
THE COMPLEMENT-ADJUNCT DISTINCTION AS GRADIENT BLENDS: THE CASE OF ENGLISH PREPOSITIONAL PHRASES

A PREPRINT

Najoung Kim*, Kyle Rawlins, Paul Smolensky

Department of Cognitive Science
Johns Hopkins University
{n.kim,kgr,smolensky}@jhu.edu

August 8, 2019

ABSTRACT

We present a novel gradient blend analysis for the complement-adjunct distinction, the nondichotomous properties of which have been a longstanding problem in linguistic theory. We use English prepositional phrases as a testing ground, where the nondichotomy is especially salient. We make a typological argument that gradience (scalar activation) and blendedness (simultaneous activations of discrete structures) are crucial in explaining the conflicts among traditionally accepted diagnostic tests. Furthermore, we provide empirical support to this claim by collecting gradient complement-adjuncthood judgments and showing that modeling [V PP] constructions as gradient blends of two discrete structures (proto-complement and proto-adjunct) offers a coherent explanation to the connection between gradient complement-adjuncthood and diagnostic acceptability.

Keywords: complement, adjunct, argumenthood, gradience, acceptability, prepositional phrases, English

An adequate linguistic theory will have to recognize degrees of grammaticality.

– Chomsky (1975)

1 Introduction

Modern linguistic theories rely heavily on acceptability judgments as a source of evidence; such judgments are traditionally treated as binary. However, judgments that are not clearly dichotomous have long been recognized and discussed (Bard et al., 1996; Sorace and Keller, 2005; Bresnan, 2007; Sprouse and Almeida, 2013; Schütze and Sprouse, 2014; Lau et al., 2017; Sprouse et al., 2018, *i.a.*). Although the existence of gradience in linguistic judgments is generally undisputed¹, the magnitude of the gradient judgments is not as often actively utilized in developing and revising theories of grammar (i.e. descriptions of competence). Frequently, the source of this variability is attributed solely to performance. In this paper, we carry out a case study in which a noncategorical model of grammar offers a better explanation of observed gradience over a categorical alternative. Then we argue in favor of formalisms more expressive than current discrete symbolic systems, developing a grammatical analysis of a systematically gradient phenomenon. Our work contributes to the recent efforts in GRADIENT SYMBOLIC COMPUTATION (GSC; Smolensky et al. 2014), which has been successfully applied to model complex phenomena in phonology (Smolensky and Goldrick, 2016),

*Corresponding author

¹The prevalent use of ?, *?, and even ** and ?? for marking comparative degrees of acceptability is a clear sign.

morphology (Rosen, 2018, 2019), and sentence processing (Cho and Smolensky, 2016; Cho et al., 2017), using partially active symbolic structures. We demonstrate the extended applicability of GSC to the syntax-semantics interface, by modeling the complement-adjunct distinction² of prepositional phrases (PPs) and its connection to traditionally accepted diagnostic tests.

1.1 Structure of the paper

We first set up the background for this paper in Section 2. We motivate our work by discussing gradient judgments and how we view their connection to competence. We then provide a survey of the literature on the issue of nondichotomous complement-adjuncthood, especially focusing on discussions about PP verbal dependents. We furthermore lay out the specifics of the terminology we adopt, stating the reasons behind the choice of particular terms and their theoretical implications. In Section 3, we propose a GRADIENT BLEND model that enables a formalization of the gradient status of PP dependents, under which a PP is a weighted mixture of simultaneously active PROTO-COMPLEMENT and PROTO-ADJUNCT structures. We claim this model provides a principled explanation of conflicting diagnostic acceptability judgments of argumenthood, which cause problems for models based on discrete representations. We make a typological argument that models which do not formally capture both gradience and blendedness (which is true of both dichotomous and probabilistic models) cannot account for the full range of observed diagnostic judgment patterns. Sections 4 and 5 describe experiments in which we elicit judgments about PP verbal dependents with varying degrees of argumenthood. We use a scaled judgment collection protocol designed for nonlinguist participants, inspired by Rissman et al. (2015) and Reisinger et al. (2015). Participants are trained with clear-cut examples of complements and adjuncts, without being taught explicit definitions, and then presented with the main task that contain more difficult examples. This protocol is first verified through a pilot study which confirms that the collected judgments correlate well with those of linguists. The magnitude of the elicited judgment is hypothesized to be affected by the blend weights, or the degree of activation of proto-complement and proto-adjunct structures. In Section 6, we discuss how the results from our experiments and acceptability judgments from traditional diagnostic tests of complement-adjuncthood fit into our theory of gradient blends. We show that the predictions of our model are consistent with the patterns in the data we collected. Finally, in Section 7, we describe a computational model for calculating the blend weights and how these weights connect to the different diagnostic tests that may produce conflicting results. We treat these as optimization problems, using the gradient variant of Harmonic Grammar (Smolensky and Goldrick, 2016).

2 Background

2.1 STABLE GRADIENCE as part of competence

Gradience in linguistic judgments clearly exist, but when do they merit an explanation outside of categorical models of grammar? Schütze (2011) argues that the existence of systematic gradience in judgments itself does not necessitate gradient models of competence; for instance, a scaled design of an experiment may bias participants towards gradient responses. However, Lau et al. (2017)’s experiments suggest there exist observable differences in the distribution of scaled judgments deriving from underlyingly categorical and gradient representations (e.g. number parity versus body weight), even under the same scaled experimental setup.

Figure 1 illustrates the contrast between gradience due to experimental noise in dichotomous judgments and STABLE GRADIENCE. This figure shows hypothetical distributions over linguistic judgments collected experimentally in a normalized space, where 0 and 1 are extrema that correspond to the traditional binary judgments. Taking noise into consideration, dichotomous judgments would be analyzable as two distributions centered around modes 0 and

²We use the terms COMPLEMENT and ADJUNCT instead of ARGUMENT and MODIFIER, or any other possible combinations of these terms. Although there seems to be a preference in the literature to use COMPLEMENT/ADJUNCT in a syntactic discussion and ARGUMENT/MODIFIER in a semantic discussion, our choice of terminology is not intended to have particular theoretical implications regarding the syntactic or semantic nature of the concepts. Our stance, although not extensively discussed here, is that if we adopt a grammar that renders the syntax-semantics connection transparent, such as Combinatory Categorical Grammar (Steedman, 2000), the syntactic and semantic aspects of the complement-adjunct status could be straightforwardly linked. This largely follows the views of Dowty (2003). We use the umbrella term DEPENDENTS to refer to both complements and adjuncts, following the tradition continuing from Tesnière (1959). The word ARGUMENTHOOD is also used at times to refer to the degree of complement- and/or adjunct-ness.

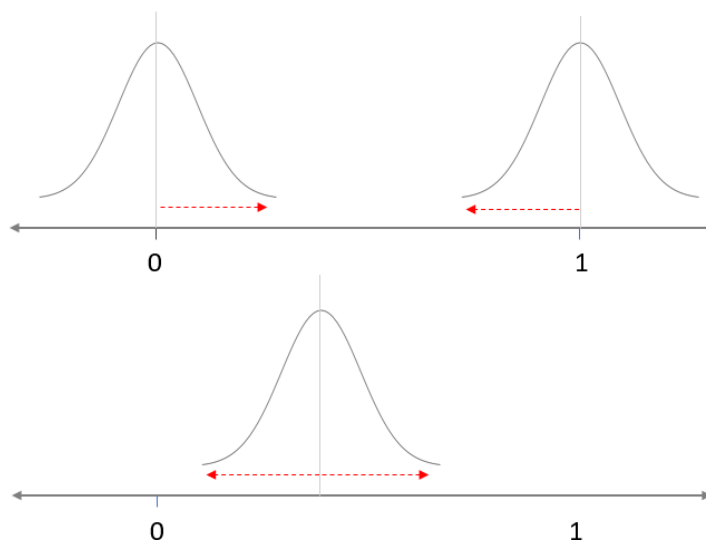


Figure 1. Gradience due to noise (top) versus stable gradience (bottom).

1 (top). However, there exists the possibility of judgments that are unimodal and centered in between (bottom), which places the distribution reliably in the gradient zone. If such a case exists, an explanation outside categorical models of competence is inevitable, whether it be from performance ‘outside the grammar proper’ (Schütze, 2011) or from gradient theories of competence.

The work we present here shows that a nondiscrete model of competence can speak to a subset of stably-gradient linguistic judgments in a principled way. We use as our testing ground stable gradience often found in argument-hood judgments about syntactically-oblique dependents such as PPs, as per many observations in the literature (Section 2.2.1). We show that stably-gradient argumenthood judgments can indeed be elicited about PP dependents (Section 4), and that a coherent explanation of their connection to diagnostic acceptability can be offered by a gradient blend model that we propose (Sections 3 and 7). This does not rule out the possibilities that gradience in PP argument-hood can be (1) reduced solely to performance factors such as processing difficulties, or (2) convincingly explained with a categorical model of competence combined with performance factors. However, our view is that these possibilities need to be concretely formulated and be supported by their demonstrable explanatory advantage over the analyses we present here. Although they are not in the scope of this paper, we would like to encourage follow-up discussions that explore alternative explanations in these directions.

2.2 Gradience in the complement-adjunct distinction

The idea that the complement-adjunct (or argument-modifier) dichotomy is problematic is by no means novel in theoretical linguistics. In Schütze (1995)’s words, ‘argumenthood is not an all-or-nothing phenomenon, but [...] it comes in degrees’. Although the contrast between complements and adjuncts seems to have psychological reality (Tutunjian and Boland, 2008) and the two concepts have been assigned a distinct status in many linguistic formalisms (Vennemann and Harlow, 1977; Chomsky, 1993; Bresnan et al., 2015, *i.a.*), researchers have struggled to pinpoint a syntactic or a semantic criterion, or even a set of criteria, for a deterministic characterization of the distinction. Some popular diagnostics employed (not as deterministic criteria but as rules of thumb) to distinguish between complements and adjuncts are well summarized in Pollard and Sag (1987) and Forker (2014); their validity and usefulness, with a focus on testing PP dependents, are discussed in Schütze (1995).

There is a substantial amount of work reporting gradience in the complement-adjunct distinction—that is, cases where it is difficult to determine the status of a verbal dependent because it patterns with both (Vater, 1978; Grimshaw, 1990; Grimshaw and Vikner, 1993; Rákosi, 2006; Toivonen, 2013, *i.a.*). Since a complement-adjunct dichotomy often does not sufficiently characterize the behavior of verbal dependents, some researchers have proposed additional subclasses. For instance, Pustejovsky (1995) suggests a four-way distinction: true argument, default argument, shadow argument,

and true adjunct, depending on their syntactic and semantic behavior. Grimshaw and Vikner (1993) introduces the notion of obligatory adjuncts to group adjuncts that play an aspectual role in the complex event structure denoted by the predicate. However, such finer-grained categorizations, or any other categorical distinctions proposed, still do not provide a satisfactory solution to the problem. None of them guarantee a deterministic partitioning of the space of verbal dependents, which is necessary for the completeness of a categorical theory.

Why would this happen, and what is the source of the unclear status of many verbal dependents? These questions are difficult to address, and often not adequately formulable, under categorical models of complement-adjuncthood. Harmonic Grammar (Legendre et al., 1990) and Stochastic OT (Boersma, 1997; Bresnan and Nikitina, 2003) attempt to remedy limitations as such by incorporating gradience and variability into their models (sometimes with different motivations), opening up much room for constructive future work. Manning (2003) specifically discusses the limitations of a binary model of complement/adjuncthood, proposing a probabilistic approach that yields better empirical predictions. This is a gradience-incorporating approach to complements and adjuncts which has more descriptive power than a dichotomous (or any categorical) model.

2.2.1 The case of English prepositional phrases

Among verbal dependents, PPs have been a popular site of investigation for gradient argumenthood because of their lexical-structural ambiguity and the multifaceted interaction between the verb, the preposition, and the complement under the preposition. The insufficiency of a simple complement-adjunct dichotomy for PPs has been discussed in many prior works, regarding instrumentals (Schütze, 1995; Donohue and Donohue, 2004; Rissman, 2010; Rissman et al., 2015), locatives (Arka, 2014; Cennamo and Lenci, 2018), *with*-PPs (Lewis, 2004), benefactives (Toivonen, 2013), among others. Examples (1a–c) from Cennamo and Lenci (2018) are a clear illustration. There are complement-like PPs such as (1a), and adjunct-like PPs such as (1c), but there are often cases like (1b) for which the distinction is murkier; the PP in (1b) seems to hold an intermediate complement-adjunct status between (1a) and (1c).

- (1) a. *John put the book [on the shelf].*
- b. *John went [to the park] after work.*
- c. *John slept [at home] last night.*

Predicting the complement/adjunct status of a PP has been discussed in works such as Aldezabal et al. (2002) and Villavicencio (2002). In particular, Merlo and Esteve Ferrer (2006) present a learning model based on the hypothesis that the distinction could be learned statistically by using semantic features, and test the hypothesis by training and testing the model on an annotated corpus. These studies provide empirical support for the idea that lexical-semantic information from the whole contextual environment affect argumenthood judgments, which has generally been assumed by theories of complements and adjuncts. Kim et al. (2019) discuss the prediction of gradient PP argumenthood, demonstrating that it can be captured by distributional and syntactic features to a non-negligible degree (Pearson’s $r = 0.624$). This suggests that gradient argumenthood displays linguistic systematicity, adding support to the view that it is a competence-related, stably-gradient phenomenon as discussed in Section 2.1.

2.3 PROTO-COMPLEMENTS and PROTO-ADJUNCTS (or CANONICAL complements and adjuncts)

We explore the questions raised in the literature on the indeterminate status of PPs by adopting a novel gradient view of complements and adjuncts, analyzing each dependent as a weighted blend of proto-complement and proto-adjunct structures. We use the terms PROTO-COMPLEMENT and PROTO-ADJUNCT to distinguish our notion from the traditional meaning of COMPLEMENT or ADJUNCT as a discrete status. The two proto-structures can be simultaneously present in a blend construction (i.e. each a subpart of the blend); each proto-structure is considered to have the same representation as typical complements and adjuncts as assigned by the linguistic formalism being used. The characteristics of each proto-construction are also understood to be similar to the canonical interpretation of complements and adjuncts. Under our model, a [V PP] structure is a blend of proto-complement and proto-adjunct structures, with different degrees of activation for each. It is the magnitude of these proto-structure activations that determines how strongly the complement-like or adjunct-like behaviors manifest. We claim that previously under-explained gradient traits are results of the different weights associated with the two subparts of the blend. This obviates the need for proposing fine-grained categories³ for handling cases that deviate from canonical complement- or adjunct-like behaviors.

³This motivation is reminiscent of Dowty (1991)’s criticism against role fragmentation.

We further characterize proto-complements and adjuncts in terms of accessibility to lexical-semantic information carried by the predicate. Assuming a lexical-semantic representation for verbs that encodes both argument structure and event structure, such as the representation scheme employed in Pustejovsky (1995) (Figure 2), we propose that only proto-complements can access, modify, or saturate argument structure; proto-adjuncts can only access and modify event structure. This rephrases the notions of lexical association and selectional restriction in terms of accessibility, and underwrites a connection between complementhood and the degree of lexical association between the predicate and the dependents (i.e. dependents that have stronger lexical association with the predicate are more complement-like than those with weaker lexical association with the predicate). By imposing different accessibility restrictions on complement and adjunct structures and analyzing each [V PP] construction as a weighted blend of the two, we aim to lay the groundwork for future work modeling the behavior of middle-ground cases such as ARGUMENT-ADJUNCTS or OBLIGATORY ADJUNCTS (Grimshaw, 1990; Grimshaw and Vikner, 1993).

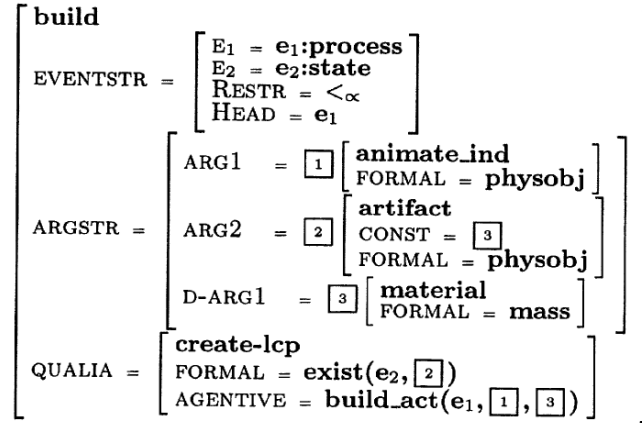


Figure 2. Example of a lexical-semantic representation of the verb *build* (from Pustejovsky 1995).

3 Model Proposal

3.1 Hypothesis: Gradient blend model

Dowty (2003) proposes a foundational *dual* approach to the analysis of PP complements and adjuncts. Dowty posits that for English learners, every PP initially receives an adjunct analysis (Figure 3, left), and then, in some cases, undergoes a complement reanalysis (Figure 3, right). Crucially, *both* analyses remain and are deployed in adulthood. One motivation for this dual analysis is that the adjunct meaning, which shares some semantic components with the complement meaning, serves as a cognitive mnemonic for the complement meaning. That is, the choice of preposition in a PP is not arbitrarily related to the complement meaning of the PP. This strategy is desirable because complement meanings typically show more idiosyncrasy (or less compositional transparency; Pollard and Sag 1987) than adjunct meanings, and thus would impose more burden on memory if the reanalysis strategy were not used and speakers had to memorize independent lexical items. Although our work focuses on [V PP] constructions, Dowty hints at the possibility of this dual analysis being applicable to all complements and adjuncts.

We take this idea as a starting point and propose that English [V PP] constructions are GRADIENT BLENDS of two structures: one a proto-adjunct structure, and one a proto-complement structure (for instance, Figure 4). Even though we use a Categorical Grammar representation here following Dowty (2003), our notion of gradient blends need not pertain to a particular formalism (i.e. given any formal grammar that assigns distinct representations to complement and adjunct structures, we can apply gradient activation levels atop). However, we believe there is an actual theoretical benefit to using Categorical Grammar for blended models and for a uniform model of syntactic and semantic argumenthood; namely the lexicalized representation of grammar and the transparent syntax-semantics interface it offers (Steedman, 2000; Steedman and Baldridge, 2011).

Let us elaborate on blendedness, the key notion advocated in this paper. A blended state is defined as simultaneous coactivation of multiple discrete structures, each with its own degree of activation (or presence). For instance, the word

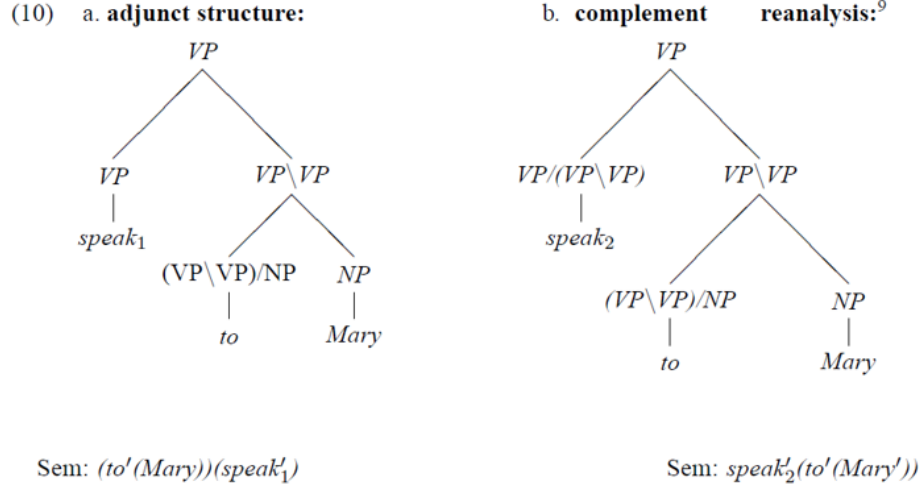


Figure 3. Complement reanalysis of *speak to Mary* (from Dowty 2003).

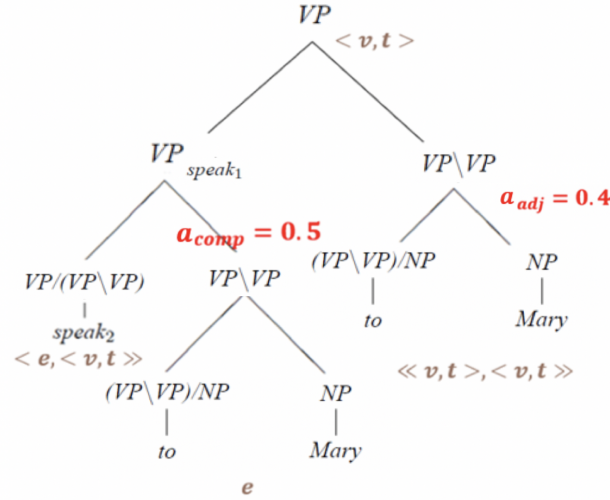


Figure 4. A possible formulation of a gradient blend analysis of *speak to Mary*, as two simultaneously active proto-structures with different activation values a_{comp} and a_{adj} (Note: it is not necessary that $a_{comp} + a_{adj} = 1$).

sequence *speak to Mary* according to our analysis, would not correspond to just the left or the right tree in Figure 3, but would be a blend (or coactivation) of the two, each present to some numerical degree (Figure 4). Parsing the [V PP] sequence under our analysis entails assigning the activation values for the two coactive structures; in unblended models, parsing entails choosing one structure over another. Note that the former is a generalization of the latter; the latter is equivalent to setting the activity of one structure to zero in the former.

Blendedness and gradience are distinct notions, though blendedness can lead to gradience in acceptability. For instance, the log-linear model proposed by Manning (2003)⁴, which assigns a conditional probability distribution over alternative symbolic structures, is a gradient model but not a blended model. The probability of a structure is a continuous measure of its strength in the probabilistic mixture, but multiple structures are never simultaneously assigned. Although the log-linear model captures gradience via probabilistic optimization over soft constraints (and it is possible that the selected structure may differ each time the constraints are evaluated due to the model's probabilistic nature), the model ultimately selects a single discrete structure. The same insight holds for Stochastic OT (Boersma, 1997). Blendedness entails that the two structures are simultaneously maintained and remain accessible, and does not enforce

⁴Equivalent to Probabilistic Harmonic Grammar (Culbertson et al., 2013).

that the weights of structures sum to one (i.e. they are not probabilities). The benefits of maintaining both sides of the blend will be argued based on the ability of blends to explain the judgment patterns that arise from the traditional diagnostic tests for argumenthood, which a discrete model or a gradience-only model cannot explain (Section 6). We once more point out that our model is not incompatible with either the discrete or the gradience-only model; it is a generalization of both, subsuming their descriptive power.

Connection to existing ideas Blendedness straightforwardly links to, and provides further theoretical grounding for the CONSTRUAL analysis advocated in recent empirical developments in preposition semantic role annotations (Hwang et al., 2017; Schneider et al., 2018). Construal analysis decomposes the semantic role of a PP into two different kinds of roles. First is FUNCTION role, which is an inherent meaning lexically encoded in a particular preposition. Second is SCENE role, the meaning contingent on the governor, which is most often the verb but can be a noun or an adjective. According to this analysis, function and scene roles could differ or overlap, but *both* roles are simultaneously assigned. Here are cases illustrating where the roles may differ, taken from Schneider et al. (2015):

- (2) a. *Put it **on/by/behind** the door.* (Inherent: LOCATION, Contextual: GOAL)
- (3) a. *Vernon works **at** Grunnings.* (Inherent: LOCATION, Contextual: Employment relation (ORGRole⁵))
- b. *Vernon works **for** Grunnings.* (Inherent: BENEFICIARY, Contextual: Employment relation (ORGRole))

The prepositions in (2) are all typically associated with the thematic role of LOCATION. However, it is also contextually correct that these PPs describe a GOAL, another traditional thematic role. This plausibility of multiple roles arises because the inherent lexical semantics of the prepositions *on/by/behind* expresses the locative meaning (i.e. the meaning that the preposition carries independent of context), whereas the goal meaning is rendered plausible by the whole expression. The goal meaning is not encoded in the prepositions themselves; it is more dependent on the argument selection of the verb and the overall situational context. The relation between (3a) and (3b) is another example. The PPs *at Grunnings* and *for Grunnings* contextually play the same role in combination with the predicate *works*, denoting employee-employer relationship between *Vernon* and *Grunnings*. However, the preposition *at* inherently encodes LOCATION, and *for*, BENEFICIARY, either of which can be used plausibly in describing the role of the PPs in (3). Under the construal analysis, these two PPs share the same scene role (employment relation) although they have different functions (LOCATION and BENEFICIARY). Of course, it need not be the case that the scene and function roles differ. Examples (4) show cases where the inherent meanings match the contextual meanings, although the prepositions used are the same as in (3):

- (4) a. *I met her **at** the park.* (Inherent: LOCATION, Contextual: LOCATION)
- b. *She did the typing **for** Thomas.* (Inherent: BENEFICIARY, Contextual: BENEFICIARY)

Although the connection is not made explicitly in the works proposing this analysis, it resembles the dual complement-adjunct analysis from Dowty (2003); the function role (inherent) corresponds to the adjunct construction and the scene role (contextual) corresponds to the complement construction. Moreover, the coassignment of both roles links to our notion of blendedness.

The distinction between inherent and contextual meanings is well-acknowledged in the literature (e.g. Fillmore 1968), but analyses that assign a nonbinary status or role to a single PP are less common. Tseng (2001) proposes an analysis that predicts a three-way distinction, one of which is an intermediate-type meaning between lexical and functional meanings. Relevant ideas are also found in the discussion of copredicating (argument sharing) PPs (Gawron, 1986), in the notion of oblique complements (Wechsler, 1995), and in the distinction between complement and pseudo-complement PPs (Verspoor, 1997), where it is assumed that the lexical content in each PP is constant but the distinction in status derives from contextual differences in which the PPs modify their verbal heads. We again emphasize that the coexistence of inherent/contextual meanings links to the notion of blendedness, and the degree of this duality (as shown by the variety of intermediate categories proposed in prior works) to the notion of gradience.

⁵The role label used in Schneider et al. (2015)'s annotation scheme for organizational membership.

3.2 Observation: failures of diagnostic tests

With the gradient blend model, we attempt to resolve the well-known *conflicts* of diagnostic tests leading to the failure in determining the complement/adjunct status of a PP dependent. We first revisit what these failures are, and then discuss how dichotomous models of complement/adjuncthood fail to offer a valid explanation.

In the literature, numerous diagnostics have been proposed as a test for argumenthood. However, because these tests do not provide necessary or sufficient conditions (singly or collectively), the results of these diagnostics are taken as tendencies rather than deterministic factors. Frequently, different diagnostic tests yield conflicting results. For example, the PP [*with acorns*] is not omissible from (5a)⁶, which is the expected result for a complement (Vater, 1978):

- (5) a. *Steve pelted Anna [with acorns].*
 b. **Steve pelted Anna.* [**Omissible (*O)*]

However, it is perfectly acceptable to pseudo-cleft the PP as in (6), which is expected for an adjunct but not a complement (Klima, 1962; Vestergaard, 1977; Takami, 1987; Needham et al., 2011):

- (6) *What Steve did [with acorns] was pelt Anna.* [Pseudo-cleftable (P)]

If a PP can only hold the status of a complement or an adjunct in a given construction, and these diagnostics deterministically differentiate between them, this conflict should not arise. However, counterexamples as above are rife. The counterexamples are found in both directions; both [**O,P*] (failing the omissibility test but passing the pseudo-cleft test) and the reverse [*O,*P*] are attested⁷. Example (5a) is a case of [**O,P*], whereas (7a) is a case of [*O,*P*]:

- (7) a. *The man strangled the victims [into a coma].*
 b. *The man strangled the victims.* [*O*]
 c. **What the man did [into a coma] was strangle the victims.* [**P*]

This conflict has long been acknowledged in theoretical linguistics (Vater, 1978; Koenig et al., 2003; Lang et al., 2003; Forker, 2014), pointing towards the potentially gradient nature of complement and adjunct status. Although a good theory of complements and adjuncts should give a principled explanation of this conflict between diagnostic tests, it remains an open question⁸. We now argue that gradience and blendedness in our proposed characterization of complements and adjuncts are together able to explain the patterns that emerge in the judgment data. Furthermore, we demonstrate that having both gradience and blendedness is crucial to the explanation.

3.3 Does gradience suffice?

We claimed that a dichotomous model of discrete complement-adjunct status cannot offer a full explanation, based on the case of two conflicting diagnostic test results on the same construction. However, we have not yet discussed the empirical necessity of representing both gradience and blendedness (for theoretical benefits, see Section 3.1). Here we discuss the descriptive power of a gradient blend model compared to a gradience-only model to show that the latter is still insufficient to capture the observed patterns.

Gradience-only analysis We first develop a GRADIENCE-ONLY hypothesis, to explore the possibility that simply recognizing gradience might lead to an account of the diagnostic conflict. This means it would be sufficient to define complement-adjunct status as a single continuum, and posit that passing each diagnostic test requires a degree of argumenthood that exceeds a test-specific threshold on this continuum. Under this analysis, conflicting results of two different diagnostic tests would be a consequence of the different thresholds targeted by each test.

⁶The judgments used throughout the paper are judgments of a nonauthor linguist, as described in Section 5.1.

⁷The judgment pattern categories are abbreviated as [**O,*P*], [**O,P*], [*O,*P*], [*O,P*], with the asterisk (*) denoting the failure of the sentence to be acceptable under the diagnostic construction. If there were no contradictions among the tests, we would only observe cases of [*O,P*] (adjunct) and [**O,*P*] (complement).

⁸One may take an extreme stance and argue that none of the diagnostic tests provide any insight into complement-adjuncthood, but we take a more moderate position that commonly used diagnostic tests are tapping into something valid. However, we also believe that a good theory must be able to explain *why* they are rules-of-thumb, rather than stopping at stating that they are.

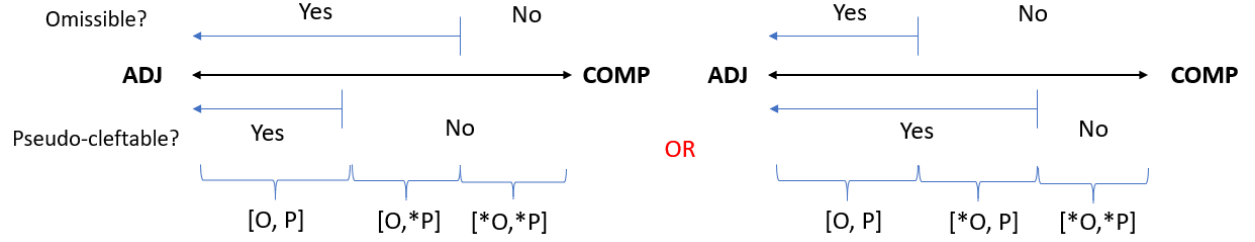


Figure 5. Possible patterns of diagnostic test results under gradience-only analysis. Passing a test corresponds to having an activation below a test-particular threshold. In either case, one out of all possible four combinations is unobservable. ([*O, P] is unobservable in the left figure and [O, *P] is unobservable in the right figure.)

We can immediately notice that this analysis is problematic because it predicts that passing a diagnostic test with a higher threshold would always entail passing another test with a lower threshold. If the P test has a higher threshold, passing P implies passing O. This means [*O, P] is predicted impossible, and the reverse holds if the O test has a higher threshold. In other words, a gradience-only model would predict that for all cases where one diagnostic fails and the other is passed, it should always be the same diagnostic that fails (i.e. if we observe a case of [O, *P], such as (7), we should never observe [*O, P] (5)–(6), because one construction would always be strictly more complement-like than another). However, we have already established that this is not the case through the coexistence of [*O, P] and [O, *P] (Examples (5)–(7)). This is not just an outlier case, since a larger set of linguist judgments we collected (Section 5.1) did contain a nontrivial number of each pattern ($n([*O, P]) = 23$, $n([O, *P]) = 94$, $n = 305$). The impossibility of observing all four patterns with two diagnostic tests with different thresholds is visualized in Figure 5. Even if we assume the coexistence of upper- and lower-bound constraints (i.e. the Yes/No are reversed for one diagnostic test in Figure 5), we cannot obtain a 4-way partitioning of the complement-adjunct scale.

Gradience-and-blendedness analysis: the proposed account Let us maintain the idea from the gradience-only account that complement and adjunct status forms a continuum. However, suppose now that the continuum is two- rather than one-dimensional; that is, we have two separate continua for complementhood and adjuncthood that can coexist. We will interpret the magnitude along the complementhood continuum as the *ACTIVITY LEVEL* of a proto-complement substructure⁹. The magnitude along the adjunct continuum is then the activity of the proto-adjunct substructure, which is simultaneously active with the proto-complement substructure. This proposal is typical of theories of linguistic representation in the GSC framework.

The unwanted prediction, of the impossibility of one out of the possible four combinations of the two diagnostic test results, is now resolvable by assuming that different diagnostics target activations of different substructures. There are diagnostics that are sensitive to proto-complement activation, which we denote a_C , requiring that a_C lie above (or below) a threshold θ_C . We will propose that the Omissibility Test is such an a_C -sensitive test, in accord with its motivation: complements are obligatory. There are also diagnostics that require proto-adjunct activation (a_A) to lie above or below a distinct threshold θ_A . We will propose that the Pseudo-cleft Test is an a_A -sensitive test. Pseudo-clefting a PP requires that its contribution to the meaning of the sentence be compositional (typical characterization of adjuncts); then this contribution can be supplied even when the PP modifies the expletive verb *do*. For example, prototypical temporal adjunct PP meanings are transparently transferrable to *do* (8), whereas in (7), the contribution of the PP *into a coma* cannot felicitously be transferred from *strangle* to *do*. See Section 6.3.1 for further discussion.

- (8) a. *The man strangled the victims [at 3pm].*
 b. *What the man did [at 3pm] was strangle the victims. [P]*

Positing a_C and a_A now gives us two independent scales to operate on, which allows for all four different types of judgment combinations (Figure 6). This fits the actual observed patterns of diagnostic test results. Note that this account does not restrict one diagnostic test to be sensitive to only one part of the blend (a_C or a_A); it could be the

⁹The activity of a substructure is its degree of presence in the overall structure (blend). Within a numerical constraint-based grammar framework, this activity level multiplies the degree of violation of constraints by the relevant substructure.

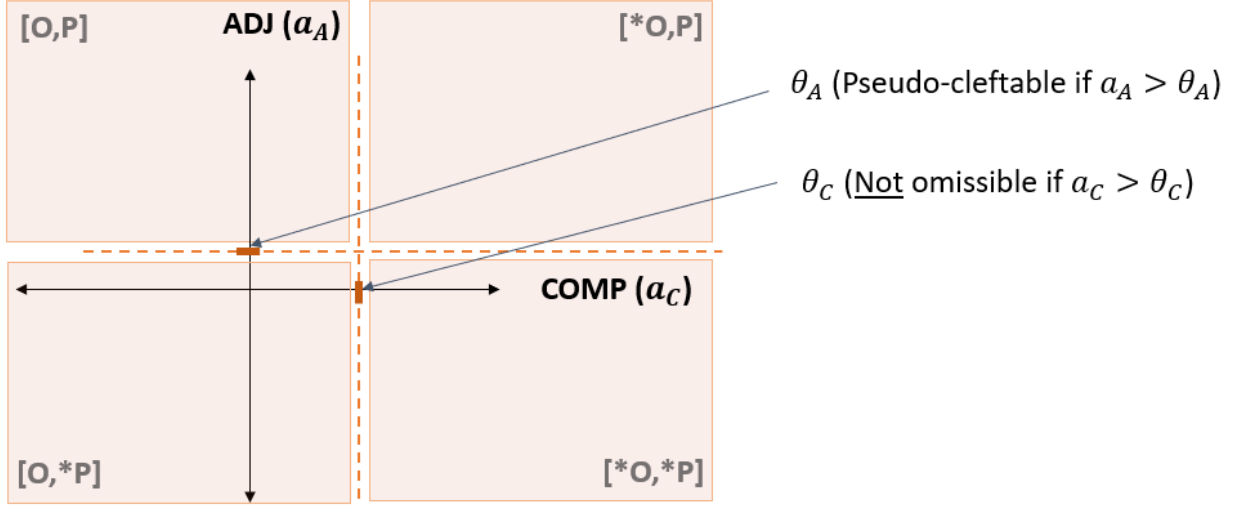


Figure 6. Possible patterns of diagnostic test results under a **gradience with blendedness** analysis. The proto-complement and proto-adjunct structures with respective activation values a_C and a_A exist independently and simultaneously.

case that one diagnostic test takes into account activations of both sides of the blend, and also to a different degree. We describe and visualize a possible partition of the judgment space in such a case in Section 7.1.1.

Blendedness-only account A final possibility is a blendedness-only account, where the complement and adjunct structures are discrete (have either zero or full activity levels), but a blended state is possible if both are active. This again only yields three possible states that the judgment patterns can be mapped to (complement, adjunct, or complement + adjunct).

Now we present two experiments in which we collect gradient argumenthood judgments, and gradient acceptability judgments of diagnostic sentences, respectively, in order to test the capability of our theory to explain the data.

4 Experiment 1: Gradient complement-adjunct judgments

4.1 Pilot: Validating complement-adjunct judgments

We are primarily interested in *degrees* of argumenthood manifested by a PP verbal dependent. Therefore, if we were to conduct an experiment to elicit related judgments, we do not want to (1) provide the participants with a dichotomous definition of complements and adjuncts as traditionally defined, and (2) instruct the participants to classify complements and adjuncts according to typically accepted diagnostic criteria. With these considerations in mind, we design our experimental protocol based upon Rissman et al. (2015)’s work on gradient representations of instruments. We also refer to Reisinger et al. (2015) for adaptation to scaled judgments. In order to validate our adaptation and confirm that our protocol is able to tap into relevant judgments, we first conduct a pilot study. The hypothesis we test in the pilot is that if our protocol is indeed tapping into some valid notion shared across native speakers, we would observe high similarity between the judgments produced by nonlinguist participants (with less exposure to theoretical priors) under our experimental setup and judgments produced by linguists on complement-adjuncthood.

To keep the instructions compact and easily comprehensible for nonlinguists, as well as avoiding direct introduction of the key concepts, we use the proxy terminology *centrality* (in contrast to *peripherality*). The centrality/peripherality of a dependent with respect to the event denoted by the predicate is expected to reflect the degree of complement/adjuncthood. Similar experimental procedures using slightly different wordings have been employed in several prior studies. For instance, the expressions *need*, *necessity* were used in Barbu and Toivonen (2015, 2016) and *important* was used in Rissman et al. (2015). Although the exact goals of these studies differ from ours, Barbu and

Toivonen (2015, 2016) report that the concept of *necessity* does seem to tap into relevant judgments (i.e. core status of dependents). Although the proxy expressions may differ, we claim that adequate training phase before the main experiments is more important than the exact wording of the questions, based on our success in replicating linguist judgments.

4.1.1 Design

Stimuli We select pairs of English sentences that differ by either the verb or the PP dependent of the verb. We construct the pairs so that the complement/adjunct contrast between the two sentences are fairly uncontroversial to linguists. The goal of this clear-contrast design is to verify whether nonlinguist participants would be able to produce sufficiently similar judgments to those of linguists, given minimal instruction using a proxy notion. Note that the linguist judgments are collected by explicitly instructing them to judge which examples are more complement-like.

PP-contrast (V-controlled) sentences only differ in their PP dependent of the main verb. The PPs are controlled by the number of words, syntactic structure, and co-occurrence frequency of the verb and the NP complement of the P head¹⁰. Here are examples of PP-contrast pairs (a contrasting pair may (9) or may not (10) have the same preposition):

- (9) a. *Paul hit his elbow [on the table].*
 b. *Paul hit his elbow [on his birthday].*
- (10) a. *I walked [to a park].*
 b. *I walked [with a friend].*

V-contrast (PP-controlled) sentences are selected from the PP-contrast set, where two sentences share the same PP but have different verbs. Therefore, V-contrast sentences are not controlled by the number of words, syntactic structure, or co-occurrence frequency. (11) is an example of a V-contrast pair:

- (11) a. *I put the eggs [on the table].*
 b. *Paul hit his elbow [on the table].*

For either contrast, the PPs in (a) sentences are intended to be more complement-like than in (b) sentences, which are more adjunct-like.

All sentences are generated based on examples from either VerbNet (Kipper-Schuler, 2005) or PropBank (Palmer et al., 2005), with some simple modifications (truncation or NP substitution) to satisfy the control conditions. Each sentence is marked as either complement-certain (CC), complement-like (CL) or adjunct-like (AL). The PP in a CC or CL sentence is either an element labeled ARG-*n* in PropBank or an element taken from example sentences of a subcategorization frame in VerbNet, and then marked CC or CL according to linguist intuition. AL sentences are manually generated by replacing the PPs of CC and CL sentences with more adjunct-like ones (again based on intuition, but PPs from other CC/CL sentences are reused wherever possible). Examples of each sentence type is given below:

- (12) Complement-Certain (CC): *I put the eggs [on the table].*
 Complement-Like (CL): *I admired him [for his honesty].*
 Adjunct-Like (AL): *We offered the paycheck [on Saturday].*¹¹

The design goal is that PPs in CC sentences should be more complement-like than those in CL sentences, and the PPs in CL sentences should be more complement-like than those in AL sentences. This gradient argumenthood should be transitive, so we would also expect PPs in CC sentences to be more complement-like than PPs in AL sentences. At least two sentences of different types are generated for 40 different verbs. The expected contrasts between the sentences are reviewed and confirmed by a nonauthor syntactician¹². Although some of the contrasts were judged to

¹⁰Brown Corpus is used for controlling co-occurrence frequency.

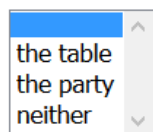
¹¹Modified from a CC sentence, *We offered the paycheck [to Amanda].*

¹²One author with a Master's degree in Linguistics and five years of graduate-level experience constructed the paired examples and marked the example that was more complement-like in a given pair. A nonauthor syntactician with a Master's degree in Linguistics and six years of graduate-level experience confirmed these judgments.

Which highlighted expression is more central to the event of **cleaning**?

Doug **cleaned** the dishes from **the table**.

Doug **cleaned** the dishes before **the party**.



is more central to the event of **cleaning** (the unchosen option is more peripheral).

Figure 7. An example of a PP-contrast question.

be not as strong as originally intended, most expected contrasts were confirmed to be present¹³ and every direction of the contrast was agreed upon.

Out of the 40 verbs, 12 verbs had all three contrast types. The remaining 28 verbs had two-way contrasts, with one verb having two different two-way contrasts (CC>CL and CC>AL). For the verbs with a three-way contrast, a binary contrast pair is created for all possible combinations (CC>CL, CC>AL, CL>AL). This gives us 66 binary contrasts (PP-contrast) for 40 verbs. An additional set of V-contrast pairs was constructed reusing sentences with shared PPs (and different verbs) as discussed previously. There were 11 such cases, yielding 77 binary contrasts in total for our pilot stimuli.

Study We ask one ternary judgment question for each of the 77 binary contrasts. We used the terms *centrality* and *peripherality* of the PP dependent with respect to the predicate, instead of providing the participants with theoretical or diagnostic definitions of complements and adjuncts.

For PP-contrast pairs, the task instruction given to the participant is to choose which NP under PP is ‘more *central* to the event of *x*’, and the participants are additionally told that the unchosen option is more peripheral. For V-contrast pairs, they are instructed to choose with which V the NP sounded ‘more *central*’. The questions are ternary rather than binary because the option ‘neither’ is available. See Figures 7 and 8 for examples of these questions.

4.1.2 Data collection

15 participants were recruited on Amazon Mechanical Turk (MTurk). The task was made available only to participants located in the United States. All participants except one self-identified as native speakers of US English, meaning: (1) they grew up speaking English in the United States, and (2) with their parents speaking English to them as children. One participant did not pass the nativity criteria (this participant answered one out of the two questions with ‘No’), and this participant’s answers were excluded from the analysis, leaving us with data from 14 participants. Nevertheless, this participant was compensated. Every participant answered all 77 questions, spending an average of 38 minutes (as reported by MTurk; so this does not distinguish active participation time from nonactive time). They were paid \$2.00 in compensation. Participants were first presented with practice questions, for which they are given feedback to familiarize themselves with the task. Questions in the main phase were presented in random order; the order of the two sentences within questions, and the response options, were also shuffled, although the option ‘neither’ was always presented last. Figure 7 shows an example of a PP-contrast question and Figure 8 shows an example of a V-contrast question.

4.1.3 Results

To confirm the hypothesis that our centrality/peripherality questions indeed target argumenthood judgments similar to those of linguists, we calculate the average accuracy of each question across all participants, taking the option chosen

¹³We discarded examples where the linguists disagreed—there were only a few (< 5) such cases.

With which **event** does **the table** sound more central?

Brenda **fought** on **the table**.

Paul **hit** his elbow on **the table**.

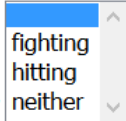
the table sounds more central with the event of  (the unchosen option is more peripheral).

Figure 8. An example of a V-contrast question.

by linguists as more complement-like as the accurate answer for the question ‘which option is more central?’. Here is a step-by-step rephrasing of our rationale for this experiment. (1) Our stimuli consist of paired sentences containing PPs that display noticeable contrast in argumenthood that was agreed upon by linguists. (2) If a large number of the participants’ answers agree with the linguists’ even without teaching them the definition of complements and adjuncts, this suggests that our centrality questions are eliciting judgments akin to those of linguists. (3) The degree of agreement with linguist judgments is measured by average accuracy on each question, taking linguists’ answer as ‘accurate’.

Table 1 shows the mean accuracy and the standard deviation for all questions, for each contrast category, and for each controlled element. The average participant accuracy across all questions was 78.5%. A binomial test indicates that accuracy of 64.3% (9/14) or higher is significantly over chance ($p = 0.02$). 65 out of the 77 questions had this accuracy or higher, meaning 84% of all questions had statistically significant above-chance accuracy. This high accuracy for most questions suggests that it is possible to elicit relevant complement-adjuncthood judgments via our protocol.

	all	CC>CL	CL>AL	CC>AL	V-contrast	PP-contrast
μ	0.785	0.703	0.796	0.821	0.805	0.781
σ	0.167	0.204	0.159	0.142	0.101	0.176

Table 1. Mean and standard deviation of accuracy by contrast category and contrasted element.

Table 1 also shows that the mean accuracy of CC>AL contrast is higher than both CC>CL and CL>AL contrasts. This aligns with our intended result, since the contrast between CC and AL was designed to be greater in magnitude than the other two contrasts (because CC>CL>AL). To confirm this effect of contrast type on accuracy (i.e. participants’ sensitivity to the intended contrast), we conduct a linear mixed-effects model analysis with contrast type as fixed effect, and control token and target tokens (the verb and the PP) as random effects. With CC>AL contrast as the reference group, the participants are less accurate at answering both CC>CL ($B = -.15, SE = .04$) and CL>AL ($B = -.07, SE = .0008$) type questions. A post-hoc Tukey’s test with Holm correction reveals that both differences are significant (both $p < .001$), and furthermore that CL>AL type questions are significantly more accurate than CC>CL type questions ($B = .08, SE = .04, p < .05$). The latter is also an intuitive result, considering that the distinction within complement-like items would be less clear than the distinction between complement-like and adjunct-like items.

Additionally, the infrequent use of the ‘neither’ option points towards the existence of the targeted contrast. The ‘neither’ option was only ever selected in 24% of all questions (i.e. questions where at least one participant selected ‘neither’). Moreover, this option was only used 34 times in the set of all answers, which comprises only 3% of the set.

From the above results, we can conclude that our protocol does elicit judgments about gradient complement and adjunct status of PP dependents that are qualitatively similar to those of linguists.

4.2 Main experiment: Scaled judgments

From the pilot experiment, we have established that our task phrased in terms of centrality and peripherality elicits judgments consistent with argumenthood judgments produced by linguists. In the main experiment, we ask the participants a scaled-judgment version of the centrality questions. To ensure that this change in format (from multiple-choice to scaled) does not affect the validity of the protocol, we include all of the sentences used in the pilot in the main experiment. The result of the pilot study is replicated in the scaled version of the experiment; we obtain high accuracy for the contrast questions in the pilot if we select the answers based on the scaled scores of each sentence in the contrasted pair (see Section 4.2.3).

4.2.1 Design

Stimuli Our stimuli consist of 305 sentences containing at least one PP¹⁴. There are 120 unique verbs, each of which is the main predicate of more than 1 sentence in the dataset. The sentences sharing the same predicate differ only by their PPs, similar to the PP-contrast stimulus-pairs in the pilot. The PPs were again controlled for the number of words, syntactic construction, and the co-occurrence frequency of the verb and the head noun, as in the pilot. Since there was no significant difference in mean accuracy between PP-contrast and V-contrast questions in the pilot ($t(75) = .44, SE = 5.47, p = .66$), we decided to drop the V-contrast. We note that even though many of the V-contrast examples did not satisfy the strict control conditions used for PP-contrast questions, the pilot accuracy was not significantly affected. Based on this observation, the control conditions are less strictly imposed in the main experiment although they are still used. For instance, if an addition of a word or changing the determiner makes the sentence more natural, we choose to make these modifications rather than adhering strictly to the control conditions.

The sentences are mainly taken from example sentences in VerbNet subcategorization frames, with simple modifications to match the control conditions. Since subcategorization frame examples are expected to be relatively more complement-like, we manually augment the dataset with more adjunct-like examples by replacing the PPs of the collected examples. As previously stated, every sentence from the pilot dataset is also included in this larger dataset to ensure the effects we saw in the pilot are replicated. The examples are intended to be diverse in their degrees of argumenthood; here are several sentences from the dataset:

He withdrew [from the trip].
They participated [as a good gesture].
Amanda shuttled the children [from home].
Nora pushed her [with the biggest smile].
Bill repaired the tractor [for a road trip].
I whipped the sugar [with cream].
It clamped [on his ankle].
The witch turned him [into a frog].
The children hid [in a hurry].

Study The participants are shown a single sentence per question, and asked to select a point on a 7-point Likert scale, according to how central they believe the highlighted NP under PP is with respect to the event denoted by the main predicate. The scale is accompanied by a help phrase ‘1 is most **peripheral** and 7 is most **central**’ (see Figure 9).

The instructions closely follow the pilot. The practice task in the training phase uses the same sentences as in the pilot, except that the participants have to give scaled answers. For the purpose of training, participants are asked to judge the centrality of two items at once, since the concept seemed easier to grasp when shown an actual contrast between items that differed significantly in their argumenthood. The participants are informed that they will be judging centrality for two different items at once only in the training phase. Figure 10 shows an example of a practice question.

4.2.2 Data collection

In the pilot, a participant spent on average 38 minutes to answer all 77 questions. Considering that our new dataset contained more than 300 sentences, we conducted the experiments in subsets of around 50 sentences to reduce the load

¹⁴Data available at: *anonymized for submission*

How central is *pliers* to the event of *bending*?

Tony *bent* the rod with *pliers*.

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7

(1 is most **peripheral** and 7 is most **central**)

[Continue](#)

Figure 9. An example question in the scaled experiment.

How central are *John* and *Sunday* to the event of *donating*?

Sam *donated* the book to *John*.

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7

Sam *donated* the book on *Sunday*.

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7

(1 is most **peripheral** and 7 is most **central**)

[Continue](#)

Figure 10. A practice question in the training phase of the scaled experiment.

and keep the participants attentive. The dataset was randomly split into 6 subsets (5 subsets of 50 sentences and 1 subset of 55 sentences), and 25 participants were recruited for each subset. The tasks were released on Amazon Mechanical Turk over 3 days, and participants were permitted to answer as many subsets as they wanted. The participants on average spent around 30 minutes on one subset (again, not distinguishing active and inactive times) and were paid \$1.50 in compensation.

Participants were restricted to those located in the United States, and the same nativity questions as in the pilot were asked. Three subsets of responses were excluded from the final results based on these nativity criteria, for which the participants were compensated nonetheless. Every participant answered every question in the given subset, and the questions were presented in random order. Two participants had technical difficulties that halted the experiment early and had to restart the task from the beginning. The retake questions were presented in random order.

4.2.3 Results

In order to account for individual variance in the use of the 7-point scale, we normalized the scores by calculating the within-subject z -score using the mean and standard deviation of each individual participant. As illustrated in Figure 11, cases of stably-gradient judgments (c.f. Section 2.1) were indeed observed, where the score distribution was centered around the midpoint of the scale with a unimodal peak.

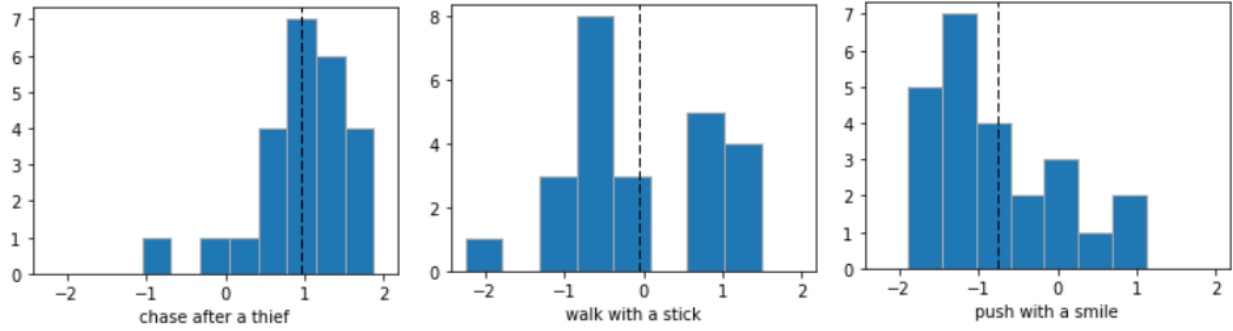


Figure 11. Distribution of scores given by individual participants ($n = 25$) for a complement-like example (left), an intermediate example (middle), and an adjunct-like example (right).

The mean z -score for each sentence is taken to be the final COMPLEMENTHOOD/CENTRALITY SCORE (C-SCORE) of the PP dependent with respect to the main predicate, which serves as our estimation of gradient argumenthood. The scores range between $[-1.435, 1.172]$, approximately centered around zero ($\mu = 1.967e-11, \sigma = .526$). As shown in Figure 12, the distribution was slightly left-skewed (right-leaning), with more positive C-score values than negative values ($z = -2.877, p < .01$).

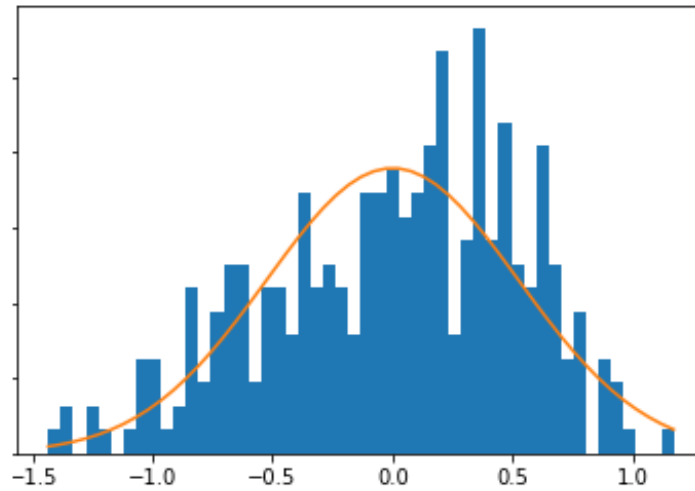


Figure 12. Distribution of C-score values in the full dataset.

The scaled scores successfully replicate the pilot results. That is, if we picked the answer of the binary contrast questions (‘which option is more central?’) based on the C-scores collected in this experiment (recall that all pilot sentences were included in the scaled dataset), we would obtain an accuracy of 88.3%, which is in fact substantially better than the aggregate accuracy reported in Table 1 (78.5%). This suggests that the scaled version of the task is actually less noisy than the binary contrast version.

Here is the same set of sentences we presented in the previous section, this time with their respective C-scores:

It clamped [on his ankle]. (0.66)
The witch turned him [into a frog]. (0.57)
He withdrew [from the trip]. (0.39)
I whipped the sugar [with cream]. (0.35)
They participated [as a good gesture]. (-0.01)
Amanda shuttled the children [from home]. (-0.10)
The children hid [in a hurry]. (-0.41)
Bill repaired the tractor [for a road trip]. (-0.71)
Nora pushed her [with the biggest smile]. (-0.76)

We refrain from assigning definitive interpretations to the absolute values of the C-scores, but how the values compare to each other should be informative of their relative difference in complement-adjuncthood (lower is more adjunct-like, higher is more complement-like). For instance, *from home* in *shuttled from home* with a score of -0.10 should be more adjunct-like than a higher-scoring construction such as *on his ankle* in *clamped on his ankle* (0.66), but is more complement-like than *hid in a hurry* with a score of -0.41 . This matches the intuition that, for a change-of-location predicate like *shuttle*, a locative PP such as *from home* would be more complement-like than a manner PP like *in a hurry* for *hide*. Nevertheless, it is still less complement-like than a more clearly complement-like PP *on his ankle* that is selected by *clamp*.

4.3 Discussion: Effect of thematic roles

It has been suggested that phrases bearing certain thematic roles manifest both complement- and adjunct-like behaviors (Toivonen, 2012); for example, benefactives (Toivonen, 2013) and instrumentals (Donohue and Donohue, 2004; Rissman et al., 2015). In particular, Donohue and Donohue (2004) present evidence from six Pacific languages that instrumentals which are ‘integral’ to the event display more ‘term-like’ (i.e. complement-like) properties than those that are not integral, which are more traditional-adjunct-like. This results in more integral instrumentals not necessarily falling in line with the standard thematic hierarchy (e.g. the one proposed in Bresnan and Kanerva 1989). Our scaled complement-adjuncthood judgment data adds empirical support to this claim. Table 2 shows that PPs bearing the same thematic role have a widely varying range of C-scores, meaning some may behave complement-like and others more adjunct-like. Table 2 lists the average C-scores of each thematic role (mostly as annotated in VerbNet) in descending order, and we can see a tendency for traditional-adjunct-like roles such as LOCATION, MANNER and TIME occupying the lower end of the scale. But we again emphasize that the variability indicated by the range of each role is very large even for these adjunct-like roles, suggesting that roles do not prohibit dependents from behaving more complement-like. We also note that there is an interesting overlap in the order of thematic roles by descending C-score with the thematic hierarchy proposed in Baker (1989); Carrier-Duncan (1985); Larson (1988); i.e. AGENT > THEME > GOAL/BENEFACTIVE/LOCATION.

	Theme	Source	Topic	Instrument	Recipient	Trajectory	Result
μ	0.426	0.393	0.341	0.276	0.225	0.190	0.181
σ	0.360	0.266	0.238	0.308	0.246	0.472	0.323
range	1.561	0.804	0.752	1.393	0.755	1.534	0.972

	Co-agent	Goal	Beneficiary	Location	Initial location	Manner	Time
μ	0.131	0.008	-0.133	-0.228	-0.333	-0.437	-0.565
σ	0.401	0.426	0.425	0.585	0.446	0.389	0.441
range	1.296	1.323	1.648	2.302	1.231	1.427	1.820

Table 2. Average C-scores by thematic role.

Further analysis of the results is presented in Section 6 jointly with the results from the second experiment.

5 Experiment 2: Diagnostic sentence acceptability judgments

Although diagnostic tests are almost always mentioned in the discussion of complements and adjuncts, no one diagnostic or set of diagnostics prove to be necessary or sufficient in determining the status of a dependent. In order to demonstrate that the gradient blend model of PPs can provide a coherent explanation of this complexity, we collect acceptability judgments for two traditional diagnostics of complement/adjuncthood. As will be discussed in the results section, a joint analysis of the judgment patterns and C-scores adds crucial empirical support for our gradient blend model. The particular choice of diagnostic tests, namely PSEUDO-CLEFTING and OMISSIBILITY, is based on the universal applicability of these tests. For instance, the ITERATIVITY test is commonly used (e.g. temporal adjuncts are iterable: *I ran [for two hours] [on a Sunday] [in March]*), but is not uniformly applicable to all constructions and requires much more creativity than the diagnostics we selected. This could lead to difficulty in quality control for the stimuli and may introduce more experimental noise. We also note that the two selected diagnostic tests are both accepted as plausible tests by Schütze (1995). Refer back to Section 3.2 for more discussion.

5.1 Linguist judgments

We recruited a trained linguist (a nonauthor syntactician with six years of graduate-level experience) to annotate all 305 sentences with whether they passed (1) the pseudo-clefting diagnostic, and (2) the omissibility diagnostic. A set of instructions was provided to her about what ‘passing’ and ‘failing’ each diagnostic test meant. We describe each test in more detail:

Pseudo-clefting To judge whether a sentence passes the pseudo-clefting diagnostic, we first transform the given sentence to the form *What X did [PP] was [...]*, where *X* is the grammatical subject and [...] is the remainder of the sentence without the extracted PP, with the predicate in its infinitive form. The intuition behind this test is that, in order for the pseudo-clefted phrase to be felicitous as a dependent of the verb *do*, it must be an adjunct rather than a complement. Here is an example of a pseudo-clefted construction:

- (13) Original: *I explained [for the hundredth time] how to do it.*
Pseudo-clefted: *What I did [for the hundredth time] was explain how to do it.*

If the pseudo-clefted version is acceptable, the original sentence passes the diagnostic. If the result is ungrammatical (7c) or the meaning of the PP is significantly altered by the extraction, it fails the diagnostic test. For instance, the linguist who produced the diagnostic judgments commented that for (14), the pseudo-clefted version is only interpretable as *Tamara* being inside *the bowl* and *pouring water*. This is inconsistent with the salient meaning of *from the bowl* in the original sentence. For such cases, the alternation was marked as unacceptable.

- (14) Original: *Tamara poured water [from the bowl].*
Pseudo-clefted: **What Tamara did [from the bowl] was pour water.*

Omissibility In the omissibility test, we remove the target PP from the original sentence and ask whether the remainder of the sentence sounds acceptable without substantially altering the meaning of the original predicate.

- (15) Original: *John conspired [with the plumber].*
Omitted PP: **John conspired.*

For both diagnostic tests, the linguist was allowed to use question marks to indicate fuzzier judgments. The results then formed a basis of comparison for this experiment as well as several points noted previously in the paper.

5.2 Scaled nonlinguist judgments

5.2.1 Design

Stimuli We use the same set of 305 sentences introduced as materials in our first experiment to generate test sentences corresponding to the two diagnostic tests described in the previous section. Here is an example:

- (16) Original: *Steve tossed the ball [for fun].*
 Omissibility-test sentence: *Steve tossed the ball.*
 Pseudo-cleftability-test sentence: *What Steve did [for fun] was toss the ball.*

Study Participants are asked to provide a scaled judgment on a 7-point Likert scale on whether the two transformed versions of a given sentence are natural. To familiarize nonlinguist participants with the acceptability task, we first present them with a set of practice questions. The sentences in the practice set are constructed such that the judgments are relatively clear, and the participants are given feedback on their performance. The practice sentences did not consist solely of the PP constructions we were interested in, in order to reduce potential bias. See Figure 13 for an example practice task.

Original sentence: Anna looked like her mother.

How natural does the following sentence sound?

What Anna did like her mother was look.

○ ○ ○ ○ ○ ○ ○
 1 2 3 4 5 6 7

(1 is most **unnatural** and 7 is most **natural**)

Original sentence: Anna danced with Elsa in the castle.

How natural does the following sentence sound?

What Anna did in the castle was dance with Elsa.

○ ○ ○ ○ ○ ○ ○
 1 2 3 4 5 6 7

(1 is most **unnatural** and 7 is most **natural**)

Original sentence: Anna looked outside through the window.

How natural does the following sentence sound?

What Anna did outside was look through the window.

○ ○ ○ ○ ○ ○ ○
 1 2 3 4 5 6 7

Figure 13. An example of a practice task for Experiment 2.

The test sentences, both in the practice and main tasks, are presented simultaneously with the original sentence, in order to create an environment similar to the process linguists would go through when applying a diagnostic test to a sentence. Figure 14 shows an example presented to the participants, asking for diagnostic judgments for the sentence *Steve tossed the ball for fun*. To ensure the semantic compatibility of the transformed sentence, the participants are asked an additional question (Figure 15) if they select values between 4 – 7.

5.2.2 Data collection

We re-recruited self-reported native US English speakers who participated in the scaled complement-adjuncthood judgment task (Section 4). We used the same six-way partitioning of the full dataset and recruited participants from matching subsets, meaning that all returning participants saw the same set of sentences that they had seen in the first

Original sentence: Steve tossed the ball for fun.

How natural does the following sentence sound?

What Steve did for fun was toss the ball.

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7

(1 is most **unnatural** and 7 is most **natural**)

Original sentence: Steve tossed the ball for fun.

How natural does the following sentence sound?

Steve tossed the ball.

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7

(1 is most **unnatural** and 7 is most **natural**)

Figure 14. An example of a diagnostic test task for the sentence *Steve tossed the ball for fun*.

Original sentence: Steve tossed the ball to the garden.

How natural does the following sentence sound?

Steve tossed the ball.

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☒ 7

(1 is most **unnatural** and 7 is most **natural**)

How compatible is the above sentence with the original sentence? (Were there any significant meaning changes, ignoring what is missing?)

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7

(1 is most **incompatible** and 7 is most **compatible**)

Figure 15. An example of a compatibility question for the omissibility diagnostic sentence *Steve tossed the ball*, with respect to *Steve tossed the ball to the garden*.

experiment. Two participants were recruited per subset. The order of the sentences in a subset and the order of the diagnostic test questions were permuted randomly for every participant.

5.2.3 Results

The scaled judgments were converted into normalized scores following the z -normalization process from Experiment 1 for each participant, to account for individual variability in use of the scale. We were interested in whether and how these judgment scores relate to the C-scores of each sentence, collected in Experiment 1. A multiple regression analysis with both diagnostic z -scores as predictors reveals that the diagnostic scores are linearly associated with C-scores ($R^2 = .024, p < .001$), but only pseudo-cleftability is a significant individual predictor ($B = -.09, SE = 0.02, p < .001$). If we also take compatibility into consideration by taking the average of compatibility and naturalness whenever the compatibility score is available (recall that the question is not displayed if the sentence is judged to be unnatural), we again observe a significant linear association ($R^2 = .037, p < 10^{-5}$), but this time with both diagnostic scores as significant predictors ($B = -.08, SE = .04, p < .05$ for O and $B = -.09, SE = .02, p < 10^{-4}$ for P). All effect directions are negative, which aligns with our expectation; lower diagnostic acceptability associates with higher C-scores and higher acceptability associates with lower C-scores. In other words, if a PP is less omissible or less pseudo-cleftable, it is more complement-like, and if a PP is more omissible or pseudo-cleftable, it is more adjunct-like. This matches the utility of these diagnostic tests as traditionally understood, but also matches the common observation in the literature that they are not decisive indicators (as suggested by the low variance explained). We provide a more in-depth analysis of these results in the following section, together with the predictions from our proposed model.

6 Analysis

6.1 Recap: failure of diagnostic tests

As described in Section 3.2, two different diagnostic tests may yield conflicting results, which renders the complement/adjunct status of a dependent indeterminable. For expository purposes, we repeat the previous examples of conflicting cases in (5)-(7), now with the z -scored acceptability values averaged across participants:

- (17) *Steve pelted Anna with acorns.*
 **Steve pelted Anna.* [*O] (mean acceptability: -0.45)¹⁵
What Steve did [with acorns] was pelt Anna. [P] (mean acceptability: 0.73)

And a case of [O,*P]:

- (18) *The man strangled the victims into a coma.*
The man strangled the victims. [O] (mean acceptability: 0.18)
 **What the man did [into a coma] was strangle the victims.* [*P] (mean acceptability: -1.87)

6.2 Recap: a typological argument for a gradient blend model

In Section 3.3, we have argued in favor of a model that incorporates both gradience (scaled activation) and blendedness (simultaneous activation of multiple structures), based on its ability to account for all four patterns of acceptability judgments that arise from two different diagnostic tests. We encourage referring back to the contrast between predictions of a gradience-only model and a gradient blend model (Figures 5 and 6, respectively) before continuing to the next section.

6.3 Predictions

With the gradient blend model, we now have adequate formal tools to hypothesize about the underlying structures that give rise to the empirical observations. If we assume that *OMISSIBILITY results from high a_C (i.e. if something is

¹⁵We provide the acceptability scores averaged across participants to illustrate how they compare to binary linguist judgments (acceptable or unacceptable). Lower scores indicate lower Likert scale numbers (i.e. constructions that were judged to be *unnatural* or *incompatible*) in our experiments.

not omissible, it is because it has high proto-complement activation) and PSEUDO-CLEFTING results from high a_A (i.e. if something can be pseudo-clefted, it is because it has high proto-adjunct activation), we can coherently explain the observed patterns. These assumptions are summarized below:

high $a_C \implies$ *OMISSIBILITY
 low $a_C \implies$ OMISSIBILITY

low $a_A \implies$ *PSEUDO-CLEFTING
 high $a_A \implies$ PSEUDO-CLEFTING

(High-ness and low-ness determined by potentially different thresholds for a_C and a_A .)

These assumptions now generate specific predictions about the correlation between C-scores (gradient argumenthood scores from Experiment 1) and binary diagnostic judgments, under the additional assumption that a C-score will be proportional to the linear combination of the activations of the two proto-structures ($C\text{-score} \propto m \cdot a_C + n \cdot a_A$; $m > 0, n < 0$). Then, roughly (when m and n are of comparable magnitude):

high a_C & low $a_A \implies$ [*OMISSIBILITY, *PSEUDO-CLEFTING] \implies high C-score
 low a_C & low $a_A \implies$ [OMISSIBILITY, *PSEUDO-CLEFTING] \implies intermediate C-score
 high a_C & high $a_A \implies$ [*OMISSIBILITY, PSEUDO-CLEFTING] \implies intermediate C-score
 low a_C & high $a_A \implies$ [OMISSIBILITY, PSEUDO-CLEFTING] \implies low C-score

6.3.1 Comparing predictions against binary judgments

It should be noted that a C-score is unlikely to be a direct realization of $m \cdot a_C + n \cdot a_A$ because it involves a conscious evaluation of the sentence after the string of words has already been parsed. At the point of assigning a C-score, although the activation values a_C and a_A are posited to be influential factors, we cannot determine exactly what other information the speakers additionally incorporate into the judgments. It is likely that the same information influencing diagnostic test judgments (e.g. lexical semantic content) would also be used in determining the argumenthood judgments, but this information may be used in different ways. Moreover, there will be performance factors and additional experimental noise that may yield deviations in the C-score away from the values predicted by a_C and a_A only. Therefore we do not expect diagnostic results to perfectly explain the variance in the C-score data. Nevertheless, we present some examples that do bear out the predictions (the binary judgments used in this section are all linguist judgments described in 5.1):

1. [*OMISSIBILITY, *PSEUDO-CLEFTING] \implies high C-score
 - *Jackie chased [after the thief].* (0.95)
 - *It was pelting [with rain].* (0.78)
 - *The witch turned him [into a frog].* (0.57)
2. [OMISSIBLE, *PSEUDO-CLEFTING] \implies intermediate C-score
 - *John collaborated with Paul [in the task].* (0.21)
 - *I learned [about the accident].* (0.11)
 - *The man strangled the victims [into a coma].* (−0.06)
3. [*OMISSIBILITY, PSEUDO-CLEFTING] \implies intermediate C-score
 - *Allison poked the needle [through the cloth].* (0.32)
 - *Cornelia lodged [with her family].* (0.10)
 - *Linda taped the picture [to the wall].* (0.03)
4. [OMISSIBLE, PSEUDO-CLEFTING] \implies low C-score
 - *Doug cleaned the dishes [before the party].* (−0.31)
 - *The children hid [in a hurry].* (−0.41)
 - *The thief stole the painting [for her boss].* (−0.67)

As mentioned above, these are cases that neatly fit into the predictions. Trends in the whole dataset is noisier, but nevertheless display the tendency predicted above—the mean C-scores of sentences that correspond to the four diagnostic pattern groups exactly align with our predicted ordering. Table 3 lists the mean and standard deviation of each diagnostic pattern group. The means of [*O,*P] and [O,P] groups are higher and lower, respectively, compared to either mean of the middle-ground groups, which is the expected result. A linear mixed-effects analysis, with the diagnostic pattern groups as a fixed effect with four levels (reference group = [O,P]) and the predicate of the sentence as random effect, reveals that [*O,*P] ($B = .40, SE = .14, p < .01$) and [O,*P] ($B = .29, SE = .06, p < .001$) imply significantly higher C-scores than [O,P] ([*O,P] has $B = .23, SE = .12, p = .051$). We follow up with a more detailed post-hoc analysis and show that significant partial differences in the distribution of the diagnostic results explain the differences in the group means (and their exact alignment to our predicted ordering).

	μ	σ
[*O, *P]	0.249**	0.382
[O, *P]	0.169***	0.485
[*O, P]	0.099†	0.386
[O, P]	-0.125	0.542

Table 3. Mean and standard deviation of C-scores, according to the binary judgment pattern groups. The order of the groups by mean is in accord with the predicted order in Section 6.3.

Figure 16 shows the kernel density estimation¹⁶ for the rankings of C-scores that pertain to each of the four possible diagnostic judgment pattern groups. This visualizes the estimation of the underlying probability distribution that likely generates the distribution we observe (datapoints shown are restricted to the observed data), which gives us a rough idea of where the differences in the group means derive from. For instance, we can visually observe that [O,*P] group is likelier to contain more sentences with higher C-scores (i.e. C-scores in the upper 25% percentile), whereas [*O,P] group is likelier to contain sentences with mid-range C-scores. We use rankings instead of absolute values, since the distribution of the values is significantly skewed (Section 4.2.3).

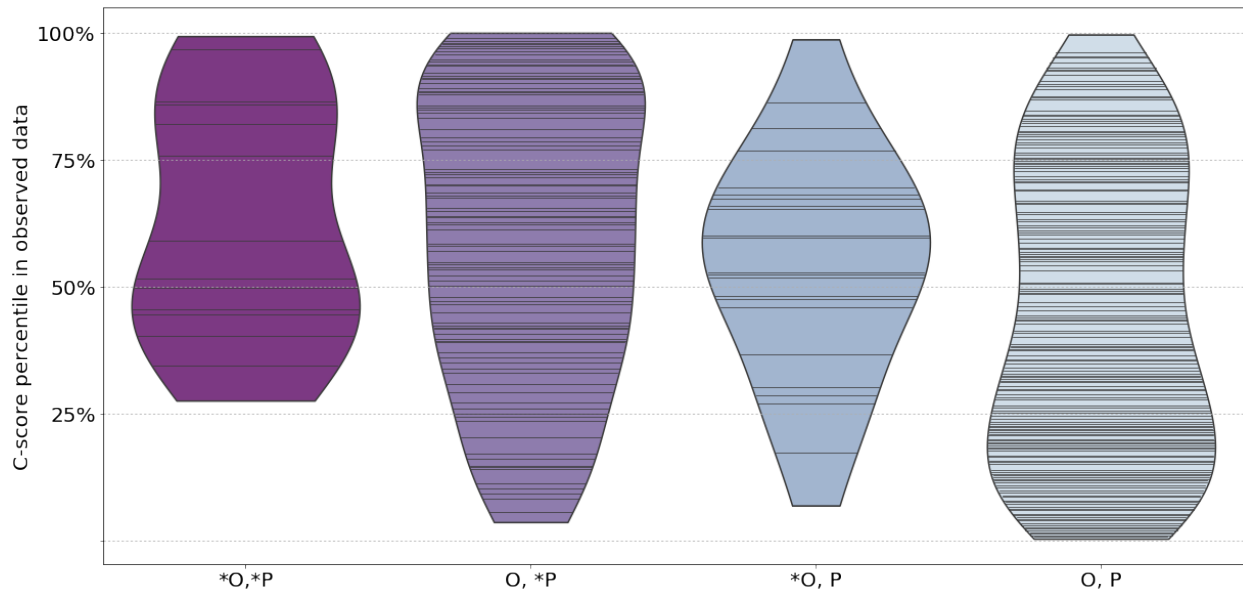


Figure 16. A kernel density estimation graph for each possible combination of diagnostic judgments (observed range only).

Based on this visualization, we hypothesize that each judgment group will affect the distribution of C-score rankings in different percentiles (e.g. the [O,*P] group will contain more sentences in the upper 25% C-score percentile than it is

¹⁶A nonparametric estimation of the probability density function, for each diagnostic group in this case.

likely by chance, and that explains why the mean C-score of this group is higher). Figure 17 shows which percentiles are significantly different from chance for the four diagnostic judgment pattern groups. The size of the circle reflects the proportional distribution of the sentences that correspond to a particular combination of the diagnostic results in the marked percentile region. The leftmost column of circles denotes the hypothetical case where the distribution is even (i.e. 25% of the cases occur in the upper and lower 25% percentiles, and 50% occur in the middle 50%)—thus, it represents the null hypothesis that the distribution is equal to chance. This is the result we would observe if a certain diagnostic pattern group picked out sentences at random. How different the distribution is in different percentiles for each of the four groups can be inspected by comparing the sizes of the circles in each column to the null hypothesis. Statistically significant deviation from the null hypothesis is marked inside the circles (z -test for one proportion, $* = p < .05$). The p values for [O,*P] were corrected for multiple comparisons using Holm correction, because the initially planned 75th and middle 50% percentile comparisons were not significant for [O,*P] and a top 50%/lower 25-50% percentile splits were additionally tested.

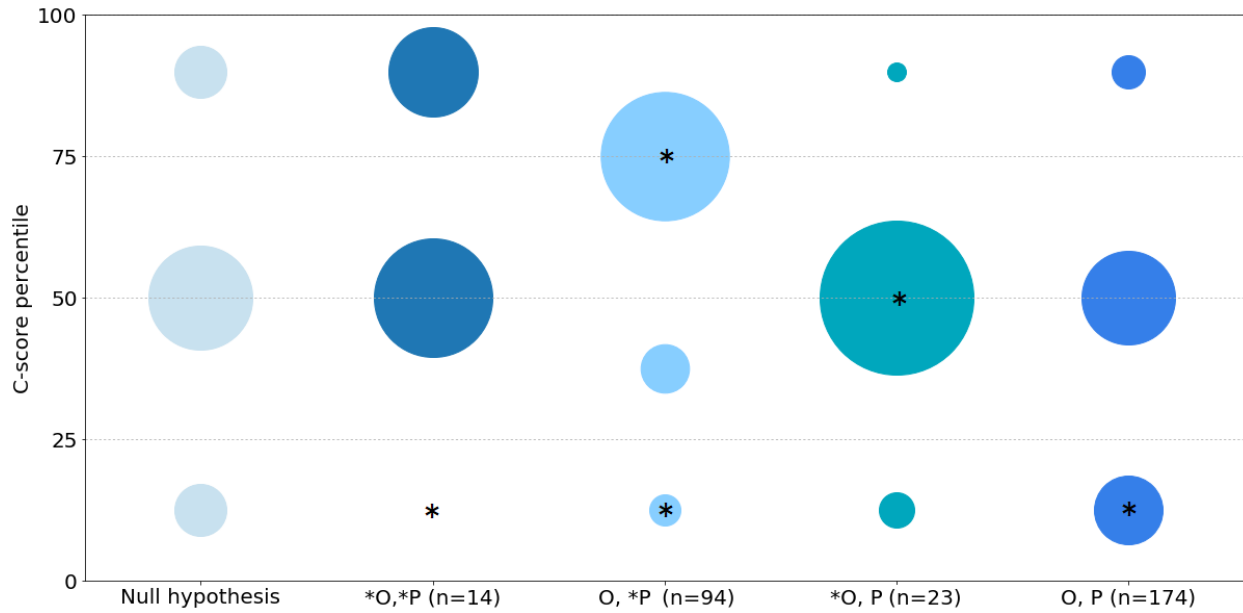


Figure 17. A percentile analysis of each possible combination of diagnostic judgments, with respect to C-score rankings. (* = $p < .05$)

Let us focus on the conflicting cases that we have been discussing, [*O,P] and [O,*P]. Figure 16 shows that these two groups, where we predicted ‘moderate’ C-scores, are visually quite different. The distribution of [O,*P] sentences is top-heavy, whereas the distribution of [*O,P] sentences is biased towards the middle. This trend is also reflected in the difference in means we observed in Table 3 (i.e. mean of [O,*P] is higher than [*O,P]).

Recall that we have previously assumed OMISSIBILITY passes if a_C is low, and PSEUDO-CLEFTING passes if a_A is high. However, the terms ‘high’ and ‘low’ are vague and should be assigned a specific value. The difference in these threshold values set for each diagnostic test is what we claim to be responsible for the disparity we observe in the C-score distribution of [O,*P] and [*O,P].

The question one may ask at this point is, what is the basis of the assumption that the two diagnostics we discussed are targeting different structural activations? As Schütze (1995) points out, it is intuitively understandable that different diagnostics would be sensitive to different structural properties of the sentence, since each test exploits distinct diagnostic environments. If so, what exactly are the different properties that the two diagnostics target? In the case of OMISSIBILITY, the main property it is probing seems to be the lexical requirements imposed by the predicate (we refrain from making the distinction between syntactic/semantic obligatoriness here). This type of information is accessed and modified via proto-complement structure (recall the discussion in Section 2.3). Therefore, if a PP dependent is not omissible (*OMISSIBILITY), it must have sufficiently high proto-complement activation. The main property targeted by PSEUDO-CLEFTING, on the other hand, seems to be the transferrability of the PP meaning. One

of the common characterization of adjuncts is that they have consistency in meaning across many predicates, whereas complements are more idiosyncratic (Pollard and Sag, 1987). The discussion in Section 3.1 is pertinent, especially regarding the duality of preposition roles. Consider the constructions we create by pseudo-clefting the PP; the PP has to modify the semantically bleached verb *do* and retain its original meaning in the unclefted sentence. Therefore, the more transparent the meaning of the PP (i.e. the less the predicate-governed idiosyncrasy), the likelier the pseudo-clefted construction will be acceptable. In other words, since attachment to *do* forces the PP to be interpreted via its inherent lexical meaning, in order for the original meaning to be transferred under pseudo-clefting, the original construction must have sufficient proto-adjunct activation.

6.4 Joint analysis of scaled complement-adjuncthood and diagnostics data

In the previous section, we presented an analysis relating the gradient judgments of argumenthood (C-scores) to traditional binary acceptability judgments produced by a linguist. Now we move on to an analysis jointly considering C-scores and the gradient diagnostic-test judgments collected in Experiment 2. We start by individually plotting C-scores against the gradient judgments for each diagnostic test, OMISSIBILITY and PSEUDO-CLEFTING.

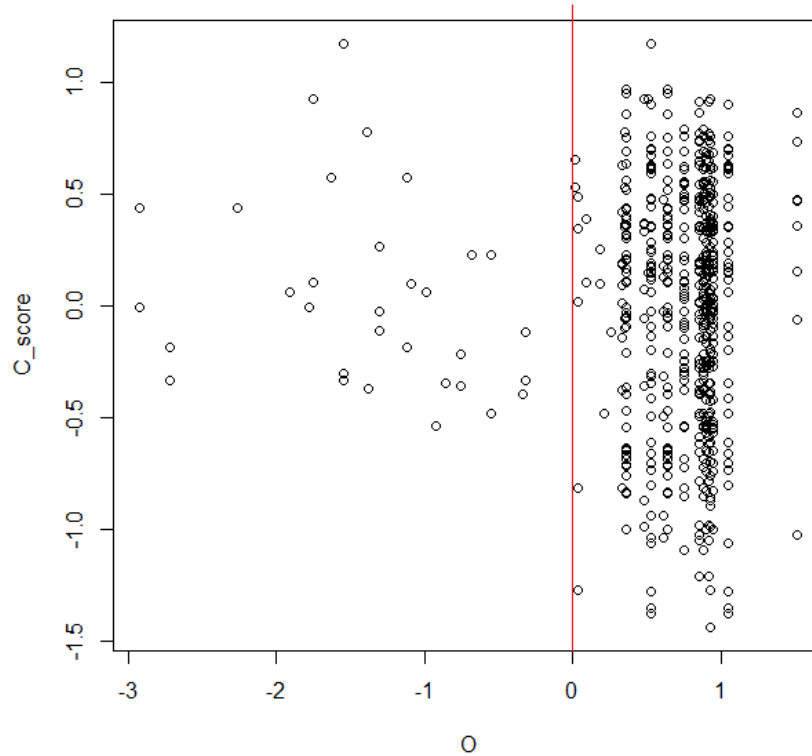


Figure 18. C-score plotted against Omissibility score.

Omissibility scores First, in Figure 18, we plot the C-score of the corresponding sentence against the normalized omissibility score (the lower the score, the less natural the sentence is without the PP)¹⁷. The linear relationship between the two is not significant ($R^2 = .001, \beta = -.03, SE = .04, p = .45$). In the plot, we see that there are two distinct regions; the plot can be vertically partitioned roughly around $O = 0$, where the datapoints on the right of the vertical line ($O > 0$) and the datapoints on the left side of the line ($O < 0$) are distributed differently. Most of the datapoints (> 90%) are concentrated in the right region, and the datapoints in the left region seem to form a linearly descending pattern although there is no significant linear fit between O and C-scores in this region ($\beta = -.26, SE = .18, p = .15$).

¹⁷We use the naturalness score here rather than the naturalness+compatibility score, since the composite measure may conflate different optimization processes (to be discussed in Section 7).

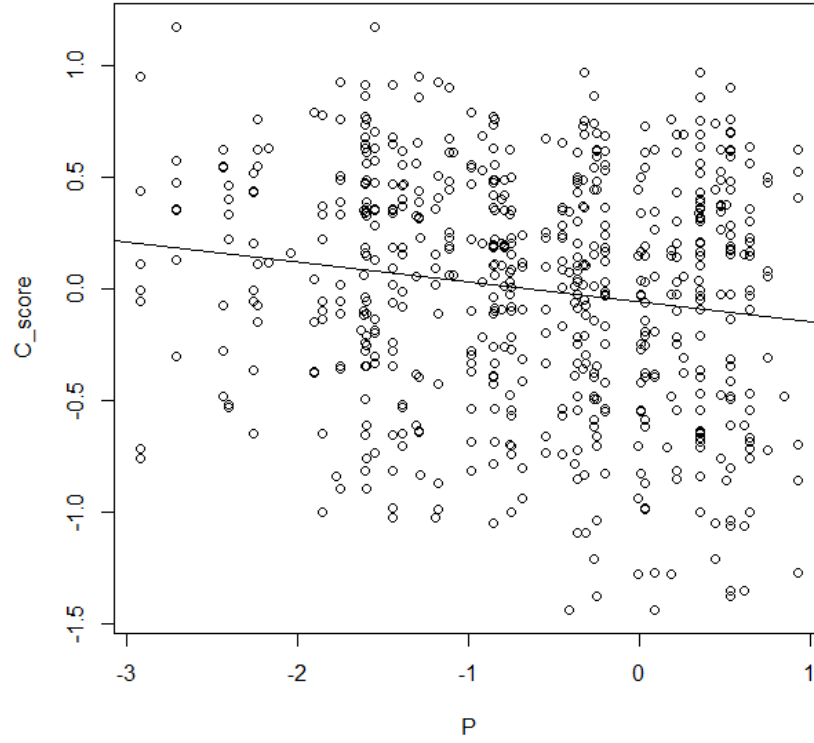


Figure 19. C-scores plotted against Pseudo-cleftability.

Pseudo-cleftability scores Figure 19 shows the relation between normalized pseudo-cleftability score (the lower the score, the less natural the sentence is when the PP is pseudo-clefted) and the C-score of the corresponding unclefted sentence. There is a very significant negative linear relationship ($R^2 = .024, \beta = -.16, SE = .04, p < 10^{-4}$). Although the slope is not very steep and the variance explained is low, this trend is consistent across the whole dataset (unlike the omissibility score plot where we see two clearly partitioned regions).

Joint analysis Recall that in Section 6.3, we hypothesized that high proto-complement activation leads to *O (i.e. a_C above a certain threshold θ_C) and high proto-adjunct activation leads to P (i.e. a_A above a potentially different threshold θ_A). This hypothesis can be restated as follows:

$$\begin{aligned} a_C > \theta_C &\implies *O \\ a_A > \theta_A &\implies P \end{aligned}$$

Moreover, we hypothesize that higher proto-complement activation (a_C) makes a (relatively) large positive contribution to the C-score, whereas higher proto-adjunct activation (a_A) makes a (relatively) small negative contribution to the C-score, based on the observations made in Section 6.3.1. This can be expressed as:

$$\begin{aligned} score &\propto w_1 a_C + w_2 a_A \quad (w_1 \geq 0, w_2 \leq 0) \\ |w_1| &> |w_2| \end{aligned}$$

Given these hypotheses, the pattern of the omissibility data shown in Figure 18 suggests that the actual value of the ‘high’ threshold for proto-complement activation (θ_C) is very large, which would correspond to a skewed partitioning of the proto-complement scale as in Figure 20.

In this case, the O region would contain examples with a wide range of proto-complement activation values. This explains why we do not see any patterns of C-scores in the O region, considering that proto-complement activation has a relatively large effect on C-scores. This also means that datapoints in the *O region are *guaranteed* to have very

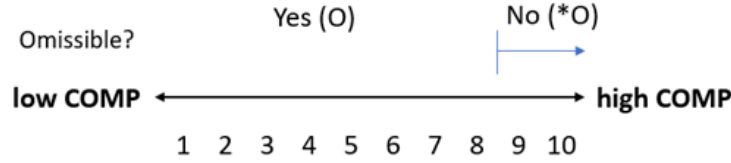


Figure 20. O and *O regions on the proto-complement scale. The numbers are arbitrarily assigned for expository purposes; here the ‘high’ threshold for *O would be $\theta_C = 8.5$.

high proto-complement activation, because θ_C is very high. However, we cannot make a claim about C-scores yet without knowing about the proto-adjunct activation, since overall argumenthood is mediated by both proto-structures. It is indeed the case that high proto-complement activation does not necessarily imply a high C-score; we see almost a full range of C-score values in the *O region ($\text{range}(\text{all}) = [-1.435, 1.172]$, $\text{range}(*O) = [-1.270, 1.172]$), and we overall did not see a significant linear relationship between omissibility and C-scores. Here are some sentences that are in the *O region but have low C-scores:

The books lean [against the door]. (−0.33)
The statue stood [on the corner]. (−0.12)
Carol cut the recipes [from the magazine]. (−0.02)

In contrast, the pseudo-cleftability scores (Figure 19) are more evenly distributed, rather than being partitioned into two distinct regions like the omissibility scores (which we attributed to the high value of θ_C above). Since this is not the pattern we observe for the pseudo-cleftability scores, the proto-adjunct scale must have less extreme threshold point θ_A , for instance as in Figure 21. The unsteep slope of the linearly descending trend matches our hypothesis that proto-adjunct activation has a small negative effect on the C-score (i.e. small $|w_2|$).

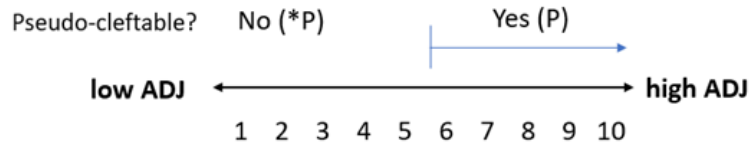


Figure 21. P and *P regions on the proto-adjunct scale. The numbers are arbitrarily assigned for expository purposes; here the threshold for P would be $\theta_A = 5.5$.

Recall that the datapoints in the *O region are guaranteed to have high proto-complement activation (a_C). Based on this observation, we can make a prediction that since a_C is high, for the datapoints in *O region, the C-score will correlate transparently with how high the proto-adjunct activation (a_A) is. This is precisely what we see when we plot C-scores against pseudo-cleftability scores for only the datapoints in the *O region (Figure 22)—there is a significant reverse linear relation between P and C ($R^2 = .16$, $\beta = -.40$, $SE = .17$, $p = .02$), the slope (standardized β) of which is much steeper than the slope between P and C for datapoints in the O region ($R^2 = .02$, $\beta = -.14$, $SE = .04$, $p < .001$).

Additionally, an interesting observation can be made about the pseudo-cleftability data with respect to our analysis. See Figure 23, which is the same as Figure 19 but with emphasis on the lower right region of the plot. We see here that the lower end of the C-score scale (the y-axis) is exclusively contained in the higher half of the P scale. This implies that in order for the C-score to be very low, it should be the case that P is high (\Leftarrow high a_A).

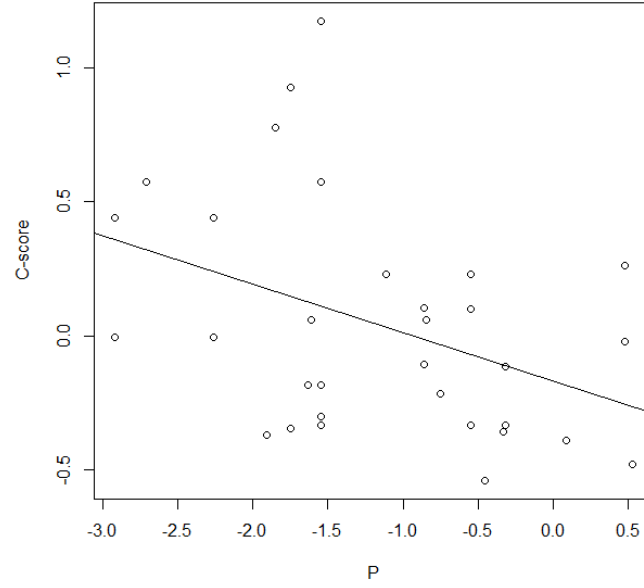


Figure 22. Pseudo-cleftability score plotted against C-score, only for datapoints in the *O region.

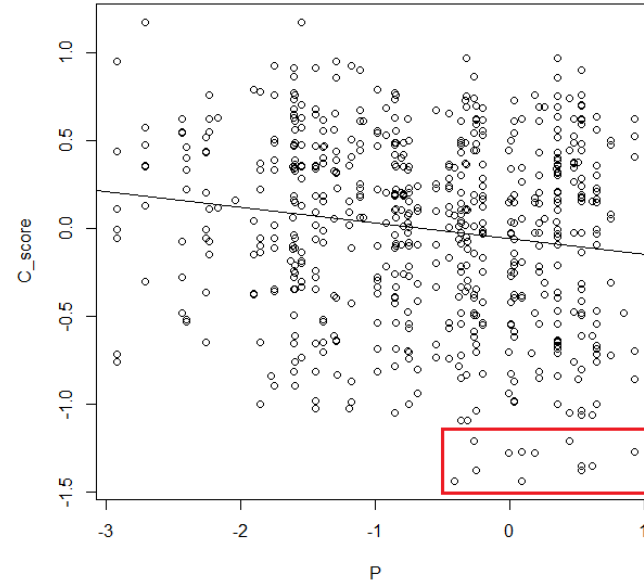


Figure 23. Pseudo-cleftability score plotted against C-score with an emphasized region of interest.

7 Computation via Gradient Harmonic Grammar

There remains a more fundamental question of why speakers need to maintain both structures simultaneously instead of converging to one. Although we direct the readers to the existing body of literature advocating parallel parsing models for general insights into this issue (Christianson et al., 2001; Cho et al., 2017; Rasmussen and Schuler, 2018; Dillon et al., 2019, *i.a.*), we highlight Dowty (2003)’s suggestion here, that inherent meaning of the adjunct form serves as a mnemonic to effectively learning and retrieving the complement meaning. For instance, it is not coincidental that the ‘directional’ preposition *to* is selected in *speak to* (‘communicate’) or *drink to* (‘toast’) over other prepositions like *in* or *on*. This sharing of labor can be made more formally explicit via simultaneously active proto-structures. We now

demonstrate how we can express the way information from both the proto-complement and proto-adjunct substructures is utilized, using constraint interactions from a gradient variant of Harmonic Grammar (gHG; Smolensky et al. 2014; Smolensky and Goldrick 2016). gHG has been successfully applied to model complex phenomena in phonology (Smolensky and Goldrick, 2016; Zimmermann, 2019) and morphology (Rosen, 2018, 2019). The neural network implementation of gHG has also been used for modeling sentence processing (Cho and Smolensky, 2016; Cho et al., 2017), sentence production (Goldrick et al., 2019) and bilingual code-mixing (Goldrick et al., 2016) using partially active symbolic structures.

7.1 Optimization with Gradient Symbolic Computation (GSC)

We use the GSC framework to calculate proto-complement and proto-adjunct activations and how they affect diagnostic acceptability judgments. We especially focus on demonstrating the capacity of our gradient blend model to generate all four possible combinations of diagnostic judgments, which we claimed was possible in Section 6. Since predicting the actual values of C-scores involve various additional factors other than the diagnostic judgments, we refer the readers to Kim et al. (2019)’s work, and provide a more focused discussion here on modeling the connection between gradient blend states and diagnostic acceptability.

When judging the acceptability of a diagnostic construction (either under a traditional setting of ‘applying’ a diagnostic test and making a binary judgment, or a scaled setting as in a trial of Experiment 2), speakers are considering a pair of sentences: a base sentence and a diagnostic sentence. We hypothesize that when speakers process the base sentence, they assign activation values to the proto-complement and proto-adjunct substructures. They then assess the acceptability of the diagnostic sentence using their parse of the base sentence as a point of comparison.

Formally expressed in gHG terms, our hypothesis is that the base sentence is assigned the blend structure that has ‘optimal’ activation values of its substructures. The optimal structure is the best-evaluated structure among potential alternatives, which may take on any activation values ranging over the non-negative real numbers. The evaluation is carried out by numerically-weighted violable constraints of Harmonic Grammar (HG) (Legendre et al., 1990). In evaluating a candidate blend structure, the penalty assessed for a substructure violating a constraint is multiplied by the activation of that substructure—its degree of presence in the overall candidate being evaluated.

We formally define the input/output variables and constraints involved in these processes as follows:

Variables.

- γ = strength of the preference ‘requires complement position to be filled’ (determined by the verb) $\gamma \geq 0$
- α = strength of the preference ‘requires adjunct position to be filled’ (determined by the whole input) $\alpha \geq 0$
- Intended construction (for diagnostic tests)

OC: Ordinary clause	NP _s V (NP _o) PP
OP: Omit PP	NP _s V (NP _o)
PC: Pseudo-cleft	What NP _s did PP was V (NP _o)
- PP attachment site activation

a_C = activation of proto-complement PP substructure	$a_C \geq 0$
a_A = activation of proto-adjunct PP substructure	$a_A > 0$

Constraints.

- $\gamma \implies a_C$ If a verb prefers a complement, follow that preference in the output.
- $\alpha \implies a_A$ If a construction prefers an adjunct, follow that preference in the output.
- FAITH-X If element X is in the input, it should be realized in the output.
- *DO-PP_C Verb *do* disprefers complement PPs.
- DO-PP_A Verb *do* prefers adjunct PPs.
- Q Quantization constraint; prefer discrete outputs (activity 0 or 1) to nondiscrete outputs.

	$(w_0 = 1)$	$(w_1 = 1)$	$(w_2 = -q)$	
$(\gamma, \alpha), \mathbf{V}, \mathbf{PP}$	$\gamma \implies a_C$	$\alpha \implies a_A$	\mathbf{Q}	\mathbf{H}
a. $\{a_C, a_A\} (a_C \geq 0, a_A > 0)$	$\gamma \cdot a_c$	$\alpha \cdot a_A$	Q	$\gamma \cdot a_c + \alpha \cdot a_A - q \cdot Q$

Table 4. Optimization of proto-complement and proto-adjunct activations (a_C, a_A) , given a [V PP].

Table 4 illustrates the optimization process during which activation values a_C, a_A are assigned to PP structures in an ordinary clause, given the verb (V), the prepositional phrase (PP), and the numerical preference strengths of the verb for taking a complement (γ) and the context for taking an adjunct (α). In HG evaluation, the optimal candidate is the one that has the maximum Harmony (H) (Equation 1; Pater 2009), which is the sum of the constraint violation or satisfaction scores (s_k) assigned to each candidate output, multiplied by the constraint weights (w_k).

$$H = \sum_{k=1}^K s_k w_k \quad (1)$$

The second to the fourth columns in Table 4 correspond to the constraints that are involved in calculating the optimal output $\{a_C, a_A\}$. The constraints $\gamma \implies a_C$ and $\alpha \implies a_A$ express the idea that ‘if the verb prefers a complement to a degree γ , and if the construction prefers an adjunct to a degree α , follow those preferences in the output’. Q is a QUANTIZATION constraint that expresses preference for discrete outputs (activity values of either 0 or 1). A structure with proto-complement activation a_C and proto-adjunct activity a_A incurs the penalty as follows (Cho and Smolensky, 2016):

$$Q = a_C^2(1 - a_C)^2 + a_A^2(1 - a_A)^2 \quad (2)$$

Note that the Q penalty is 0 if both activity values are either 0 or 1. The constant $-q$ indicated in Table 4 is the strength of this quantization constraint.

The goal of the optimization is to find the weights a_C, a_A that maximize H (which is $\gamma \cdot a_c + \alpha \cdot a_A - q \cdot Q$ when $w_0 = w_1 = 1$). To illustrate, let us set $q = 1$, and let us posit a construction in which the adjunct preference strength is $\alpha = 0.5$ and a verb for which the complement preference strength is $\gamma = 0.5$. For these values of γ, α , we get maximum H with gradient, coactive output structures with activity values $a_C = 1.26, a_A = 1.26$. This process is independent of any diagnostic test environment; we assume this is how a_C, a_A are calculated whenever a speaker encounters a [V PP] construction.

The optimization processes involved in the diagnostic tests are distinct from Table 4, in that they use as inputs the already-optimized blend weights a_C, a_A (i.e. the outputs of Table 4) of the proto-complement C and the proto-adjunct A . Furthermore, in the two diagnostic tests, a different set of constraints is utilized in the optimization process because different linguistic constructions are exploited in each test. Tables 5 and 6 are examples of how omissibility and pseudo-cleftability diagnostics may share the same values a_C, a_A in their inputs but differ in terms of the constraints involved. We claim that this difference is responsible for the diagnostic conflicts. The two diagnostic acceptability tests involve a conscious comparison of a certain diagnostic construction (Tables 5 and 6, rows 1b, 2b) with the original expression being tested (rows 1a, 2a). This supports the need for a separate optimization process for calculating diagnostic judgments, independent of the process that assigns a_C, a_A to the original expressions (this also predicts that the a_C, a_A values will not exactly translate into diagnostic test acceptability).

Table 5 for the omissibility test expresses the idea that whether a PP is omissible depends on the outcome of the competition between two faithfulness constraints; whether the output is faithful to our intention of omitting the PP (the intention of the diagnostic test) versus whether the output maintains the complement PP in the input (because of a general preference for avoiding complement omission). Table 6 for the pseudo-cleftability test expresses that whether a PP is pseudo-cleftable depends on the faithfulness-to-the-intent constraint, and also on the preferences of a newly introduced verb *do*, which disprefers a complement and prefers an adjunct.

	$(w_0 = -m_0)$	$(w_1 = -m_1)$	
(a_C, a_A) , OmitPP	FAITH-OMITPP	FAITH- a_C	H
1a. V PP (=O): $a_C + a_A > 0$	$-m_0(a_C + a_A)$		$-m_0(a_C + a_A)$
1b. V (=O): $a_C = a_A = 0$		$-m_1 a_C$	$-m_1 a_C$

Table 5. Optimization for the omissibility diagnostic test, formulated as a competition between two structures: O, with omitted PP, and *O, with the PP preserved.

	$(w_0 = -k_0)$	$(w_1 = -k_1)$	$(w_2 = k_2)$	
(a_C, a_A) , Pseudo-cleft	FAITH-PSUEDOCLEFT	*DO-PP _C	DO-PP _A	H
2a. V PP (=P): $a_C + a_A > 0$	$-k_0$			$-k_0$
2b. what X did PP [...] (=P): $a_C + a_A > 0$		$-k_1 a_C$	$k_2 a_A$	$-k_1 a_C + k_2 a_A$

Table 6. Optimization for the pseudo-cleftability diagnostic test, formulated as a competition between two structures: P, with pseudo-clefted PP, and *P, with the PP unclefted.

Now we can calculate what range of values a_C, a_A and constraint weights m, k correspond to which patterns of judgments¹⁸. According to the tableaux, *O is optimal (1a has higher harmony than 1b) if $-m_0(a_C + a_A) > -m_1 a_C$, or equivalently, if $m_0 < m_1 \frac{a_C}{a_C + a_A}$. Similarly, *P is optimal (2a has higher harmony than 2b) if $-k_0 > -k_1 a_C + k_2 a_A$, or equivalently when $k_0 < k_1 a_C - k_2 a_A$. We illustrate in the next section how this model represents and accounts for the four observed patterns of diagnostic judgments.

Furthermore, this two-step optimization process using the a_C, a_A values calculated from Table 4 as inputs to another optimization process provides a unifying explanation for the complex relationship between argumenthood (quantified by a_C and a_A), diagnostic judgments, and the C-scores we collected. That is, the diagnostic tests are optimization processes that take a_C and a_A as parts of the input, and the same a_C, a_A are predictors (but not the only predictors) of C-scores. The relation between a_C, a_A and C-scores is such that if a_C is high, the C-score is high and if a_A is high, the C-score is low. However, the existence of other factors (for instance, Kim et al. (2019) show that the presence of a direct object in the test sentence negatively affects C-scores) prevents the C-scores from being a function (e.g. linear combination) of only a_C and a_A . Nevertheless, we observe some interesting connections between the C-scores and the diagnostic tests O and P as discussed in Sections 6.3 and 6.4, because computing the C-score and diagnostic acceptability (Tables 5 and 6) both involve values a_C and a_A .

7.1.1 Simulation: how are the four patterns of judgments captured under our analysis?

To further illustrate how our proposed two-step optimization procedure represents the four patterns ([O,P], [O,*P], [*O,P], [*O,*P]), we present a simulation study. First, we find values of (a_C, a_A) that maximize H for 1000 randomly sampled pairs of (γ, α) between zero and one. We formulate this as a bounded optimization problem where $a_C \geq 0, a_A > 0$, using the L-BFGS-B algorithm from *scipy* (the initial guess is also randomly selected each time from values between zero and one). Figure 24 shows plots of (a_C, a_A) for four different values of q , which represents the strength of the quantization constraint (see Cho et al. (2018) for discussions of q as a degree of ‘commitment’).

According to the OMISSIBILITY and PSEUDO-CLEFT tableaux (Tables 5 and 6), *O wins if $m_0 < m_1 \frac{a_C}{a_C + a_A}$ and *P wins if $k_0 < k_1 a_C - k_2 a_A$. In Figure 25, we plot the right sides of the two inequalities on the x and y axes, respectively, in order to visualize the regions corresponding to the four judgment patterns [O,P], [O,*P], [*O,P], and [*O,*P]. a_C, a_A values when $q = 1$ from Figure 24 are used. The figures show that we can indeed get optimal a_C, a_A outputs that are gradient, and that the four judgment patterns can be attributed to the values of constraint weights

¹⁸Note that under this formulation, each diagnostic test is sensitive to both a_C and a_A . We can also formulate the diagnostic judgments as purely- a_C or - a_A sensitive as we described in Section 6 by adopting a different set of constraints. For instance, Table 6 can have a different constraint in place of *DO-PP_C that is sensitive only to a_A , in which case it will be a purely a_A -sensitive test. Using purely a_C - and a_A -sensitive versions of the diagnostic tests does not contradict any findings reported.

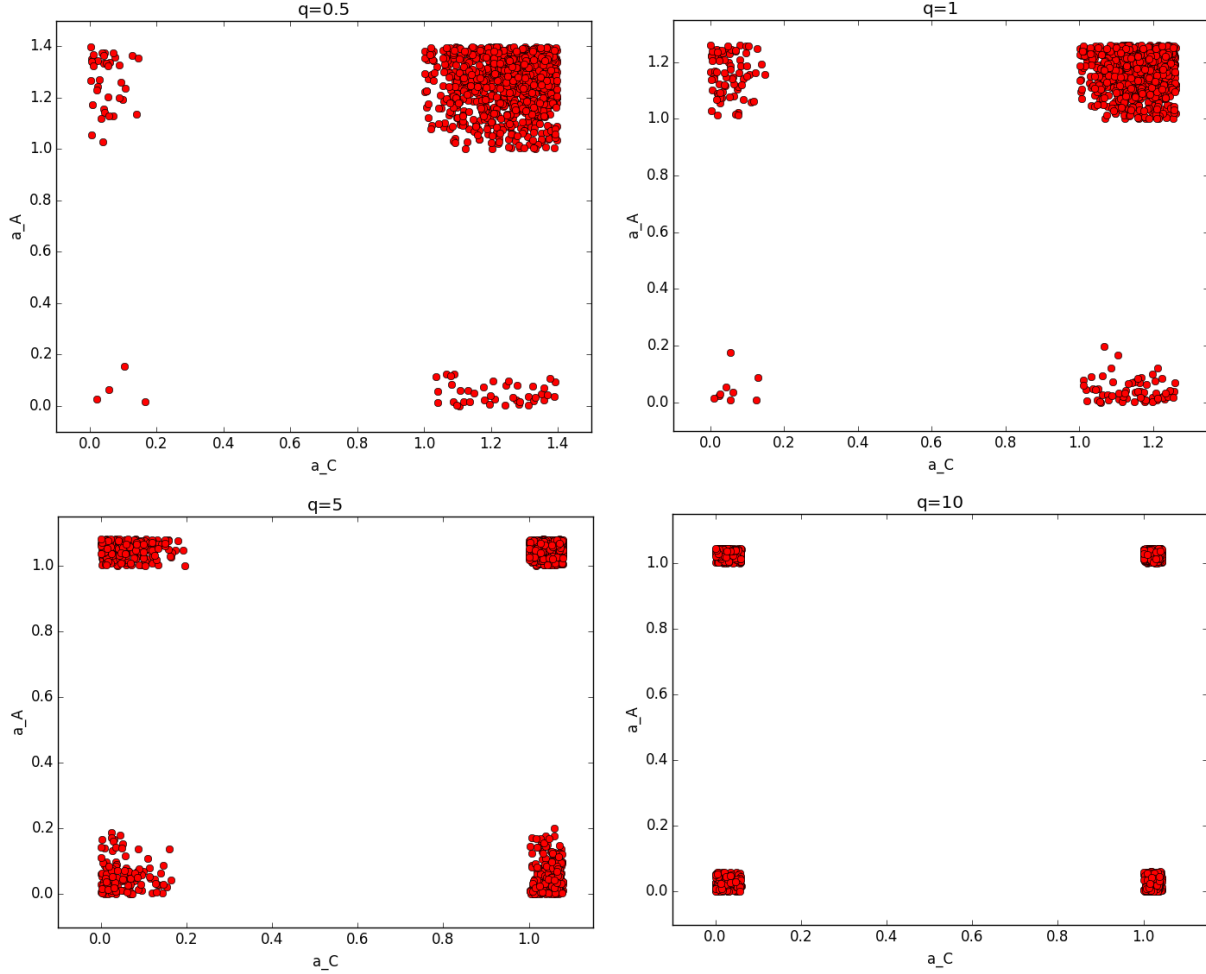


Figure 24. Optimal values of (a_C, a_A) for 1000 randomly sampled of γ, α between zero and one, for four different values of q .

(i.e. how the constraint weights create a linear four-way partitioning of the Harmony space such that each partition contains a datapoint). Using categorical representations (using 0/1 for a_C and a_A) do not give rise to such a partition; they reduce to two single points in the space (i.e. using a gradient model of diagnostic judgments and a categorical model of complement-adjuncthood does not suffice).

8 Conclusion

We proposed a novel gradient blend analysis of [V PP] constructions in English with respect to the complement-adjunct status of the PP, claiming that [V PP] constructions in English are blends of gradiently-active proto-complement and proto-adjunct structures. We emphasize that our gradient blend analysis is in fact not incompatible with the traditional discrete symbolic view of complements and adjuncts, but further generalizes it, since discrete complement and adjunct structures can be viewed as blends of proto-structures with zero activation of one part of the blend. We believe that for many linguistic phenomena, a discrete analysis could serve as a good approximation, as evident from the progress made in linguistics research under symbolic frameworks. However, we should also acknowledge the existence of a non-negligible set of linguistic phenomena that are better explained by models that move beyond discrete symbolic representations.

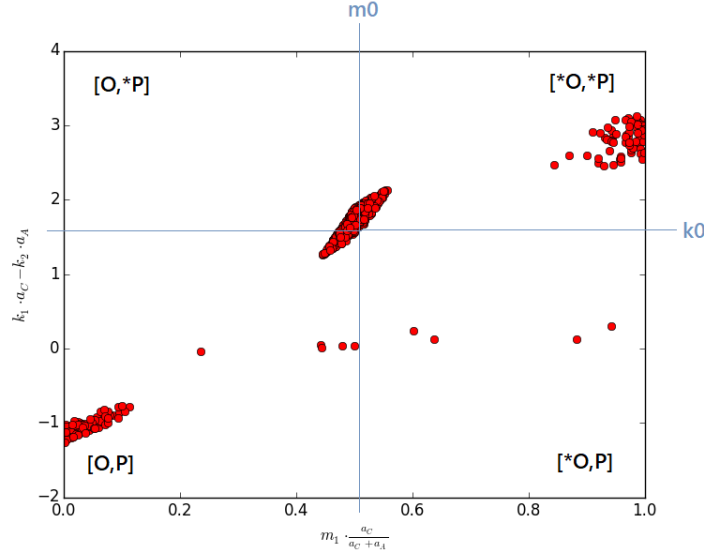


Figure 25. A possible partition of the Harmony space, when $q = 1$, $m_1 = k_1 = k_2 = 1.0$, $m_0 = 0.5$, $k_0 = 1.5$.

This work is our attempt at empirically justifying the necessity of gradient symbolic blend models, based on a long-standing problem of the complement-adjunct status of verbal dependents. First, we collected gradient judgments of complement-adjuncthood for sentences that contain a [V PP] construction. Then we additionally collected acceptability judgments for two frequently used diagnostic tests: OMISSIBILITY and PSEUDO-CLEFTABILITY, which verified the existence of the problematic conflict across different diagnostic tests. We made a typological argument that both gradience and blendedness are crucial to explain this seemingly incoherent diagnostic test results that have long been problematic in the literature. Namely, we have shown that in order to explain all four patterns of judgments from the two different diagnostic tests, a discrete or a gradience-only model do not suffice. We demonstrated how our gradient blend model is able to offer a principled explanation about the data and provide an optimization-based formal description about the connection between argumenthood and diagnostic tests. Additionally, we presented a simulation study using Gradient Harmonic Grammar, in which it is possible to obtain a four-way partitioning of the Harmony space based on the constraints governing the optimization process, that correspond to each of the four observed diagnostic judgment patterns.

8.1 Future work

The first natural extension of this work would be investigating whether our gradient blend model extends to verbal dependents other than PPs. For instance, can implicit objects and how their omissibility ties into their argumenthood, be subject to a similar analysis to what we have presented? We also hope to clarify further our view on syntactic and semantic argumenthood under our analysis, which was not in the scope of this work.

Second, working towards confirming the psychological reality of blend states is a promising future direction. There exist numerous works in the literature that could serve as experimental guidelines. The idea that humans may maintain multiple coactive parses in parallel during language processing has been an active topic of discussion, and prior research has shown psycholinguistic (Dillon et al., 2019) and simulation-based (Cho et al., 2017) evidence. We believe experimental protocols as such can be adapted to test whether speakers do maintain blend states of complements and adjuncts.

Finally, recent breakthroughs in Natural Language Processing (NLP) that are based on gradient, distributed underlying representations have produced dramatic improvements in NLP tasks that are purely discrete in nature (LeCun et al., 2015; Peters et al., 2018; Devlin et al., 2019, *i.a.*). This raises a question of whether distributed representations could speak to theoretical problems in linguistics. We hope to productively combine recent methodological developments in NLP and investigate whether and to what degree they might aid our theoretical understanding of language.

Acknowledgements

We thank the NSF for partially supporting this research (NSF INSPIRE BCS-1344269). We also thank C. Jane Lutken, Geraldine Legendre, Lilia Rissman, Benjamin Van Durme, Pyeong Whan Cho, Matt Goldrick, Giulia Cappelli, the audience at CiALT2 and the Acceptability Judgments Workshop, and members of the Neurosymbolic Computation Lab and Semlab at JHU for their helpful comments and feedback.

References

- ALDEZABAL, IZASKUN, MAXUX ARANZABE, KOLDO GOJENOLA, KEPA SARASOLA, and AITZIBER ATUTXA. 2002. Learning argument/adjunct distinction for Basque. In *Proceedings of the ACL Workshop on Unsupervised Lexical Acquisition*, volume 9, 42–50. Association for Computational Linguistics.
- ARKA, I WAYAN. 2014. Locative-related roles and the argument-adjunct distinction in Balinese. *Linguistic Discovery* 12.56–85.
- BAKER, MARK C. 1989. Object sharing and projection in serial verb constructions. *Linguistic Inquiry* 20.513–553.
- BARBU, ROXANA-MARIA, and IDA TOIVONEN. 2015. Arguments and adjuncts: at the syntax-semantics interface. *Florida Linguistics Papers* 3.
- BARBU, ROXANA-MARIA, and IDA TOIVONEN. 2016. Event participants and linguistic arguments. In *Proceedings of the 38th Annual Meeting of the Cognitive Science Society, 1961–1966*.
- BARD, ELLEN GURMAN, DAN ROBERTSON, and ANTONELLA SORACE. 1996. Magnitude estimation of linguistic acceptability. *Language* 32–68.
- BOERSMA, PAUL. 1997. How we learn variation, optionality, and probability. In *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam*, volume 21, 43–58. Amsterdam.
- BRESNAN, JOAN. 2007. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. *Roots: Linguistics in search of its evidential base* 96.77–96.
- BRESNAN, JOAN, ASH ASUDEH, IDA TOIVONEN, and STEPHEN WECHSLER. 2015. *Lexical-functional syntax*, volume 16. John Wiley & Sons.
- BRESNAN, JOAN, and JONNI M. KANERVA. 1989. Locative inversion in chicheŵa: a case study of factorization in grammar. *Linguistic inquiry* 1–50.
- BRESNAN, JOAN, and TATIANA NIKITINA. 2003. The gradience of the dative alternation. *Ms., Stanford University*, <http://www-lfg.stanford.edu/bresnan/download.html>.
- CARRIER-DUNCAN, JILL. 1985. Linking of thematic roles in derivational word formation. *Linguistic Inquiry* 1–34.
- CENNAMO, MICHELA, and ALESSANDRO LENCI. 2018. Gradience in subcategorization? Locative phrases with Italian verbs of motion. *Studia Linguistica*.
- CHO, PYEONG WHAN, MATTHEW GOLDRICK, RICHARD L. LEWIS, and PAUL SMOLENSKY. 2018. Dynamic encoding of structural uncertainty in gradient symbols. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*, 19–28.
- CHO, PYEONG WHAN, MATTHEW GOLDRICK, and PAUL SMOLENSKY. 2017. Incremental parsing in a continuous dynamical system: Sentence processing in gradient symbolic computation. *Linguistics Vanguard* 3.
- CHO, PYEONG WHAN, and PAUL SMOLENSKY. 2016. Bifurcation analysis of a gradient symbolic computation model of incremental processing. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, 1487–1492.
- CHOMSKY, NOAM. 1975. *The logical structure of linguistic theory*. Plenum Press, New York.
- CHOMSKY, NOAM. 1993. *Lectures on government and binding: The Pisa lectures*. Number 9. Walter de Gruyter.
- CHRISTIANSON, KIEL, ANDREW HOLLINGWORTH, JOHN F. HALLIWELL, and FERNANDA FERREIRA. 2001. Thematic roles assigned along the garden path linger. *Cognitive Psychology* 42.368–407.
- CULBERTSON, JENNIFER, PAUL SMOLENSKY, and COLIN WILSON. 2013. Cognitive biases, linguistic universals, and constraint-based grammar learning. *Topics in Cognitive Science* 5.392–424.

- DEVLIN, JACOB, MING-WEI CHANG, KENTON LEE, and KRISTINA TOUTANOVA. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, volume 1, 4171–4186.
- DILLON, BRIAN, CAROLINE ANDREWS, CAREN M. ROTELLO, and MATTHEW WAGERS. 2019. A new argument for co-active parses during language comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 45.1271.
- DONOHUE, CATHRYN, and MARK DONOHUE. 2004. On the special status of instrumentals. In *Proceedings of the LFG04 Conference, On-line proceedings*, ed. by Miriam Butt and Tracy Holloway King. Stanford: CSLI Publications.
- DOWTY, DAVID. 1991. Thematic proto-roles and argument selection. *Language* 67.547–619.
- DOWTY, DAVID. 2003. The dual analysis of adjuncts/complements in categorial grammar. *Modifying adjuncts* 33.
- FILLMORE, CHARLES J. 1968. *The case for case*. New York: Holt, Rinehart & Winston.
- FORKER, DIANA. 2014. A canonical approach to the argument/adjunct distinction. *Linguistic Discovery* 12.
- GAWRON, JEAN MARK. 1986. Situations and prepositions. *Linguistics and Philosophy* 9.327–382.
- GOLDRICK, MATTHEW, LAUREL BREHM, CHO PYEONG WHAN, and PAUL SMOLENSKY. 2019. Transient blend states and discrete agreement-driven errors in sentence production. In *Proceedings of the Society for Computation in Linguistics (SCiL)*, 375–376.
- GOLDRICK, MATTHEW, MICHAEL PUTNAM, and LARA SCHWARZ. 2016. Coactivation in bilingual grammars: A computational account of code mixing. *Bilingualism: Language and Cognition* 19.857–876.
- GRIMSHAW, JANE. 1990. *Argument structure*. MIT Press.
- GRIMSHAW, JANE, and STEN VIKNER. 1993. Obligatory adjuncts and the structure of events. *Knowledge and Language* 2.143–155.
- HWANG, JENA D., ARCHNA BHATIA, NA-RAE HAN, TIM O’GORMAN, VIVEK SRIKUMAR, and NATHAN SCHNEIDER. 2017. Double trouble: the problem of construal in semantic annotation of adpositions. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM)*, 178–188.
- KIM, NAJOUNG, KYLE RAWLINS, BENJAMIN VAN DURME, and PAUL SMOLENSKY. 2019. Predicting the argumenthood of English prepositional phrases. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI-19)*.
- KIPPER-SCHULER, KARIN. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. University of Pennsylvania dissertation.
- KLIMA, EDWARD S. 1962. Structure at the lexical level and its implication for transfer grammar. In *International Conference on Machine Translation of Languages and Applied Language Analysis*, volume 1, 98–108.
- KOENIG, JEAN-PIERRE, GAIL MAUNER, and BRETON BIENVENUE. 2003. Arguments for adjuncts. *Cognition* 89.67–103.
- LANG, EWALD, CLAUDIA MAIENBORN, and CATHRINE FABRICIUS-HANSEN. 2003. *Modifying adjuncts*, volume 4. Walter de Gruyter.
- LARSON, RICHARD K. 1988. On the double object construction. *Linguistic Inquiry* 19.335–391.
- LAU, JEY HAN, ALEXANDER CLARK, and SHALOM LAPPIN. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science* 41.1202–1241.
- LECUN, YANN, YOSHUA BENGIO, and GEOFFREY HINTON. 2015. Deep learning. *Nature* 521.436.
- LEGENDRE, GERALDINE, YOSHIRO MIYATA, and PAUL SMOLENSKY. 1990. Harmonic grammar—a formal multi-level connectionist theory of linguistic well-formedness: Theoretical foundations. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, 388–395. Lawrence Erlbaum.
- LEWIS, HEATHER. 2004. The *with*-phrase theme in english: Argument or adjunct. Master’s thesis, University of Canterbury.

- MANNING, CHRISTOPHER D. 2003. Probabilistic syntax. *Probabilistic Linguistics* 289–341.
- MERLO, PAOLA, and EVA ESTEVE FERRER. 2006. The notion of argument in prepositional phrase attachment. *Computational Linguistics* 32.341–378.
- NEEDHAM, STEPHANIE, IDA TOIVONEN, MIRIAM BUTT, and TRACY HOLLOWAY KING. 2011. Derived arguments. In *Proceedings of the LFG11 Conference*, 401–421. CSLI.
- PALMER, MARTHA, DANIEL GILDEA, and PAUL KINGSBURY. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics* 31.71–106.
- PATER, JOE. 2009. Weighted constraints in generative linguistics. *Cognitive Science* 33.999–1035.
- PETERS, MATTHEW, MARK NEUMANN, MOHIT IYER, MATT GARDNER, CHRISTOPHER CLARK, KENTON LEE, and LUKE ZETZLEMOYER. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, volume 1, 2227–2237.
- POLLARD, CARL, and IVAN SAG. 1987. *Information-Based Syntax and Semantics*, volume 1. CSLI.
- PUSTEJOVSKY, JAMES. 1995. *The Generative Lexicon*. MIT Press.
- RÁKOSI, GYÖRGY. 2006. On the need for a more refined approach to the argument-adjunct distinction: the case of dative experiencers in hungarian. *Proceedings of the LFG06 Conference*.
- RASMUSSEN, NATHAN E., and WILLIAM SCHULER. 2018. Left-corner parsing with distributed associative memory produces surprisal and locality effects. *Cognitive Science* 42.1009–1042.
- REISINGER, DEE ANN, RACHEL RUDINGER, FRANCIS FERRARO, CRAIG HARMAN, KYLE RAWLINS, and BENJAMIN VAN DURME. 2015. Semantic proto-roles. *Transactions of the Association for Computational Linguistics* 3.475–488.
- RISSMAN, LILIA. 2010. Instrumental *with*, locatum *with* and the argument/adjunct distinction. In *LSA Annual Meeting Extended Abstracts*, volume 1, 23–1.
- RISSMAN, LILIA, KYLE RAWLINS, and BARBARA LANDAU. 2015. Using instruments to understand argument structure: Evidence for gradient representation. *Cognition* 142.266–290.
- ROSEN, ERIC. 2018. Predicting semi-regular patterns in morphologically complex words. *Linguistics Vanguard* 4.
- ROSEN, ERIC. 2019. Learning complex inflectional paradigms through blended gradient inputs. *Proceedings of the Society for Computation in Linguistics (SCiL)* 2.102–112.
- SCHNEIDER, NATHAN, JENA D. HWANG, VIVEK SRIKUMAR, JAKOB PRANGE, AUSTIN BLODGETT, SARAH R. MOELLER, AVIRAM STERN, ADI BITAN, and OMRI ABEND. 2018. Comprehensive supersense disambiguation of English prepositions and possessives. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 1, 185–196. Association for Computational Linguistics.
- SCHNEIDER, NATHAN, VIVEK SRIKUMAR, JENA D. HWANG, and MARTHA PALMER. 2015. A hierarchy with, of, and for preposition supersenses. In *Proceedings of The 9th Linguistic Annotation Workshop*, 112–123.
- SCHÜTZE, CARSON T. 1995. PP attachment and argumenthood. *MIT working papers in linguistics* 26.151.
- SCHÜTZE, CARSON T. 2011. Linguistic evidence and grammatical theory. *Wiley Interdisciplinary Reviews: Cognitive Science* 2.206–221.
- SCHÜTZE, CARSON T., and JON SPROUSE. 2014. Judgment data. *Research methods in linguistics* 27–50.
- SMOLENSKY, PAUL, and MATTHEW GOLDRICK. 2016. Gradient symbolic representations in grammar: The case of french liaison. *Rutgers Optimality Archive* 1286.
- SMOLENSKY, PAUL, MATTHEW GOLDRICK, and DONALD MATHIS. 2014. Optimization and quantization in gradient symbol systems: a framework for integrating the continuous and the discrete in cognition. *Cognitive Science* 38.1102–1138.
- SORACE, ANTONELLA, and FRANK KELLER. 2005. Gradience in linguistic data. *Lingua* 115.1497–1524.
- SPROUSE, JON, and DIOGO ALMEIDA. 2013. The empirical status of data in syntax: A reply to Gibson and Fedorenko. *Language and Cognitive Processes* 28.222–228.

- SPROUSE, JON, BERACAH YANKAMA, SAGAR INDURKHYA, SANDIWAY FONG, and ROBERT C. BERWICK. 2018. Colorless green ideas do sleep furiously: gradient acceptability and the nature of the grammar. *The Linguistic Review* 35.575–599.
- STEEDMAN, MARK. 2000. *The syntactic process*, volume 24. MIT Press.
- STEEDMAN, MARK, and JASON BALDRIDGE. 2011. Combinatory categorial grammar. *Non-Transformational Syntax: Formal and Explicit Models of Grammar*.
- TAKAMI, KEN-ICHI. 1987. Adjuncts and the internal structure of VP. *English Linguistics* 4.55–72.
- TESNIÈRE, LUCIEN. 1959. *Eléments de syntaxe structurale*. Klincksieck.
- TOIVONEN, IDA. 2012. Between arguments and adjuncts. In *International conference of Nordic and General Linguistics*.
- TOIVONEN, IDA. 2013. English benefactive NPs. In *Proceedings of the LFG13 Conference*, ed. by Miriam Butt and Tracy Holloway King, 503–523.
- TSENG, JESSE L., 2001. *Representation and selection of prepositions*. University of Edinburgh dissertation.
- TUTUNJIAN, DAMON, and JULIE E. BOLAND. 2008. Do we need a distinction between arguments and adjuncts? Evidence from psycholinguistic studies of comprehension. *Language and Linguistics Compass* 2.631–646.
- VATER, HEINZ. 1978. On the possibility of distinguishing between complements and adjuncts. *Valence, semantic case and grammatical relations* 1.21–45.
- VENNEMANN, THEO, and RAY HARLOW. 1977. Categorical grammar and consistent basic VX serialization. *Theoretical linguistics* 4.227–254.
- VERSPOOR, CORNELIA M, 1997. *Contextually-dependent lexical semantics*. University of Edinburgh dissertation.
- VESTERGAARD, TORBEN. 1977. *Prepositional phrases and prepositional verbs: a study in grammatical function*. Number 161. de Gruyter Mouton.
- VILLAVICENCIO, ALINE. 2002. Learning to distinguish PP arguments from adjuncts. In *Proceedings of the 6th Conference on Natural Language Learning (CoNLL)*, volume 20, 1–7. Association for Computational Linguistics.
- WECHSLER, STEPHEN. 1995. *The semantic basis of argument structure*. CSLI.
- ZIMMERMANN, EVA. 2019. Gradient symbolic representations and the typology of ghost segments. In *Proceedings of the Annual Meetings on Phonology (AMP)*, volume 7.