

Measuring Context as Structural Organization in Natural Language

A DSRP-Based Empirical Analysis Using WinoGrande and SQuAD 2.0

Derek Cabrera^{a,b,@} and Laura Cabrera^{b,@}

a: Cabrera Lab

b: Cornell University

@:dac66@cornell.edu

ABSTRACT: Context is central to language interpretation, yet it is rarely defined in a way that permits direct measurement. In linguistics, cognitive science, and artificial intelligence, context is often treated informally as surrounding text, background knowledge, or situational framing. This study proposes a structural operationalization of context grounded in DSRP theory, in which context is defined as the organization of information relative to a target element, $C = O(I \mid t_0)$. Using 600 items from WinoGrande and SQuAD 2.0, we coded structural context with DSRP-derived features and computed context scores mechanically from those features. Inter-rater reliability ranged from $\kappa = 0.42$ to $\kappa = 0.83$. Structural context scores were significantly associated with interpretation accuracy ($r = 0.347, p < 0.001$). Accuracy increased from 4.7% in the low-context bin to 64.0% in the mid-high-context bin, a 13.7-fold increase. Information quantity alone did not significantly predict interpretation accuracy. In addition, low-context items frequently exhibited underconstrained interpretation spaces, suggesting that structural context may also function as a measure of interpretive determinacy and benchmark ambiguity. These findings support the view that interpretation depends not simply on how much information is present, but on how that information is organized, and that weak contextual organization leaves interpretation less determinate.

KEYWORDS: context | complexity | structural organization | DSRP | natural language interpretation | systems thinking | cognition | AI interpretability | WinoGrande | SQuAD 2.0 | Love Reality Loop | natural language processing (NLP) benchmarks | NLP interpretation tasks

Cite this paper: Cabrera, D. and Cabrera, L.. Measuring Context as Structural Organization in Natural Language: A DSRP-Based Empirical Analysis Using WinoGrande and SQuAD 2.0. *Journal of Systems Thinking*. 6(1), 1–19. <https://doi.org/10.54120/jost.v25i1.dc6mcnl6>

1. Introduction

Language interpretation depends heavily on context. Humans routinely rely on contextual structure to resolve ambiguity, infer meaning, and understand references. Yet despite its central role in linguistics,

cognitive science, and artificial intelligence, context is still often treated *descriptively rather than as a measurable variable*. In practice, context is commonly reduced to surrounding text, background knowledge, or situational framing. These descriptions are useful, but they rarely produce *an operational definition that can be tested empirically*.

A systems perspective suggests a different starting point. Instead of asking only how much information is present, systems thinking asks *how information is organized*. Recent work in DSRP theory and O-Theory [1] argues that organization is not a secondary feature added to information after the fact, but a *fundamental operator underlying both cognition and reality*. In that formulation, mental models and structured reality are both *organized systems* [2,3] governed by the same basic law: *information becomes meaningful through organization*. The core formalism is given by the Law of Structural Cognition, $\mathbb{M} = \text{I} \cdot \text{O}$ [4], and the parallel Law of Structural Reality, $\mathbb{R} = \text{I} \cdot \text{O}$ [4], with *model–world alignment represented as the progressive convergence of mental and real organization over time*.

Within this framework, organization is not vague or metaphorical. It is specified through four universal patterns: Distinctions, Systems, Relationships, and Perspectives (DSRP). Distinctions organize identity and other; Systems organize part–whole structure; Relationships organize action–reaction dynamics; and Perspectives organize point–view framing. O-Theory treats these not as separate “buckets,” but as *simultaneous and co-implicated structural mechanics* that recur across cognitive, material, and computational systems [5], *formally described within the unified calculus* [6]. DSRP therefore sits within a broader empirical research program documenting the *existence* [7–11] and *effect* [12–16] of Distinctions, Systems, Relationships, and Perspectives across both mind and nature.

This matters for language because interpretation is not generated by information alone, but by information as organized relative to a target element. Two passages may contain comparable informational content yet yield different interpretations because the surrounding information is organized differently with respect to the same focal text. Thus, in both written and spoken language, there is always both text and context: a target or focus text, and a surrounding textual field that functions as its contextual other. Context is not non-text; it is the remainder of the text insofar as it structures the interpretation of the target. It distinguishes the target from alternatives, situates it within part–whole and role structures, relates it causally or associatively to other elements, and frames it from particular points of view. In DSRP terms, meaning depends on how distinctions are drawn, what systemic roles are activated, what relationships constrain candidate interpretations, and what perspective organizes the target. From this standpoint, context is best understood not as mere textual adjacency, but as the structural organization of information around and relative to what is being interpreted.

Unlike many traditional treatments in linguistics and natural language processing, context is often implicitly treated as additional information—for example, surrounding text, background knowledge, or situational cues added to the focal utterance [17–25]. While such descriptions are descriptively useful, they tend to frame context as *more content rather than as an organizational property of content*.

This paper adopts a different view. Context is not simply additional text appended to the focal element. Instead, context is the surrounding textual field that functions as the structural other of the focal text, organizing and constraining how that focal element can be interpreted. In this sense, context is not

external to text but relationally defined: it consists of the remainder of the textual field insofar as it structures the meaning of the target element.

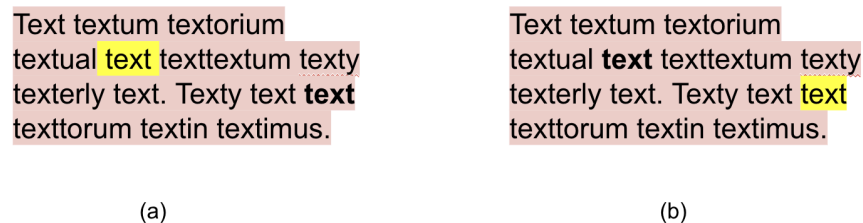
From this perspective, interpretation depends not merely on the presence of additional information but on how that information is organized relative to a focal target. The same informational content can produce different interpretations when the organization of surrounding elements changes. This view aligns with systems-theoretic approaches to cognition in which meaning emerges from the interaction of information and organization rather than from information alone [4].

Accordingly, this paper defines context as the structural organization of information relative to a target element:

$$C = O(I \mid t_0)$$

where I is the available information, O is its organization, and t_0 is the target element whose interpretation depends on that organization. This definition follows directly from the broader O-Theory claim that outcomes depend multiplicatively on the joint interaction of information and organization, not on information quantity alone.

Figure 1. Text and Context as Target and Surrounding Textual Field. The figure illustrates how the same informational content can produce different interpretations depending on which element is treated as the target text. In each panel, the highlighted word represents the focal text, while the surrounding words constitute the contextual field that organizes its meaning. Although the informational content remains identical, shifting the focal element changes the contextual structure surrounding it, demonstrating that context is not additional information but the surrounding textual organization relative to the target element. Yellow highlighting indicates the focal text, red indicates context, and bold indicates the focal text as a and b swap roles as text and context.



The present study translates that theoretical claim into an empirical test. Using WinoGrande and SQuAD 2.0, we operationalize structural context through DSRP-derived features and ask four questions: Can context be measured structurally? When information is approximately constant, do organizational differences matter? Does interpretation accuracy increase with structural context? And does information quantity alone fail to explain interpretation? We also examine whether low structural context corresponds to underconstrained interpretation spaces in which multiple plausible answers remain viable. In doing so, the paper aims to provide a measurable, testable account of context that bridges systems thinking, cognition, and AI interpretability.

2. Literature Review and Theoretical Background

Research on language interpretation routinely acknowledges the importance of context, but context is often defined only loosely—as surrounding text, prior discourse, background knowledge, or situational framing. These descriptions are useful, but they do not by themselves yield a measurable structural variable. The present study addresses that gap by drawing on O-Theory and the broader DSRP research program, which argue that meaning depends not only on information but on *the organization of information*.

O-Theory extends an empirically validated systems framework by formalizing organization as the invariant operator common to cognition and reality. In that formulation, the mind organizes information according to the Law of Structural Cognition, $\mathbb{M} = \text{I} \cdot \text{O}$, while reality is governed by the parallel Law of Structural Reality, $\mathbb{R} = \text{I} \cdot \text{O}$. Their ratio defines model–world alignment, expressed in the Love Reality Loop (LRL), with cognition and reality asymptotically converging toward structural equivalence over time. O-Theory therefore treats organization not as a descriptive add-on, but as *a general law linking mind, matter, and meaning across mental, material, and computational systems*.

Within this framework, organization is specified through the simultaneous field of four universal patterns: Distinctions, Systems, Relationships, and Perspectives (DSRP). Distinctions organize identity and other; Systems organize parts and wholes; Relationships organize actions and reactions; and Perspectives organize points and views. O-Theory emphasizes that these patterns are not independent analytic “buckets,” but *co-arising, co-implicated manifestations of a single structural field*. Activating one implicates the others through simultaneity, recursive containment, and fractal organization. This matters for language because interpretation is rarely resolved by one isolated cue; it is constrained simultaneously by boundaries, systemic roles, relations, and framing.

The broader DSRP literature cited in O-Theory positions these patterns as empirically supported across both cognition and nature. O-Theory summarizes this evidence as converging on three empirical pillars: existence, effect, and universality, arguing that the same structural operators observed in cognition are instantiated in natural systems and produce measurable effects when trained. The referenced body of work includes empirical studies of Distinctions, Systems, Relationships, and Perspectives in mind and nature, along with studies showing that metacognitive awareness of DSRP can significantly increase cognitive complexity and systems thinking performance. O-Theory also cites work on the Systems Thinking and Metacognitive Inventory (STMI), which supports the claim that systems thinking and metacognition can be measured as structured skills rather than treated as vague dispositions.

A central implication of this literature is that information alone is insufficient for understanding. O-Theory states explicitly that *understanding is not about more information, but about how information is organized*, and that better learning and decision-making occur when the structure of thought more closely mirrors the structure of reality. The same logic appears in the related Unified Calculus paper, which argues that information alone is inert and that meaning, function, and form arise from organization, with DSRP serving as the canonical manifestation of this patterned organization. This provides the theoretical basis for the present study’s core equation:

$$C = O(I \mid t_0)$$

where context is defined not as text volume or background knowledge in general, but as the structural organization of information relative to a target element t_0 .

This theoretical perspective also has direct relevance for artificial intelligence. O-Theory argues that most current AI systems rely heavily on statistical association, whereas a more adequate model of intelligence would require recursive structural organization and model–world alignment. The implications section further identifies AI as a domain in which O-Theory provides a “*universal model–world alignment engine*” for *structural cognition and self-organizing AI*. From this standpoint, failures of interpretation are not simply failures of information access; they are *failures of structural organization*.

The present study translates these theoretical claims into an empirical test. If O-Theory is correct, then context should be measurable as a structural variable, interpretation accuracy should increase with structural context, and information quantity alone should not reliably explain interpretation outcomes. By operationalizing context through DSRP-derived features in WinoGrande and SQuAD 2.0, this paper tests whether the structural law proposed in O-Theory can be observed in natural language interpretation tasks.

3. Hypotheses

The study evaluated four hypotheses about the role of structural context in interpretation. Together, these hypotheses test whether context can be operationalized as a measurable structural variable, whether organization matters independently of information quantity, and whether increases in structural context improve interpretation accuracy.

Table 1. Study hypotheses and symbolic form. The four hypotheses tested in the study, stated verbally and in symbolic form.

Hypothesis	Name	Verbal Statement	Symbolic Logic
H1	Measurability (O-measurable)	Context can be measured as a structural variable derived from DSRP features.	$C = f(D_c, S_c, R_c, P_c, K_c)$
H2	Organization Sensitivity (O-sensitive)	When information remains approximately constant, differences in organization are expected to influence interpretation.	$I_1 \approx I_2, O_1 \neq O_2 \rightarrow \hat{Y}_1 \neq \hat{Y}_2$
H3	Context–Accuracy Relationship (O-accuracy)	Interpretation accuracy increases as structural context (O) increases.	$\frac{dA}{dC} > 0$

H4	Information Insufficiency (O-necessity or I-insufficiency)	Information quantity alone does not reliably predict interpretation accuracy.	$\frac{dA}{dI} \approx 0$
----	--	---	---------------------------

Where:

C = context score

D_c, S_c, R_c, P_c, K_c = coded DSRP-derived structural features

I = information quantity

O = organization

$\hat{Y}$ = interpretation

A = interpretation accuracy

In the present study, H2 is evaluated indirectly through differences in interpretation success and interpretive determinacy rather than through exhaustive semantic comparison of all possible interpretations. These hypotheses follow directly from the underlying theoretical claim that meaning depends on both information and organization, rather than on information alone.

4. Methods

4.1 Datasets

Two widely used natural language benchmarks were selected. *WinoGrande*: WinoGrande is a commonsense reasoning dataset designed to test contextual interpretation of ambiguous pronouns. The dataset requires identifying which candidate entity fits a sentence based on contextual constraints. *SQuAD 2.0*: SQuAD 2.0 is a reading comprehension dataset containing both answerable and unanswerable questions about passages.

4.2 Sample

A random sample of 600 items was drawn.

Table 2. Study sample. Number of sampled items drawn from each dataset.

Dataset	Items
WinoGrande	300
SQuAD 2.0	300
Total	600

Sampling used a fixed seed (42) to ensure reproducibility.

4.3 Structural Context Coding

Structural context was coded using five DSRP-derived variables.

Table 3. Structural context coding variables. DSRP-derived variables used to compute structural context.

Variable	Meaning
D_c	Distinction constraint
S_c	System/role constraint
R_c	Relational/causal constraint
P_c	Perspective/frame constraint
K_c	Structural convergence

Each feature was scored as follows:

Table 4. Feature scoring scale. Ordinal coding scale used for each DSRP-derived feature.

Score	Meaning
0	absent
0.5	partial
1	strong

4.5 Context Score

Structural context was calculated as:

$$\bar{S} = (D_c + S_c + R_c + P_c)/4$$

$$C = \bar{S} \cdot K_c$$

An optional reporting scale was used:

$$C_{(0-4)} = 4C$$

4.6 Feature Extraction

Feature extraction was performed using AI-assisted raters under a locked coding protocol. AI raters were used only to extract structural features; context scores were computed deterministically from the coded features.

4.7 Reliability

A subset of 40 items was independently coded by two AI raters.

Table 5. Inter-rater reliability

Caption: Agreement and Cohen's kappa for each structural feature.

Feature	Agreement	κ
D_c	0.875	0.704
S_c	0.900	0.707
R_c	0.775	0.424
P_c	0.850	0.616
K_c	0.900	0.827

4.8 Interpretation Accuracy

Interpretation accuracy was computed differently for each dataset. *WinoGrande*: Correct interpretation required selecting the benchmark-labeled correct candidate entity. *SQuAD 2.0*: Answers were evaluated using normalized exact match and token-level F1 scoring. For SQuAD 2.0, interpretation outcomes were evaluated using benchmark answer spans. Because some items may permit multiple logically plausible interpretations, benchmark mismatch should not always be interpreted as semantic error. In the present design, such cases are treated as informative indicators of weak contextual convergence.

5. Results

5.1 Dataset Summary

The two datasets occupied markedly different regions of the information–organization space. SQuAD 2.0 items contained substantially more information on average (154.247 tokens) but produced near-zero benchmark accuracy (0.003), whereas WinoGrande items contained far less information (21.143 tokens) but much higher interpretation accuracy (0.660). Mean structural context was also substantially higher in WinoGrande (0.890) than in SQuAD 2.0 (0.470). Taken together, these descriptive results already suggest that interpretation is more closely related to structural organization than to information quantity alone.

Table 6. Dataset statistics. Mean context score, interpretation accuracy, and information quantity by dataset.

Dataset	Mean Context Score	Accuracy	Tokens
SQuAD 2.0	0.470	0.003	154.247
WinoGrande	0.890	0.660	21.143

5.2 Context–Accuracy Correlation

Structural context showed a significant positive relationship with interpretation accuracy.

$$r = 0.347, p < 0.001$$

This result indicates that items with greater structural context were more likely to be interpreted correctly. In other words, as contextual organization increased, interpretation accuracy increased as well.

5.3 Context Gradient

Items were divided into four structural context levels.

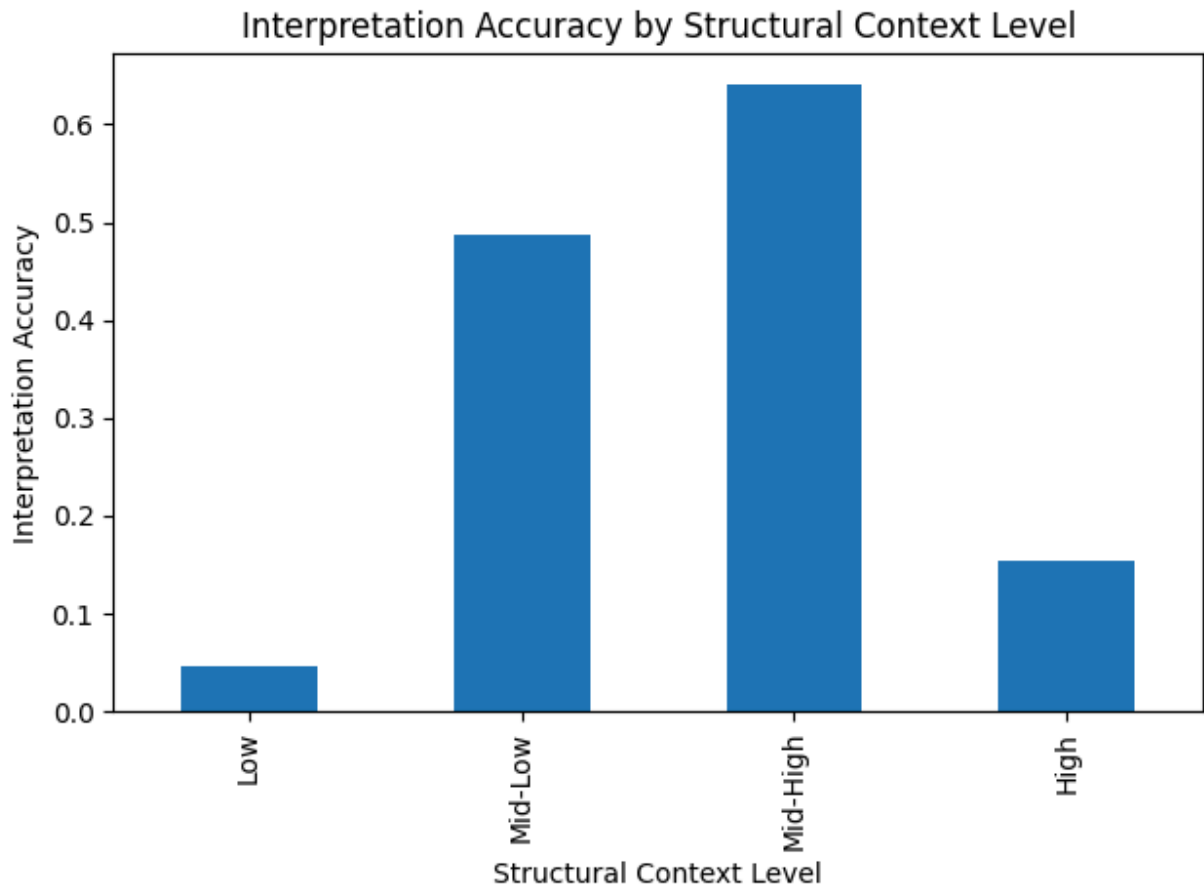
Table 7. Interpretation accuracy by structural context level. Mean interpretation accuracy across four context bins.

Context Level	Accuracy
Low	0.047
Mid-Low	0.487
Mid-High	0.640
High	0.153

Accuracy increased from 4.7% in the low-context bin to 64.0% in the mid-high-context bin, representing a 13.7× increase. This provides strong evidence that structural context constrains interpretation. The decline in the highest bin is best interpreted as a composition effect produced by the dataset mix rather than as evidence against the theory. Specifically, the highest bin included a disproportionate share of SQuAD 2.0 items, for which benchmark scoring remained near zero despite the presence of substantial text.

Figure 2. Interpretation Accuracy by Structural Context Level. Mean interpretation accuracy across four structural context bins. Accuracy rises sharply from low to mid-high context, indicating that stronger structural organization substantially improves interpretive success. The decline in the highest

bin reflects the mixed dataset composition of that bin rather than a reversal of the underlying context effect.



5.4 Context Gradient by Dataset

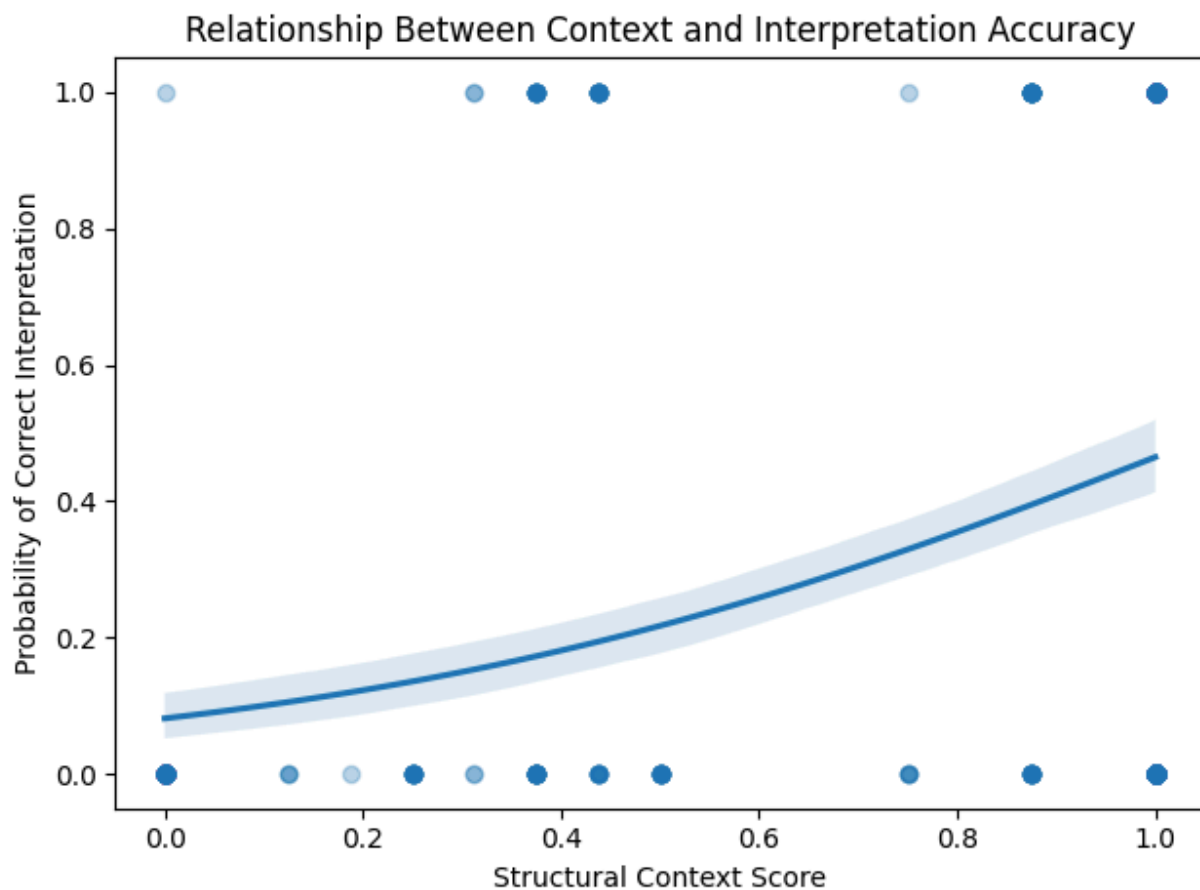
The dataset-specific results reveal two distinct interpretive regimes. WinoGrande showed consistently high interpretation accuracy across context bins, reflecting the fact that many items already contained strong structural constraints. Because most WinoGrande items were high-context items, variation across bins was compressed.

Table 8. Interpretation accuracy by dataset and structural context level. Mean interpretation accuracy across context bins separately for each dataset.

Dataset	Low	Mid-Low	Mid-High	High
SQuAD 2.0	0.006993	0.000000	NaN	0.000000
WinoGrande	0.857143	0.663636	0.640000	0.696970

SQuAD 2.0, by contrast, produced near-zero benchmark accuracy across bins. However, this should not be interpreted solely as model failure. Inspection of individual items indicates that some SQuAD 2.0 questions were structurally underconstrained relative to the annotated answer span, leaving multiple candidate interpretations viable. In such cases, benchmark scoring penalized responses that were logically consistent with the passage but did not match the labeled span. This pattern is consistent with the structural account proposed here: weak contextual convergence reduces interpretive determinacy, so benchmark “error” may partly reflect ambiguity in the item rather than failure of interpretation.

Figure 3. Relationship Between Context and Interpretation Accuracy. Logistic fit relating structural context score to probability of correct interpretation. The positive slope indicates that higher structural context is associated with higher interpretation accuracy overall, even though the underlying datasets occupy different contextual regimes.



5.5 Regression Analysis

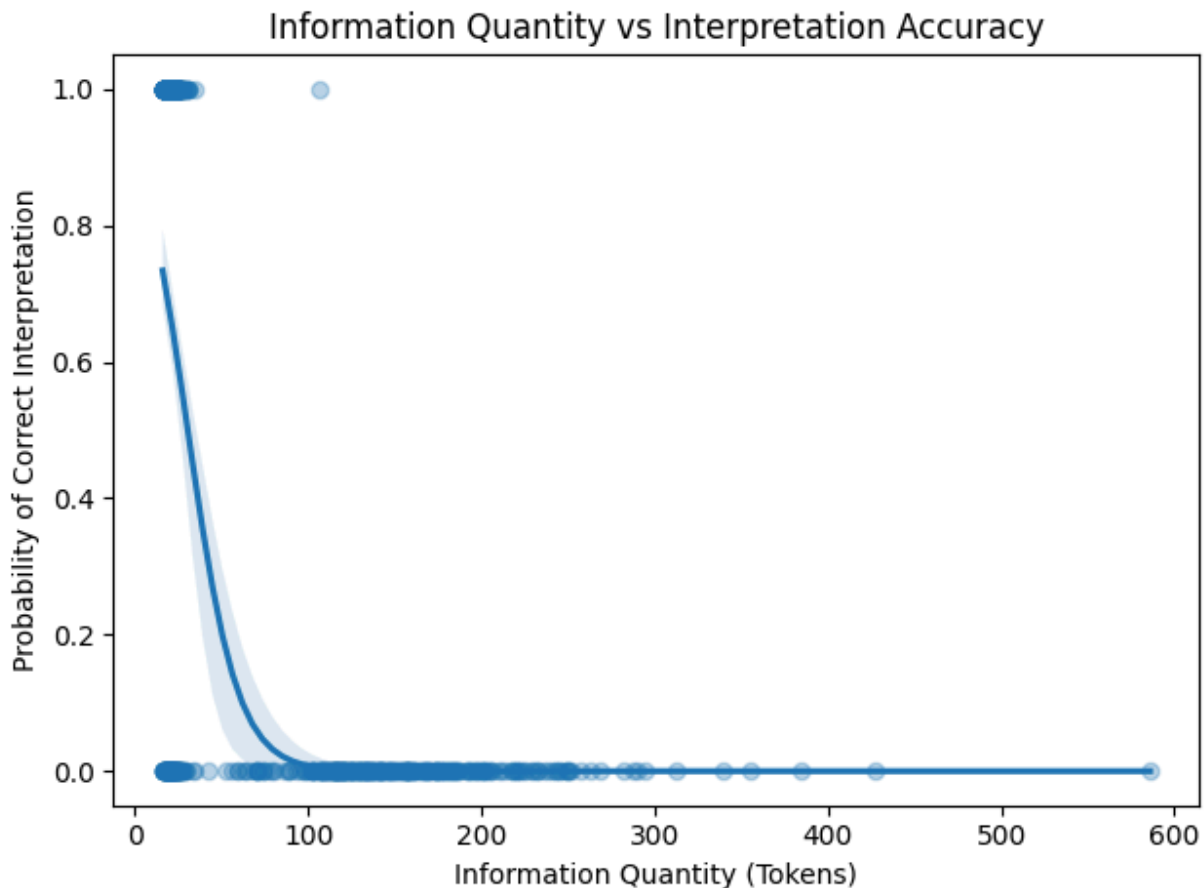
Logistic regression models were used to evaluate predictors of interpretation accuracy.

Model 1:

Interpretation $\sim$ Context + Information + Dataset

Results indicated that context was not statistically significant after controlling for dataset differences, and information quantity was not a significant predictor. This pattern suggests that the main effect of context is partly expressed through dataset structure rather than through a simple uniform linear increase across all items. More importantly, information quantity alone did not explain interpretation accuracy, reinforcing the central claim that organization matters more than raw textual volume.

Figure 4. Information Quantity vs Interpretation Accuracy. Logistic fit relating information quantity to probability of correct interpretation. The near-flat relationship across most of the observed range indicates that information quantity alone does not reliably predict interpretation accuracy.



5.6 Structural Context and Interpretive Determinacy

An additional result of the present study is that structural context appears to predict not only interpretation accuracy, but also *interpretive determinacy*. To test this directly, ambiguity risk was modeled as the inverse of interpretation correctness under benchmark scoring. Logistic regression showed that structural context significantly reduced the probability of interpretive ambiguity or benchmark error ($\beta = -2.28, p < 0.001$). The model significantly improved fit relative to the null model (likelihood-ratio $p = 1.25 \times 10^{-19}$) and achieved a McFadden pseudo- R^2 of 0.108. These results indicate that as structural context increases, the interpretation space becomes more constrained and benchmark ambiguity declines.

In plain terms, low-context items were much more likely to leave multiple interpretations viable, whereas higher-context items more often constrained interpretation toward a single outcome. Structural context therefore appears to be not only a predictor of whether an answer is benchmark-scored as correct, but also a measure of how strongly the surrounding textual field converges on one interpretation.

Figure 5. Interpretive Ambiguity vs Structural Context. Logistic regression showing the probability of interpretive ambiguity (benchmark error) as a function of structural context score. Ambiguity risk declines as structural context increases, indicating that weak contextual organization produces underconstrained interpretations.

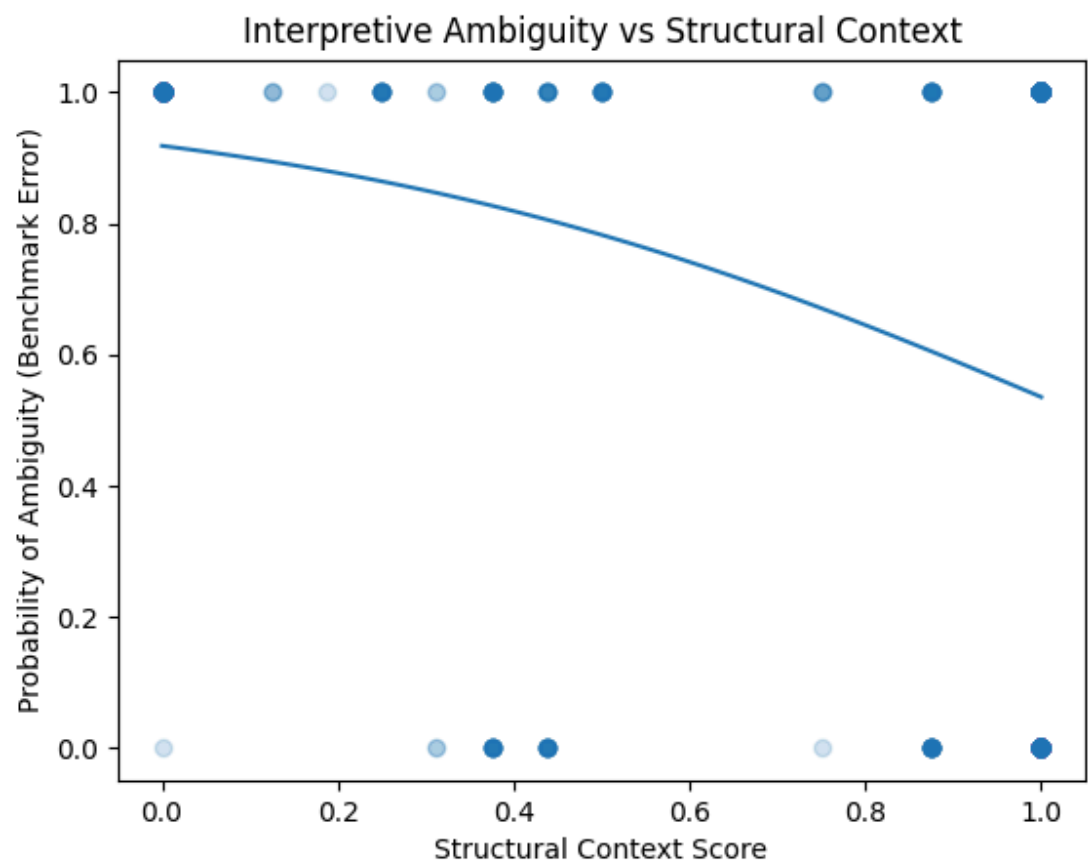


Table 9. Ambiguity risk by dataset and structural context level. Number of items and mean ambiguity risk across context bins for each dataset. Ambiguity risk is defined as 1 – interpretation correctness under benchmark scoring.

Dataset	Context Bin	Items	Ambiguity Rate
SQuAD 2.0	Low	143	0.993007
SQuAD 2.0	Mid-Low	40	1.000000
SQuAD 2.0	Mid-High	0	NaN

SQuAD 2.0	High	117	1.000000
WinoGrande	Low	7	0.142857
WinoGrande	Mid-Low	110	0.336364
WinoGrande	Mid-High	150	0.360000
WinoGrande	High	33	0.303030

Table 9 shows that ambiguity risk was consistently high in SQuAD 2.0 and substantially lower in WinoGrande. More importantly, across the combined dataset, lower structural context was associated with higher ambiguity risk. This pattern supports the interpretation that weak contextual organization leaves the target underconstrained, allowing multiple candidate interpretations to remain viable under benchmark scoring.

5.7 Structural Underspecification in SQuAD

Closer inspection of individual SQuAD 2.0 items suggests that some benchmark errors arise not from failed interpretation, but from structural underspecification in the item itself. In such cases, the passage and question do not converge strongly enough on a single determinate answer, even though benchmark scoring permits only one annotated span.

Table 10 shows three representative examples drawn directly from the analyzed dataset. Each item belongs to the low structural context bin and illustrates situations in which the interpretation space remains weakly constrained.

Table 10. Representative low-context SQuAD 2.0 items exhibiting interpretive ambiguity. Passage excerpts are included to show how weak structural convergence leaves the question underconstrained relative to the benchmark answer (emphasis added).

Passage Excerpt	Question	Gold Answer	Model Interpretation	Context Score
"The paper showed hostility to the EU and approval of public spending cuts, tax cuts, and promotion of right-wing ministers to the cabinet."	"What's another thing the paper showed hostility to?"	promotion of right-wing ministers to the cabinet	The paper showed hostility to the EU	0.25 (low)
The Caribbean initially provided England's most important and lucrative colonies, but not before several attempts at colonisation failed. An	When did England first establish a colony in Guiana?	1604	The answer is not explicitly stated in the passage but an attempt	0.00 (low)

attempt to establish a colony in Guiana in 1604 lasted only two years, and failed in its main objective to find gold deposits..."			was made in 1604	
"Guest judges may occasionally be introduced. In season two , guest judges such as Lionel Richie and Robin Gibb were used, and in season three ..."	In what season was Lionel Richie a guest judge on American Idol?	season two	Lionel Richie was a guest judge in season two	0.00 (low)

In the first example, the passage lists several political positions associated with the newspaper. The question asks for "another thing the paper showed hostility to," while the sentence simultaneously references hostility to the EU and approval of several policy positions. Under benchmark scoring, only the annotated span "promotion of right-wing ministers to the cabinet" is accepted as correct. However, the model selects "the EU," which is also a semantically coherent target within the sentence structure.

The second example illustrates a different form of weak contextual convergence. The passage mentions that an *attempt* to establish a colony in Guiana occurred in 1604 (not the establishment of one), but the model predicts that the answer is not explicitly stated because it distinguished between *attempt* and *establishment*. This indicates that the question–passage linkage is sufficiently diffuse that the model fails to resolve the relevant distinction.

The third example shows a case where the model produces the correct interpretation ("Lionel Richie was a guest judge in season two"), yet the structural context score remains extremely low because the surrounding textual field provides minimal constraints linking the question and the answer span (only "season two").

From the perspective of the present framework, these cases represent low contextual convergence. The available distinctions, systems, relationships, and perspectives do not converge strongly enough to force a single determinate interpretation. Under such conditions, multiple candidate interpretations may remain viable within the interpretation space.

Benchmark scoring protocols, however, require a single labeled span. When contextual organization is weak, responses that are semantically coherent with the passage may still be scored as incorrect. Structural context therefore provides a principled way to identify when a benchmark item is not simply difficult, but structurally underconstrained.

Taken together with the quantitative results in Section 5.6, these examples suggest that structural context predicts not only interpretive success, but also interpretive determinacy. When contextual organization is strong, interpretation converges toward a single answer and benchmark agreement increases. When contextual organization is weak, multiple interpretations may remain plausible, and benchmark "errors" may partly reflect ambiguity in the item rather than failed comprehension alone.

It is also important to note that SQuAD 2.0 contains a substantial number of questions labeled as unanswerable. In these cases, the benchmark requires the model to predict that no answer span exists in the passage. In the analyzed sample, 115 of 142 SQuAD 2.0 items (81.0%) fell into this category. When a model extracts a span from the passage in response to such questions, the benchmark necessarily scores the prediction as incorrect, even if the extracted span is semantically plausible or correct. These cases therefore represent a distinct evaluation condition rather than the type of interpretive ambiguity illustrated in Table 10. The examples discussed above focus specifically on answerable items in which multiple candidate interpretations remain viable under weak contextual convergence.

6. Discussion

The results support a structural interpretation of context. Most importantly, they support the broader O-Theory claim that outcomes depend on the joint interaction of information and organization, not on information alone. In O-Theory, cognition and reality are governed by the same dual law— $\mathbb{M} = \mathbb{I} \cdot \mathbb{O}$ and $\mathbb{R} = \mathbb{I} \cdot \mathbb{O}$ —with model–world alignment defined by their recursive convergence through the Love Reality Loop (LRL). The present study operationalized this claim in natural language interpretation by defining context as structural organization relative to a target element, $C = O(I \mid t_o)$, and by measuring that organization using DSRP-derived features.

First, the study shows that context can be measured as a structural variable. Inter-rater reliability was moderate to strong across all five coded features, indicating that DSRP-based structural constraints are not merely philosophical constructs but extractable and reproducible empirical variables. This finding is consistent with O-Theory’s argument that DSRP forms the generative grammar of organization and that these same structural operators are instantiated across cognitive, material, and computational systems. In this sense, the paper does not merely use DSRP as a metaphor for context; it treats DSRP as a measurable formal basis for contextual organization.

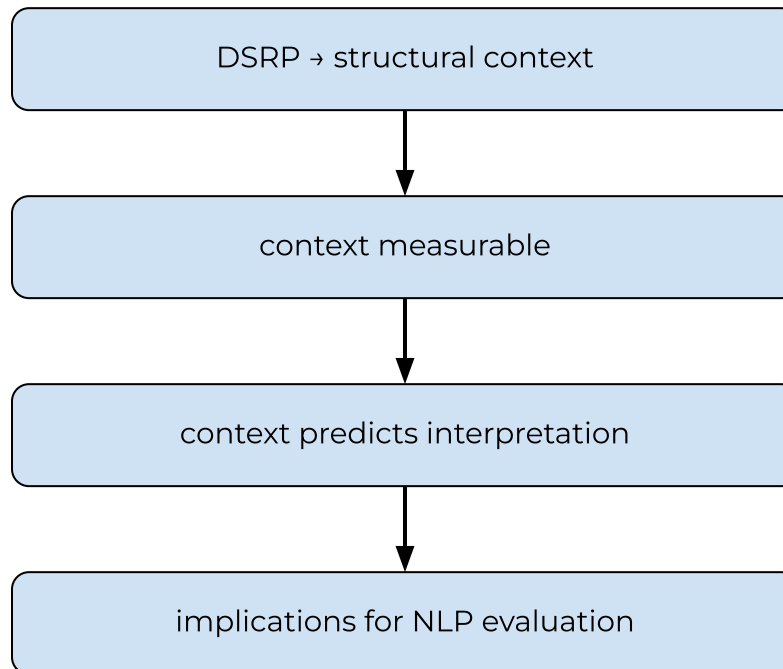
Second, the study provides evidence that interpretation accuracy increases with structural context. Structural context scores were significantly correlated with interpretation accuracy, and the context-gradient analysis showed a substantial rise in accuracy from the low-context bin to the mid-high-context bin. This is exactly the kind of relationship one would expect if meaning depends on organization. O-Theory states this principle directly: understanding is not a function of “more information,” but of how information is organized, and better outcomes occur when the structure of thought more closely mirrors the structure of reality. The present findings extend that claim into natural language interpretation by showing that structurally richer contexts constrain meaning more effectively than structurally weaker ones.

Third, information quantity alone did not reliably predict interpretation accuracy. This result is theoretically important because it cuts against a common implicit assumption in both NLP and educational practice: that more text, more tokens, or more data should naturally improve understanding. O-Theory and the Unified Calculus argue instead that information alone is inert and that meaning, function, and form arise from organization. The present results are consistent with that claim. Token count by itself was not a significant predictor, whereas structural context showed a meaningful relationship to

interpretive success. Because:

$$\text{Interpretation} \neq f(I) \quad \text{but} \quad f(I, O)$$

Or visually,



The results also point to something more specific: structural context appears to predict not only interpretation accuracy, but also *interpretive determinacy*. The ambiguity-risk regression showed that as structural context increased, the probability of benchmark ambiguity or benchmark error declined sharply. This suggests that structural context is not merely associated with getting an answer “right” under benchmark scoring; it is also associated with how strongly the surrounding textual field converges on a single interpretation. Put simply, low-context items tend to leave more than one answer alive, whereas higher-context items narrow the interpretation space.

This matters because benchmark datasets often collapse interpretation into a single labeled answer, even when the text may support more than one plausible response. The SQuAD 2.0 examples examined here suggest that some benchmark “errors” are better understood as cases of structural underspecification. In such items, the available distinctions, systems, relationships, and perspectives do not converge strongly enough on a single target interpretation. Under those conditions, multiple answers may remain semantically coherent with the passage even though benchmark scoring permits only one annotated span. From the perspective of the present framework, these are not simply failures of comprehension; they are signs of low contextual convergence.

These findings also matter for systems thinking more broadly. O-Theory treats DSRP not as four independent analytic steps but as simultaneous, co-implicated manifestations of a single structural field linking cognition and reality. That matters for this study because natural language interpretation is rarely

resolved by one cue in isolation. A candidate meaning becomes more or less interpretable depending on how distinctions are drawn, what system roles are activated, what relationships constrain interpretation, and what perspective frames the target. Context, in this view, is not “more surrounding text”; it is the dynamic structural field that organizes possible interpretations and, in so doing, determines how narrow or wide the interpretation space remains.

The results further connect to the Love Reality Loop. The LRL describes how cognition iteratively aligns with reality through cycles of feedback and fitback, reducing the discrepancy between mental and real structure over time. In language terms, interpretation can be understood as a local model–world alignment problem: the interpreter must organize available information so that the resulting interpretation fits the structure of the underlying referent, event, or relation. When contextual organization is weak, ambiguity remains high and interpretive fit degrades. When contextual organization is strong, the alignment problem becomes easier because structural constraints converge on a smaller set of plausible interpretations. The present findings suggest that contextual convergence is one of the key mechanisms through which interpretive fit is achieved.

The study also has implications for AI interpretability. O-Theory argues that artificial intelligence requires a model–world alignment engine grounded in structural cognition, while the Unified Calculus argues that most current AI systems still operate largely on statistical association rather than structured cognition. The present results support that concern. WinoGrande, which contains relatively strong structural constraints, produced much higher interpretation accuracy than SQuAD 2.0 under the evaluated benchmark protocol. More importantly, the ambiguity analysis suggests that current benchmark performance may reflect not only model competence, but also the degree to which contextual structure converges on a determinate answer. From that standpoint, improving AI interpretability may require systems that represent and manipulate structural organization more explicitly, rather than relying mainly on token co-occurrence or surface statistical regularities.

At the same time, the results should be interpreted with appropriate caution. The decline in the highest context bin is best understood as a dataset-composition artifact rather than evidence that very high structural organization reduces interpretability. The combined sample includes two different task regimes: WinoGrande, where contextual structure is generally strong and accuracy is relatively high, and SQuAD 2.0, where a number of items appear to admit multiple plausible interpretations under benchmark scoring. Thus, the present evidence strongly supports the claim that structural context matters, but it does not yet support a general claim that all high-context tasks are uniformly easier or that benchmark “error” always reflects semantic failure rather than ambiguity in the item itself.

Taken together, the findings support the central systems-thinking claim that meaning depends on organization. They suggest that context can be treated not as an informal background condition, but as a measurable structural property of information. More specifically, the study suggests that structural context constrains both *interpretive success* and *interpretive determinacy*. In that sense, the paper provides an empirical bridge between systems thinking, natural language interpretation, and AI interpretability: it shows that interpretive performance is shaped less by how much information is present than by how that information is structurally organized.

7. Limitations

Two limitations should be noted.

First, the datasets used in this study occupy different regions of structural context. WinoGrande items generally contain strong contextual constraints, whereas many SQuAD 2.0 items exhibit comparatively weak contextual convergence. As a result, the combined dataset does not provide a smooth distribution across the full spectrum of structural context.

Second, benchmark scoring protocols—particularly in extractive question-answering datasets such as SQuAD 2.0—require a single annotated answer span even when multiple interpretations may remain structurally plausible. In such cases, responses that are logically consistent with the passage can still be scored as incorrect. Accordingly, benchmark error in low-context items should be interpreted cautiously, as some apparent errors may reflect interpretive ambiguity rather than failed comprehension.

Importantly, these limitations do not alter the central findings of the study. Structural context was measurable, structural context significantly predicted interpretation outcomes, and information quantity alone did not reliably explain interpretation accuracy.

Future research should apply this framework to datasets that span a wider range of contextual convergence and to evaluation protocols that explicitly account for interpretive ambiguity.

8. Conclusion

This study demonstrates that context can be operationalized as structural organization.

Using DSRP-derived features, structural context scores predicted interpretation accuracy and produced a clear context–accuracy gradient. Information quantity alone did not predict interpretation outcomes. These findings provide empirical support for a structural account of context and suggest that interpretation depends not simply on how much information is present, but on how that information is organized.

The results further suggest that structural context may function not only as a predictor of interpretation accuracy, but also as a diagnostic of interpretive determinacy. When contextual cues are weak or insufficiently convergent, multiple plausible interpretations may remain viable even under benchmark evaluation. In this sense, the present framework offers a way to measure not only contextual strength, but also ambiguity and underspecification in natural language tasks.

Taken together, these findings support the broader O-Theory claim that meaning emerges from the joint interaction of information and organization. They also suggest that a structural account of context can help bridge systems thinking, language interpretation, and AI evaluation, and may function not only as a predictor of interpretation accuracy, but also as a diagnostic of interpretive determinacy and benchmark ambiguity.

9. References

1. Cabrera DCL. From DSRP Theory to O-Theory: The Science of Organization. *Journal of Systems Thinking*. 2025;5: 1–8.
2. Cabrera D, Cabrera L. The grammar of reality: The organizing principles of structure and emergence. *Journal of Systems Thinking*. 2025;5: 1–18. doi:[10.54120/jost.2025-000003](https://doi.org/10.54120/jost.2025-000003)
3. Cabrera D. DSRP as Universal Ontology: The Structural Foundation of All Systems. *Journal of Systems Thinking*. 2025;5: 1–40. doi:[10.54120/jost.v25i1.dc6sfas6](https://doi.org/10.54120/jost.v25i1.dc6sfas6)
4. Cabrera D, Cabrera L. The structure of thought and reality: empirical evidence for the three universal laws linking mind, matter, and meaning. *Journal of Systems Thinking*. 2025;5: 1–20. doi:[10.54120/jost.v25i1.dc6smio6](https://doi.org/10.54120/jost.v25i1.dc6smio6)
5. Cabrera D. Enclosure: The Raw Structural Materials Necessary and Sufficient for Any Identity to Exist. *Journal of Systems Thinking*. 2025;5. doi:[10.54120/jost.v25i1.dc6deei6](https://doi.org/10.54120/jost.v25i1.dc6deei6)
6. Cabrera D. The Unified Calculus of organization, structure, emergence, and meaning: The patterns that connect mind and nature. *Journal of Systems Thinking*. 2025;5: 1–41. doi:[10.54120/jost.2025-000002](https://doi.org/10.54120/jost.2025-000002)
7. Cabrera D, Cabrera L, Cabrera E. Distinctions Organize Information in Mind and Nature: Empirical Findings of Identity–Other Distinctions (D) in Cognitive and Material Complexity. *Systems*. 2022;10. doi:[10.3390/systems10020041](https://doi.org/10.3390/systems10020041)
8. Cabrera D, Cabrera L, Cabrera E. Systems Organize Information in Mind and Nature: Empirical Findings of Part-Whole Systems (S) in Cognitive and Material Complexity. *Systems*. 2022.
9. Cabrera D, Cabrera L, Cabrera E. Relationships Organize Information in Mind and Nature: Empirical Findings of Action-Reaction Relationships (R) in Cognitive and Material Complexity. *Systems*. 2022.
10. Cabrera D, Cabrera L, Cabrera E. Perspectives Organize Information in Mind and Nature: Empirical Findings of Point-View Perspective (P) in Cognitive and Material Complexity. *Systems*. 2022.
11. Cabrera D, Cabrera LL, Cabrera E. A Literature Review of the Universal and Atomic Elements of Complex Cognition. *Journal of Systems Thinking*. 2023; 1–85. doi:[10.54120/jost.0000032](https://doi.org/10.54120/jost.0000032)
12. Cabrera D, Cabrera L, Cabrera E. The “Fish Tank” Experiments: Metacognitive Awareness of Distinctions, Systems, Relationships, and Perspectives (DSRP) Significantly Increases Cognitive Complexity. *Systems*. 2022;10. doi:[10.3390/systems10020029](https://doi.org/10.3390/systems10020029)
13. Cabrera D, Cabrera L. The Pareto structure of thought: empirical discovery of the six foundational mental moves. *J Syst Think*. 2025;2025: 1–41. doi:[10.54120/jost.v25i1.dc6par6](https://doi.org/10.54120/jost.v25i1.dc6par6)
14. Aljrarri W 'ad, Bustamante J, Ingwell A, Labrada T, Nadig B, Shin J, et al. New Hope for Policy Schools: Examining Systems Thinking Solutions to Improve Policy Education. *Journal of Systems Thinking*. 2024; 1–XX. doi:[10.54120/jost.00000101](https://doi.org/10.54120/jost.00000101)
15. Di Marco G, Cabrera D. Common biases in systems thinking. *Journal of Systems Thinking Preprints*. 2024. doi:[10.54120/jost.pr000037.v1](https://doi.org/10.54120/jost.pr000037.v1)
16. Cabrera D, Das SKVB, Cabrera L. The thinking quotient (TQ): Further validation of a skill-metric for systems thinking. *Journal of Systems Thinking*. 2025;5: 1–18. doi:[10.54120/jost.v25i1.dc6tqsm6](https://doi.org/10.54120/jost.v25i1.dc6tqsm6)
17. Halliday MAK, Hasan R. *Cohesion in English*. London: Longman; 1976.
18. Clark HH. *Using Language*. Cambridge: Cambridge University Press; 1996.

19. Jurafsky D, Martin JH. *Speech and Language Processing*. 3rd ed. 2023.
20. Wittgenstein L. *Philosophical Investigations*. Oxford: Blackwell; 1953.
21. Firth JR. A Synopsis of Linguistic Theory 1930–1955. *Studies in Linguistic Analysis*. Oxford: Philological Society; 1957. pp. 1–32.
22. Fillmore CJ. Frame Semantics. In: of Korea TLS, editor. *Linguistics in the Morning Calm*. Seoul: Hanshin Publishing; 1982. pp. 111–137.
23. Levinson SC. *Pragmatics*. Cambridge: Cambridge University Press; 1983.
24. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT*. 2019. pp. 4171–4186.
25. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems (NeurIPS)*. 2020. pp. 1877–1901.