**Using statistical classification to discover cross-linguistic semantic prototypes: The causation domain**

Jürgen Bohnemeyer, University at Buffalo – SUNY jb77@buffalo.edu https://www.acsu.buffalo.edu/~jb77/

**1. Introduction:** I present a study of the semantic/pragmatic ecology of the causative domain based on primary data collected from speakers of 13 languages (12 genera) spoken on four continents (cf. Table 1). I show that machine learning algorithms can facilitate the semantic/pragmatic mapping of a domain (here: causation) by generating hypotheses regarding semantic prototypes.

**2. Motivation:** 50 years of typological research on causatives has focused on the broad division of labor between simple and complex causative constructions (Dixon 2000; Shibatani & Pardeshi 2002; Song 1996; *inter alia*) and the role of iconicity in its motivation (Haiman 1983; Haspelmath 2008). What has been largely lacking to date are detailed and comprehensive examinations of the resource ecology of the causative domain based on primary data (but see Levshina 2022). This is where the present study aims to contribute.

**3. Methods:** The data was collected using an innovative combination of production and acceptability rating data. In a first phase, descriptions of 43 video clips were collected from speakers of each language. The video clips featured everyday causal chains and were designed to systematically vary *causer type* (CRType; intentional actor vs. accidental actor vs. natural force), *causee/affectee type* (CEAFType; controlled vs. psychologically impacted vs. physically impacted vs. inanimate), *mediation* (the presence vs. absence of an intermediate subevent/participant between cause and effect, often referred to as 'directness' in the literature), and further variables not included in the analysis presented here. Each language's causative constructions were extracted from the responses and new descriptions were created in consultation with speakers that crossed the stimulus scenes with the extracted response types. Every construction type was applied to every clip to the extent this was possible without making up new lexical items. This procedure resulted in an average of 381 descriptions of the clips per language. Ratings of the goodness of fit of these verbal stimuli as descriptions of the clips were then collected from 12+ speakers per language.

**4. Analysis:** Only response types that had been tested against more than 30 clips were included in the analysis. These response types are listed in Table 1. Three types of analyses of performed.

**4.1. Cluster analysis:** Figure 3 shows the result of a cluster analysis that compares the language-specific constructions in terms of the rating vectors they elicited (one rating per stimulus description and video clip). Although the analysis has no morphosyntactic information as its input, it finds three top-level clusters corresponding broadly to, from left to right, (i) various types of adverbial modifier and causal connective constructions (Cluster 1 in Table 1), (ii) lexical causative verbs, including derived causatives of limited productivity (Cluster 2); and (iii) periphrastic causatives and fully productive morphological causatives (Cluster 3).

**4.2. Predictive models:** Two types of classifiers were applied to the data, conditional inference trees (Hothorn et al. 2006) and random forests (Breiman 2001), using the Party and Ranger packages in R. The models predict ceiling rating of a given description for a given clip, indicating well-formedness, truth, and pragmatic appropriateness. The predictors were the clip variables (CRType, CEAFType, Mediation). Figures 1 and 2 show a conditional inference tree modeling the acceptability of the Zauzou periphrastic causative construction and the corresponding variable importance ranking.

**4.3. Multi-dimensional scaling (MDS):** MDS analyses were performed on matrices comparing rating vectors by participants and response types. Figure 4 illustrates for Zauzou (alpha-numeric codes representing participants, colors response types).

**5. Results and discussion – 5.1. Cluster analysis:** The cluster analysis confirms that different morphosyntactic types of causative constructions differ in their semantic and pragmatic properties. Furthermore, the three clusters differ broadly in morphosyntactic complexity, with Cluster 1 comprising the most complex and loosely integrated types, Cluster 2 consisting almost exclusively of simplex and complex lexical items, and Cluster 3 hosting most of the intermediate-complexity periphrastic causatives.

**5.2. Predictive models:** The predictive models can be interpreted as heuristics informing hypotheses regarding semantic prototypes. E.g., Figure 1 suggests that Zauzou periphrastic causatives have two discrete semantic prototypes: one for intentional human causers (*IHCr*) combined with controlled second participants (*ContrHCEAF*; 96.9% acceptability) and the other for natural force causers in the absence of controlled second participants (*NFCr*; 58.3% acceptability). Each prototype boosts acceptability to a local maximum, and the two are not contiguous in terms of the predictor variables, since they occur in distinct, complementary daughters of the highest node. The random forest model (Figure 2) confirms the dominance of the control exerted by the causee and the type of causer involved (intentional, accidental (*AHCr*), natural force) over the other predictor variables. Mediation emerges as the dominant predictor of the use of simplex causatives in most (though not all) sample languages. In contrast, surprisingly, for the vast majority of more complex causatives, the intentionality and control of the participants in the causal chain prove more important than mediation.

**5.3. Multi-dimensional scaling:** Figure 4 shows that the Zauzou construction type with the greatest amount of variation is the periphrastic causative construction with its multiple prototypes. This pattern recurs across the sample languages: the mid-level-complexity constructions show the greatest amount of inter-speaker variation.

**6. Conclusions:** Whereas most lexical causatives have unmediated causation as their unique semantic prototype, in line with what is predicted by the literature, 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).

**References:** Breiman, L. (2001). Random forests. *Machine Learning* 45: 5-32. * Dixon, R. M. W. (2000). A typology of causatives: Form, syntax and meaning. In R. M. W. Dixon & A. Y. Aikhenvald (eds.), *Changing valency: Case studies in transitivity* (pp. 30–83). Cambridge: Cambridge University Press. * 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. * Hothorn, T., K. Hornik, & A. Zeileis. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics* 15 (3): 651–674. * Levshina, N. (2022). Semantic maps of causation: New hybrid approaches based on corpora and grammar descriptions. *Zeitschrift für Sprachwissenschaft* 41(1): 179-205. * Shibatani, M., & P. Pardeshi. (2002). The causative continuum. In M. Shibatani (ed.), *The grammar of causation and interpersonal manipulation*. Amsterdam: Benjamins. 85–126. * Song, J. J. (1996). *Causatives and causation: A universal-typological perspective*. London: Longman.

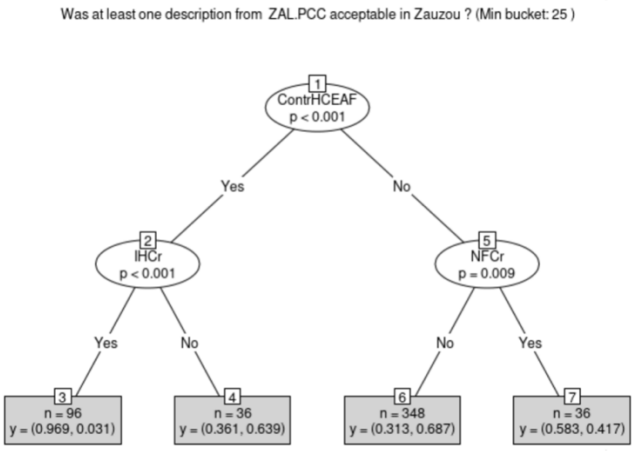


**Figure 1.** *Conditional inference tree based on acceptability ratings for the Zauzou periphrastic causative construction*

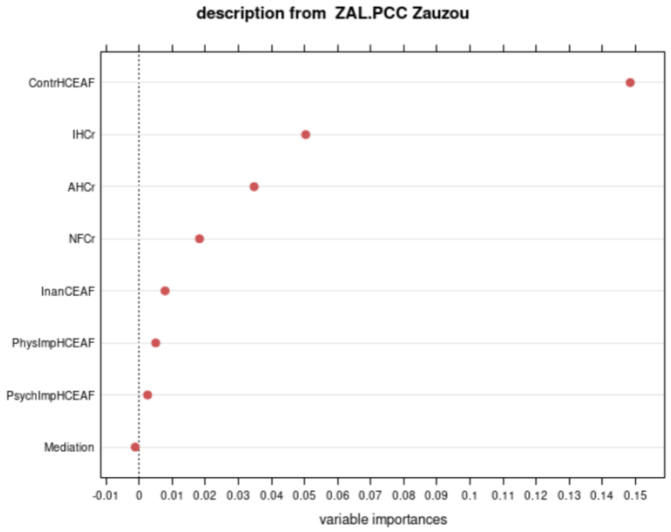


**Figure 2.** *Variable importance plot of random forest model based on acceptability ratings for the Zauzou periphrastic causative construction*

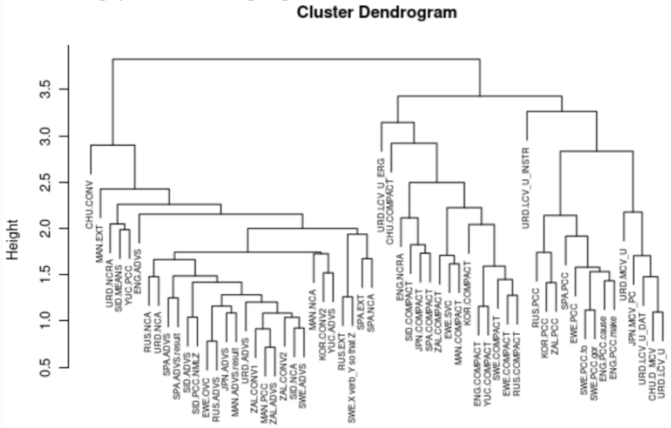


**Figure 3.** *Cluster dendrogram based on matrix comparing language-specific constructions in terms of the ratings vectors they elicited*

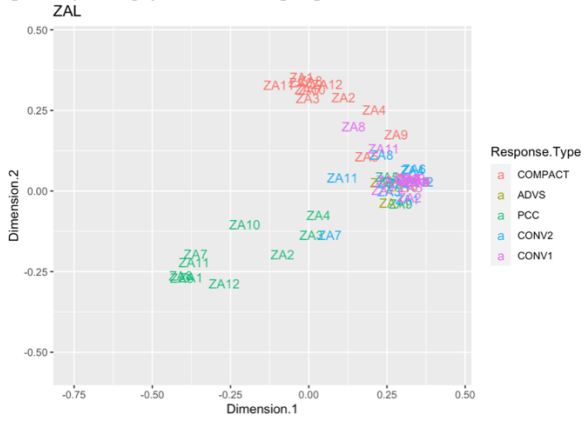


**Figure 4.** *Multi-dimensional scaling plot based on matrix comparing rating vectors for Zauzou participants × response types*

**Table 1**. *Sample languages by genus, field site, and response types included in the analysis, sorted by the output of the cluster analysis (Figure 3). ADVS – causal clause; ADVS.result – result clause; COMPACT – lexical causative, incl. not fully productive morphological causatives; CONV – converb construction; D_MCV – double morphological causative; EXT – extent ('so X that Y') construction; LCV – light verb construction; MCV – fully productive morphological causative; MEANS – means construction ('by pushing'); NCA – non-sentential cause adjunct ('because of the woman's push'); NCRA – non-sentential causer adjunct ('because of the woman'); OVC – 'overlapping' clause construction; PCC – periphrastic causative construction; PCC.NMLZ – periphrastic causative construction with cause nominalization; SVC – serial verb construction*

| Language | Genus | Field site | Cluster 1 | Cluster 2 | Cluster 3 |
|---|---|---|---|---|---|
| Chuvash | Turkic | Russia | CONV | COMPACT | D_MCV, MCV |
| English | Germanic | United States | ADVS | COMPACT, NCRA | PCC (2 types) |
| Ewe | Tano | Ghana | OVC | COMPACT, SVC | PCC |
| Japanese | Japanese | Japan | ADVS | COMPACT | MCV |
| Korean | Korean | South Korea | CONV | COMPACT | PCC |
| Mandarin | Chinese | China | ADVS.result, EXT, NCA, PCC | COMPACT | |
| Russian | Slavic | Russia | ADVS, EXT, NCA | COMPACT | PCC |
| Sidaama | Highland East Cushitic | Ethiopia | ADVS, MEANS, PCC.NMLZ, NCA | COMPACT | |
| Spanish | Romance | Spain | ADVS, ADVS.result, EXT, NCA | COMPACT | PCC |
| Swedish | Germanic | Sweden | ADVS, ADVS.result | COMPACT | PCC (2 types) |
| Urdu | Indic | Pakistan | ADVS, NCA, NCRA, | LCV | LCV (3 types), MCV |
| Yucatec | Mayan | Mexico | ADVS, PCC | COMPACT | |
| Zauzou | Loloish | China | ADVS, CONV (2 types) | COMPACT | PCC |