea Ariño-Bizarro,² Emanuel Bylund,^{3,4}
Jing Du,⁵ James Essegbey,⁶ Stephanie Evers,¹ Saima Hafeez,¹ Iraide Ibarretxe-
Antuñano,² Pia Järnefelt,⁷ Kazuhiro Kawachi,⁸ Yu Li,⁹ Thomas Fuyin Li,¹⁰ Tatiana Nikitina,¹¹
Sang-Hee Park,¹² Anastasia Stepanova,¹ and Guillermo Montero-Melis^{4,13}



SYNOPSIS

- ▶ The pragmatic ecology of causation
- ▶ A new study design for semantic typology
- ▶ Variables and stimuli: the CAL Clips
- ▶ The language sample
- ▶ Cluster analysis
- ▶ Predictive models
- ▶ Inter-speaker variation
- ▶ Summary and discussion

THE PRAGMATIC ECOLOGY OF CAUSATION

- ▶ we map the semantics and pragmatics of the causative domain in 13 languages from 12 genera
- ▶ based on primary data
collected from 12+ speakers per language
- ▶ using an innovative combination of production and acceptability judgment elicitation

- ▶ background: the expressions that form a semantic domain in a particular language are pragmatically related
 - ▶ the speaker aims to choose from among them the one that best fits the situation and her communicative intent
 - ▶ this idea has been expressed invoking notions of
 - ▶ opposition and markedness (structuralists)
 - ▶ conversational maxims (Gricean pragmatics)
 - ▶ ecology and system theory (evolutionary linguistics)
- ▶ these perspectives are not mutually exclusive

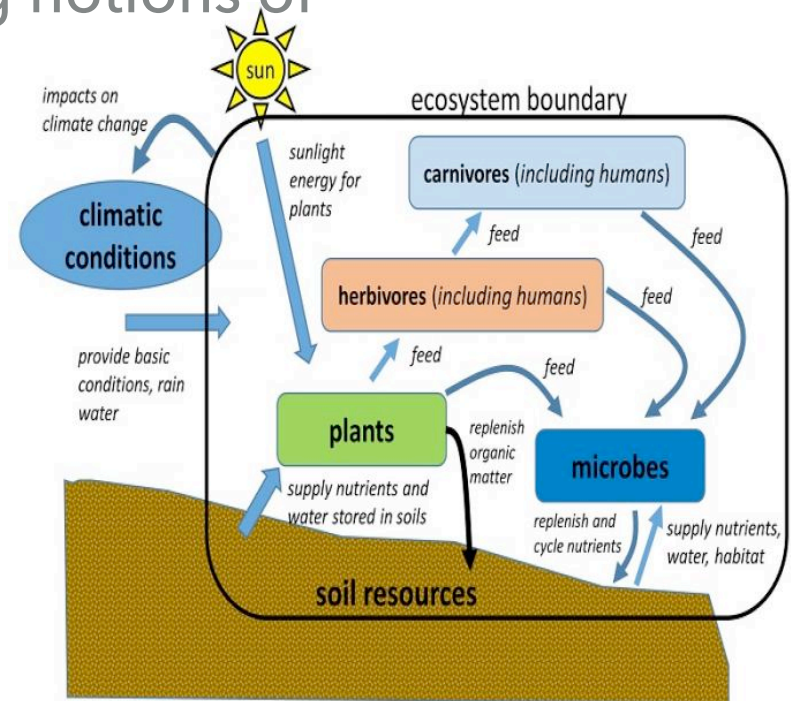


Figure 1.1. A system ecology (e-education.psu.edu)

- ▶ example: the domain of causation
 - ▶ simple 'direct' causal chains
favor simple causative constructions

(1.1) Le=máak=o' t-u=**nik**-ah le=bàaso-s-o'b=o'
 YUC DEF=person=D2 PRV-A3=scatter-CMP(B3SG) DEF=cup-PL-PL=D2
 'The man, he scattered the cups'



Figure 1.2. HO5_cuptower

- ▶ more complex constructions/descriptions
are preferred for more complex, 'indirect' chains

(1.2) a. #Le=x-ch'úupal=o' t-u=**nik**-ah le=bàaso-s-o'b=o'
 YUC DEF=female:child=D2 PRV-A3=shatter+slap-APP-CMP(B3SG) DEF=cup-PL-PL=D2
 'The girl, she scattered the cups'

b. Le=x-ch'úupal=o' t-u=**mèet**-ah
 DEF=F-female:child=D2 PRV-A3=make-CMP(B3SG)
u=nik-ik le=bàaso-o'b le=máak=o'
 A3=scatter-INC(B3SG) DEF=cup-PL DEF=person=D2
 'The girl, she made the man scatter the cup'



Figure 1.3. HUO2_cups

- ▶ 50 years of typological research on causatives has focused on the broad division of labor
 - ▶ between simple and complex causatives
- ▶ particularly the iconicity it involves and the underlying causes of this iconicity
 - ▶ Bohnemeyer et al (2010); Comrie (1981); Dixon (2000); Haiman (1983); Haspelmath (2008); Kemmer & Verhagen (1994); Levin & Rappaport-Hovav (1995); Levshina (2015), (2016), (2017); McCawley (1976, 1978); Shibatani ed. (1976); Shibatani & Pardeshi (2002); Talmy (1976); Verhagen & Kemmer (1997); Wolff (2003); *inter alia*

- ▶ mostly missing so far: a comprehensive typological examination of the causative ecology based on primary data
 - ▶ yielding a semantic map of the domain for each language
- ▶ exceptions
 - ▶ Bohnemeyer et al. (2010) (pilot study, data from just four languages; highly unbalanced stimulus set)
 - ▶ Levshina (2022) (movie subtitle data from 22 languages (13 Indo-European))
- ▶ our goal: contribute toward closing this gap based on a new methodology for semantic typology

SYNOPSIS

- ▶ The pragmatic ecology of causation
- ▶ A new study design for semantic typology
- ▶ Variables and stimuli: the CAL Clips
- ▶ The language sample
- ▶ Cluster analysis
- ▶ Predictive models
- ▶ Inter-speaker variation
- ▶ Summary and discussion

A NEW STUDY DESIGN FOR SEMANTIC TYPOLOGY

- ▶ a new approach

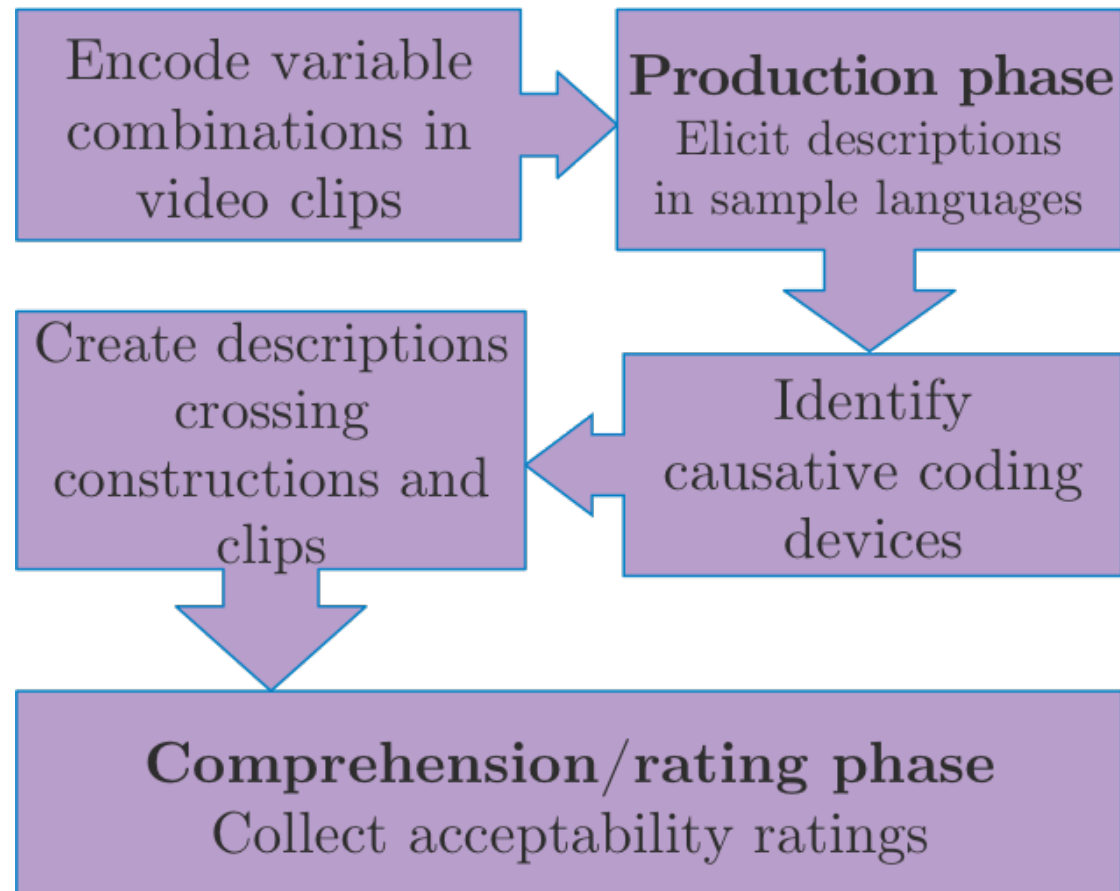


Figure 2.1. *A hybrid study design for semantic typology*

- ▶ advantages of this hybrid design type
 - ▶ vis-à-vis corpus studies
 - ▶ applicable to languages for which (large) corpora are unavailable
 - ▶ provides both positive and negative evidence
 - ▶ gives direct access to the scene being described
 - ▶ vis-à-vis traditional elicited production studies (the staple in contemporary semantic typology)
 - ▶ allows rapid data collection and analysis from a larger number of speakers
 - ▶ provides both positive and negative evidence

- ▶ the rating scale
 - ▶ after some experimentation,
we settled on a four-point qualitative scale
 - ▶ we trained the participants with the help of additional stimuli to distinguish among
 - ▶ ungrammatical utterances (1)
 - ▶ well-formed but inaccurate descriptions (2)
 - ▶ accurate but misleading descriptions (3)
 - ▶ accurate and appropriately informative descriptions (4)

SYNOPSIS

- ▶ The pragmatic ecology of causation
- ▶ A new study design for semantic typology
- ▶ Variables and stimuli: the CAL Clips
- ▶ The language sample
- ▶ Cluster analysis
- ▶ Predictive models
- ▶ Inter-speaker variation
- ▶ Summary and discussion

VARIABLES AND STIMULI: THE CAL CLIPS

- variables that have been shown to impact causative choice

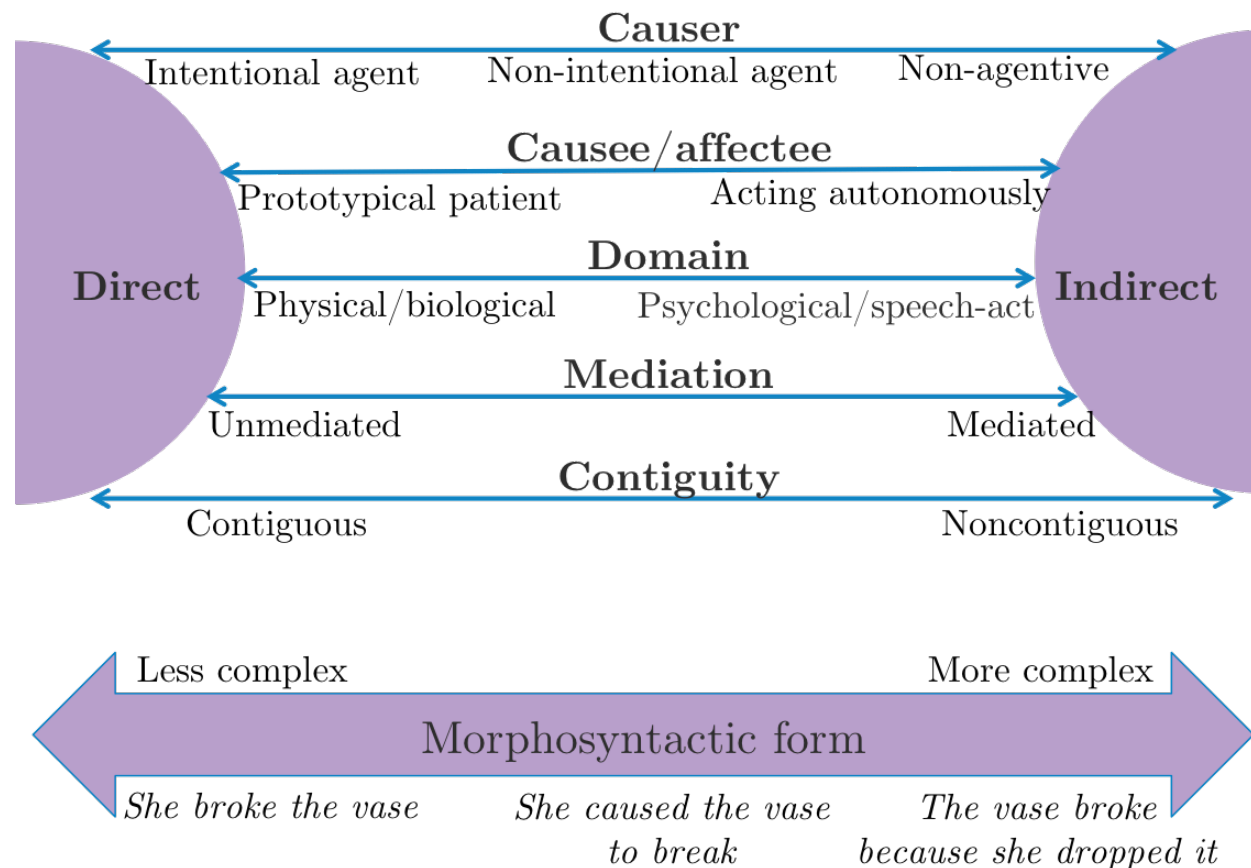


Figure 3.1. A multidimensional continuum model of causation directness

- ▶ design: E. Bellingham; J. Bohnemeyer
- ▶ 58 short video clips featuring everyday causal chains
 - ▶ most staged/enacted, a few found on the internet
- ▶ variables manipulated
 - ▶ **causer (CR)** type: volitional vs. accidental vs. force
 - ▶ **causee (CE;** = intermediate participant in the chain) type
 - ▶ volitional/controlled
 - ▶ vs. involuntary response to psychological impact
 - ▶ vs. involuntary response to mechanical impact
 - ▶ vs. no CE



- ▶ **affectee (AF) type**
 - ▶ volitional/controlled
 - ▶ vs. involuntary response to psychological impact
 - ▶ vs. involuntary response to mechanical impact
 - ▶ vs. physical object
- ▶ **resulting event type**
physical state change vs. location change vs. process
- ▶ **force dynamics**
 - ▶ causation (43 core + 10 sup.) vs. letting (5 sup. scenes)

- ▶ stimuli: the CAL Clips (cont.)
 - ▶ examples
 - ▶ CR = force; CE = none; AF = mechanically impacted; resultant event = location change; FD = causation

Figure 3.1. NM2_reporter



- ▶ stimuli: the CAL Clips (cont.)
 - ▶ examples (cont.)
 - ▶ CR = accidental; CE = volitional/controlled; AF = object; resultant event = location change; FD = letting



Figure 3.2. UCO1_ball

- ▶ stimuli: the CAL Clips (cont.)
 - ▶ examples (cont.)
 - ▶ CR = volitional; CE = psychologically impacted; AF = object; resultant event = physical change; FD = letting

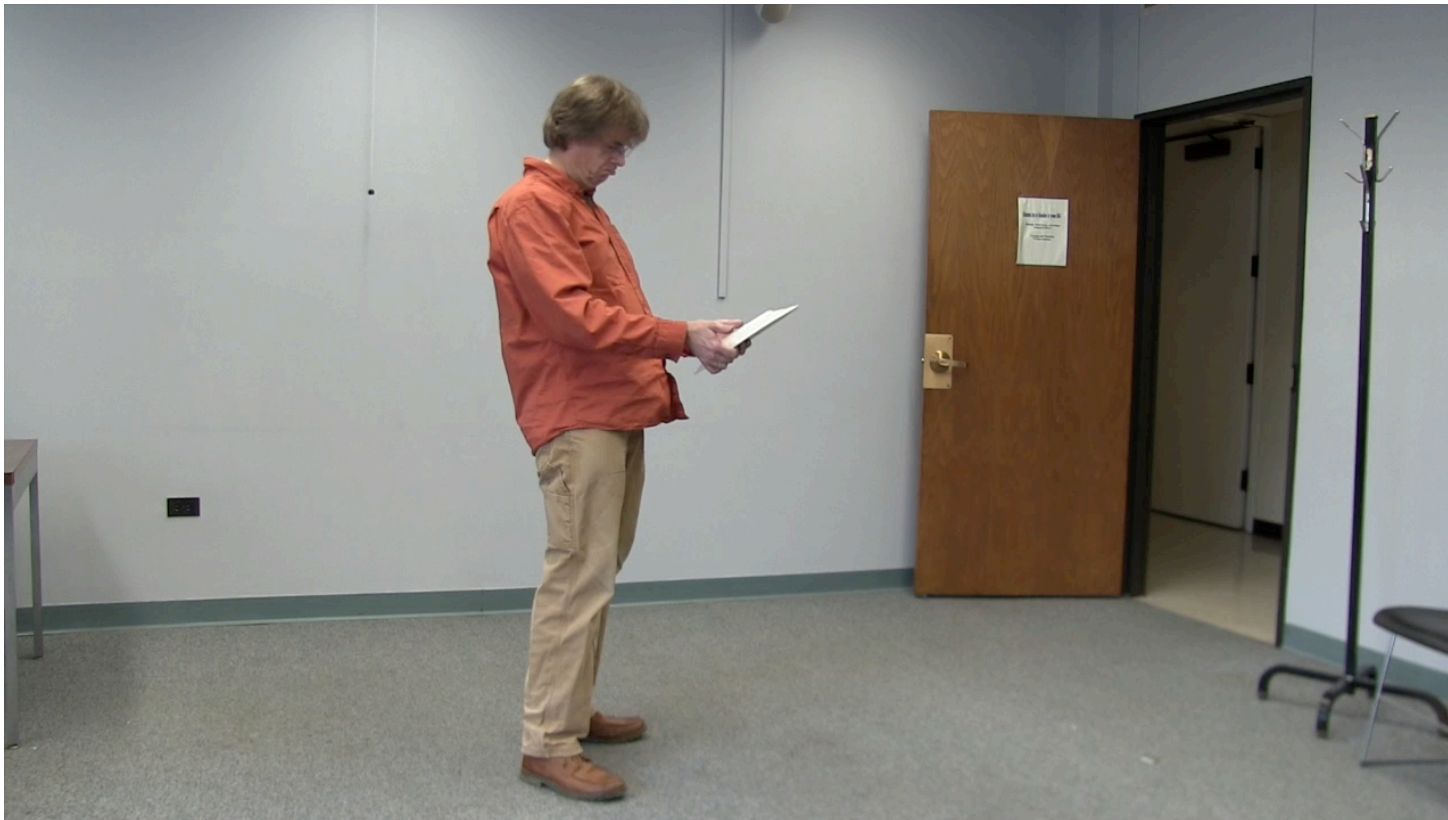


Figure 3.3. HUO1_plate

- ▶ stimuli: the CAL Clips (cont.)
 - ▶ examples (cont.)
 - ▶ CR = volitional; CE = volitional/controlled; AF = object; resultant event = process; FD = causation



Figure 3.4. HCOproc1_swing

SYNOPSIS

- ▶ The pragmatic ecology of causation
- ▶ A new study design for semantic typology
- ▶ Variables and stimuli: the CAL Clips
- ▶ The language sample
- ▶ Cluster analysis
- ▶ Predictive models
- ▶ Inter-speaker variation
- ▶ Summary and discussion

THE LANGUAGE SAMPLE

- the languages from which data has been collected for the Semantic Typology subproject so far

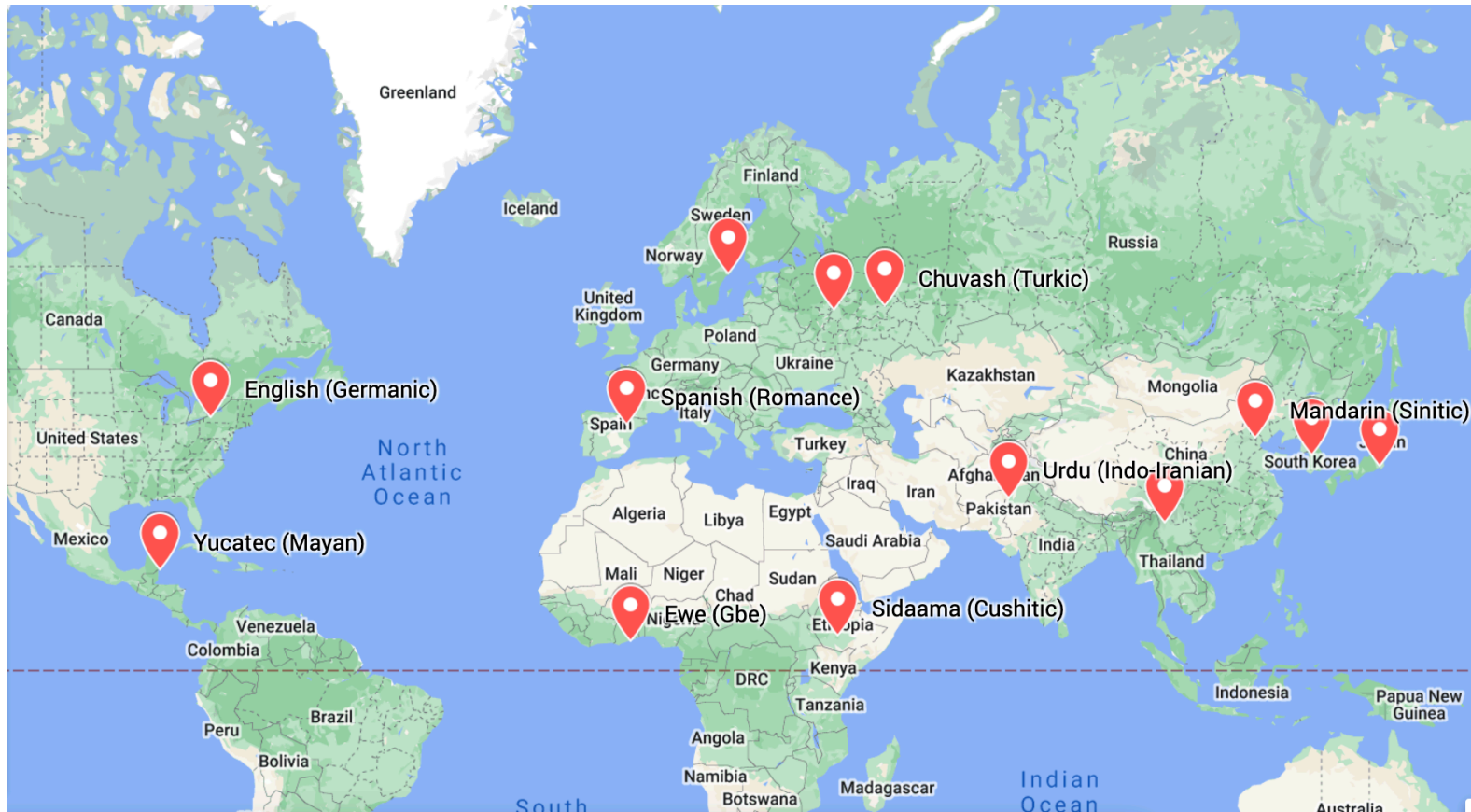
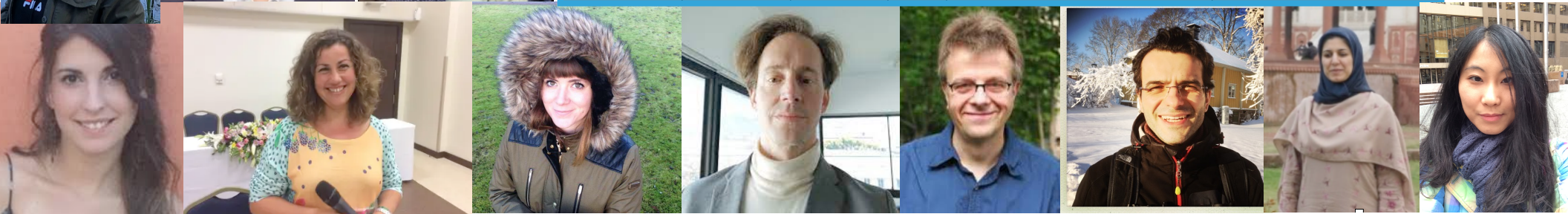


Figure 4.1. *The current sample of the CAL Semantic Typology subproject (widgets marking approximate field sites)*

► populations and researchers

			Language	Genus	Field site	N	Researcher	Affiliation
			Chuvash	Turkic	Russia	12	T. Nikitina	CNRS
			English	Germanic	U.S.A.	13	E. Bellingham, S. Evers	U at Buffalo
			Ewe	Kwa	Ghana/ U.S.A	12	J. Essegbey	U of Florida
			Japanese	Japonic	Japan	15	K. Kawachi	Keio U
			Korean	Isolate	R.O.K.	12	S. Park	Kyung Hee U
			Mandarin	Chinese	China	12	J. Du, T. F. Li	UCAS, Beihang U
			Russian	Slavic	Russia	12	A. Stepanova	U at Buffalo
			Sidaama	Cushitic	Ethiopia	12	K. Kawachi	Keio U
			Spanish	Romance	Spain	13	A. Ariño, I. Ibarretxe Antuñano	U of Zaragoza
			Swedish	Germanic	Sweden	12	P. Järnefelt, G. Montero- Melis, E. Bylund	Stockholm U, MPI for Psycholinguistics
			Urdu	Indic	Pakistan	12	S. Hafeez	U at Buffalo
			Yucatec	Mayan	Mexico	12	J. Bohnemeyer	
			Zauzou	Lolo- Burmese	China	12	Y. Li	Wuhan U

Table 4.1. The current sample
of the CAL Semantic Typology
subproject



► causative expressions included in the analysis

Table 4.2. *Causative coding devices in the sample languages that were included in the analysis*

Construction	Chu- vash	Eng- lish	Ewe	Japa- nese	Ko- rean	Man- darin	Rus- sian	Sidaa- ma	Spa- nish	Swe- dish	Urdu	Yuca- tec	Zauzou
Lexical & not fully productive morphological causatives	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓
Light verb constructions											✓		
Serial verb constructions			✓										
Fully productive morphological causatives	✓			✓							✓		
Periphrastic causatives		✓	✓		✓	✓	✓	✓	✓	✓		✓	✓
Non-sentential causer adjunct		✓									✓		
Non-sentential cause adjuncts						✓	✓	✓	✓		✓		
Clause-layer serialization			✓										
Causal converb constructions	✓				✓								✓
Causal clause constructions		✓		✓		✓	✓	✓	✓	✓	✓	✓	✓
Extent ('So X that Y') constructions							✓		✓				
Means construction								✓					

SYNOPSIS

- ▶ The pragmatic ecology of causation
- ▶ A new study design for semantic typology
- ▶ Variables and stimuli: the CAL Clips
- ▶ The language sample
- ▶ Cluster analysis
- ▶ Predictive models
- ▶ Inter-speaker variation
- ▶ Summary and discussion

CLUSTER ANALYSIS

- ▶ this and the following analyses are based on data from the 43 core scenes of the CAL Clips
- ▶ for each language-specific response type (RT, i.e., causative construction type), a rating vector was calculated
 - ▶ one dimension per stimulus clip
 - ▶ coordinates represent the proportion of speakers who rated the stimulus description acceptable for the clip
 - ▶ i.e., well-formed, accurate, and appropriately informative
- ▶ where multiple descriptions were tested for a given RT, the ratio was incremented if a least one description was rated acceptable

- ▶ a cluster analysis was performed over all 60 RT vectors

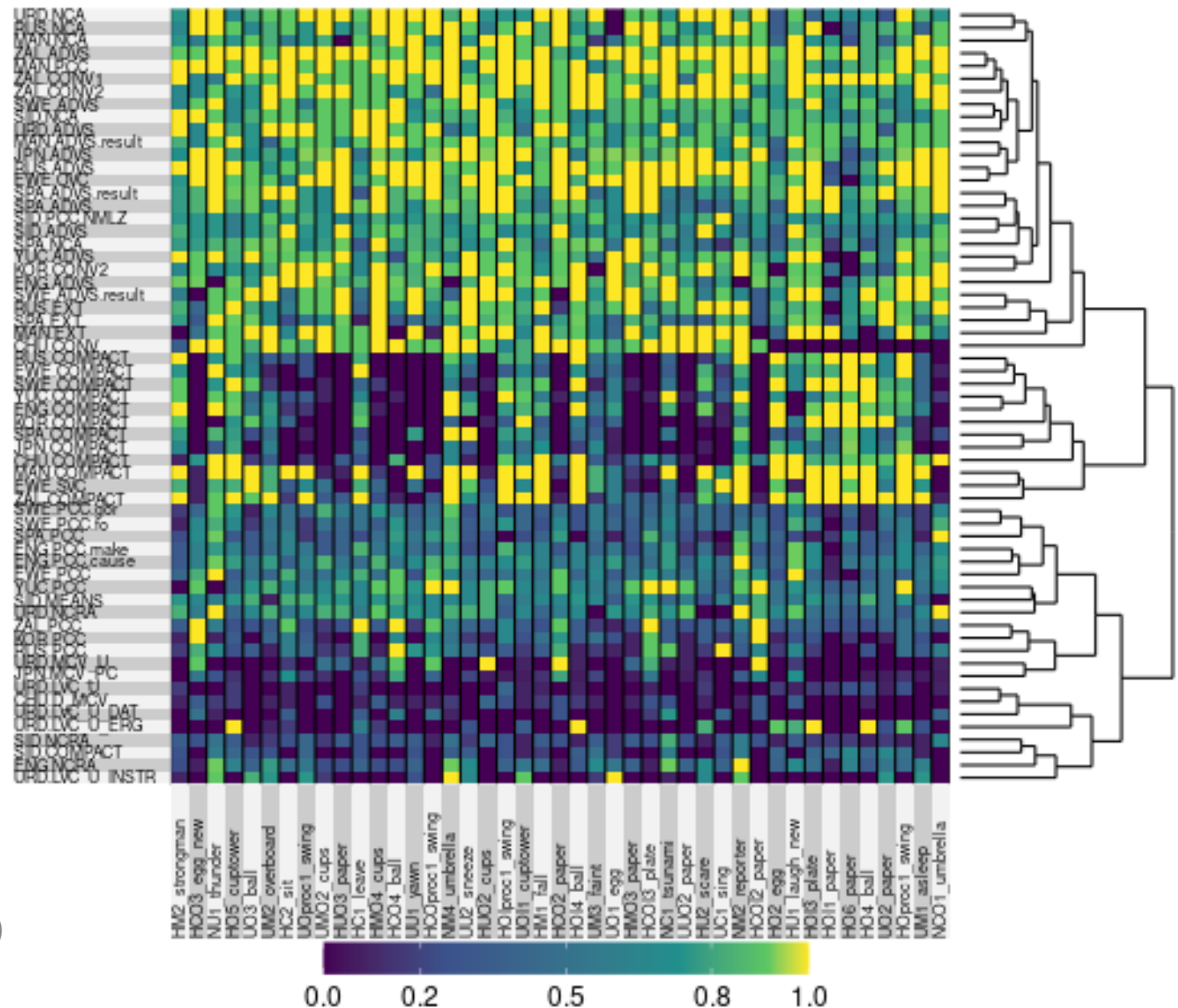


Figure 5.1. Heat map and cluster dendrogram of the rating vectors associated with the 60 language-specific response types (RTs) included in the analysis (x-axis: stimulus clips; y-axis: language-specific RTs)

- ▶ discussion
 - ▶ the rating vectors solely reflect the acceptability ratings
 - ▶ the model had no access to morphosyntactic information
 - ▶ remarkably, the model nevertheless was able to group
 - ▶ lexical and not fully productive morphological causatives
 - ▶ periphrastic (= analytical/syntactic) and fully productive morphological causatives
 - ▶ adverbial modifier constructions such as causal clause and converb constructions
 - ▶ suggesting that each construction type has a unique semantic/pragmatic profile

- ▶ discussion (cont.)
 - ▶ fully productive morphological causatives such as those of Chuvash, Japanese, and Urdu
 - ▶ behave semantically and pragmatically like periphrastic causatives in other languages
 - ▶ confirming Shibatani (1973)

- ▶ discussion (cont.)
 - ▶ exceptions
 - ▶ Mandarin periphrastic causatives in the adverbial cluster
 - ▶ Sidaama compact causatives in the periphrastic cluster
 - ▶ Urdu light verb constructions in the periphrastic cluster
 - ▶ 'non-sentential causer adverbials' (English, Sidaama, Urdu) in the periphrastic cluster

(5.1) *The man knocked over the cups **because of the woman***

SYNOPSIS

- ▶ The pragmatic ecology of causation
- ▶ A new study design for semantic typology
- ▶ Variables and stimuli: the CAL Clips
- ▶ The language sample
- ▶ Cluster analysis
- ▶ Predictive models
- ▶ Inter-speaker variation
- ▶ Summary and discussion

PREDICTIVE MODELS

- ▶ not all semantic predictor variable level combinations could be instantiated with equal frequency in the CAL Clips
- ▶ so to discover the effects of the predictor variables, we used machine learning classifiers instead of regression models
- ▶ all lexical and not fully productive morphological (= '**compact**') causatives showed a single rating maximum involving
 - ▶ absence of mediation
(no intervening subevents or participants)
 - ▶ affectees/patients with no control over the caused event
 - ▶ intentional causers
- ▶ as predicted by the literature

▶ example: English

Was at least one description from ENG.COMPACT acceptable in English ? (Min bucket: 25)

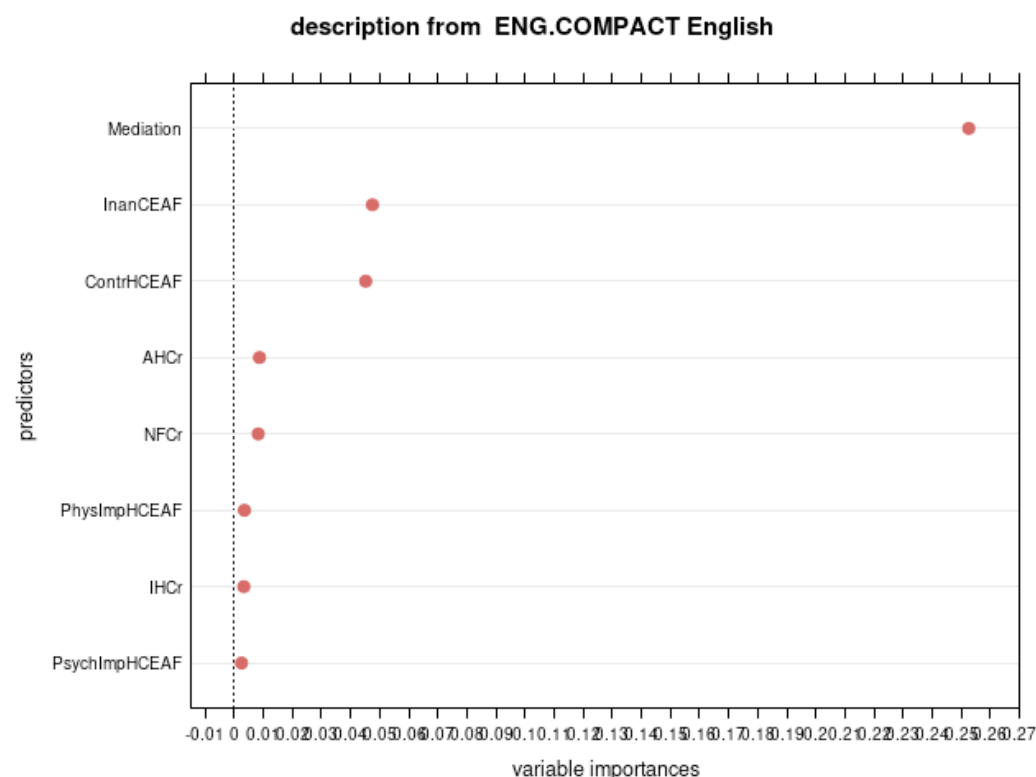
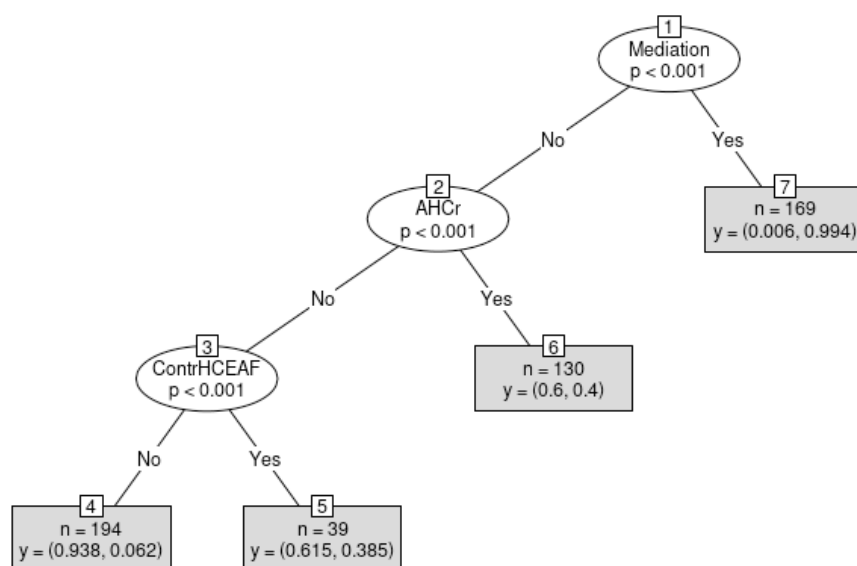


Figure 6.1. Conditional inference tree and variable importance plot based on a random forest model of the English 'compact' causative construction (i.e., base-transitive causative verbs). (AHCr - Accidental human causer; ContrHCEAF - Causee/affectee with control over the caused event; InanCEAF - Inanimate causee/affectee; NFCr - Natural force causer; PhysImpHCEAF - Physically impacted causee/affectee; IHCr - Intentional human causer; PsychImpHCEAF - Psychologically impacted causee/affectee)

- ▶ mediation proved generally the top variable for compact causatives
- ▶ one exception: ergative-marked causer NPs entail intentionality with compact Urdu causatives

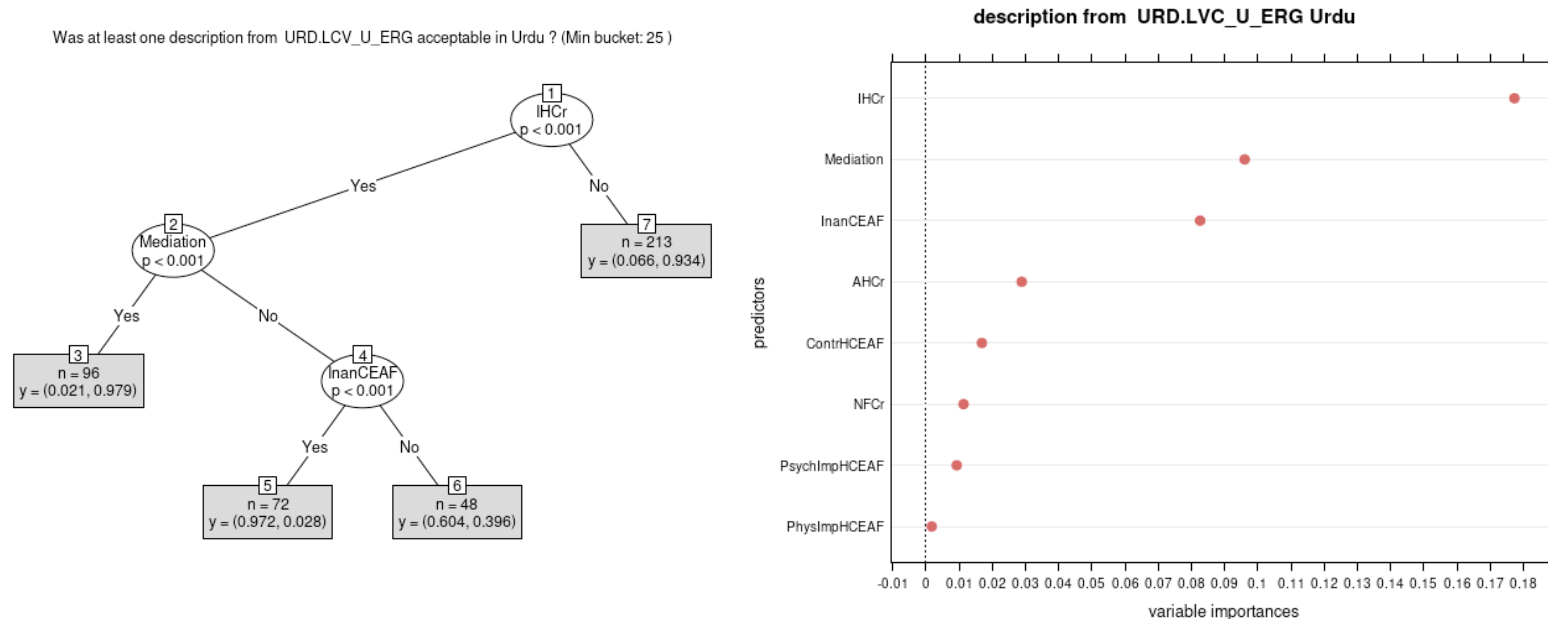


Figure 6.2. Conditional inference tree and variable importance plot based on a random forest model of the Urdu light verb construction with ergative causer NP. (IHCr - Intentional human causer; InanCEAF - Inanimate causee/affectee; AHCr - Accidental human causer; ContrHCEAF - Causee/affectee with control over the caused event; NFCr - Natural force causer; PhysImpHCEAF - Physically impacted causee/affectee; PsychImpHCEAF - Psychologically impacted causee/affectee)

- the surprise: the semantic prototypes of complex causatives aren't simply complementary to those of compact causatives
 - periphrastic causatives in particular often show multiple discrete prototypes, one of which involves natural forces
 - example: Zauzou (Loloish, Yunan Province, PRC)

Was at least one description from ZAL.PCC acceptable in Zauzou ? (Min bucket: 25)

Rating maximum I:
Intentional causer
and controlled
causee/affectee

Rating maximum II:
Natural force causer

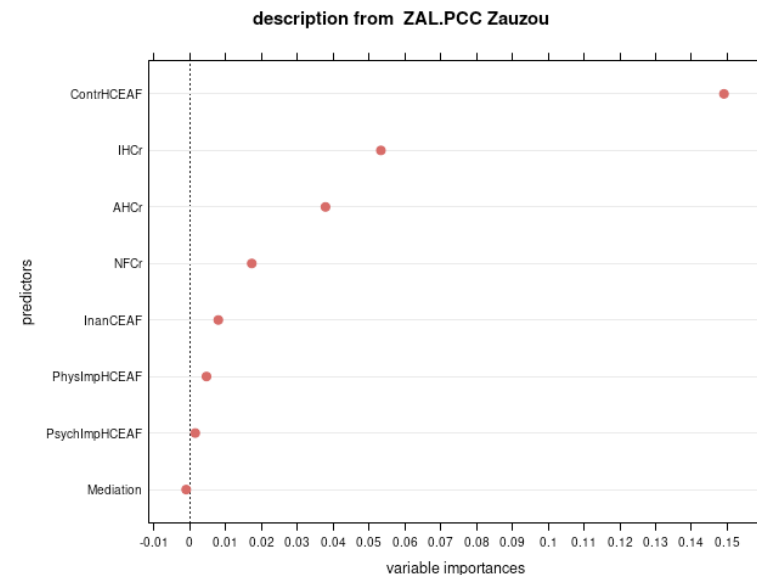
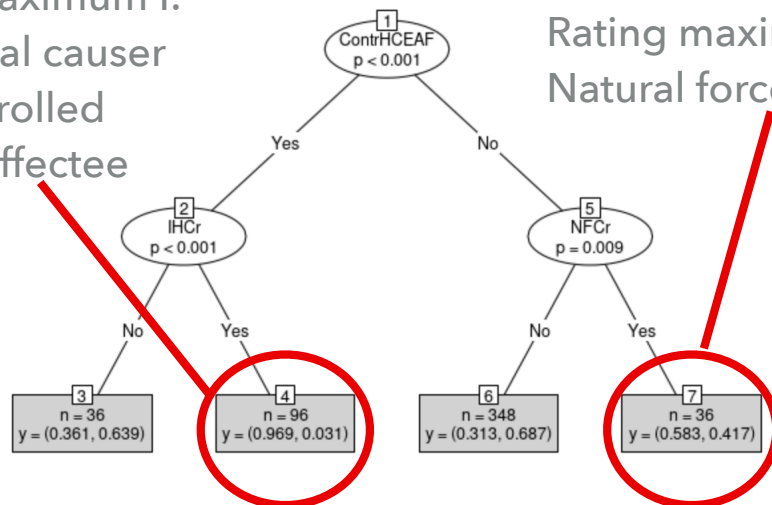


Figure 6.3. Conditional inference tree and variable importance plot based on a random forest model of the Zauzou periphrastic causative construction. (ContrHCEAF - Causee/affectee with control over the caused event; IHCr - Intentional human causer; NFCr - Natural force causer; AHCr - Accidental human causer; InanCEAF - Inanimate causee/affectee; PhysImpHCEAF - Physically impacted causee/affectee; PsychImpHCEAF - Psychologically impacted causee/affectee)

- ▶ overall, of 11 periphrastic causative constructions
 - ▶ 6 show evidence of multiple prototypes
 - ▶ 7 show evidence of natural force causer prototypes
- ▶ in contrast, the fully productive morphological causatives of Japanese and Urdu show a single prototype
 - ▶ involving mediation and intentional causers
- ▶ the fully-productive morphological causative of Chuvash elicited low acceptability across the board

SYNOPSIS

- ▶ The pragmatic ecology of causation
- ▶ A new study design for semantic typology
- ▶ Variables and stimuli: the CAL Clips
- ▶ The language sample
- ▶ Cluster analysis
- ▶ Predictive models
- ▶ Inter-speaker variation
- ▶ Summary and discussion

INTER-SPEAKING VARIATION

- ▶ to assess inter-speaker variation, we computed separate rating vectors for each participant and response type
- ▶ and generated multi-dimensional scaling plots of their Hamming distances by language

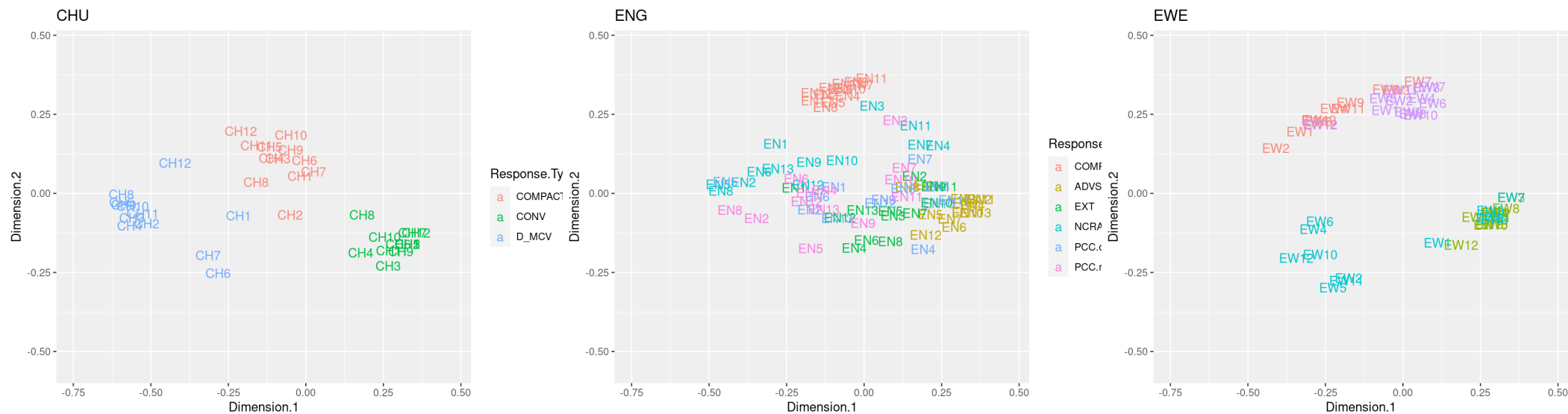


Figure 7.1. Plotting the first two dimensions of a multi-dimensional scaling model of the rating vectors by participant and response type for Chuvash, English, and Ewe

- ▶ in every language,
inter-speaker variation is minimal with compact causatives
- ▶ and maximal with periphrastic
and fully productive morphological causatives



Figure 7.2. Plotting the first two dimensions of a multi-dimensional scaling model of the rating vectors by participant and response type for Japanese, Korean, and Mandarin

- discussion
 - inter-speaker agreement with compact causatives is consistent with them having unique prototypes

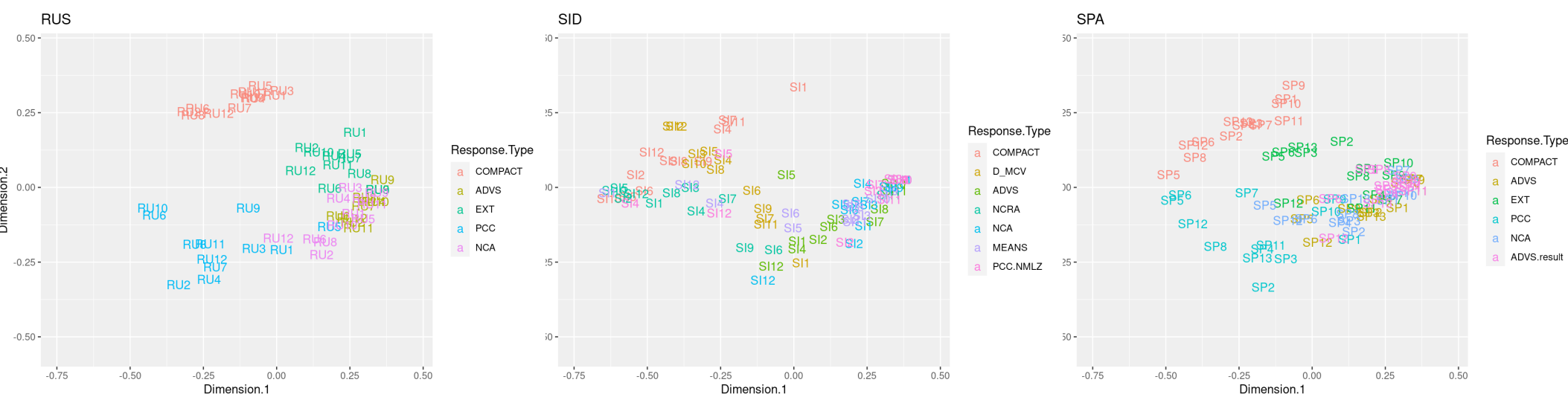


Figure 7.3. Plotting the first two dimensions of a multi-dimensional scaling model of the rating vectors by participant and response type for Russian, Sidaama, and Spanish

- ▶ discussion (cont.)
 - ▶ adverbial modifier constructions show relatively high acceptability across the board

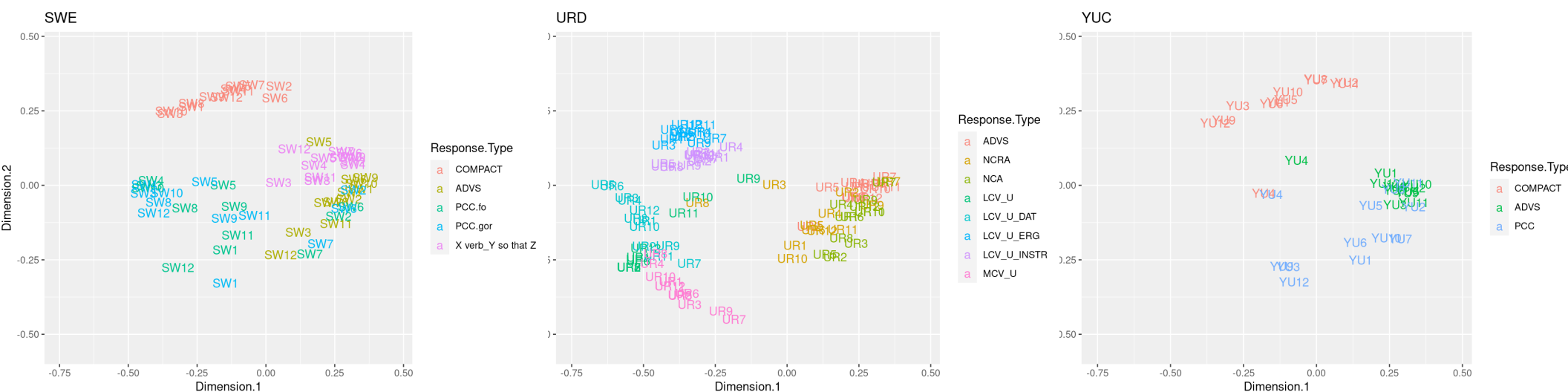


Figure 7.4. Plotting the first two dimensions of a multi-dimensional scaling model of the rating vectors by participant and response type for Swedish, Urdu, and Yucatec

- ▶ discussion (cont.)
 - ▶ intermediate-complexity constructions are “caught in the middle”
 - ▶ lacking both unique semantic prototypes and across-the-board acceptability

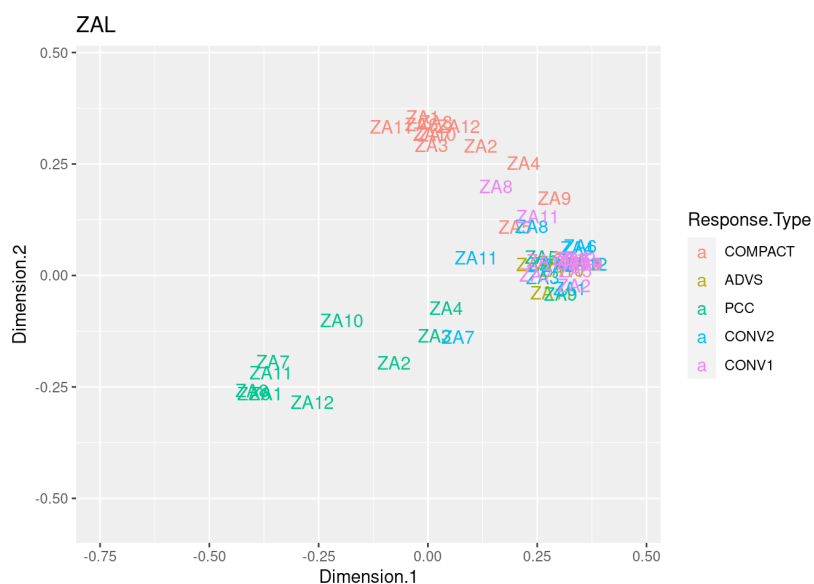


Figure 7.6. Plotting the first two dimensions of a multi-dimensional scaling model of the rating vectors by participant and response type for Zauzou

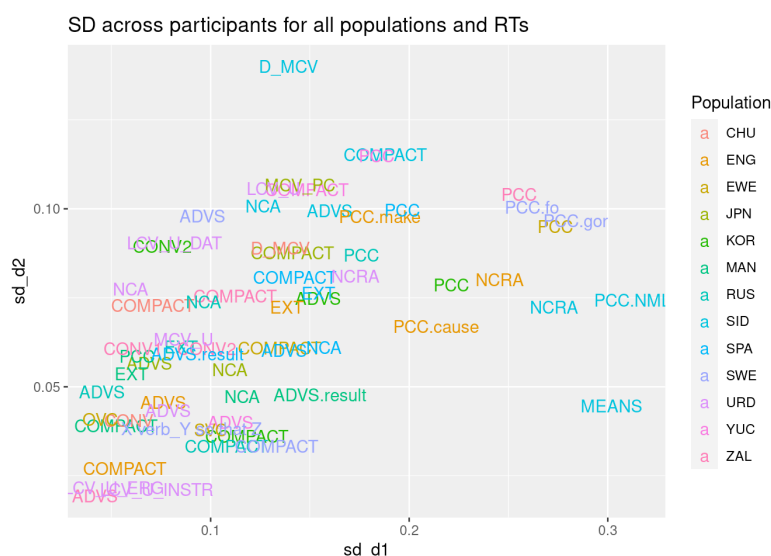


Figure 7.7. Plotting the standard deviation of the first and second dimension of a multi-dimensional scaling model of the rating vectors by response type (labels) and language (colors)

SYNOPSIS


- ▶ The pragmatic ecology of causation
- ▶ A new study design for semantic typology
- ▶ Variables and stimuli: the CAL Clips
- ▶ The language sample
- ▶ Cluster analysis
- ▶ Predictive models
- ▶ Inter-speaker variation
- ▶ Summary and discussion

SUMMARY AND DISCUSSION

- ▶ new hybrid approach to gather primary typological data on semantics and pragmatics
- ▶ new method for inferring semantic prototypes from acceptability rating data using machine learning models

- ▶ most lexical causatives have unmediated causation as their unique semantic prototype
 - ▶ in line with what previous research suggests
- ▶ however, the semantics and pragmatics of complex causatives turns out to be more diverse
 - ▶ both crosslinguistically and in terms of inter-speaker variation
 - ▶ and also more diffuse in the sense of having multiple prototypes or no clear prototype at all
- ▶ this is consistent with complex constructions being used much less frequently (Haspelmath 2008)

ACKNOWLEDGMENTS

- ▶ epic thanks to
 - ▶ the study participants
 - ▶ colleagues who have provided advice:
Dare Baldwin; Dedre Gentner; Beth Levin; Gail Mauner;
Eric Pederson; Robert D. Van Valin, Jr., Phillip Wolff
 - ▶ all of whom shall be held blameless for any foolish and harebrained claims in this presentation
- ▶ our sponsor 
 - ▶ the material presented here is based upon work supported by the National Science Foundation under Grant No. BCS153846 and BCS-1644657, 'Causality Across Languages'; PI J. Bohnemeyer.

A Newton's cradle with five spheres is shown in a grayscale image. The cradle is positioned over a piece of paper that contains text. The word "Thanks!" is overlaid in blue on the cradle. The paper has text that is partially obscured by the cradle's frame. The text on the paper includes "ISAACO NEWTON", "Populi Communis", "Pp. Thoma Le Sire & Fournier", and "Ex Gallia M...".

Thanks!

REFERENCES

- Bohnenmeyer, J., N.J. Enfield, J. Essegbey, & S. Kita. (2010). The Macro-Event Property: The segmentation of causal chains. In *Event representation in language: Encoding events at the language-cognition interface*, eds. Jürgen Bohnemeyer and Eric Pederson, 43–67. Cambridge: Cambridge University Press.
- Breiman, L. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth International Group.
- Breiman, L. (2001). Random forests. *Machine Learning* 45(1): 5-32.
- Comrie, B. (1981). *Language universals and linguistic typology: Syntax and morphology*. Chicago: University of Chicago Press.
- Dixon, R.M. (2000). A typology of causatives: form, syntax and meaning. In *Changing valency: Case studies in transitivity*, eds. Robert M. W. Dixon and Alexandra Y. Aikhenvald, 30--83. Cambridge: Cambridge University Press.
- Escamilla Jr, R.M. (2012). *An updated typology of causative constructions: Form-function mappings in Hupa (Californian Athabaskan), Chungli Ao (Tibeto-Burman) and Beyond*. PhD Dissertation, University of California, Berkeley.
- Haiman, J. (1983). Iconic and economic motivation. *Language* 59(4):781–819.
- Haspelmath, M. (2008). Frequency vs. iconicity in explaining grammatical asymmetries. *Cognitive Linguistics* 19(1): 1-33.
- Kemmer, S. & A. Verhagen. (1994). The grammar of causatives and the conceptual structure of events. *Cognitive Linguistics* 5(2):115–156.

REFERENCES (CONT.)

- Levshina, N. (2015). European analytic causatives as a comparative concept: Evidence from a parallel corpus of film subtitles. *Folia Linguistica* 49(2): 487–520.
- Levshina, N. (2016). Why we need a token-based typology: A case study of analytic and lexical causatives in fifteen European languages. *Folia Linguistica* 50(2): 507–542.
- Levshina, N. (2017). Measuring iconicity: A quantitative study of lexical and analytic causatives in British English. *Functions of Language* 24(3): 319–347.
- Levshina, N. (2022). Semantic maps of causation: New hybrid approaches based on corpora and grammar descriptions. *Zeitschrift für Sprachwissenschaft* 41(1): 179-205.
- McCawley, J. (1976). Remarks on what can cause what. In *Syntax and Semantics VI: The grammar of causative constructions*, ed. Masayoshi Shibatani, 117–129. New York, NY: Academic Press.
- McCawley, J. (1978). Conversational implicature and the lexicon. In *Syntax and semantics IX: Pragmatics*, ed. Peter Cole, 245-258. New York, NY: Academic Press.
- Shibatani, M. (ed.) (1976). *The grammar of causative constructions*. New York: Academic Press (Syntax and Semantics; 6).
- Shibatani, M. & P. Pardeshi. (2002). The causative continuum. In *The grammar of causation and interpersonal manipulation*, ed. Masayoshi Shibatani, 85–126. Amsterdam: Benjamins.
- Talmy, L. (1976). Semantic causative types. In Masayoshi Shibatani (ed.), *Syntax and semantics, vol. 6: The grammar of causative constructions*, 43-116. New York: Academic Press.
- Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science* 12:49-100.
- Verhagen, A. & S. Kemmer. (1997). Interaction and causation: Causative constructions in modern standard Dutch. *Journal of Pragmatics* 27:61–82.
- Wolff, P. 2003. Direct causation in the linguistic coding and individuation of causal events. *Cognition* 88(1): 1–48.