A Qualia-Based Audit of Procedural Event Annotations

Kyeongmin Rim and Marc Verhagen and James Pustejovsky
Brandeis University
Waltham, MA, USA
{krim, verhagen, jamesp}@brandeis.edu

Abstract

Procedural event annotations record *what changed* but not the semantic relevance or grounding of the change: whether the annotated entity is the kind of thing whose state matters for the domain. We present Entity Qualia Structure (EQS), a per-entity sortal-type categorization (coarsened from Generative Lexicon's type system to three categories: natural, artifactual, instrument) extracted from existing lexical resources. Applied to the OpenPI food domain, EQS reaches 84.7% coverage of the 518-item entity vocabulary; across 9367 transformation annotations, only 51.1% concern food entities themselves, while 30.2% record state changes of instruments, entities whose sortal type places them outside the food-state task. In a three-way comparison against existing cleanup efforts, EQS uniquely flags 15.6% of annotations that neither human re-annotation (OpenPI-C) nor LLM salience scoring (OpenPI 2.0) catches. Analysis of the AGENTIVE quale reveals that 93% of agentive-positive annotations involve instruments rather than food: entity creation can only be detected when the agentive feature is paired with the associated verb's event semantics.

1 Introduction

Crowdsourcing the annotation of entity state changes in procedural text to anonymous online workers has enabled datasets at large scale; however, in resources such as OpenPI (Tandon et al., 2020), which give crowd annotators free-text (open-vocabulary) input, the resulting annotations do not distinguish semantically central changes from incidental ones. Consider a recipe step *"Bake the cake for 30 minutes."* An annotator might record that the oven became "hot," the timer gone to "0," and the cake went from "raw batter" to "baked." All three are factually correct, but only the last tracks food state; the others describe instrument and timer state. The question is not *what changed*, but whether the entity undergoing change is *the kind of thing whose state matters* for the procedure's domain.

OpenPI's annotation quality has prompted independent cleanup efforts: Wu et al. (2023) produced OpenPI-C via three-stage human re-annotation, filtering $\sim$32% of state changes as not reliably inferrable from the input; Zhang et al. (2024) apply LLM-based clustering and per-step salience scoring (OpenPI 2.0). Both approaches address quality empirically, by re-annotating or filtering with human or LLM judgment. We offer a complementary symbolic angle: defining incidental annotations by the entity's sortal type, grounded in Generative Lexicon theory rather than annotator or model agreement. Our analysis shows that this uniquely catches 15.6% of annotations the empirical methods miss (§4).

Generative Lexicon (GL) theory (Pustejovsky, 1995) provides a formal basis for this distinction. An entity's FORMAL quale determines its sortal type, which we coarsen to natural, artifactual, or instrument for the audit; its AGENTIVE quale records whether the entity has an origin event with an associated agent (typically a maker). The contrast between *bake a potato* and *bake a cake* illustrates why both qualia matter: the same verb and event topology yield different outcomes because potato (a natural kind, no creation origin) is transformed, while cake (an artifact with a baking origin) is created, a phenomenon GL calls *co-composition*, where the event semantics is determined by the verb and its arguments' qualia jointly.

In this paper, we extract these qualia features from existing lexical resources and use them as a symbolic audit of OpenPI annotations. Our contributions are: (1) a cascade method that builds Entity Qualia Structure (EQS) data capturing coarsened GL sortal types from noun-focus language resources, covering 84.7% of the OpenPI food-domain vocabulary with a 32.2% cross-resource disagreement rate as a built-in quality diagnostic;

Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications, pages 49–57

(2) an audit showing that nearly half of OpenPI food annotations are not about food at all, with instruments alone accounting for 30.2%, a mismatch directly readable from the entity's sortal type; and (3) an analysis of the AGENTIVE quale showing that 93% of agentive-positive annotations involve instruments, confirming that this theoretically motivated feature requires verb-side composition before it becomes discriminative. EQS provides the argument-side input for a compositional account of entity-state semantics; the complementary predicate-side analysis appears in Rim and Pustejovsky (2026).

2 Related Work

2.1 Entity State Annotation and Dataset Quality

The annotation of entity states in procedural text has evolved from tracking textual mentions to capturing implicit argument structures. Early work grounded entity state tracking in Semantic Role Labeling (Palmer et al., 2005) and qualia-based semantic tagging (GLML; Pustejovsky et al., 2009). ProPara (Dalvi et al., 2018) introduced entity tracking in procedural paragraphs with a closed set of state labels (created, destroyed, moved). Subsequent datasets have addressed implicit arguments (RISeC; Jiang et al., 2020), bridging relations under state transformation (RecipeRef; Fang et al., 2022), and entity identity and coreference using GL event models (CUTL; Rim et al., 2023). Kazeminejad et al. (2021) used the VerbNet semantic parser to automatically annotate entity existence and location states on ProPara, illustrating the value of symbolic, lexical-resource-grounded approaches for procedural entity-state work.

OpenPI (Tandon et al., 2020) scaled to open-vocabulary coverage across different procedural domains, but at the cost of the subeventual and ontological constraints found in the earlier literature. The cleanup efforts of Wu et al. (2023) and Zhang et al. (2024), discussed in §1, address the resulting reliability issues with human and LLM judgment respectively; we instead ground our audit in lexical-semantic resources.

2.2 Lexical-Semantic Resources for Events

Following GL's dual-aspect view of event semantics, we develop Entity Qualia Structure (EQS), an entity-side qualia representation that complements predicate-side resources such as VerbNet-GL (Brown et al., 2022). EQS builds on the GL tradition of computational lexicon construction: the SIMPLE ontology (Bel et al., 2000), which standardized GL-native semantic types across 12 European languages; the Brandeis Semantic Ontology (BSO; Pustejovsky et al., 2006; Havasi et al., 2007), an English lexicon informed by SIMPLE, publicly released alongside this work; CoreLex (Buitelaar, 1998), which derives systematic polysemy classes from WordNet (Fellbaum, 1998); and the Principle of Type Ordering (Pustejovsky, 2001), which formally justifies EQS's minimal feature set (sortal type + agentive quale availability) as sufficient for co-composition. The audit operationalizes the two qualia that surface directly in BSO entries: coarsened FORMAL (sortal type) and AGENTIVE (creation availability); TELIC enters the cascade indirectly through the type hierarchy.

2.3 Direct Precedents for Qualia Annotation

Prior work on annotating or extracting GL qualia informs the EQS cascade design. Pustejovsky et al. (2010) introduced the SemEval-2010 GLML task on argument selection and coercion, annotating whether the TELIC or AGENTIVE quale of a noun was activated in verb-argument context; this is the closest methodological ancestor to EQS, though scoped to verbal argument positions rather than discourse-level procedural entities. Yamada and Baldwin (2004) demonstrated automatic acquisition of TELIC and AGENTIVE roles from syntactic patterns and noted that domain-shifting lemmas (e.g., cooking entities appearing in natural-kind and artifact roles across documents) require token- or type-level resolution—the challenge EQS addresses through type-level coarsening. Bouillon et al. (2012) report annotator agreement on qualia annotation in Italian and French complex nominals, finding AGENTIVE relations reliably annotated by trained linguists, which empirically supports AGENTIVE as one of EQS's two operative qualia.

3 EQS: Extraction and Audit Method

OpenPI provides open-vocabulary state annotations across multiple procedural domains; we focus on the food slice for three reasons. First, open-vocabulary annotation is the property that makes annotation quality a research problem worth addressing with symbolic typing, and food is the slice where this property is best exercised by the existing data. Second, food preparation interleaves natu-

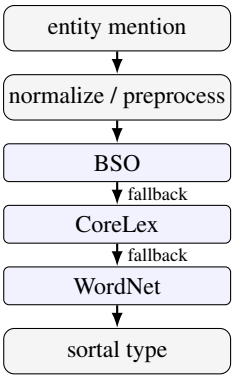

Figure 1: Sortal cascade pipeline. Entity mentions are normalized, then resolved against three lexical resources in priority order (first-resource-wins). The output is a coarsened sortal type that fills FORMAL.TYPE_CAT. Other EQS fields (FORMAL.INDIVIDUATION, AGENTIVE) are populated by direct single-resource lookup, not shown.

ral ingredients, artifactual dishes, and instruments within the same step, exercising exactly the FORMAL contrast that motivates EQS. Third, BSO's type hierarchy is densest in the food domain (its *Nourishment* subtree is the most developed branch), making food the natural first target for cascade evaluation.

3.1 Source Resources and Field Resolution

An EQS record captures two independent qualia features per entity type, grounded in Generative Lexicon theory: FORMAL (sortal type, coarsened to natural, artifactual, or instrument) and AGENTIVE (`yes` if the entity's type has a specific lexicalized creation activity in BSO, unspecified otherwise). These features are extracted automatically from existing lexical resources: the sortal type via the cascade described below, other fields via direct single-resource lookup. Figure 2 shows the AVM schema for two contrastive examples: *cake* (artifactual with a specific creation activity) and *knife* (instrument that is also GL-correctly agentive but whose creation is irrelevant to a recipe context).

The instrument-vs-artifact discrimination is the operative concern that motivates the choice of GL over simpler ontologies. WordNet's `noun.artifact` lumps knives, bowls, and baked goods together; ConceptNet (Speer et al., 2017) has no systematic instrument tag; Wikidata's (Vrandečić and Krötzsch, 2014) Q-types are too granular to coarsen reliably. In BSO, lexicalized TELIC roles are stored per-type, but the existence of a TELIC role is also encoded in the non-terminal

$$\left[\begin{array}{ll} \textit{eqs} & \\ \text{ENTITY} & \textit{cake} \\ \text{FORMAL} & \textit{artifactual} \\ \text{AGENTIVE} & \textit{yes} \end{array}\right]$$

$$\left[\begin{array}{ll} \textit{eqs} & \\ \text{ENTITY} & \textit{knife} \\ \text{FORMAL} & \textit{instrument} \\ \text{AGENTIVE} & \textit{yes} \end{array}\right]$$

Figure 2: EQS schema for two contrastive examples. Cake's agentive=yes (*Bake Activity* in BSO) licenses a creation reading under co-composition with verbs whose subevent matches; knife's agentive=yes (*Create Material Entity*) is GL-correct but operationally inert in recipe context.

node names of the type hierarchy (e.g., *Material Object with InstrumentTelic*). The cascade recovers the instrument category by ancestor walking the hierarchy alone, exploiting these node-name annotations; the per-type lexicalized values are reserved for finer-grained reasoning where ancestor walking is insufficient (e.g., N-N compound resolution; see §3.2).

The FORMAL field is resolved by a cascade of three lexical resources, applied in priority order with first-resource-wins per field. The BSO serves as the primary resource: entity stems are looked up in BSO's ∼50k-entry GL type hierarchy.[1] A set of ancestor-matching rules (hand-crafted for the food domain, external to BSO) walks the type tree to classify each sense (e.g., *Meat* → natural, *Artifactual Food* → artifactual, *Material Object with InstrumentTelic*[2] → instrument). For polysemous stems, all senses are checked; rule priority is set by their order in the domain profile, and the cascade returns the sense whose ancestor chain triggers the earliest-listed rule. CoreLex provides a polysemy-class fallback for entities outside BSO's coverage, mapping lemmas to basic types derived from systematic WordNet polysemy patterns, though it cannot distinguish instruments from general artifacts. WordNet supersenses serve as the last layer, mapping synsets to coarse type categories (`noun.plant` → natural, `noun.artifact`

[1]The cascade consumes the BSO release accompanying this paper at `http://brandeis-llc.github.io/bso`. BSO has been refined incrementally since its original publication (Pustejovsky et al., 2006), kept internal until this release; food-domain classifications are identical to the 2006 release used in earlier development of the cascade.

[2]Specific BSO type labels are subject to revision in future releases; the example rules here document the cascade's rule table at submission. The instrument-vs-artifact discrimination is structural in the type hierarchy and survives label changes.

Field	Value	Count	Source
FORMAL (coverage: 84.7%)	natural	205	BSO→CL→WN
	artifactual	118	
	instrument	116	
AGENTIVE (coverage: 18.0%)	yes	93	BSO core qualia
	None	425	

Table 1: EQS field resolution and FORMAL distribution on OpenPI Food vocabulary (518 entity types; 79 unresolved across five backend-limitation categories). Fields are independent: different resources, different GL layers.

→ artifactual); `noun.food` is skipped as ambiguous between natural and artifactual. Lexicalized food compounds (e.g., *ice cream*, *peanut butter*, *olive oil*) fall in WordNet's ambiguous `noun.food` class and remain unresolved by the cascade; these surface as one of the five backend-limitation categories discussed in §6.

The AGENTIVE field is resolved independently, by BSO core qualia lookup only. It is populated when BSO's type hierarchy annotates a specific creation activity on the entity's type, for example, *Bake Activity* on *Baked Good*, or *Stew Food Activity* on *Stew*. Generic placeholders (e.g., *Prepare Food Activity* on the *Food* supertype) are excluded, as they indicate the type *can* have origins, not that it has a specific lexicalized one. BSO stores the specific activity value, but EQS binarizes it to `yes/none`; the specific activity becomes relevant only at co-composition time, when the verb's subevent structure is available for matching (§5).

Our target corpus is the OpenPI "Food and Entertaining" subset restricted to concrete-goal documents across all three splits: 169 documents (train 150 / dev 12 / test 7), 840 steps, and 9367 transformation annotations. Table 1 summarizes coverage and distribution. Of 518 entity types in this vocabulary, 439 (84.7%) are resolved on FORMAL; the 79 unresolved entities fall into five backend-limitation categories (`noun.food` compounds, productive N-N compounds, brand names and rare lemmas, conjunctions, surface-form variants), discussed in §6.

3.2 Validation and Evaluation

Cross-resource disagreement. FORMAL resolution has no external gold standard at the type level. As an internal sanity check, every lemma is resolved against all three cascade backends unconditionally (bypassing the first-wins shortcut) and the resulting type categorizations are compared across resources; cross-resource disagreement serves as a built-in quality metric. Of 518 entities, 167 (32.2%) show cross-resource disagreement on FORMAL, falling into three patterns: (1) BSO=instrument vs. CL/WN=artifactual (∼85%; CL/WN lack an instrument category; BSO correct); (2) BSO=natural vs. CL=artifactual (polysemy in CoreLex; BSO correct); (3) 18 entities with no BSO coverage where CL and WN disagree, adjudicated against the SIMPLE OWL ontology (Toral and Monachini, 2007). The pattern analysis confirms first-wins resolution is correct in the large majority of cases.

Sample	n	correct	precision
precision-instrument	49	45	0.918
precision-food	40	25	0.625
recall sample (FN spot-check)	41	31	—

Table 2: FORMAL evaluation on a stratified sample. Estimated overall recall 0.575 (FN rate 0.244 extrapolated to the non-instrument population), giving F1 0.707.

Manual evaluation (FORMAL). We complement cross-resource disagreement with a stratified manual spot-check. A trained linguist familiar with GL theory adjudicated three samples: 50 instrument-predicted entities (precision-instrument), 50 food-predicted entities (precision-food), and 50 non-instrument entities mentioned in OpenPI annotations (recall, sampled false-negative-style). Adjudications use cascade evidence (BSO type chain, CoreLex polysemy class, WordNet supersense) for the prototype-level type judgment. We exclude 19 rows whose decision was sourced from outside the cascade backends (e.g., from CoreLex-only or WordNet-only fallback) since they cannot test cascade behavior directly; their adjudications remain valid type judgments and are reported separately as cascade-blind-spot evidence. Effective sample sizes are 49 / 40 / 42. Table 2 reports per-sample precision; the recall sample supports an estimated recall of 0.575 and F1 of 0.707 by extrapolating the false-negative rate to the non-instrument population.

Error analysis (N-N compounds). The dominant error pattern on precision-instrument is noun-noun compounds with a food-noun modifier and a container-noun head: *vodka bottle*, *vanilla extract bottle*, *marshmallow package*, *strawberry custard dish*. All four precision-instrument errors on the 50-entity sample come from this pattern. The cas-

Sub-sample	n	correct	rate
agentive=yes & food	12	11	0.917
agentive=yes & instrument	25	0	0.000
agentive=None (FN check)	25	8	0.320

Table 3: AGENTIVE evaluation on three sub-samples. "Correct" indicates whether the AGENTIVE label correctly reflects *recipe-relevant* creation: for agentive=yes rows, the entity is something the recipe creates (cake, bread); for agentive=None rows, the entity is not created in the recipe (pre-existing tools, ingredients). For instruments, every instance is GL-correct (the instrument *was* manufactured) but not recipe-relevant. 17/25 agentive=None cases are false negatives where BSO misses an applicable creation activity.

cade's head-noun fallback consistently selects the container reading; adjudicators selected the content reading because the modifier is food. This is the canonical N-N compound polysemy problem: morphology alone cannot distinguish [vodka bottle] (the contents) from [glass bottle] (the container). Ye et al. (2025) address this problem with a neural approach: LLM textual enrichment that surfaces qualia-role binding through prompt augmentation. A complementary symbolic route would consult the head noun's TELIC quale (encoded in BSO but not currently invoked by the cascade), matching the modifier's sortal type against the container's functional content type; this is a localized cascade extension rather than a methodological revision.

Manual evaluation (AGENTIVE). We additionally evaluate AGENTIVE on three sub-samples reflecting the field's three operational states (Table 3). The 12 agentive=yes food entities show 91.7% precision: when EQS asserts a specific creation activity for a food entity, the assertion is almost always correct. The 25 agentive=yes instrument entities show 0% operational relevance: every instance is GL-correct (the instrument *was* manufactured) but the creation history is irrelevant in recipe context. The agentive=None sample (25 entities) shows that BSO misses creation activities for 17 of 25 (68%); BSO's qualia coverage is sparser than the operational landscape suggests.

3.3 Auditing OpenPI

We operationalize *instrument* as an artifact whose procedural role is functional-use, not being transformed (state-changed): containers, tools, appliances, surfaces, and measurement gear. The operational test is whether the recipe *produces or transforms* the entity, or whether it *uses* the entity to do work on food; the latter is instrument.

The audit itself is a cross-reference: for each of the 9367 OpenPI food-domain annotations, we look up the entity's EQS FORMAL value and partition annotations by entity category. Entities classified as `instrument` are predicted to be peripheral to food-state tracking—their annotations record tool and container state (bowl weight, knife cleanness, oven temperature) rather than the event's food-level output, though such annotations may be informative for other purposes (e.g., workflow modeling). Entities classified as `natural` or `artifactual` are predicted to carry the food-state signal.

This is deliberately simple: a single symbolic feature (FORMAL) applied without any verb-side analysis, role binding, or co-composition. The contribution is showing how much incidental annotation one entity-level symbolic classification can detect. A richer consistency taxonomy (distinguishing *consistent* annotations (food entity, food-state change) from *lexically-underspecified* ones: food entity in a process event showing state change that the verb's subevent structure does not predict; cf. Generalized Result Role, Jezek and Melloni 2011; Rim et al. 2023) requires pairing the EQS classification with predicate-side subevent structure; Rim and Pustejovsky (2026) provide the complementary verb-side analysis.

4 Results: Incidental-Annotation Analysis

Table 4 presents the primary result. Cross-referencing EQS FORMAL against 9367 OpenPI food-domain annotations reveals that 51.1% involve entities classified as food (natural or artifactual). The largest incidental category is instrument tracking (30.2%): annotations on entities such as *bowl*, *knife*, *spoon*, and *blender*, classified as `instrument` by the cascade, whose identity persists through the event regardless of what attribute changes annotators recorded. The 18.7% unresolved category includes entities not in the EQS vocabulary, falling across five backend-limitation categories (§6). A complementary keyword-based analysis identifies 49% of annotations as low-value via attribute-name matching (location 21.5%, weight 7.6%, cleanness 7.3%); the two methods are complementary, as EQS catches incidental annotations from non-food entities entirely while keyword matching catches incidental annotations *within* food-entity records (e.g., location-only changes on

Category	Trans.	%
Food signal (natural + artifactual)	4786	51.1
Instrument (incidental)	2829	30.2
Unresolved / not in EQS	1752	18.7
Total	9367	100.0

Table 4: OpenPI food annotations by EQS entity category. 51.1% involve actual food entities; 30.2% is incidentally tracked instrument state detected by FORMAL, not surface keywords.

Flagged by	Trans.	% of 9367
EQS only	1458	15.6
OpenPI-C only	2760	29.5
OpenPI 2.0 (local-salience) only	286	3.1
All three	222	2.4
None (kept by all three)	3109	33.2
EQS total	2829	30.2
OpenPI-C total	4350	46.4
OpenPI 2.0 total	1055	11.3

Table 5: Three-way comparison: EQS instrument-flag vs. OpenPI-C re-annotation vs. OpenPI 2.0 per-step low local-salience (≤ 2). EQS uniquely catches 15.6% of annotations the empirical methods miss. We do not use V2's entity clusters or paraphrase expansion because those layers have sub-50% F1 against unreliable gold.

food items).

Comparison with empirical cleanup methods. Table 5 compares EQS's instrument-flag against the two empirical cleanup methods reviewed in §2: OpenPI-C's three-stage human re-annotation (Wu et al., 2023) and OpenPI 2.0's per-step LLM salience scoring (Zhang et al., 2024). EQS uniquely flags 15.6% of annotations that neither OpenPI-C nor OpenPI 2.0's local salience catches: instrument-typed entities whose state changes the empirical methods retain as relevant. Only 2.4% of annotations are flagged by all three methods simultaneously; conversely, 33.2% are retained as signal by every method, providing a high-confidence consensus subset. The three filters are complementary: they ground their decisions on different evidence (entity ontology, annotator agreement, LLM-derived salience) and catch different incidental categories.

5 Discussion

5.1 When Argument Qualia Become Operative

The FORMAL field drives the incidental-annotation analysis in §4, but GL theory predicts that the AGENTIVE quale should provide a finer distinction: whether an entity undergoes *creation* (its origin process is instantiated by the event) or merely *transformation* (its identity is preserved). We test this prediction by cross-referencing the AGENTIVE field against OpenPI annotations.

Of 2626 annotations on AGENTIVE=yes entities, 93.1% involve instruments such as *bowl*, *blender*, *pan*, and *knife* (Table 6). These are GL-correctly agentive: a knife *was* manufactured, and BSO records a *Create Activity* on its type. But a knife's origin process is irrelevant in a recipe context; the recipe does not create knives. Only 12 food entity types (182 annotations, 6.9%) have genuinely operative AGENTIVE qualia: *cookies*, *bread*, *cake*, *crust*, *wine*, *beer*, among others, whose creation process (baking, brewing) may actually be instantiated by the procedural verb.

The AGENTIVE quale answers a type-level question *"does this entity have a lexicalized origin process?"* not a compositional one *"does* this event *instantiate that origin?"* The distinction between a knife's irrelevant manufacturing and a cake's operative baking emerges only when the verb's semantic structure is available for matching: *bake* instantiates cake's AGENTIVE quale (*Bake Activity*); *move* does not. BSO stores the specific activity value (not just yes/none), making this matching feasible in principle, but it requires the predicate side: a co-composition operator that is beyond the scope of the present work.

This finding mirrors a complementary result from the predicate side: Rim and Pustejovsky (2026) show that VerbNet-GL's verb-only prediction achieves only 29.4% overall accuracy for entity identity change, because the verb's subevent structure cannot determine the outcome without argument qualia. Our agentive analysis confirms the converse: argument qualia cannot determine the outcome without the verb. The 76% of process-event steps that show state changes in the OpenPI data despite VN-GL predicting no result state provide further context: EQS's FORMAL can distinguish which of these are legitimate (food entities undergoing implicit transformation) from incidental (instrument state tracking), but the full resolution (predicting *what kind* of change each food entity undergoes, e.g., whether AGENTIVE licenses creation or transformation) requires both sides of the composition.

Category	Trans.	% of ag=yes
Instrument	2444	93.1
Food (natural + artifactual)	182	6.9
Total agentive=yes	2626	100.0

Table 6: Breakdown of AGENTIVE=yes transformations. 93.1% involve instruments, GL-correctly agentive (manufactured), but whose creation is irrelevant to the procedural context. The feature becomes operative only when composed with the verb's subevent structure.

5.2 Repairing the Symbolic Backbone

The audit exposes lexicographic coverage gaps that the symbolic backbone alone cannot resolve. A representative case: BSO has *platter* with senses {Food, Music Artifact} but no Tableware/Dish sense, so the food-domain rule priority picks the Food reading and the contextually-correct tableware reading is unreachable; sibling lemmas (*bowl*, *plate*, *serving bowl*) resolve correctly because a Dish sense *is* present, and absence is silent.

These gaps live in the data layer, not in the A-V extraction method, so the natural place for repair sits outside our pipeline. Manual lexicographic curation is the canonical option, but it scales poorly against a resource the size of BSO. Considering recent advancements in language models, one promising route is LLM-assisted, corpus-rooted reevaluation, where model proposals surface candidate senses or relations and the symbolic backbone serves to validate them against the existing type hierarchy: a division of labor that uses each side for what it does best. The cross-resource disagreement signal we use for validation (§3.2) is a useful place to begin, since it already flags the lemmas where coverage is contested.

5.3 Domain Specificity of the Work

Domain specificity in our pipeline is concentrated at two points, both surfaced through a single `DomainProfile` object in the implementation: the BSO ancestor rules (the paper's explicit food-domain contribution, a small reviewable GL-grounded artifact), and the WordNet `noun.food` supersense skip. The most immediate generalization target is OpenPI's Home and Garden topic, the second-largest concrete-procedure slice within the same corpus (146 concrete documents vs. Food's 169; 5552 concrete transformations vs. Food's 9367) and conceptually adjacent, requiring only `DomainProfile` changes rather than architecture changes. Procedural domains farther from cooking (medical protocols, industrial workflows, scientific procedures) use the same attach point but additionally require sourcing domain-appropriate corpora beyond OpenPI.

A related concern is *within-resource bias*: each backend brings its own design choices that bias classification. Polysemy resolution illustrates this: lemmas like *oil* (BSO senses Fat, Painting, Combustible, Ointment) or *dressing* (Clothing Artifact, Seasoning) are resolved by walking each sense against the BSO ancestor rules, and for food-domain entities this typically lands on the intended sense, but the mechanism is an implicit domain prior rather than contextual disambiguation. Comparable choices in WordNet supersense priority and CoreLex polysemy classes contribute their own biases. These are bias *by design* in the food domain; the same architecture, configured differently, would express different biases in another domain.

A complementary kind of adaptability concerns the task rather than the domain: the framing of an entity as "incidental" is itself task-relative. Instrument annotations are peripheral to food-state tracking but signal for workflow- or tool-modeling tasks, where tracking pan temperature and knife cleanness IS the point. EQS's symbolic typing offers a re-orientable filter rather than a fixed noise/signal partition; swapping the relevance class re-uses the same cascade output.

6 Conclusion

We have presented three contributions. First, EQS: a per-entity qualia representation automatically extracted from noun-level lexical resources, achieving 84.7% coverage on the OpenPI food-domain vocabulary with a 32.2% cross-resource disagreement rate that serves as a built-in quality diagnostic. Second, an incidental-annotation analysis showing that 51.1% of OpenPI food annotations track genuine food-state change, with 30.2% constituting incidental instrument tracking detectable by entity type alone. Third, an agentive analysis demonstrating that argument qualia—even when GL-correctly assigned—are not operative without semantic counterpart: 93% of agentive-positive annotations involve instruments whose creation is irrelevant to the procedural context.

A primary limitation is coverage: the 18.7% unresolved category is not exclusively low-frequency entities, but falls into five backend-limitation cat-

egories, each pointing to a different repair locus. Two sit at the lexical-data layer. `noun.food` compounds (e.g., *ice cream*, *peanut butter*, *olive oil*) fall into WordNet's ambiguous food class that the cascade deliberately skips because it conflates natural and artifactual readings; resolving them needs either a curated lexicalized-compound list or the kind of LLM-assisted sense disambiguation discussed in §5. Brand names and rare lemmas are out-of-vocabulary in all three backends and require either resource extension or surface-form heuristics (capitalization, brand databases).

Two sit at the preprocessing layer. Surface-form variants (plurals, inflections, casing) fall through because BSO is keyed on singular stems, addressable by adding a better lemmatization at lookup time or by extending the stem index with surface variants. Conjunctions (*salt and pepper*, *cream and sugar*) decompose into individually-classifiable components, but conjoined mentions are not currently split before lookup.

Productive N-N compounds (*vodka bottle*, *strawberry custard dish*) sit at the compositional layer and are the theoretically interesting case; their resolution requires reasoning over both modifier and head qualia, as outlined in the N-N error analysis (§3.2).

More broadly, these results suggest that symbolic qualia structure can provide an ontological audit layer that purely human-judgment and LLM-judgment approaches to annotation quality lack. Future work will complete the compositional picture by matching the verb-side semantics against the EQS representation; address the domain-specificity limitation discussed in §5 by scaling the audit to other procedural domains, starting with OpenPI's immediately adjacent Home and Garden slice and extending to medical, industrial, and scientific procedure corpora; and feed the resulting typed entity layer into operational entity-state pipelines that require sortal-type filtering at input.

References

Nuria Bel, Federica Busa, Nicoletta Calzolari, Elisabetta Gola, Alessandro Lenci, Monica Monachini, Antoine Ogonowski, Ivonne Peters, Wim Peters, Nilda Ruimy, Marta Villegas, and Antonio Zampolli. 2000. SIMPLE: A general framework for the development of multilingual lexicons. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Pierrette Bouillon, Elisabetta Jezek, Chiara Melloni, and Aurélie Picton. 2012. Annotating qualia relations in Italian and French complex nominals. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*.

Susan Windisch Brown, Julia Bonn, Ghazaleh Kazeminejad, Annie Zaenen, James Pustejovsky, and Martha Palmer. 2022. Semantic Representations for NLP Using VerbNet and the Generative Lexicon. *Frontiers in Artificial Intelligence*, 5.

Paul Buitelaar. 1998. *CoreLex: Systematic Polysemy and Underspecification*. Ph.D. thesis, Brandeis University.

Bhavana Dalvi, Lifu Huang, Niket Tandon, Wen-tau Yih, and Peter Clark. 2018. Tracking state changes in procedural text: a challenge dataset and models for process paragraph comprehension. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1595–1604, New Orleans, Louisiana. Association for Computational Linguistics.

Biaoyan Fang, Timothy Baldwin, and Karin Verspoor. 2022. What does it take to bake a cake? the RecipeRef corpus and anaphora resolution in procedural text. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3481–3495, Dublin, Ireland. Association for Computational Linguistics.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Catherine Havasi, Anna Rumshisky, and James Pustejovsky. 2007. An evaluation of the Brandeis Semantic Ontology. In *Proceedings of the Fourth International Workshop on Generative Approaches to the Lexicon (GL2007)*.

Elisabetta Jezek and Chiara Melloni. 2011. Nominals, polysemy, and co-predication. *Journal of Cognitive Science*, 12(1):1–31.

Yiwei Jiang, Klim Zaporojets, Johannes Deleu, Thomas Demeester, and Chris Develder. 2020. Recipe instruction semantics corpus (RISeC): Resolving semantic structure and zero anaphora in recipes. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 821–826, Suzhou, China. Association for Computational Linguistics.

Ghazaleh Kazeminejad, Martha Palmer, Tao Li, and Vivek Srikumar. 2021. Automatic entity state annotation using the VerbNet semantic parser. In *Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 123–132.

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

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press, Cambridge, MA.

James Pustejovsky. 2001. Type construction and the logic of concepts. In Pierrette Bouillon and Federica Busa, editors, *The Language of Word Meaning*, pages 91–123. Cambridge University Press.

James Pustejovsky, Catherine Havasi, Jessica Littman, Anna Rumshisky, and Marc Verhagen. 2006. Towards a generative lexical resource: The Brandeis semantic ontology. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

James Pustejovsky, Jessica Moszkowicz, Olga Batiukova, and Anna Rumshisky. 2009. GLML: Annotating argument selection and coercion. In *Proceedings of the Eight International Conference on Computational Semantics*, pages 169–180, Tilburg, The Netherlands. Association for Computational Linguistics.

James Pustejovsky, Anna Rumshisky, Alex Plotnick, Elisabetta Jezek, Olga Batiukova, and Valeria Quochi. 2010. SemEval-2010 task 7: Argument selection and coercion. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval)*, pages 27–32.

Kyeongmin Rim and James Pustejovsky. 2026. Subevent structure as a predictor of entity identity change in procedural text. In *Proceedings of the Workshop on Structured Linguistic Data and Evaluation (SLiDE)*.

Kyeongmin Rim, Jingxuan Tu, Bingyang Ye, Marc Verhagen, Eben Holderness, and James Pustejovsky. 2023. The coreference under transformation labeling dataset: Entity tracking in procedural texts using event models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12448–12460, Toronto, Canada. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, pages 4444–4451.

Niket Tandon, Keisuke Sakaguchi, Bhavana Dalvi, Dheeraj Rajagopal, Peter Clark, Michal Guerquin, Kyle Richardson, and Eduard Hovy. 2020. A dataset for tracking entities in open domain procedural text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6408–6417, Online. Association for Computational Linguistics.

Antonio Toral and Monica Monachini. 2007. SIMPLE-OWL: A Generative Lexicon ontology for NLP and the semantic web. In *Proceedings of the workshop on GL2007*.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Xueqing Wu, Sha Li, and Heng Ji. 2023. OpenPI-C: A better benchmark and stronger baseline for open-vocabulary state tracking. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7213–7222, Toronto, Canada. Association for Computational Linguistics.

Ichiro Yamada and Timothy Baldwin. 2004. Automatic discovery of Telic and Agentive roles from corpus data. In *Proceedings of the 18th Pacific Asia Conference on Language, Information and Computation (PACLIC)*.

Bingyang Ye, Jingxuan Tu, and James Pustejovsky. 2025. Enhanced noun-noun compound interpretation through textual enrichment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25896–25911, Suzhou, China. Association for Computational Linguistics.

Li Zhang, Hainiu Xu, Abhinav Kommula, Chris Callison-Burch, and Niket Tandon. 2024. OpenPI2.0: An improved dataset for entity tracking in texts. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 166–178, St. Julian's, Malta. Association for Computational Linguistics.