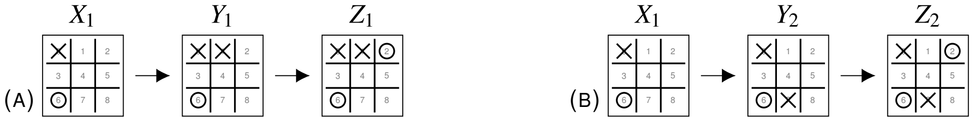
GRADED CAUSATIVES

Introduction. Semanticists have long been interested in how concepts present in causal relationships are lexicalized (C&H'15; B&S'21; N&S'22; L'00; S'11; S'76). The predominant approach to analyzing verbs of causing has been to argue that they convey some version of *sufficiency*, which is measured given parameters of a causal situation. Here, we provide experimental evidence for a differentiating and multi-faceted semantics of three causing verbs using explicitly-defined causal models, which represent how participants reason about the stimuli. This approach enables us to quantify concepts including sufficiency and use them as predictors.

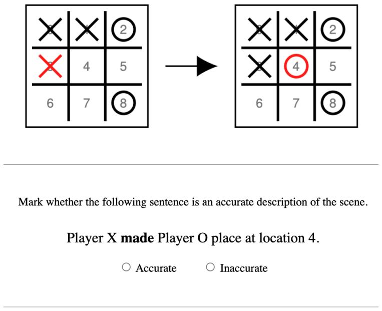
Contribution. We focus on the constructions *C caused/made/forced E* and argue that **H1.** *cause*, *make*, and *force* are in an asymmetric entailment relation, and that **H2**. this entailment relation is structured not by sufficiency, intentionality, or alternatives alone, but by an interaction of these three. Our experiment uses tic-tac-toe (ttt) sequences defined using structural causal models (SCMs; P'09). The use of SCMs enable us to make predictions about verb selection by defining probability distributions across counterfactual scenarios.



Possible scales. We postulate that graded causatives have a semantics built around threshold values on a continuous scale, similar to gradable adjectives. We consider three measures that are relevant features of causal relationships: ALT, INT, and SUF. Firstly, previous work (F'69; P'00) argues that *the number of alternative actions available to the causee* can distinguish between causal relationships in which the causer is (or is not) culpable for the action taken by the causee. This feature is also of interest for differentiating the semantics of causal verbs, since it provides the contrast in (1) *The child was* $\{$*made/?forced*$\}$ *to get into the car, although she could've chosen to do otherwise*. So, our first measure ALT quantifies the number of alternative actions available to the causee. This postulates that w.r.t. (1), the threshold ALT for *force* is less than the threshold ALT for *make*. In three-state ttt sequences as in Fig. (A), ALT is measured as the number of empty squares in the third state. So, ALT$(Y_1) = 5$. Secondly, the notion of *intention* is also strongly related to alternatives (W&M'06) and relevant for distinguishing causal situations (C'18). For example, consider that the pirate's intention is what distinguishes (2) *The pirate* $\{$*intentionally/?accidentally*$\}$ *forced the prisoner down the plank*. Building on this intuition, our second model (INT) is based on the 'degree of intention' proposed by H&K-W'18 (see their paper for details), which is roughly the probability of reaching the goal state given the current action versus given alternative actions. This is why (3) *Player O placing at location 2* is more intentional in Z_1 than Z_2. Specifically, any alternative to *Player O placing at location 2* in (A), e.g. *Player O placing at location 5*, would make it highly probable that *Player X* wins at the next time-step, thereby largely decreasing the probability of reaching the goal-state of *Player O*. The same is not true for (B). Thirdly, the notion of *causal sufficiency* has been well-represented in previous literature on causal verb selection – G'23 argues that *cause* entails local sufficiency, while L&N'18 and N&L'20 argue that *make* conveys (non-probabilistic) causal sufficiency. Intuitively, this distinguishes between causing and enabling verbs – in (4a/b) *The pirate* $\{$*made/let*$\}$ *the prisoner walk down the plank*, we can say that likely *the prisoner walks down the plank* in (4a) while it is less clear whether this result comes about in (4b). Thus, our third model (SUF) is P'18's 'probability of sufficiency', which is defined as the probability that the event $X = 1$ would be sufficient to produce outcome $Y = 1$. Descriptively, SUF denotes the capacity of C to produce the outcome E in situations where the agent of C did some action other than the one encoded in C. Intuitively, *Player X placing at location 1* in Y_1 is more sufficient in bringing about *Player O placing at location 2*, than *Player X placing at location 7* in Y_2 is for bringing about the same. This is because in sequences where settings Y_1 and Y_2 don't result in *Player O placing at location 2* at the next time-step, it is more likely that Y_1 will *eventually* lead to *Player O placing at location 2* to block X's clear three-in-a-row than Y_2, which does not present that danger to *Player O*. To conclude our measurements, observe that our definitions have been applied to ttt sequences, which can be defined as partial setting of a SCM. This means that given some setting of variables in a SCM, we can apply functions ALT, INT, and SUF that output a numerical value. Minimally, a probabilistic SCM has a set of exogenous variables with an associated probability distribution, a set of endogenous variables, and a set of deterministic functions that assigns a value to each exogenous variable given values of some subset of exogenous and endogenous variables (see C-et-al'18 for technical detail). This framework can encode any causal process as a directed acyclic graph (DAG). Thus, we choose the game of ttt as experimental stimuli, since an entire game-tree can be efficiently stored as a DAG (and consequently defined as a SCM). In this way, an endogenous board-state variable is stored as a conjunction of location-demarcation assignments. So, given a statement such as *Player X placing at location 3 made Player O place at location 5*, we can measure the number of possible alternative actions that Player O could have taken besides placing at location 5, the degree of intention of that Player X had for bringing about the event of Player O placing at location 5, and the probability that Player X placing at location 3 would bring about Player O placing at location 5.

Experiment. Our stimuli consist of 30 two-frame ttt sequences, filtered from 21 full games to represent the range of possible ALT, INT, and SUF values. Participants were asked to rate whether sentences such as *Player X forced Player O to place at location 3* are accurate in describing the stimulus (see example in Fig. 2). We recruited 109 L1 English participants, of which 19 were excluded for failing attention check(s).

Results/Analysis. We find that holding the set of stimuli constant, participants were less likely to determine *made* than *caused* as accurate in describing a scenario, and less likely to determine *forced* than *made* as accurate (Fig. 3). This supports **H1**, since the semantic interpretations of weaker predicates are entailed by the use of stronger ones (M-et-al'10). Regarding **H2**, we fit (I) an initial Bayesian linear regression using participant judgements as the outcome variable and model such using a Bernoulli distribution. The predictors include the verb used in the sentence presented to participants, the ALT, INT, and SUF value of the associated stimuli as fixed effects, as well as their interactions. The results (full model results in Tab. 1) provide evidence that besides the different levels of verb, SUF and the three-way interaction of ALT:INT:SUF has a non-zero effect on the response variable. We then fit a second regression (II) that predicts judgements using only verb and SUF. We find that $WAIC(I) = 1941.93$ ($SE = 32.24$), $WAIC(II) = 2003.09$ ($SE = 28.18$), and $WAIC(I) - WAIC(II) = -61.15$ ($SE = 17.24$), indicating that (I) is the better fit, and that our results are better explained by including all three predictors and their interactions, than by SUF alone. Next, we fit follow-up regressions similar to (I), except without INT and all of its interactions (III), and without ALT and all of its interactions (IV). Comparing (I) to (III), we get $WAIC(III) = 1961.96$ ($SE : 30.69$) and $WAIC(I) - WAIC(III) = -20.02$ ($SE = 10.42$), indicating that since (III) does reliably worse, the predictor INT does matter despite including 0 in its CrI in regression (I). Comparing (I) to (IV), we get $WAIC(IV) = 2001.33$ ($SE = 28.57$) and $WAIC(I) - WAIC(IV) = -59.40$ ($SE = 16.79$), indicating that since (IV) does reliably worse, the predictor ALT also matters (despite also including 0 in its CrI in regression (I)). To conclude, our Bayesian analysis demonstrate that all three features – ALT, INT, and SUF – have reliable effects on participant judgements of *cause*, *make*, and *force*. This work demonstrates that these causatives not only encode information about sufficiency, but also intention and possible alternative actions.



Mark whether the following sentence is an accurate description of the scene.

Player X **made** Player O place at location 4.

○ Accurate ○ Inaccurate

FIGURE 2. Example of experiment question.

	Estimate	Est.Error	l-95% CI	u-95% CI
Intercept	-3.96	0.75	-5.42	-2.49
verbmade	-0.35	0.13	-0.61	-0.09
verbforced	-0.62	0.14	-0.90	-0.36
SUF	5.97	1.48	3.10	8.88
INT	0.19	1.92	-3.60	3.85
ALT	0.32	0.19	-0.07	0.68
SUF:INT	-4.97	3.49	-11.74	1.89
SUF:ALT	0.08	0.50	-0.90	1.07
INT:ALT	-0.25	0.49	-1.21	0.71
SUF:INT:ALT	2.72	1.17	0.34	5.02

TABLE 1. Full model results for (I).

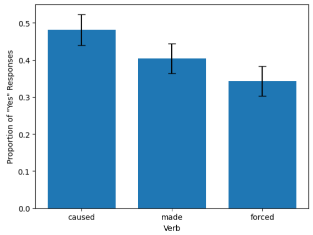


FIGURE 3. Proportion of "Yes" w/ 95% CIs.