Aspectual similarity predicts sense similarity

Aaron Steven White,[1] Scott Grimm,[1] & Lelia Glass[2*]

Abstract. A long-standing assumption in the lexical-aspect literature is that telicity and durativity are sense-level features: once a verb's sense is fixed in a given syntactic frame, its event descriptions are aspectually homogeneous. We show that this assumption requires revision. Using data from two experiments aimed at measuring telicity and durativity for 40 verb forms, we derive per-sentence probabilities of telic and durative interpretation that partial out the standard compositional and pragmatic predictors of aspect, and find substantial within-sense, within-syntactic-frame residual variability, even for verb classes prior work treats as telicity-stable. In a paired-similarity task on the same materials, we find that this within-sense variability is predictive: two same-sense tokens that match aspectually are reliably rated as more similar than two that do not, with the within-sense aspectual-matching effect larger than the across-sense one on both telicity and durativity. Aspect thus varies meaningfully within a sense, and that variation registers in how speakers judge similarity.

Keywords. telicity; durativity; lexical aspect; word sense

1. Introduction. A standard view in the lexical-aspect literature treats telicity and durativity as features of a verb sense: once a verb's sense is fixed in a given syntactic frame, its event descriptions are taken to be aspectually homogeneous, with any remaining variation attributed to compositional or pragmatic factors of the description itself (Verkuyl 1972, Dowty 1979, Verkuyl 1989, Krifka 1998, Kratzer 2004, Rothstein 2004, Tenny 1994, Beavers et al. 2010, Beavers 2011, Moens & Steedman 1988, Pustejovsky 1995, Filip 2012). (1)–(3) illustrate that the view is not obviously right: *hit* uniformly has a sense involving contact in all three, all three are transitive with definite arguments in the same syntactic context, yet (1) is telic, (2) is atelic, and (3) can be either.

(1) The arrow hit the target {in, #for} one second. (*under durative interpretation of adverb*)
(2) The boxer hit his opponent {#in, for} two minutes. (*under durative interpretation of adverb*)
(3) The hurricane hit the coastline {in, for} 10 hours.

This paper asks whether the sense-level view holds up in two places where it makes testable predictions. First, if telicity and durativity are sense-level features, then within-sense, within-syntactic-frame residual variability in those properties should be small once the standard compositional and pragmatic predictors are controlled for. In Sections 2 and 3, we combine a cloze task measuring per-sentence preferences for temporal modifiers headed by *in* or *for* with a proportion task that diagnoses the two known confounds of those modifiers (the inchoative interpretation of *in* and the result-state interpretation of *for*). From the two, we derive per-sentence probabilities of telic and durative interpretation (Section 4). Consistent with the observation about (1)–(3), we find that substantial residual variability remains, even for verb classes that the literature treats as telicity-stable.

Second, if telicity and durativity were exhausted by the sense, then once a sense is fixed, two same-sense event descriptions that match aspectually should be rated no more similar than two that do not (the sense label has already done the work). In a paired-similarity task on the same materials,

*Author affiliations: [1]University of Rochester, [2]Georgia Institute of Technology.

Semantic field	*Manner*/result pairs	Source
Cleaning	*wipe*/clear, *scrub*/clean	Levin & Rappaport Hovav (1991)
Contact caused change	*hit*/break, *jab*/shatter, *bump*/chip, *kick*/split	Fillmore (1970), Levin (1993)
Contact caused motion	*push*/open, *pull*/close, *pound*/fold	Fillmore (1970), Levin (2015, 2020)
Violence	*stab*/kill, *shoot*/murder, *fight*/destroy	Beavers & Koontz-Garboden (2012), Levin (2015)
Moving liquid	*wash*/empty, *pour*/fill	Levin & Rappaport Hovav (1991), Levin (2015)
Moving substance	*smear*/cover, *shake*/combine	Gentner (1978), Levin & Rappaport Hovav (1991), Levin (2015)
Deliberating	*consider*/conclude, *investigate*/discover, *ponder*/decide, *debate*/prove	Fillmore (2006)

Table 1: Manner/result pairs grouped by semantic field.

we find the opposite in Section 5: aspectual matching predicts perceived similarity reliably, and the contribution of aspectual matching is sharper within a sense than across one. Aspect therefore varies meaningfully within a sense, and that variation registers in similarity judgments.

2. Experiment 1a: A cloze diagnostic for telicity. Experiment 1a measures the per-sentence preference for associating an event description with a temporal modifier headed by *in* versus *for* (Vendler 1957, Dowty 1979), our initial measure of telicity (refined in Experiment 1b). We use a cloze task in which the response options are constrained to a temporal modifier of that form.

2.1. MATERIALS. Experiments 1a and 1b share a set of event descriptions, constructed around 40 transitive verbs in 20 manner–result pairs spanning a variety of semantic domains (Table 1). The two members of each pair are semantically similar except that one is classified in the literature as a manner verb and the other as a result verb. Deliberation verbs, which have seen less discussion in the literature, are divided using FrameNet frames (Fillmore 2006): manner verbs (e.g. *consider*) fall into the 'cogitation' frame, result verbs (e.g. *conclude*) into 'coming to believe'.

For each verb we construct items at three points along a single GENERALITY cline that ranges from maximally specific to maximally general event descriptions. *Contentful* sentences sit at the specific end: their arguments are concrete definite NPs that pick out a narrow set of event-tokens compatible with one annotated sense. *Templatic* sentences sit at the general end: their arguments are variables (*X*, *Y*) that pick out the maximal set of event-tokens compatible with any sense of the verb. *Bleached* sentences sit between them: their arguments are definite NPs headed by general nouns that pick out a wide but bounded set of event-tokens compatible with a subset of the verb's senses, determined by which general noun heads the bleached argument.

Varying the size of an item's compatible-sense set in this way allows us to evaluate the sense-level view of aspect: if telicity and durativity were properties of the sense, an item should behave aspectually like an aggregation (mean, max, sample, etc.) over items at a narrower generality level whose senses fall inside its compatible-sense set: bleached items at the aggregation of their compatible-sense contentful items and templatic items at the verb-form aggregate. The cline thus stratifies the data along three units of analysis: verb TYPE (manner vs. result), verb SENSE (e.g. *shoot.01* vs. *shoot.02*), and verb TOKEN (the individual event description). We find that aspect varies at each level. We model aggregation as a mean over an empirically derived probability distribution over verb senses for each verb. As a proxy for the true verb sense set for each verb, we use that verb's PropBank rolesets. We stress that these rolesets are merely a proxy and are likely too fine-grained in some places and too coarse-grained in others. We return to this point in

Section 6 but note here that our results throughout the paper suggest they are a reasonable proxy.

To estimate a probability distribution over these senses, we use the English portion of OntoNotes 5.0 (Weischedel et al. 2013) to obtain sense frequencies. For each verb we count how often each PropBank roleset is assigned to the lemma, and normalize within verb to obtain $p(\text{sense} \mid \text{verb})$. These per-verb sense distributions feed the multi-membership weighting of bleached and templatic items in this section and Section 3 as well as the same-sense predictor in Section 5.

Contentful items. For each verb we gathered transitive PropBank senses (Palmer et al. 2005) (e.g. `hit.01` *strike*, `hit.02` *reach*), set aside highly idiomatic senses (`hit.03` *go to*) and senses favoring non-*the* determiners (`fold.02` *fold hands*), and labeled each as concrete or abstract.

Senses supporting both interpretations were split, as in (4). For each (sub)sense, we hand-wrote two contentful sentences with definite NP subjects and objects in the past tense.

(4) a. The batter hit the ball. (`hit.01-concrete`)
b. The idea hit the researcher. (`hit.01-abstract`)

Bleached items. For each contentful sentence we constructed a bleached counterpart by replacing the contentful arguments with definite NPs headed by general nouns (primarily *person* and *thing*, with *organization*, *event*, *time*, *place*, and *stuff* as fallbacks; White & Rawlins 2016, 2020). When both arguments would share a head noun, we inserted *other* into the object NP, as in (6). A bleached item is in general compatible with multiple senses of its verb. We restrict its licensed senses to the concrete variants by default, and to the abstract variants only when the bleached subject or object is itself inherently abstract (in our materials specifically *time* or *event*, as in (5)).

(5) a. The event filled the time.
b. The person shot the event.

(6) a. The arrow hit the target.
b. The thing hit the other thing.

This default reflects both the experimenters' modal intuitions and a convergent literature on dual-coding (Paivio 1991) and on Conceptual Metaphor Theory (Lakoff & Johnson 1980), which treats concrete senses as the source domains from which abstract senses are systematically extended.

Templatic items. At the maximally general end of the cline we generated, for each verb, a templatic sentence in which the subject is realized as *X* and the object as *Y*, after Kane et al. (2022). Templatic items pick out the full set of the verb's senses with no restriction from the argument heads. The per-item sense distribution is therefore the OntoNotes-derived prior over all senses.

2.2. PROCEDURE. On each trial, the participant saw a sentence containing three in-line dropdowns: *in* or *for* (PREPOSITION), a numeral from 1–100 in increments of 1 from 1–10 and increments of 5 form 10–100 (NUMERAL), and a time unit from *milliseconds*, *seconds*, *minutes*, *hours*, *days*, *weeks*, *months*, *years*, *decades*, *centuries*, or *millennia* (TIME UNIT), as in (7).

(7) The child hit the pillow {in, for} {1, . . . , 100} {milliseconds, . . . , millennia}.

Participants were asked to fill the dropdowns with the values they thought most natural, and were advised that multiple answers could make sense. We allowed participants to choose NUMERAL and TIME UNIT, whose concatenation we refer to as *duration*, because we no independent norm for the distribution of durations for the events described by our descriptions and wanted to avoid inadvertently biasing preposition choice. This approach turns out to be important because there is a strong relationship between longer durations and higher proportions of *for* responses (longer durations, we believe, encourage habitual interpretations).

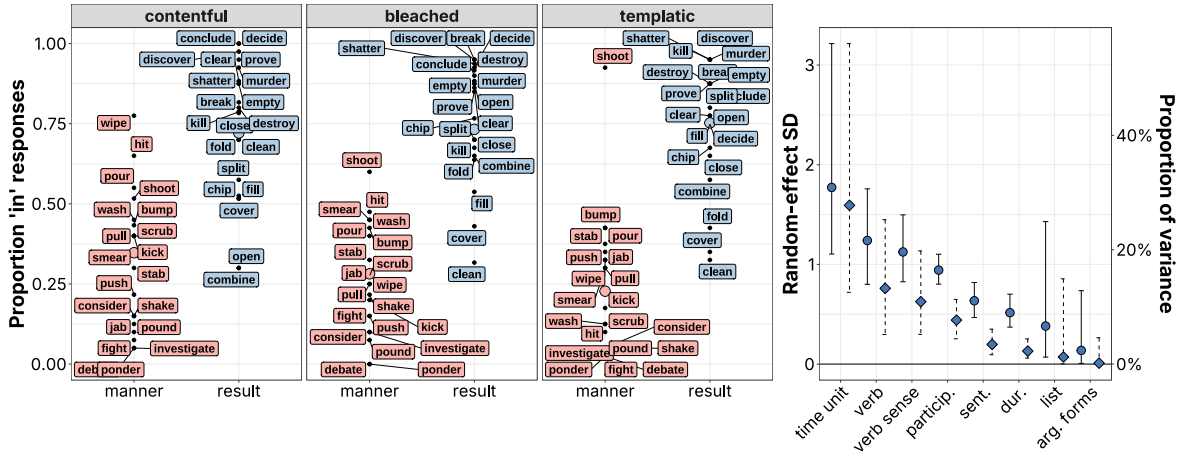

Figure 1: Cloze *in*-rates by VERB TYPE × GENERALITY (left; small points = individual verbs, large points = cell-level posterior means) and per-group random-effect SDs (right; circles, left axis) with the share of total variance they account for (diamonds, right axis).

2.3. PARTICIPANTS. We recruited 120 participants through Prolific. All participants were self-identified native English speakers located in the United States. No participant was permitted to participate in more than one of the studies reported here. The average payment rate was $14.31/hour.

2.4. ANALYSIS. We fit a Bayesian mixed-effects logistic regression in `brms` (Bürkner 2017) predicting PREPOSITION from VERB TYPE (sum-coded) crossed with GENERALITY (successive-difference-coded (Venables & Ripley 2002) so the two coefficients estimate the per-step movement along the cline). The random-effects structure absorbs the standard formal predictors of telicity (argument shape, modifier form, sense, and the design factors), so that what remains at the by-SENTENCE and by-VERB SENSE level cannot be attributed to them (see the analysis code for the full model specification). Bleached and templatic items contribute fractionally to each of their compatible (sense, concreteness) combinations via the multi-membership `mm()` term, with weights set by the OntoNotes-derived $p(\text{sense} \mid \text{verb})$ described above under add-1 smoothing.

2.5. RESULTS. A robust verb-class difference emerges (Figure 1): result verbs prefer *in* substantially more than manner verbs do ($\beta_{\text{VERB TYPE}} = 1.57$ log-odds, $[1.09, 2.08]$), reliable across all levels of GENERALITY; the GENERALITY effect and its interaction with VERB TYPE are small and unreliable. The by-verb, by-verb-sense, and by-sentence random-effect SDs (Figure 1, right) are jointly large enough that per-item variability swamps the verb-class effect for many individual items. The within-verb variability illustrated by (1)–(3) is therefore not an outlier phenomenon: substantial variation in telicity remains across event descriptions even after the standard formal predictors of telicity are absorbed. We use the per-sentence cloze *in*-rate as an initial measure of telicity, refining it and complementing it with a measure of durativity using Experiment 1b.

3. Experiment 1b: A temporal-overlap diagnostic for durativity. The *in–for* diagnostic is subject to systematic noise from alternative interpretations (Filip 1999, Kearns 2007): *in* can describe a delay before the event begins, as in (8a), and *for* can describe the duration of an event's result state, as in (8b). Both interpretations are non-durative in the relevant sense and contaminate the cloze diagnostic. Experiment 1b diagnoses them by asking participants directly what proportion

of the modifier's time span the event was occurring.

(8) a. The water was bubbling in two minutes. (*delay interpretation of in*)
b. The woman opened the window for five minutes. (*result-state interpretation of for*)

3.1. MATERIALS. The materials reuse Experiment 1a's verb set, sense inventory, and three-point GENERALITY cline (contentful, bleached, templatic). Each base sentence was presented with a fixed temporal modifier whose numeral-by-time-unit combination was the modal Experiment 1a response to that sentence under each preposition.

3.2. PROCEDURE. On each trial the participant saw a sentence with a temporal modifier headed by *in* or *for*, followed by a question asking what proportion of the modifier's time span the event was happening: *none of* TIME, *less than half of* TIME, *about half of* TIME, *more than half of* TIME, or *the entirety of* TIME. The response options for (9b) would have TIME = *those five minutes*.

(9) a. The child hit the pillow for five minutes.
b. For what proportion of those five minutes was the child hitting the pillow?

A *none of* response signals a non-overlap interpretation (under *in*, the delay interpretation; under *for*, the result-state interpretation or some other modifier-event mismatch), while the other four signal overlap to varying degrees (and thus a durative interpretation for the adverb).

3.3. PARTICIPANTS. We recruited 250 participants through Prolific, subject to the same eligibility and exclusion criteria as Experiment 1a.

3.4. ANALYSIS. We binarized the response as OVERLAP=1 when the participant chose any option other than *none of* (OVERLAP=0), then fit a Bayesian mixed-effects logistic regression in `brms` predicting OVERLAP from a three-way interaction of PREP(OSITION(PREF(ERENCE) (the per-sentence cloze in-rate from Experiment 1a, scaled to $[-1, 1]$), PREPOSITION, and GENERALITY. The random-effects structure mirrors Experiment 1a, with by-grouping PREPOSITION slopes added where the design supports them (see the analysis code for details).

3.5. RESULTS. Figure 2 shows the proportion of OVERLAP responses against the per-sentence cloze *in*-rate, separately for the two prepositions. We observe a reliable interaction between PREP PREF and PREPOSITION ($\beta = 0.60$, $[0.17, 1.06]$): sentences that prefer *in* in the cloze task show more overlap responses when paired with *in*, while sentences that prefer *for* show more overlap when paired with *for*. The participants' overlap judgments track the cloze preferences in the predicted direction, validating cloze preposition preference as an initial aspectual diagnostic. We believe the high overall proportion of overlap responses is a task effect (four of five response options indicate overlap) that we control for in the durativity measure developed in Section 4.

4. Deriving per-sentence telicity and durativity. The Experiment 2 analysis predicts pairwise similarity from a per-sentence telicity and a per-sentence durativity. For each sentence, the cloze fit gives a posterior over its preference for *in* versus *for*, and the proportion fit gives a posterior over how readily the sentence licenses overlap under either preposition. We derive per-sentence telicity from the first and per-sentence durativity from the second by recasting each posterior as the probability that the sentence belongs to a latent binary class (telic vs. atelic for telicity, durative vs. punctual for durativity), given what the corresponding fit says about it.

Each per-sentence prediction is evaluated at the time unit at which the sentence's empirical

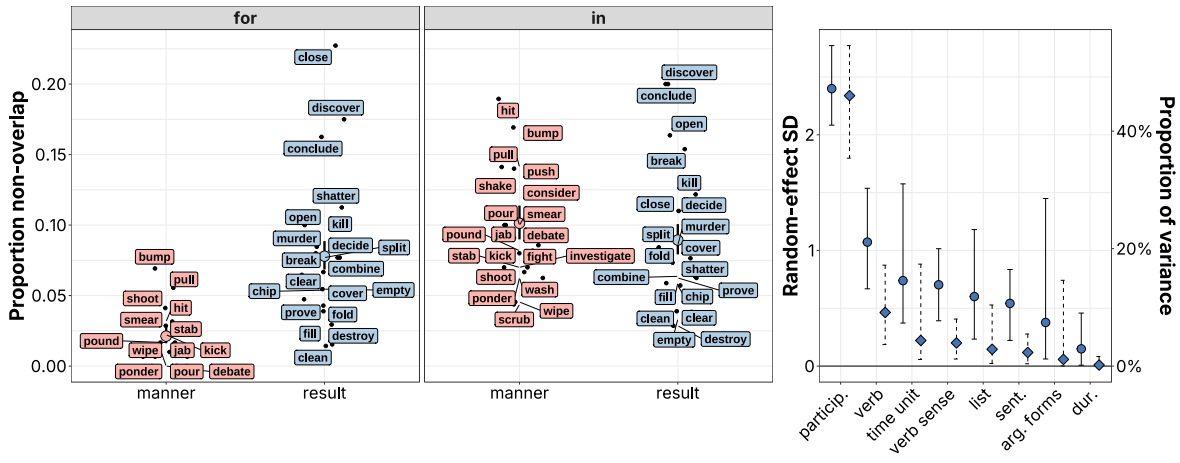

Figure 2: Per-verb non-overlap rates by preposition (left; small points = individual verbs, large points = cell-level posterior means) and per-group random-effect SDs (right; circles, left axis) with the share of total variance they account for (diamonds, right axis).

responses cluster. We identify that time unit by converting each response's discrete DURATION to a continuous duration in minutes. We then map this value to log-minutes, which puts the time units on a roughly equally-spaced scale. We apply a Gaussian kernel density estimator over the per-sentence log-minute durations, taking the mode, and snapping to the time unit t whose canonical (numeral $= 1, t$) duration sits closest to the mode in log-minutes.[1] The resulting evaluation time unit falls in $\leq$days for 83% of sentences (firmly episodic for the verbs we use) and $\geq$weeks for the remaining 17%. Some of the latter have salient stative interpretations (10) though not all (11).

(10) The insurance covered the damage. (11) The detective investigated the crime.

Per-sentence telicity. Experiment 2 requires a per-sentence probability that the sentence is telic (a quantity in $[0, 1]$ we can plug in as a covariate). The cloze fit's natural per-sentence signal is on a different scale: at the canonical evaluation configuration $\theta_0(i)$ (numeral $= 1$, time unit $= t_i$, participant and list random effects marginalized via the standard new-level trick), the per-sentence linear predictor $\ell_i^c = \text{logit}\, P_{\text{cloze}}(\text{in} \mid i, \theta_0(i))$ lives on an unbounded logit scale and has its own predictive uncertainty. Telic-leaning sentences sit high on the logit scale; atelic-leaning ones low; the predictive uncertainty tells us how sharply the model has pinned the sentence down.

Recasting as a latent class. The predictive distribution of ℓ_i^c conflates two distinct things: where the sentence sits relative to the cloze fit's grand intercept α_c (the evidence that it is telic), and how diffusely the fit localized it (the noise on that evidence). We collapse these into the single probability we want using a standard recipe from Bayesian cognitive modeling (Lee & Wagenmakers 2014): treat ℓ_i^c as a noisy indicator of a latent binary class and apply Bayes' rule. Posit a latent $\tau_i \in \{-1, +1\}$ (with $\tau_i = +1$ for telic) generating ℓ_i^c through a Gaussian model $\ell_i^c \sim \alpha_c + \gamma_c \tau_i + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}\big(0, \sigma_{c,i}^2\big)$, where $\sigma_{c,i}^2$ is the predictive variance of ℓ_i^c and γ_c is the per-class shift on the logit scale, i.e. the per-sentence offset we would expect if the class were observed. Under a flat prior on τ_i, Bayes' rule gives the per-sentence posterior class probability $\text{tel}_i = P(\tau_i = +1 \mid \ell_i^c) = \sigma\big(2\,\gamma_c\,(\mu_{c,i} - \alpha_c)/\sigma_{c,i}^2\big)$, where $\mu_{c,i}$ is the predictive mean of ℓ_i^c. An

[1] We do not take the mode over DURATION directly because there is substantial variability in the NUMERAL used with a particular TIME UNIT; the kernel density estimator smooths over this variability.

extreme logit far from α_c that the fit localized sharply pushes the sentence close to 0 or 1; a diffuse predictive distribution, or one that sits near the population mean, stays near $1/2$.

Per-sentence durativity. We construct the analogous quantity from the proportion fit: a per-sentence probability that the sentence is durative. A durative sentence is one for which an overlap response is readily licensed under at least one preposition, so the natural per-sentence signal is the prep-marginal overlap probability, $\ell_i^d \equiv \text{logit}\ \frac{1}{2}\big(P_{\text{prop}}(\text{overlap} \mid i, \text{in}, \theta_0(i)) + P_{\text{prop}}(\text{overlap} \mid i, \text{for}, \theta_0(i))\big)$, again at $\theta_0(i)$, and with `pref_in` held at its grand mean so the cloze-driven asymmetry between the two preposition cells does not re-enter through the three-way interaction.

Recasting as a latent class. The raw $\frac{1}{2}(p_{\text{in}} + p_{\text{for}})$ does not work directly: four of the five proportion-task response options indicate some overlap, so the marginal sits well above $1/2$ for almost every sentence (the task effect flagged in Section 3). What discriminates durative from punctual is not the raw level but how high a sentence sits relative to that baseline. The same latent-class recasting as for telicity, with a latent $\delta_i \in \{-1, +1\}$ (durative when $+1$) generating ℓ_i^d through a Gaussian model with class spacing γ_d and baseline α_d, gives by Bayes' rule $\mathsf{dur}_i = \sigma\big(2\,\gamma_d\,(\mu_{d,i} - \alpha_d)/\sigma^2_{d,i}\big)$, with $\mu_{d,i}$ and $\sigma^2_{d,i}$ the predictive mean and variance of ℓ_i^d. Sentences that sit high above the prep-marginal baseline (where overlap is robust under at least one preposition) land near 1. Sentences whose marginal sits at or below it (where overlap is suppressed by both the result-state interpretation and the delay interpretation) land near 0.

Per-pair match scores. The Experiment 2 analysis uses a per-pair notion of telicity match and durativity match. For each pair (i, j) and aspectual axis $\mathsf{a} \in \{\mathsf{tel}, \mathsf{dur}\}$, we treat the per-sentence probability a_i as the parameter of an independent Bernoulli draw and define the match score as the probability that the two draws agree: $\mathsf{match}^{\mathsf{a}}_{ij} = \mathsf{a}_i\,\mathsf{a}_j + (1 - \mathsf{a}_i)(1 - \mathsf{a}_j)$.

Face-validity check. The lowest-durativity sentences are uncontroversially achievement-like, as in (12); the highest-durativity sentences are extended changes of state, as in (13).

(12) a. The dean bumped the meeting.
b. The teacher bumped the deadline.

(13) a. The people destroyed the stuff.
b. The things filled the other thing.

Figure 3 plots each contentful or bleached sentence in $(\mathsf{tel}_i, \mathsf{dur}_i)$ space, faceted by verb. Verbs with tight hulls occupy aspectually contiguous regions of their family; verbs with diffuse hulls span aspectually discontiguous sub-regions.

5. Experiment 2: Sense similarity. Experiment 2 asks how the per-sentence telicity and durativity measures derived in Section 4 relate to perceived sense similarity, and how that relationship is modulated by sense identity and by position on the generality cline introduced in Section 2. Each pair in Experiment 2 therefore differs along several dimensions: (a) how much aspectual structure the two descriptions share; (b) the probability that they pick out the same sense; and (c) where they sit on the generality cline relative to one another.

5.1. PROCEDURE. Each trial paired two event descriptions from Experiment 1 sharing a verb, as in (14) (within sense) and (15) (across sense). The temporal modifier was removed; participants rated the similarity of the bare descriptions on a continuous slider labeled *completely identical* on the left and *extremely different* on the right. Throughout, we work in the slider's coding so lower values mean greater similarity. The slider always began at the left edge; participants had to move

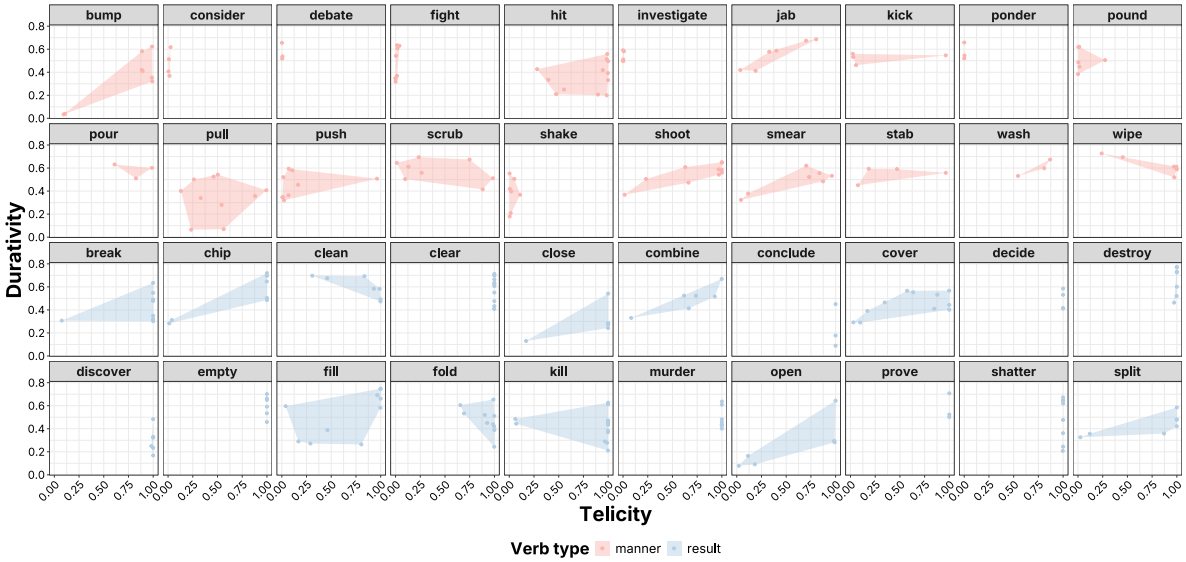

Figure 3: Per-sentence $(\text{tel}_i, \text{dur}_i)$ for contentful and bleached items, faceted by verb (manner first, then result, alphabetical within each block); convex hulls shaded by verb type.

it to submit. Calibration items with known very-similar and very-different pairs were interspersed and excluded from analysis.

(14) a. The analyst cleaned the data.
b. The housekeeper cleaned the kitchen.

(15) a. The cat killed the mouse.
b. The joke killed the evening.

5.2. PARTICIPANTS. We recruited 150 participants through Prolific, subject to the same eligibility and exclusion criteria as Experiments 1a and 1b.

5.3. HYPOTHESES. We consider four hypotheses about how aspectual similarity, sense identity, and position on the generality cline jointly predict perceived similarity.

H1 Aspectual similarity predicts perceived similarity in its own right: matching on either telicity or durativity should predict greater perceived similarity even after sense identity is controlled for.

H2 Pairs likely to share a verb sense should be judged more similar even after aspectual match is controlled for. Under both, the corresponding main effects on dissimilarity are negative.

H3 The contribution of aspectual similarity to perceived similarity is amplified within a sense. The idea is that pairs likely to share a sense belong to the same regular-polysemy family, where their relevant variation is already constrained to a coherent semantic neighborhood. Within such a neighborhood, fine-grained aspectual distinctions are not swamped by sense identity.

H4 Position on the generality cline shifts perceived similarity in its own right. Pairs whose two descriptions sit at different positions on the cline differ in the set of event tokens each description picks out, and may be judged systematically more or less similar as a result.

5.4. ANALYSIS. We fit a Bayesian mixed-effects ordered-beta regression (Kubinec 2023) in `brms` (via `ordbetareg`), predicting the similarity rating from four pair-level predictors and their pairwise interactions: telicity match $\mathsf{match}^{\mathsf{tel}}_{ij}$ and durativity match $\mathsf{match}^{\mathsf{dur}}_{ij}$ from Section 4; same-sense s_{ij}, the dot product of the two items' multi-membership sense distributions from Section 2

(concreteness marginalized); and generality pair g_{ij}, indicating whether the two descriptions sit at the same position on the generality cline (both contentful) or at different positions (one contentful, one bleached). The continuous predictors are centered and SD-scaled; generality pair is sum-coded. The random-effects structure is the maximal one the design supports for the main effects: by-VERB and by-PARTICIPANT random intercepts and slopes for each main-effect predictor, plus a by-SENTENCE PAIR random intercept. (See the analysis code for the full specification.)

5.5. RESULTS. We find that **H1–4** are confirmed. **H1** holds on both aspectual axes. Matching on telicity contributes $\beta = -0.134$ ($[-0.253, -0.007]$, $P_{\text{sign}} = .98$) and matching on durativity contributes $\beta = -0.125$ ($[-0.269, +0.008]$, $P_{\text{sign}} = .97$): in both cases nearly all posterior mass sits on the hypothesized direction, with the durativity-match posterior placing $\approx 3\%$ of its mass on the opposite sign. **H2** and **H4** are supported as well: same-sense contributes $\beta = -0.510$ ($[-0.656, -0.365]$, $P_{\text{sign}} > .999$), the largest main effect in the model, and generality pair contributes $\beta = -0.304$ ($[-0.385, -0.222]$, $P_{\text{sign}} > .999$), with contentful-contentful pairs rated systematically more dissimilar than contentful-bleached pairs after aspect and sense are controlled for. Finally, **H3** holds on both aspectual axes. The telicity-by-same-sense interaction is $\beta = -0.092$ ($[-0.159, -0.026]$, $P_{\text{sign}} > .997$) and the parallel durativity-by-same-sense interaction is $\beta = -0.063$ ($[-0.134, +0.005]$, $P_{\text{sign}} = .97$): both posteriors place nearly all their mass on the predicted direction, with telicity contributing a larger posterior-mean effect than durativity. Pairs likely to share a sense show sharper aspectual-match-to-similarity coupling on both axes.

Three of the four remaining interactions in the model, for which none of our hypotheses make a prediction, have near-uniform posteriors over sign (P_{sign} between $.51$ and $.63$ for whichever direction is more probable). The fourth, telicity match $\times$ generality pair, stands out: $\beta = +0.047$ ($[-0.014, +0.108]$, $P_{\text{sign}} = .94$ for the positive direction). The positive sign suggests that the telicity-match effect may be weaker in contentful-bleached pairs than in contentful-contentful pairs. That is, two fully specified descriptions reward aspectual matching more than a fully specified description paired with a general one. One reason for this asymmetry might be that, if bleached items (by picking out a wider set of compatible event-tokens) carry diffuser aspectual signals, the per-pair telicity match is a noisier cue to actual aspectual agreement on the pair. The effect's posterior is not as well-localized as the H1–H3 effects, but it is a natural target for a follow-up experiment powered specifically to detect generality-by-aspect interactions.

6. Discussion. Experiment 1's residual variability adds something to the compositional tradition (which has held since at least Verkuyl 1972 that telicity is contributed at the event-description level) by showing that variability after controlling for known factors remains substantial even for verb classes (quantized scalar-change roots and two-point-scale predicates) that prior work treats as telicity-stable (Rappaport Hovav 2008, Beavers 2013, Beavers & Koontz-Garboden 2017, 2020).

Experiment 2 shows that the residual is structured by sense similarity in a particular way. Within a verb form, pairs of descriptions that share telicity and durativity are rated as more similar in meaning (**H1**), and the regular-polysemy amplification hypothesis (**H3**) is supported on both aspectual axes: aspectual matching predicts greater perceived similarity at all levels of same-sense, but its contribution is amplified for pairs likely to share a sense. Perceived similarity therefore respects two qualitatively different kinds of structure in the lexicon: the sense partition as a categorical feature respected across senses, and continuous variation in aspectual agreement within

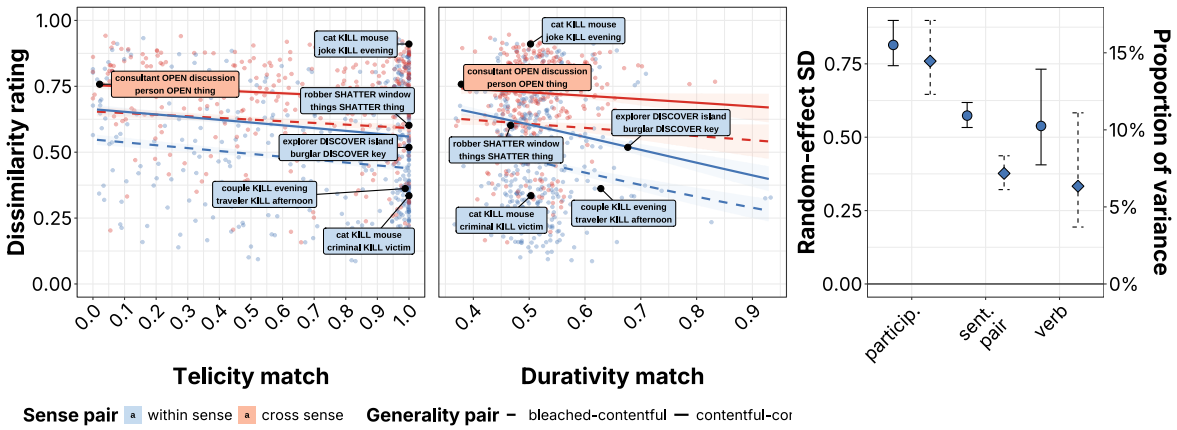

Figure 4: Per-pair dissimilarity against telicity match (left) and durativity match (right), colored by same-sense (within = blue, cross = red); colored smooths are beta-regression fits per same-sense level with 95% CIs. Six evocative pairs labeled.

a sense. This pattern has a potentially interesting consequence in that it suggests a diagnostic for sense individuation: where within-verb aspectual matching stops predicting within-verb similarity is where the verb's sense partition cuts, and so this pattern can potentially be used to locate sense boundaries empirically without presupposing a sense inventory.

There is also substantial verb-level variation: the by-VERB random-slope SDs are comparable across telicity match, durativity match, and same-sense, and the entropy of a verb's $p(\text{sense} \mid \text{verb})$ distribution is negatively associated with its same-sense slope (posterior median $r = -0.22$, 95% CI $[-0.45, +0.02]$, $P(r < 0) = .97$): verbs whose sense mass is spread evenly across multiple senses show sharper same-sense amplification than verbs concentrated on a dominant sense.

Stepping back: the three-point generality cline our materials are built around lets us separate verb TYPE, verb SENSE, and verb TOKEN, and all three carry weight. Type and sense matter in the expected way: manner and result verbs differ in their preference for *in* versus *for*, and pairs of descriptions likely to share a sense are rated as more similar. The contribution of this paper is the third: tokens matter too. Within a sense, aspectual variation survives the standard predictors, and two tokens that match aspectually are rated as more similar than two that do not. A verb form is not a single fixed point but a small family of related event concepts, and where a description sits within that family shapes how speakers perceive the event it describes.

References

Beavers, John. 2011. On affectedness. *Natural Language & Linguistic Theory* 29(2). 335–370. 10.1007/s11049-011-9124-6.

Beavers, John. 2013. Aspectual classes and scales of change. *Linguistics* 51(4). 681–706. 10.1515/ling-2013-0024.

Beavers, John & Andrew Koontz-Garboden. 2012. Manner and Result in the Roots of Verbal Meaning. *Linguistic Inquiry* 43(3). 331–369. 10.1162/LING_a_00093.

Beavers, John & Andrew Koontz-Garboden. 2017. Result verbs, scalar change, and the typology of motion verbs. *Language* 93(4). 842–876. 10.1353/lan.2017.0060.

Beavers, John & Andrew Koontz-Garboden. 2020. *The roots of verbal meaning* (Oxford Studies in Theoretical Linguistics 74). Oxford: Oxford University Press. 10.1093/oso/9780198855781.001.0001.

Beavers, John, Beth Levin & Shiao Wei Tham. 2010. The typology of motion expressions revisited. *Journal of Linguistics* 46(2). 331–377. 10.1017/S0022226709990272.

Bürkner, Paul-Christian. 2017. brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software* 80(1). 1–28. 10.18637/jss.v080.i01.

Dowty, David R. 1979. *Word Meaning and Montague Grammar: The Semantics of Verbs and Times in Generative Semantics and in Montague's PTQ*, vol. 7 Synthese Language Library. Dordrecht: D. Reidel. 10.1007/978-94-009-9473-7.

Filip, Hana. 1999. *Aspect, Eventuality Types, and Nominal Reference* Outstanding Dissertations in Linguistics. New York: Garland Publishing.

Filip, Hana. 2012. Lexical aspect. In Robert I. Binnick (ed.), *The Oxford handbook of tense and aspect*, 721–751. Oxford: Oxford University Press. 10.1093/oxfordhb/9780195381979.013.0025.

Fillmore, Charles J. 2006. Frame semantics. In Dirk Geeraerts (ed.), *Cognitive Linguistics: Basic readings*, vol. 34, 373–400. Berlin & New York: Mouton de Gruyter.

Fillmore, Charles John. 1970. The grammar of hitting and breaking. In R.A. Jacobs & P.S. Rosenbaum (eds.), *Readings in English Transformational Grammar*, 120–133. Waltham, MA: Ginn.

Gentner, Dedre. 1978. On relational meaning: The acquisition of verb meaning. *Child Development* 49(4). 988–998. 10.2307/1128738.

Kane, Benjamin, Will Gantt & Aaron Steven White. 2022. Intensional Gaps: Relating veridicality, factivity, doxasticity, bouleticity, and neg-raising. *Semantics and Linguistic Theory* 31. 570–605. 10.3765/salt.v31i0.5137.

Kearns, Kate. 2007. Telic senses of deadjectival verbs. *Lingua* 117(1). 26–66. 10.1016/j.lingua.2005.09.002.

Kratzer, Angelika. 2004. Telicity and the meaning of objective case. In Jacqueline Guéron & Jacqueline Lecarme (eds.), *The syntax of time* Current Studies in Linguistics, 389–423. Cambridge, MA: MIT Press. 10.7551/mitpress/6598.003.0017.

Krifka, Manfred. 1998. The Origins of Telicity. In Susan Rothstein (ed.), *Events and Grammar* Studies in Linguistics and Philosophy, 197–235. Dordrecht: Springer Netherlands. 10.1007/978-94-011-3969-4_9.

Kubinec, Robert. 2023. Ordered beta regression: A parsimonious, well-fitting model for continuous data with lower and upper bounds. *Political Analysis* 31(4). 519–536. 10.1017/pan.2022.20.

Lakoff, George & Mark Johnson. 1980. *Metaphors We Live By*. Chicago: University of Chicago Press.

Lee, Michael D. & Eric-Jan Wagenmakers. 2014. *Bayesian Cognitive Modeling: A Practical Course*. Cambridge: Cambridge University Press. 10.1017/CBO9781139087759.

Levin, Beth. 1993. *English Verb Classes and Alternations: A preliminary investigation*. Chicago: University of Chicago Press.

Levin, Beth. 2015. Semantics and pragmatics of argument alternations. *Annual Review of Linguis-*

tics 1(1). 63–83. 10.1146/annurev-linguist-030514-125141.

Levin, Beth. 2020. Resultatives and constraints on concealed causatives. In *Perspectives on causation* Jerusalem Studies in Philosophy and History of Science, 185–217. Springer International Publishing. 10.1007/978-3-030-34308-8_6.

Levin, Beth & Malka Rappaport Hovav. 1991. Wiping the slate clean: A lexical semantic exploration. *Cognition* 41(1-3). 123–151. 10.1016/0010-0277(91)90034-2.

Moens, Marc & Mark Steedman. 1988. Temporal Ontology and Temporal Reference. *Computational Linguistics* 14(2). 15–28. https://aclanthology.org/J88-2003/.

Paivio, Allan. 1991. Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology* 45(3). 255–287. 10.1037/h0084295.

Palmer, Martha, Daniel Gildea & Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics* 31(1). 71–106. 10.1162/0891201053630264.

Pustejovsky, James. 1995. *The Generative Lexicon*. Cambridge, MA: MIT Press. 10.7551/mitpress/3225.001.0001.

Rappaport Hovav, Malka. 2008. Lexicalized meaning and the internal temporal structure of events. In Susan Rothstein (ed.), *Theoretical and crosslinguistic approaches to the semantics of aspect* (Linguistik Aktuell/Linguistics Today 110), 13–42. Amsterdam: John Benjamins. 10.1075/la.110.03hov.

Rothstein, Susan. 2004. *Structuring events: A study in the semantics of lexical aspect* Explorations in Semantics. Malden, MA and Oxford: Blackwell. 10.1002/9780470759127.

Tenny, Carol L. 1994. *Aspectual Roles and the Syntax-Semantics Interface*, vol. 52 Studies in Linguistics and Philosophy. Dordrecht: Springer Netherlands. 10.1007/978-94-011-1150-8.

Venables, W. N. & B. D. Ripley. 2002. *Modern Applied Statistics with S*. New York: Springer 4th edn. 10.1007/978-0-387-21706-2.

Vendler, Zeno. 1957. Verbs and Times. *The Philosophical Review* 66(2). 143–160. 10.2307/2182371.

Verkuyl, H. J. 1989. Aspectual classes and aspectual composition. *Linguistics and Philosophy* 12(1). 39–94. 10.1007/BF00627398.

Verkuyl, Henk J. 1972. *On the Compositional Nature of the Aspects*, vol. 15 Foundations of Language. Dordrecht: D. Reidel. 10.1007/978-94-017-2478-4.

Weischedel, Ralph, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin & Ann Houston. 2013. OntoNotes Release 5.0. 10.35111/xmhb-2b84.

White, Aaron Steven & Kyle Rawlins. 2016. A computational model of S-selection. *Semantics and Linguistic Theory* 26. 641–663. 10.3765/salt.v26i0.3819.

White, Aaron Steven & Kyle Rawlins. 2020. Frequency, acceptability, and selection: A case study of clause-embedding. *Glossa: a journal of general linguistics* 5(1). 105. 10.5334/gjgl.1001.