Towards an empirically-grounded typology of unacceptability judgments

Alexandre Cremers[1] **& Wataru Uegaki**[2]
[1]Cognistats.eu and Vilnius University, alexandre.cremers@gmail.com
[2]School of Philosophy, Psychology, and Language Sciences, University of Edinburgh

Abstract

Theoretical linguistics commonly distinguishes between syntactic ungrammaticality, semantic anomaly, and pragmatic infelicity, yet there is no established behavioral method for empirically separating these categories. We report two experiments aimed at deriving an empirically grounded typology of unacceptability judgments using multi-dimensional behavioral data. Participants evaluated sentences from a wide range of phenomena against prompts targeting naturalness, truth, interpretability, or contingency. Across both experiments, unsupervised clustering revealed that a stable classification emerges when multiple judgment dimensions are considered and when dimensions which further divide the semantic category are controlled. The resulting typology robustly separates syntactic, semantic, and pragmatic violations, and consistently groups several meaning-driven phenomena, including unlicensed negative polarity items and existential *there* constructions, with core syntactic violations rather than with semantic or pragmatic ones.

Keywords: unacceptability; ungrammaticality; infelicity; meaning-driven unacceptability; syntax-semantics interface; logicality of language; presuppositions; implicatures; NPIs; behavioral measures.

Background

It is commonly assumed in theoretical linguistics that speakers' unacceptability judgments for a given sentence can be categorized into different types. These include *syntactic ungrammaticality*, *semantic violations (such as truth-conditional falsity and contradictions)* and *pragmatic infelicity in a context* (Chomsky, 1957; Grice, 1975; Matthewson, 2004).

Neural imaging methods, such as EEG, MEG, and fMRI have been used to distinguish between syntactic and semantic phenomena (e.g., Friederici 2002; Kutas and Hillyard 1980; Kutas et al. 2006; Osterhout and Holcomb 1992; Pylkkänen 2019; Schell et al. 2017; cf. Shain et al. 2024). However, existing neurolinguistic research does not provide categorization for several classes of unacceptable sentences that play important roles in the theoretical literature, such as Magri cases (Magri, 2009) or Maximize Presupposition violation (Heim, 1991). (These classes will be explained below.) At the same time, there is currently no established *behavioral* method that allows categorization of unacceptability judgments.[1]

[1]There is of course extensive work on the empirical validation of acceptability judgments (e.g., Marty et al., 2020; Sprouse and Almeida, 2012, 2013) or truth-value judgments (Cremers et al., 2023), but no work directly comparing the type of behavioral response triggered by syntactic vs. semantic or pragmatic violations.

This constitutes an important gap in the current research, as some theoretical proposals in linguistics crucially rely on the assumption that relevant unacceptable sentences are classified in a certain way. Of particular interest is the literature on the so-called meaning-driven ungrammaticality (also known as *logicality* cases), such as existential-*there* constructions and unlicensed Negative Polarity Items (NPIs):

(1) Existential-*there* unacceptable with some determiners
   a. There is **a** smiling cat.
   b. *There is **every** smiling cat.

(2) NPI *any* unacceptable without a licenser e.g., negation
   a. The musician has **not** won *any* Grammy Award.
   b. *The musician has won *any* Grammy Award.

It has been argued that unacceptability of sentences such as (1b) and (2b) come from logical triviality at the level of compositional semantics (e.g., Barwise & Cooper, 1981; Chierchia, 2013). Assuming that the relevant unacceptability judgment is that of ungrammaticality rather than semantic anomaly, some authors have furthermore argued that these cases motivate the presence of a cognitive mechanism that filters out a certain class of semantic logical triviality from the set of grammatical sentences (Chierchia 2013; Del Pinal 2019; Gajewski 2002; cf. Qing and Uegaki 2025).

In order to evaluate these theoretical claims, it is crucial that we empirically evaluate the assumption that the relevant cases of unacceptability can be categorized as ungrammatical rather than semantic anomaly. In this paper, we report on an experiment that is aimed at providing an empirically grounded typology of unacceptability, using simple judgment data.

Experiment 1

Methods

120 US-residing English monolinguals were recruited on Prolific and paid £1.75. Each participant answered one of the 6 prompts in (3) about 75 sentences, using a slider. Prompts and associated instructions were between-subjects, but items were the same for all subjects. The sentences spanned a range of phenomena argued to lead to unacceptability, from syntax, semantics, and pragmatics, as well as controls intended to trigger a positive response to each prompt, all presented in Table 1. Since no context was provided, all sentences were designed so their truth could be evaluated against common knowledge. A PCIbex demo is available here (the last digit in

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

the URL determines the prompt). Stimuli, data, and scripts are available on OSF. This experiment was exploratory, and the analyses we present were decided post-data collection.

(3) Prompts:

Naturalness	Does the sentence sound natural?
Truth	Do you think the sentence is true?
Falsity	Do you think the sentence is false?
Interpretability	Do you understand what the speaker meant?
Contingency_1	Can you imagine someone genuinely believing this?
Contingency_2	Could you imagine this sentence being true if the world was different?

Results

24 participants were excluded for rejecting the controls at a rate 1SD above the mean. We also excluded responses which took more than 20s (2.0% of the data). For each combination of prompt and item, we extracted 3 variables. The first corresponded to the average slider position among participants. The second measured disagreement between participants.[2] The third measure, rRT, was obtained by taking the residuals of a linear mixed-effects model fitted on response times, with sentence length and prompt as predictors, and by-subject random slopes. These variables were z-scored to yield 18 features for each item (3 variables × 6 prompts). We defined the center of each category by taking the median of each feature. A model-based clustering with 4 spherical equal-volume components fitted with `mclust::Mclust` (Scrucca et al., 2023) identified the following clusters: C_1 = {controls}, C_2 = {contradictions, false cases, presupposition failures}, C_3 = {A violations, $\overline{\text{A}}$ violations, existential *there*, NPIs, anti-rogativity}, and C_4 = {MP violations, Magri cases, tautologies, processing, factive islands}. In short, we retrieve the expected categories of acceptable sentences (C_1), syntactic violations (C_3), semantic violations (C_2), and pragmatic violations (C_4), with the exceptions of hard-to-process sentences, which we initially expected to involve syntax. Logicality cases all fall into the Syntax cluster, except for factive islands, which are closer to processing difficulties. Fig. 1 shows the similarity between individual sentences, grouped by class, and shows the clustering tree on classes on the left.

Looking at individual sentences, we see that some classes are relatively homogeneous, while others contain clear outliers (e.g., presupposition failures, existential *there*). This, together with the low number of items per class, motivated the use of median centers for clustering instead of centers of mass or direct clustering on individual items.

Next, we ran a series of ablation tests to determine what is or isn't essential for the distinction between syntactic, semantic, and pragmatic phenomena to emerge. We found that only prompt Contingency_2 could be dropped (in most other cases, distinctions within the semantic category would dominate the distinction between syntactic and pragmatic phenomena, resulting in poor classification). We also found that removing either rRT, disagreement, or both, greatly reduced our ability to identify natural categories. Lastly, we ran a principal component analysis on the 18 features. Only the first 4 components explained more than 5% of the variance, and together, they explained 89% of the total. PC1 primarily tracked truth/falsity, which separates semantic violations from the rest. PC2 combined high naturalness and truth, low disagreement, and faster responses. It isolated controls and some pragmatic cases from the rest. PC3 tracked low naturalness and interpretability, which separate most syntactic and logicality cases from the rest, while PC4 indicated differences in disagreement between Naturalness and Contingency_1. Since PC4 was primarily responsible for distinctions within the semantic category, and not distinction between categories, we reasoned that keeping only the first three components would yield a cleaner classification. We confirmed this by rerunning the model-based clustering with only these 3 components. We obtained the same clustering as with all features, but the clustering tree was "cleaner" in that the distances within category (and in particular among the semantic phenomena) were greatly reduced.

Finally, we tested whether the classification could be done directly on individual sentences. Letting the model pick the best number of clusters resulted in 5 clusters which were not particularly informative. Many categories were split, in ways that didn't make much sense from a theoretical perspective.

Discussion

Exp. 1 demonstrated that it is possible to separate the standard categories of unacceptability discussed in the theoretical linguistic literature (syntactic, semantic, and pragmatic) using only behavioral measures. On the theoretical side, logicality cases fell closer to syntactic violations than either semantic or pragmatic ones, confirming the standard assumption.

This said, the results do not fully align with our expectations. Hard-to-process sentences, which we expected to involve syntax, clustered with pragmatic violations. Note however that these sentences are technically grammatical, and that they all clearly violate Grice's maxim of manner ("Be clear!"); this could explain their affinity with pragmatic violations. Factive island also patterned with pragmatic violations (in fact, very close to hard-to-process sentences), which may not be surprising given that the source of their unacceptability has been debated (see Liu et al., 2022 for a recent review).

The results also revealed some limitations in our design. First, the categorical clustering is not perfectly stable: removing any prompt except Contingency_2 leads to a different classification. A closer look at the data showed that the instability comes from the distance between different semantic classes, which is of the same order as the distance between syntactic and pragmatic phenomena. A PCA on the features we defined identified 4 useful components. PC1 separates semantic violations from the rest, while PC2 isolates acceptable controls.

[2] Because slider positions are bound to $[0, 1]$, the variance is capped at $\mu(1-\mu)$, where μ is the mean. We used the normalized standard deviation $s = \sigma/\sqrt{\mu(1-\mu)}$ to define a measure of disagreement as independent of the mean response as possible.

	Class	Explanation	Example
	Control	Acceptable and true sentences	*Most people eat breakfast in the morning*
Syntax	A-violation	Violations involving local dependencies	*Empire the Roman collapsed*
	$\overline{\text{A}}$-violation	Violations involving non-local dependencies	*Librarians usually know which author which famous novel wrote*
	Processing	Garden-path sentences, center-embedding…	*Fat people eat accumulates in their bodies*
Semantics	False	Grammatical but false sentences	*The Pacific Ocean is smaller than Lake Michigan*
	Presup. failure	Sentences whose presuppositions are not satisfied	*Dogs walk on both their legs*
	Contradiction	Sentences that express a logical contradiction	*An empty swimming pool is full*
Pragmatics	Tautology	Sentences that are logically true but uninformative	*A full swimming pool isn't empty*
	Magri case	Cases where deriving a scalar implicature results in a contradiction (Magri, 2009)	*LeBron James is sometimes tall*
	MP violation	Cases where a contextually equivalent alternative presupposes more content (Heim, 1991)	*Most people can hold a pound with all their hands*
Logicality	exist.-*there*	Use of a strong determiner in the existential-*there* construction (Milsark, 1974, 1977)	*There are most motorized cars*
	NPIs	Unlicensed negative polarity items *any, ever…* (Chierchia, 2013; Krifka, 1994; Ladusaw, 1979)	*Humans have ever been to the moon*
	Anti-rogativity	Selection violations: *believe/hope* with a *whether*-clause (Theiler et al., 2019; Uegaki & Sudo, 2019)	*Parents usually hope whether their children pass important exams*
	Factive island	Extraction from factive environments (Abrusán, 2014; Schwarz & Simonenko, 2018)	*It is clear which atom the scientists know is the smallest*

Table 1: Phenomena tested in Experiment 1, with examples. We included 10 controls and 5 sentences of each other class.

Perhaps more interesting is PC3, which combines naturalness and interpretability, and allowed us to separate syntactic from pragmatic violations, which otherwise rate similarly high on truth and low on naturalness. PC4 is harder to interpret since it mostly correlates with a difference in agreement between naturalness and contingency, but it separates our classes in three groups: on the low end, contradictions and tautologies, on the high end, plain false sentences, and in the middle everything else, including presupposition failures.[3] PC4 is therefore responsible for the low cohesiveness of the semantic category, and probably for the instability of the classification overall.

Experiment 2

The main goal of Exp. 2 was to address some of the shortcomings in the design of Exp. 1 and test more rigorously the analysis plan we adopted. To do so, we updated and expanded the stimuli of Exp. 1 to make classes more consistent, and we pre-registered the experiment with its analysis script before collecting the data. All details can be found on the same OSF repository. A secondary goal was to further investigate logicality cases, with a focus on some puzzling divergences within single classes in Exp. 1.

[3]What happened here is that false sentences rate high on naturalness, with high agreement, while participants disagree on the naturalness of presupposition failures, contradictions, and tautologies. By contrast, false sentences are close to the middle of the scale for contingency$_1$ and give rise to the highest disagreement of any class on any prompt, while contradictions are at the low end and tautologies at the high end, both with high agreement. Presupposition failures are close to the middle of the scale, but with relatively low disagreement.

Methods

Exp. 2 followed the same design as Exp. 1, but dropped the sixth prompt (Contingency$_2$), which ablation tests showed to be superfluous in Exp. 1. The stimuli were updated. Controls meant to match syntactic violations turned out to be too complex and were reclassified into Processing. We expanded our presupposition failures class from 5 items representing 3 triggers to 15 items representing 6 triggers, as previous studies have found empirical differences between triggers (Cremers et al., 2017; Cummins et al., 2012; Domaneschi et al., 2018; Göbel, 2022, a.o.). The *A*-violation class was expanded to 10 items illustrating 5 types of violations (agreement, case, unlicensed reflexive, unlicensed negation, verb inflection). A new class was introduced for ungrammatical word-order. Garden-path sentences were moved to a new class as they differed markedly from other hard-to-process sentences. We split Magri cases into those involving quantifiers (discussed in the literature) and those involving disjunctions (which, as far as we know, haven't been discussed, and seemed to behave differently in Exp. 1). We also split existential *there* items into two categories: *there* + quantifier (*all, every, each, most, both*) and *there* + definite article (Exp. 1 included one such item, which diverged from the rest). We also addressed issues with items that had been identified as outliers within their class in Exp. 1.[4] All these changes were meant to make classes more

[4]For instance, some presuppositions were too easy to accommodate ("Dogs walk on both their legs" is not a presupposition failure if *legs* is interpreted as 'hind legs'; it was replaced with "Spiders walk on both their legs"). A typical Magri sentence like "Some Canadians come from a cold country" can be accommodated if we consider

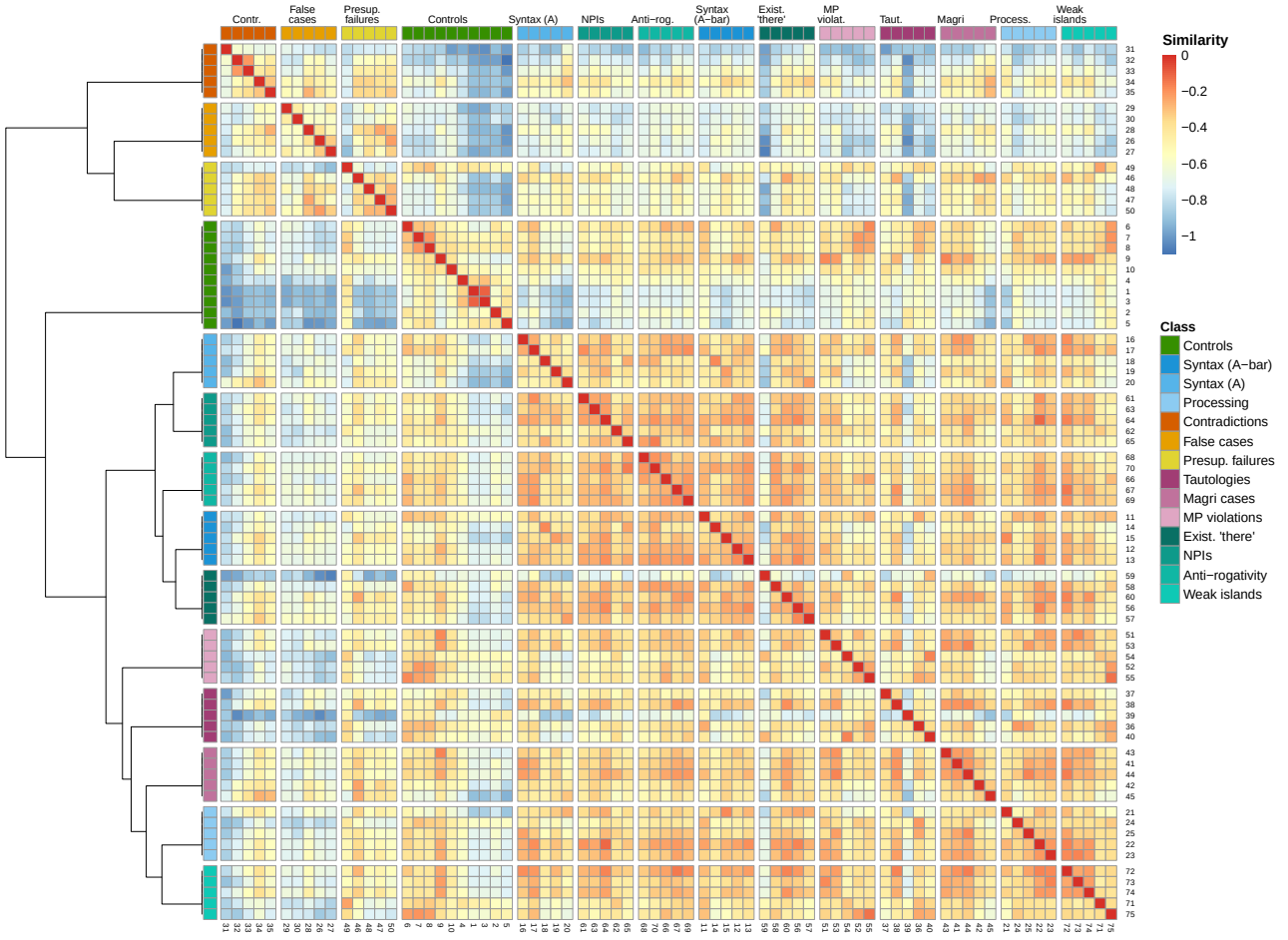

Figure 1: Exp. 1: Similarity between individual items, organized hierarchically by similarity within and between classes.

consistent by removing outliers or splitting them into independent classes when we suspected that the underlying phenomenon may not be uniform. The total number of items rose to 110. 100 US-residing English monolinguals were recruited on Prolific (20 per prompt, as in Exp. 1) and paid £2.20.

Results

8 participants were excluded for rejecting controls at a rate 1SD above the mean, and we excluded responses which took more than 20s (again 2.0% of the data).

We applied the same model-based clustering as in Exp. 1, but this time it resulted in a single cluster for all syntactic and pragmatic violations, and two clusters for semantic violations (one for contradictions, the other for false sentences and presupposition failures). This confirmed our concern regarding the instability of a clustering done with all features, because distances between semantic classes can dominate the distance between syntactic and pragmatic violations. We omitted the pre-registered ablation tests, as the main clustering failure rendered them uninformative.

We followed up with the analysis on the first 3 components of a PCA, which had shown promising results in Exp. 1 and was pre-registered as an exploratory analysis. Despite dropping the last prompt and updating the simuli, the PCA results were surprisingly similar to those of Exp. 1. The first 4 components were again the only ones to explain more than 5% of variance, and together explained 89% of it. Their composition closely matched what we saw in Exp. 1 (modulo sign). PC4 again tracked differences in disagreement between prompts, but this time placed more weight on Falsity than Contigency. This time too, restricting to the first 3 components reduced the distance between semantic classes, and returned the standard classification for model-based clustering. The membership probabilities are presented in Fig. 2, and the detailed by-item similarities (also in the PCA space) are in Fig. 3.

Finally, our pre-registered exploratory analyses also included the clustering on individual sentences. As in Exp. 1, this didn't work well, as only semantic violations were assigned a consistent cluster.

naturalized Canadians with foreign origins; it was replaced by "Some residents of Canada live in a cold country". For sentences with *there* and a definite, we avoided numerals, as we suspected that participants may inadvertently repair sentences like "there are the 50 states" by ignoring the definite.

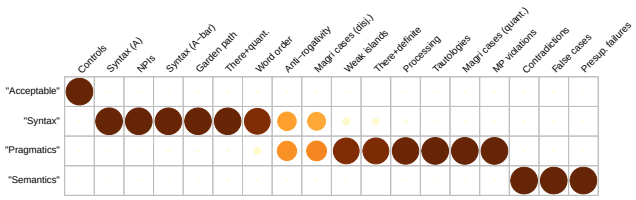

Figure 2: Exp. 2: Membership probability in each cluster for the model-based clustering based on the first 3 PCA components. The names of the clusters are purely descriptive (they are assigned post-hoc since the classification is unsupervised).

Discussion

Exp. 2 largely replicates and strengthens the main conclusions of Exp. 1, while also clarifying the conditions under which a behavioral typology of unacceptability can be recovered. Most importantly, as suspected in Exp. 1, when the analysis is restricted to the first three principal components, we again obtain a clear separation between syntactic, semantic, and pragmatic violations. This convergence is notable given several substantial changes to the design: the removal of one prompt, the expansion and restructuring of stimulus classes, and the pre-registration of the analysis plan. It suggests that the typology uncovered in Exp. 1 is not an artifact of stimuli or analytic flexibility, but reflects stable structure in speakers' judgments.

At the same time, Exp. 2 confirms an important limitation already hinted at in Exp. 1: clustering based on the full feature space is not stable. When all features are included, semantic distinctions again dominate the clustering solution, collapsing syntactic and pragmatic violations into a single cluster. The replication of this limitation and the fact that the PCA components themselves are highly similar across experiments suggest that this is not a statistical accident, but reflects a genuine property of the judgment space. There are certainly meaningful differences between the classes we placed in the semantic category, which would deserve further investigation. Yet, if our goal is to recover the standard tripartite distinction between syntactic, semantic, and pragmatic phenomena, it is crucial to eliminate the principal component which encodes these distinctions between semantic phenomena.

Turning to the question of Logicality, Exp. 2 confirms that these phenomena mostly fall within the syntax category, with the exception of factive islands which are clearly closer to pragmatic violations. After splitting the existential *there* class however, we find a clear difference between definites (which fall in the pragmatic category) and quantifiers (which remain with syntax). Perhaps surprisingly, anti-rogativity now falls halfway between syntactic and pragmatic phenomena. Note that this class is extremely homogeneous, so this unclear classification cannot be due to an internal split between items.

Beyond these core findings, Exp. 2 also reveals several smaller-scale patterns that are worth noting, although they should be interpreted cautiously. First, a subset of presupposition failure items and good controls cluster closely together. Closer inspection suggests that these items involve presuppositions that are relatively easy to accommodate, such as uniqueness from definites or duality from 'both', and control items which are somewhat debatable or prone to world-knowledge uncertainty ("The sun rises in the East"). This convergence therefore seems accidental and does not undermine the broader classification of presupposition failures as semantic violations, but it highlights the sensitivity of this category to accommodation, and the importance of carefully constructing unambiguously true and natural good controls. Interestingly, the rest of the presupposition class was surprisingly homogeneous (in contrast with previously cited results). Second, two tautologies were closer to good controls, including "When it rains, it rains". Finally, some A-violations (agreement and case) patterned more closely with pragmatic violations. Since all of these exceptions concern a minority of items within their class, they do not affect our clustering, which is based on medians.

We also confirm a finding of Exp. 1 regarding disjunctive Magri cases. These sentences involve a symmetric relation, which makes 'or' equivalent to 'and' (e.g., "Newyorkers or Bostonians agree with each other that the Amtrak is too slow"), just like 'some' becomes equivalent to 'all' in familiar cases. To the best of our knowledge, they have not been discussed in the literature. We found that quantifier cases, as a class, were closer to pragmatic classes, but had low internal cohesiveness (with one item even sharing similarity with contradictions). By contrast, disjunctive cases were homogeneous, but did not fall cleanly in either pragmatics or syntax.

General discussion

Together, our experiments confirm that a stable, empirically grounded typology of unacceptability judgments can emerge from behavioral data, provided that (i) judgments are probed along several prompts, some of which are standard (naturalness, truth, falsity), some of which are new (interpretability, contingency), (ii) raw values are complemented by derived measures of (dis)agreement among participants **and** response time, and (iii) some distinctions within the semantic category are appropriately factored out. The replication of the PCA structure and the resulting clustering across two experiments lends credibility to the proposed methodology and its theoretical implications.

The application of our methodology to logicality cases demonstrates that its fruitfulness: it confirmed the consensus in the theoretical literature that many of these phenomena have been grammaticalized, with the exception of factive islands (the nature of which has been debated) and existential *there* with definites (while quantifier cases are definitely in the syntax category).

Another important result concerns the nature of pragmatic violations. In both experiments, they were much closer to syntactic violations than to semantic ones. This runs counter to the common assumption that semantics and pragmatics are closely related, with syntax as a distinct module. Instead, our results suggest that, from the perspective of speakers' judg-

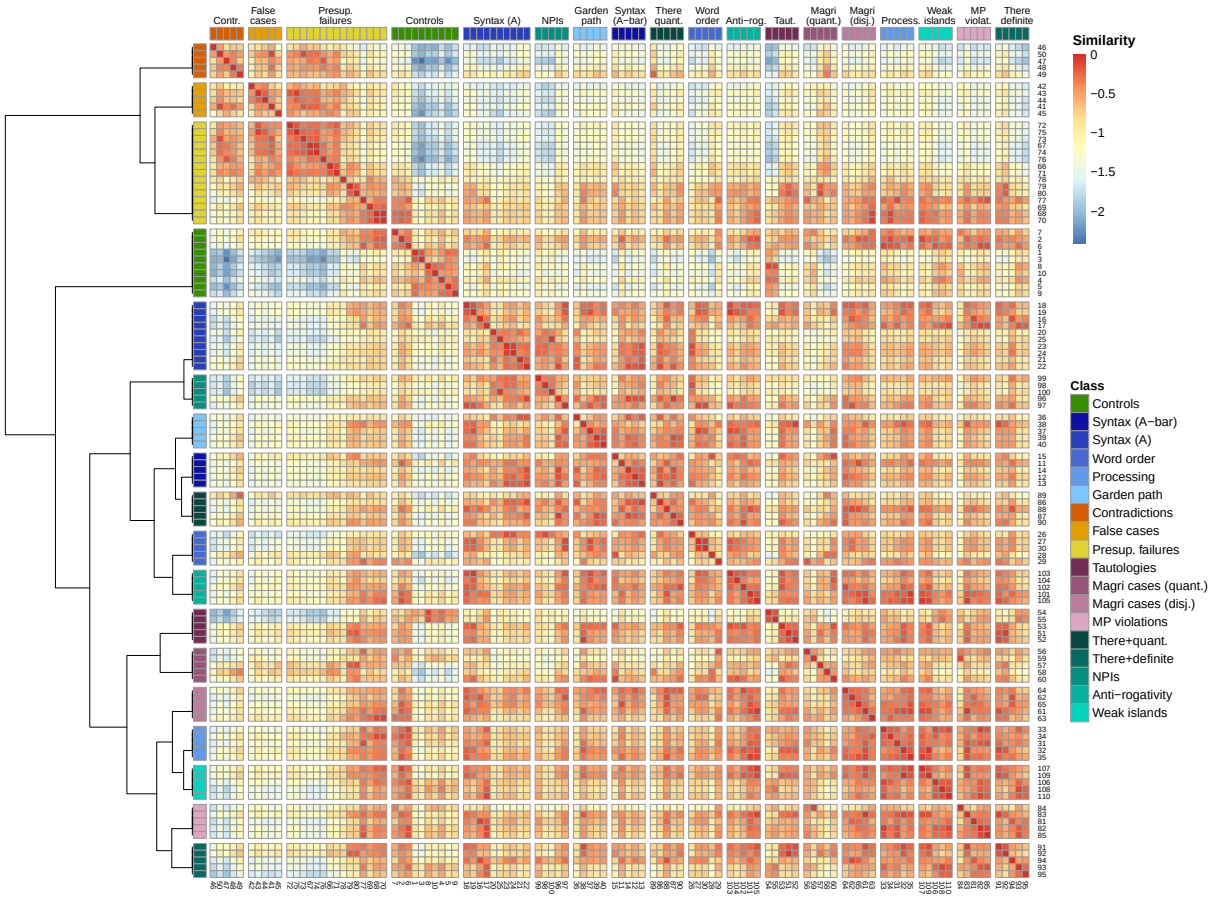

Figure 3: Exp. 2: Similarity between individual items in the space defined by the first 3 PCA components, organized hierarchically by similarity within and between classes.

ments, violations that do not directly affect truth-conditions are more difficult to distinguish from each other than either is to semantic anomaly. This observation invites reexamination of how the syntax–semantics–pragmatics distinction is operationalized in experimental work, and cautions against assuming that theoretical adjacency necessarily predicts similarity in judgment behavior.

Lastly, we found that clustering could only be reliably done on classes' median centers, not on individual sentences, due to excessive within-class variance. There are two possible sources for this variance. First, each theoretical phenomenon must be instantiated by concrete sentences, and it can be difficult to design sentences with enough variety while remaining representative of the intended category. Aggregating via medians helps isolate the "typical" behavior of the phenomenon and abstract away from irrelevant differences between particular instantiations. Indeed, linguistic research questions are formulated in terms of classes, not individual sentences. A more problematic source of variation arises when a single label from the theoretical literature in fact encompasses multiple distinct phenomena. There, medians risk obscuring possibly meaningful variation (as was the case with *there*+definite and disjunctive Magri cases in Exp. 1). A close inspection of individual items is necessary to diagnose such issues.

Our recommendations to anyone planning to use our method are: (i) design at least 10 instantiations of each phenomenon, aiming to cover as many sub-cases as possible (the use of medians mitigates the impact of outliers), (ii) record response times (ablation tests showed all three measures are necessary), (iii) run a PCA to identify and eliminate the component that distinguishes between semantic subcategories, (iv) make sure to inspect individual items to diagnose potential meaningful splits within classes hidden by medians.

To conclude, we offer a new behavioral method to diagnose different types of unacceptabilities, validated on a range of linguistic phenomena. Our results also raise new questions about some specific cases and the nature of pragmatic violations.

Acknowledgment

We would like to thank Laura Vilkaitė-Lozdienė, Davide Castiglione, Hiromu Sakai and Kenny Smith for helpful discussions, as well as 3 anonymous reviewers for their precious feedback. This research was supported by UKRI Future Leaders Fellowship for the project 'Logic in Semantic Universals' (MR/V023438/1; APP75291).

References

Abrusán, M. (2014). *Weak island semantics* (Vol. 3). OUP.

Barwise, J., & Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy*, *4*(2), 159–219.

Chierchia, G. (2013). *Logic in grammar: Polarity, free choice, and intervention*. OUP Oxford.

Chomsky, N. (1957). Logical structure in language. *Journal of the American Society for Information Science*, *8*(4), 284.

Cremers, A., Fricke, L., & Onea, E. (2023). The importance of being earnest: How truth and evidence affect participants' judgments. *Glossa Psycholinguistics.*, *2*(1), 1–17.

Cremers, A., Križ, M., & Chemla, E. (2017). Probability judgments of gappy sentences. In *Linguistic and psycholinguistic approaches on implicatures and presuppositions* (pp. 111–150). Springer.

Cummins, C., Amaral, P., & Katsos, N. (2012). Experimental investigations of the typology of presupposition triggers. *Humana.Mente Journal of Philosophical Studies*, *5*(23), 1–15.

Del Pinal, G. (2019). The logicality of language: A new take on triviality,"ungrammaticality", and logical form. *Noûs*, *53*(4), 785–818.

Domaneschi, F., Canal, P., Masia, V., Vallauri, E. L., & Bambini, V. (2018). N400 and p600 modulation in presupposition accommodation: The effect of different trigger types. *Journal of Neurolinguistics*, *45*, 13–35.

Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in cognitive sciences*, *6*(2), 78–84.

Gajewski, J. R. (2002). *On analyticity in natural language* [Ms., MIT].

Göbel, A. (2022). On the role of focus-sensitivity for a typology of presupposition triggers. *Journal of Semantics*, *39*(4), 617–656.

Grice, H. P. (1975). Logic and conversation. In *Speech acts* (pp. 41–58). Brill.

Heim, I. (1991). Artikel und definitheit. In A. von Stechow & D. Wunderlich (Eds.), *Semantik: Ein internationales handbuch derzeitgenössischen forschung/semantics: An international handbook of contemporary research* (pp. 487–535). de Gruyter.

Krifka, M. (1994). The semantics and pragmatics of weak and strong polarity items in assertions. *Semantics and linguistic theory*, 195–219.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203–205.

Kutas, M., Van Petten, C. K., & Kluender, R. (2006). Psycholinguistics electrified ii (1994–2005). In *Handbook of psycholinguistics* (pp. 659–724). Elsevier.

Ladusaw, W. A. (1979). *Polarity sensitivity as inherent scope relations.* [Doctoral dissertation, The University of Texas at Austin].

Liu, Y., Winckel, E., Abeillé, A., Hemforth, B., & Gibson, E. (2022). Structural, functional, and processing perspectives on linguistic island effects. *Annual Review of Linguistics*, *8*, 495–525.

Magri, G. (2009). A theory of individual-level predicates based on blind mandatory scalar implicatures. *Natural language semantics*, *17*(3), 245–297.

Marty, P., Chemla, E., & Sprouse, J. (2020). The effect of three basic task features on the sensitivity of acceptability judgment tasks. *Glossa: a journal of general linguistics (2021-...)*, *5*(1), 72.

Matthewson, L. (2004). On the methodology of semantic fieldwork. *International journal of American linguistics*, *70*(4), 369–415.

Milsark, G. (1974). *Existential sentences in english* [Doctoral dissertation, Massachusetts Institute of Technology].

Milsark, G. (1977). Toward an explanation of certain peculiarities of the existential construction in english. *Linguistic Analysis*, *3*, 1–29.

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of memory and language*, *31*(6), 785–806.

Pylkkänen, L. (2019). The neural basis of combinatory syntax and semantics. *Science*, *366*(6461), 62–66.

Qing, C., & Uegaki, W. (2025). Semantic triviality leads to ungrammaticality through iterated learning. *Proceedings of Sinn und Bedeutung*, *29*, 1318–1335.

Schell, M., Zaccarella, E., & Friederici, A. D. (2017). Differential cortical contribution of syntax and semantics: An fmri study on two-word phrasal processing. *Cortex*, *96*, 105–120.

Schwarz, B., & Simonenko, A. (2018). Factive islands and meaning-driven unacceptability. *Natural Language Semantics*, *26*(3), 253–279.

Scrucca, L., Fraley, C., Murphy, T. B., & Raftery, A. E. (2023). *Model-based clustering, classification, and density estimation using mclust in r*. Chapman; Hall/CRC.

Shain, C., Kean, H., Casto, C., Lipkin, B., Affourtit, J., Siegelman, M., Mollica, F., & Fedorenko, E. (2024). Distributed sensitivity to syntax and semantics throughout the language network. *Journal of Cognitive Neuroscience*, *36*(7), 1427–1471.

Sprouse, J., & Almeida, D. (2012). Assessing the reliability of textbook data in syntax: Adger's core syntax. *Journal of Linguistics*, *48*(3), 609–652.

Sprouse, J., & Almeida, D. (2013). The empirical status of data in syntax: A reply to Gibson and Fedorenko. *Language and Cognitive Processes*, *28*(3), 222–228.

Theiler, N., Roelofsen, F., & Aloni, M. (2019). Picky predicates: Why believe doesn't like interrogative complements, and other puzzles. *Natural Language Semantics*, *27*(2), 95–134.

Uegaki, W., & Sudo, Y. (2019). The **hope-wh* puzzle. *Natural Language Semantics*, *27*(4), 323–356.