Projecting attitudes*

Aaron Steven White
Johns Hopkins University

Valentine Hacquard
University of Maryland

Jeffrey Lidz
University of Maryland

Abstract

This paper explores the granularity with which a word's semantic properties are recoverable from its syntactic distribution, taking propositional attitude verbs (PAVs), such as *think* and *want*, as a case study. Three behavioral experiments aimed at quantifying the relationship between PAV semantic properties and PAV syntactic distribution are reported. Experiment 1 gathers a measure of PAV syntactic distributions using an acceptability judgment task. Experiments 2 and 3 gather measures of semantic similarity between those same PAVs using a generalized semantic discrimination (triad or "odd man out") task and an ordinal (likert) scale task, respectively. Two kinds of analyses are conducted on the data from these experiments. The first compares both the acceptability judgments and the semantic similarity judgments to classifications of PAVs derived from traditional distributional analysis. The second kind compares the acceptability judgments to the semantic similarity judgments directly. Through these comparisons, we show that there is quite fine-grained information about PAV semantics in PAV syntactic distributions—whether one considers the sorts of discrete qualitative classifications that linguists traditionally work with or the sorts of continuous quantitative classifications that can be derived experimentally.

1 Introduction

Words have meanings. Those meanings must be learned. Learning the meaning of some words seems like it could be quite easy. For the moment, assume that learning meanings—e.g. the meaning of the word *dog*—involves pairing some concept or set—e.g. the dog concept or set of dogs, call either DOG—with some linguistic symbol: *dog*. How one goes about doing this, the intuitive story goes, is by noticing that utterances involving the word *dog* cooccur with instantiations of the concept/set DOG quite often, and thus an association between the word and the concept is built. And just so, the meaning of *dog* is learned. In this scenario, the learner's ability to discover correlations between the language and the *nonlinguistic context*—the perceivable objects and events surrounding the hearer—is paramount.

Left unelaborated, this story has well-known problems (cf. Goodman, 1955; Quine, 1960; Kripke, 1982): why not consider subparts of the dog (TAIL, HEAD)? Superordinate categories properly containing the dogs (MAMMAL, ANIMAL)? Or DOG at the time of utterance, CAT every other time? Nonetheless, few would deny that nonlinguistic context plays a major role in learning the meanings of at least some—maybe most—words. How else *would* a learner figure

*We are grateful to audiences at the University of Maryland, Johns Hopkins University, UC Santa Cruz, and NELS 43 as well as Philip Resnik, Naomi Feldman, Hal Daumé III, Alexander Williams, Norbert Hornstein, Colin Phillips, Erin Eaker, Shevaun Lewis, Kaitlyn Harrigan, Naho Orita, Rachel Dudley, Morgan Moyer, Tom Grano, Pranav Anand, Jane Grimshaw, Kyle Rawlins, Ben Van Durme, Jill de Villiers, John Grinstead, Chris Kennedy, Keir Moulton, Paul Portner, Aynat Rubinstein, Meredith Rowe, Florian Schwarz, Toby Mintz, Florian Jaeger, Noah Goodman, Jesse Snedeker, and Elizabeth Bogal-Allbritten for useful comments on the theory, experiments, and modeling contained in this paper. This work was supported by NSF BCS DDRIG grant 1456013, NSF BCS grant 1124338, and NSF DGE IGERT grant 0801465. This manuscript is currently under review at *Cognitive Science*. Comments are welcome and should be addressed to aswhite@jhu.edu.

out that *dog* means DOG? It has become clear that solving these problems requires understanding both the nature of human conceptual understanding—how the learner conceptualizes the nonlinguistic context—and the structure of the mechanism that learners use to link words with concepts—how the learner extracts information from nonlinguistic context. Within the latter vein, there have been many interesting proposals: some involving empirically motivated learning biases (cf. Carey and Bartlett, 1978; Markman and Hutchinson, 1984; Markman and Wachtel, 1988; Merriman and Bowman, 1989; Markman, 1990, a.o.) and others that rely on more general properties of inductive reasoning (cf. Xu and Tenenbaum, 2007; Frank et al., 2009, a.o.). For instance, maybe, as Markman and Wachtel (1988) suggest, children prefer to map words to whole objects—DOG is more salient as a meaning for *dog* than TAIL or HEAD—or, as Tenenbaum and Griffiths (2001) and Xu and Tenenbaum (2007) suggest, maybe they assume that concepts with smaller extensions should be preferred to ones with larger extensions—a sort of "soft" Subset Principle (Berwick, 1985).

1.1 Two problems

But even if learners are equipped with these principles and reasoning mechanisms, learning the meanings of other words is probably quite a bit harder. For example, how do learners acquire those words whose meanings are not obviously linked with features of the nonlinguistic context—or more precisely, conceptualizations of these contexts? The parade case of such words—what Gleitman (1990) refers to as words with meanings "closed to observation" and which Gleitman et al. (2005) dub the *hard words*—are those that refer to abstract objects/concepts (*liberty*, *tyranny*), mental states (*think*, *know*), preferences (*want*, *prefer*), authorizations (*allow*, *forbid*), etc. Many of these hard words are verbs involving *propositional attitudes*, which express relations to ways the world might be, in fact is, would be best if it were, etc. It is these hard words that this paper focuses in on.

1.1.1 The problem of observability

One problem with hard words is that one cannot very well see, hear, or feel propositional attitudes like thinkings or wantings, so it is quite unclear how the learner pairs up words for these attitudes—*think* or *want*—with the appropriate concepts—for now, call them THINK and WANT—under an account where correlations between particular words and nonlinguistic context are the primary (or only) data (Landau and Gleitman, 1985; Gleitman, 1990).

There is now a wealth of experimental results evidencing the magnitude of this problem. One particular instance of this can be found in work within the Human Simulation Paradigm (HSP; Gillette et al. 1999; Snedeker and Gleitman 2004). In one instantiation of this paradigm, adult participants are given videos of parents playing with their children. In these videos the sound has been removed, with the idea that this partially replicates the learner's nonlinguistic context. A beep is then placed where a target word was uttered, and participants are asked to say what the word is. Accuracy is quite high in recovering concrete nouns, like *dog*, but essentially zero in recovering mental state verbs, like *think* or *want*.

What this suggests is that desires and beliefs are just not salient as potential word meanings from the nonlinguistic context alone—even for adults, who are constantly talking about desires and beliefs. And indeed, further work within this paradigm suggests that, even if scenes are constructed to make propositional attitudes salient, gains from nonlinguistic context alone are only modest at best (Papafragou et al., 2007).

1.1.2 The problem of multi-faceted meanings

The problem of observability is sharpened by the fact that hard words also tend to have meanings that are multifaceted. For instance, note that hopings seem to involve wantings—if (1b)

VERB	BELIEF	DESIRE
want		✓
hope	✓	✓
think	✓	

Table 1: An example of multi-faceted propositional attitude verb meanings.

is true, (1a) must also be true. But *hope* seems to have an extra facet of its meaning over and above the component it shares with *want*. If one first utters (1), it is fine to follow with (1a) but odd to follow with (1b).

(1) Bo believes he'll never dance with Jo, but...
   a. Bo nonetheless wants to dance with Jo.
   b. #Bo nonetheless hopes to dance with Jo.

Thus, *hope* appears to have two facets to its meaning. Like *want*, it involves desire; but unlike *want*, it also involves belief—namely, it requires that the hoper believe that the state of affairs they want to come about is also possible (Portner, 1992; Scheffler, 2009; Anand and Hacquard, 2013; Hacquard, 2014; Harrigan, 2015).[1]

This gives rise to a potentially general learning problem: even if the nonlinguistic context is sufficient to learn the meaning of, e.g., *want* or *think*, a learner who also posits the meaning of *want* for the meaning of *hope* might very well find herself with a subset problem (Wexler and Hamburger, 1973; Baker, 1979; Berwick, 1985; Pinker, 1989); *want* will always be true in the contexts that *hope* is true in, so an account for how a learner selects the correct meaning of, e.g., *hope* is necessary.

1.2 Solving the two problems

But then how *do* learners acquire this constellation of facts about verbs like *want* and *hope*? There are various possible answers to this question. The one we pursue in this paper is that learners need to move beyond nonlinguistic context as their only source of evidence for word-learning; they need to furthermore incorporate a word's *linguistic context*. To understand what this means, it is useful to step back and consider a rough but useful division that exists in the literature between two broad types of linguistic context: *lexical context* and *syntactic context*.

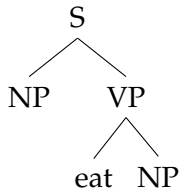
1.2.1 The importance of syntactic context

To a first approximation, lexical context encompasses words that cooccur with the one being learned, and syntactic context encompasses the types of abstract structures the word is found in. For instance, suppose a learner heard (2). The lexical context of *eat* might be represented as in (2a). And assuming the learner can parse the string in an adult-like way, its syntactic context might be represented as in (2b).

(2) The men eat apples.
   a. {men, the, apples}

[1] In her seminal paper on presupposition, Heim (1992) gives a denotation for *want* that explicitly encodes belief, so whether *want* involves belief or not is something of an open question. In any case, there is a clear contrast between (1a) and (1b) as follow-ups to (1)—a contrast whose source must be learned.

b.



If linguistic context is necessary for learning words like *want*, *hope*, *think*, and *know*—i.e. those whose meanings are not associated with sensory correlates—which subtype of information—lexical context or syntactic context—might be used? The likely answer to this is that both are necessary in interaction, to different extents, for different kinds of verbs. However, authors differ on whether syntactic context could ever be necessary independent of lexical context. For instance, Pinker (1994) and Grimshaw (1994) both argue that the only sense in which syntactic context might be useful is in interaction with lexical context. Knowing that *eat* tends to take NPs referring to edibles, like *apples*, in object position could plausibly help a learner figure out that *eat* means EAT. But most of the information in this case is coming from a semantic generalization—*commonly occurs with words referring to edibles*—and so it is unclear what further work, if any, the syntactic context is doing here. Indeed, if the learner can (i) figure out what role the referent of each noun phrase plays in the event named by the verb—via only the semantics of those noun phrases—and (ii) there are some general constraints regarding where each argument may be syntactically situated given its entailments (cf. Baker, 1988; Grimshaw, 1990; Dowty, 1991), the syntactic context may be doing little work beyond highlighting the lexical material on which the learning mechanism should make its inferences (cf. Connor et al., 2013).

We take this to be a reasonable position for verbs like *eat*, whose syntactic contexts likely contain very little information about its meaning beyond that it involves two participants. (In fact, due to the presence of intransitive uses such as *the men ate*, unilateral reliance on the syntax might even lead learners astray.) Less clear is whether this strategy of seeding inference with only lexical context could be extended to all verbs—especially propositional attitude verbs like *think*, *know*, and *want*. For instance, it seems unlikely that a learner could glean much at all about the meaning of *want* from the distribution of nouns it occurs with, since *want* imposes few to no restrictions on its direct object's meaning (cf. Resnik, 1996, p. 138, Table 1). And this is not specific to *want*; many propositional attitude verbs that allow NP direct objects do not constrain the semantics of those objects.

(3) Bo wants {an apple, a toy, a back rub}.

(4) Bo {knows, remembers, needs, demands} {a doctor, a story, the time}.

Thus, it seems that, for at least some distinctions among propositional attitude verbs, neither nonlinguistic context nor lexical context are likely to help in drawing fine-grained distinctions among propositional attitude verbs, and so a final possibility remains within the rough taxonomy: the sort of distinctions we have been discussing must be learned using syntactic context. How might the learner do this?

The answer that Landau and Gleitman (1985) and Gleitman (1990) propose under the heading *syntactic bootstrapping* is that children use the syntactic contexts that a word cooccurs with—its *syntactic distribution*—to deduce its meaning.[2] The proposal moves forward in the following

[2]This is at least the accepted genealogy of the idea. As a historical note, this was actually proposed earlier in Lasnik 1989, a paper in the proceedings of the 1982 University of Western Ontario Learnability Workshop. The relevant quote:

> ...there appears to be a tacit assumption that the meaning of, e.g., a verb, can be presented and apprehended in isolation. But this seems implausible. Rather, verbs are presented in grammatical sentences which, therefore, explicitly display subcategorization properties. In fact, one might consider reversing the whole story: subcategorization is explicitly presented, and the child uses that information to

way. Suppose learners make the following assumption: the degree to which two verbs overlap in their syntactic contexts correlates with the degree to which their meanings overlap. They then note that the syntactic contexts of *want* and *hope* only partially overlap. *Want*, but not *hope*, allows noun phrase objects (5a); both *want* and *hope* allow subjectless infinitival complements (5b) (control complements); and *hope*, but not *want*, allows tensed subordinate clause complements (5c).

(5) a. Bo {wants, *hopes} an apple.
b. Bo {wants, hopes} to have an apple.
c. Bo {*wants, hopes} that he will have an apple.

Finally, based on these data and the premise that overlap in syntactic contexts correlates with overlap in meaning, they might then infer that *want* and *hope* may share some facet(s) of their meanings, but not others.

Now, this on its own is insufficient. Knowing that there is an overlap in meaning does not yet indicate what that particular overlap is. Thus, this story needs to be augmented to explain not only how to find out whether there's an overlap, but also to say how that overlap gets labeled with the appropriate facet or feature of the words' meanings. The learner needs to know not only that *want* and *hope* share a meaning feature, but furthermore what that shared feature is. We refer to the first of these problems—finding out that there is an overlap—as the *clustering problem* because it deals with finding out that particular words cluster together with respect to the syntactic contexts they occur in. We refer to the second of these problems—finding out what facet of the meaning a particular clustering of verbs corresponds to—as the *labeling problem* because it deals with labeling the clusters of words that are found.[3]

The traditional solution to both problems within the syntactic bootstrapping literature is to assume that whatever mechanism solves the clustering problem simultaneously solves the labeling problem—by associating particular syntactic contexts with particular semantic features.[4] Thus, under a standard syntactic bootstrapping account, the learning mechanism comes "pre-built" with "...grammatical knowledge [that] includes principles that provide a systematic mapping between semantic and syntactic structures" (Lidz et al., 2004), and this systematic mapping is deployed to label clusters associated with particular syntactic features. The upshot of this account is that a syntactic bootstrapping mechanism's efficacy is directly tied to (i) the "granularity" of the semantic structures that these principles make reference to and (ii) the amount of information preserved in the mappings.

1.2.2 Syntactic bootstrapping and linguistic theory

Linguists have a long-standing interest in these questions of granularity and information preservation—an interest that Zwicky (1971) distills quite elegantly in the introduction to his classic squib on manner-of-speech verbs.

> To what extent is it possible to predict certain properties of words (syntactic, semantic, or phonological), given others? [And] insofar as there are such dependencies among properties, what general principles explain them? (ibid., p. 223)

deduce central aspects of the meaning of verbs. (ibid., p. 195, fn. 12)

[3]The terminological choices *clustering problem* and *labeling problem* belie a particular view of the problem a learner faces—namely, that at base, the learner must discover symbolic relationships among verbs' meanings. This does not preclude a learning model that imputes continuous representations to the learner, as long as those representations are somehow linked to symbolic representations.

[4]As noted by Kako (1997) as well as Lidz et al. (2004), this might be cashed out in a couple of different ways—either by imbuing syntactic contexts themselves with semantic content (the *Frame Semantic Hypothesis*) or by imbuing them with semantic content inherited from the verb's semantic features (the *Lexical Projection Hypothesis*).

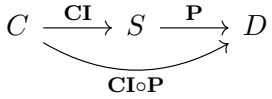


Figure 1: The mapping $\mathbf{CI} \circ \mathbf{P}$ from the concept space C to syntactic distribution space D, mediated by the grammatically relevant semantic features S.

Indeed, it has long been recognized that questions regarding which semantic distinctions are morphosyntactically (grammatically) relevant are the only ones linguists can claim propriety over; distinctions in meaning beyond those predictable from other linguistic properties fall equally well into the domain of the lexicographer (Fillmore, 1970)—or in modern times, the computer scientist (Kilgarriff, 1997). Embedded in this view is the idea that a lexical item's linguistic contexts are responsive to only some conceivable contrasts in meaning and that a linguistic theory of the link should speak to exactly which these are and why other conceivable contrasts are excluded (cf. Jackendoff, 1972; Grimshaw, 1979; Pinker, 1989; Levin, 1993). As Zwicky puts it, the question for the linguist is "what sorts of word classes are there, and why these and not others?" (ibid., p. 223)

An example of the distinction between grammatically relevant and linguistically irrelevant semantic distinctions comes from Pesetsky (1991). Following Zwicky, he notes that, though "verbs of manner of speaking"—e.g. *holler* and *whisper*—and "verbs of content of speaking"—e.g. *say* and *propose*—are distributionally distinguishable, "verbs of loud speech"—e.g. *holler* and *shout*—and "verbs of soft speech"—e.g. *whisper* and *murmur*—do not seem to be. (For example, verbs of content of speaking "resist adjunct extraction and allow complementizer deletion" (Pesetsky, 1991, p. 14).) That is, the manner-content distinction has consequences for the syntax, whereas the loud-soft contrast does not.

In fact, the generalization extends beyond predicates that refer to speech sounds to predicates that refer to sounds in general. The volume, pitch, resonance, and duration of the relevant sound do not seem to have bearing on its distributional properties, but the mode of generation (internally v. externally caused) does (Levin and Rappaport Hovav, 2005). This suggests that, whatever constitutes the nature of the connection between a word's semantic properties and its syntactic distribution, it is blind to certain possible conceptual distinctions—in this case, sonic properties.

Thus, though nonlinguistic meanings—i.e. concepts—may be distinguishable to a very fine grain-size, linguistic meanings may not be. In this sense, the linguistic system can be conceived of as a filter on the properties of the conceptual system's objects, retaining some properties while discarding others. To introduce a convention we use throughout the paper, suppose $\mathbf{c}_i$ is some representation of concept i, then $\mathbf{s}_i$ is some representation that encodes all and only the grammatically relevant features of $\mathbf{c}_i$.[5] At a high level of abstraction, then, (part of) the interface between language and other areas of cognition—in particular, the Conceptual-Intentional (CI) interface—might be viewed as a mapping $\mathbf{CI}$ from objects in the concept space C to their syntactically relevant features in the semantic feature space S.[6]

[5]We use the following conventions throughout the remainder of the paper: italicized capital letters stand for representational spaces—i.e. possible representations; normal capital letters refer to mappings between these spaces; bolded lower-case letters, which will tend to be subscripted, refer to particular instantiations of the corresponding space, and bolded upper-case letter refer to collections of these specific instantiations. The bolding convention in particular is used because representational instantiations are cashed out as vectors and their collections as matrices or tensors, and bolding is standard in linear algebra and related disciplines for representing vectors and their generalizations.

[6]This presupposes a contentious point about the complexity of lexical items (cf. Fodor and Lepore, 1998, 1999). While Fodor and Lepore's arguments are serious, we believe that a reader who abides by the dictum that lexical items be represented atomically might still find use in this paper. We take it as *prima facie* reasonable that representations of the regularities in a word's syntactic distribution may be complex in a way that the representation of its content may not be, since of course, one needs to explain C-selection somehow (cf. Grimshaw, 1979; Pesetsky,

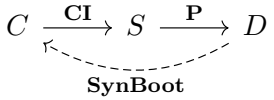


Figure 2: The function of the syntactic bootstrapping mechanism is to "reverse" **CI** and **P**.

It is worth stressing the following implication: $\mathbf{s}_i$ may not exhaust the semantic representation of word i; indeed, $\mathbf{s}_i$ may not even be semantic in any important sense. For instance, it might be conceived of as a (structured) index into subsets/subspaces of concepts. Thus, knowing $\mathbf{s}_i$ for a word i would likely be insufficient for fixing that word's corresponding concept $\mathbf{c}_i$.

By definition, however, $\mathbf{s}_i$ would be sufficient for determining various linguistic properties of word i, such as its syntactic distribution $\mathbf{d}_i$. To say that the syntactic distribution $\mathbf{d}_i$ of word i can be determined from its grammatically relevant semantic features $\mathbf{s}_i$ is to say that there is some mapping from the space of possible semantic representations S to the space of syntactic distributions D. In standard models of the syntax-semantics interface, this mapping—call it **P**—is determined by a set of *projection rules* (cf. Gruber, 1965; Carter, 1976; Chomsky, 1981; Pinker, 1989; Grimshaw, 1990; Levin, 1993; Hale and Keyser, 2002).

Putting these two components together—the mapping from the conceptual space C to the (grammatically relevant) semantic feature space S and the mapping from the semantic feature space S to the syntactic distribution space D—the abstraction in Figure 1 results. If this model is correct, the upshot for a theory of verb-learning that relies on syntactic context—e.g. syntactic bootstrapping—is that there is likely a limit on the meaning properties that syntactic context could be used to learn even in principle. Why? Suppose the learner has access to the syntactic distribution $\mathbf{d}_i$ for some word i and that their job is to infer the concept $\mathbf{c}_i$ association with word i. That is, they need to "reverse" both the projection rules **P** and the mapping from the conceptual space to semantic features **CI**, schematized in Figure 2.[7]

A learner's ability to perform this reversal from the syntax alone will necessarily be bound by the information lost due to the mapping from semantic features to syntactic distributions (**P**)—i.e. the projection rules—and the information lost due to the mapping from concepts to semantic features (**CI**). This gives Zwicky's question new force.

The traditional methodology for approaching this question is to assume knowledge of pairings $(\mathbf{c}_i, \mathbf{d}_i)$ for many words i and then to attempt to (qualitatively) infer both the space of (grammatically relevant) semantic representations S and the projection rules **P**. Thus, making the simplifying assumption that learners have access to the same conceptual space C and the syntactic distributions of some words **D**, the linguist and the learner look very much alike. Under this model, the only difference between the two is that the learner does not have access to the pairing of the concept $\mathbf{c}_i$ and syntactic distribution $\mathbf{d}_i$ for word i; rather, they have the syntactic distribution $\mathbf{d}_i$ of word i and a space of possible concepts C. Looked at from a machine learning perspective, the linguist carries out some supervised learning algorithm to infer S and the learner carries out some unsupervised—or perhaps, *distantly supervised* (Snow et al., 2004; Mintz et al., 2009)—learning algorithm.

Casting the learner and the linguist within the same framework in this way is useful, since it suggests a way of approaching the question of what can be learned from syntactic context: whatever grammatically relevant semantic features the linguist discovers could in principle be deducible by a learner with access to the linguistic features that those semantic features are relevant to—e.g. syntactic distributions. To get a handle on propositional attitude verb learning, then, it is useful to review what is known from the syntax-semantics literature about

1982). One way to cash this out might be to view S as a space of (structured) indices into subsets/subspaces of (atomic) concepts.

[7] In the remainder of this paper, we use the following convention: solid lines represent theoretical, or *computational-level* (Marr, 1982), relationships; dashed lines represent algorithms that map between representations—as in the case of syntactic bootstrapping, possibly utilizing the computational-level relationships.

the semantic features that are grammatically relevant. We carry out this review in the next section.

1.3 Propositional attitude verb syntax and semantics

In this section, we present a brief overview of the literature on the syntax and semantics of propositional attitude verbs. The upshot of this section is that, though there appear to be some possibly useful correlations between attitude verb syntax and semantics—correlations that are useful for constructing our experiments and conducting analysis—it remains unclear how robust these correlations actually are. This uncertainty about the robustness of such correlations forms the impetus for the experiments and computational modeling carried out in the remainder of the paper.

1.3.1 Representationality

Perhaps the most well-known semantic distinction among propositional attitude verbs is that between verbs that express beliefs—or represent "mental pictures" or "judgments of truth" (Bolinger, 1968)—and those that express desires—or more generally, orderings on states of affairs induced by, e.g. commands, laws, preferences, etc. (Bolinger, 1968; Stalnaker, 1984; Farkas, 1985; Heim, 1992; Villalta, 2000, 2008; Anand and Hacquard, 2013, a.o.). Within the first class, which we henceforth refer to as the representationals, fall verbs like *think* and *know*; and within the second class, which we henceforth refer to as the preferentials, fall verbs like *want* and *order*.

There appear to be various aspects of the syntactic distribution that roughly track this distinction in English. One well-known case is finiteness: representationals tend to allow finite subordinate clauses (6a) but not nonfinite ones (6b); preferentials tend to allow nonfinite subordinate clauses (7b) but not finite ones (7a).[8]

(6) a. Bo thinks that Jo went to the store.
b. *Bo thinks Jo to go to the store.

(7) a. *Bo wants that Jo went to the store.
b. Bo wants Jo to go to the store.

There are two important things to note about this distinction. First, though the representationality distinction is often talked about as though it were mutually exclusive, some verbs appear to fall into both categories, and suggestively, show up in both frames. For instance, as noted in the last section, *hope p* involves both a desire that *p* come about and the belief that *p* is possible (Portner, 1992; Scheffler, 2009; Anand and Hacquard, 2013; Hacquard, 2014; Harrigan, 2015, but see also Portner and Rubinstein 2013), and it occurs in both finite (8a) and nonfinite (8b) syntactic contexts.

(8) a. Bo hopes that Jo went to the store.
b. Bo hopes to go to the store.

Second, the link between representationality and finiteness is just a tendency. Some verbs plausibly classed as representationals allow nonfinite subordinate clauses (9a)/(9b), and others plausibly classed as preferentials allow subordinate clauses that look finite (9c).[9]

[8] This correlation is computationally corroborated to some extent in Barak et al. 2013, 2014, though caution is required here since Barak et al. investigate only a small set of verbs and ignore various complications, discussed in the text, inherent to this distinction.

[9] Whether (9c) involves a finite subordinate clause is to some extent dependent on whether what is often called the English subjunctive involves tense. On the one hand, the complementizer *that* is the same one that occurs with tensed subordinate clauses, but on the other, the verb shows up in its base (untensed) form.

The roughness of this correlation is perhaps not surprising since not all languages track representationality with tense: for instance, various Romance languages track the distinction with mood—representationals tending to take indicative mood and preferentials tending to take subjunctive mood (Bolinger, 1968; Hooper, 1975; Farkas, 1985; Portner, 1992; Giorgi and Pianesi, 1997; Giannakidou, 1997; Quer, 1998; Villalta, 2000, 2008, a.o.).

(9) a. Bo believes Jo to be intelligent.
b. Bo claims to be intelligent.
c. Bo demanded that Jo go to the store.

But though the correlation between representationality and tense is imperfect, even in English, finiteness does not appear to be the only associated syntactic (distributional) property. Also relevant appears to be a distinction in whether the verb's subordinate clause can be fronted—or in Ross's (1973) terms, S-lifted.[10] At least some representationals' subordinate clauses (10) appear to be able to undergo S-lifting, but many preferentials' subordinate clauses (11) cannot (Bolinger, 1968). We return to this question of cross-linguistic variation in Section 4.

(10) Jo already went to the store, I {think, believe, suppose, hear, see}

(11) a. *Bo already went to the store, I {want, need, demand}.
b. *Bo to go to the store, I {want, need, order}.

(Not all representationals allow S-lifting. This is likely because the availability of S-lifting for a particular verb is conditioned by other semantic and pragmatic properties it has, so we defer further discussion of which verbs allow it until distinctions beyond representationality have been discussed.)

1.3.2 Factivity

The representationality distinction is cross-cut by another common distinction: factivity (Kiparsky and Kiparsky, 1970; Karttunen, 1971; Horn, 1972; Hooper, 1975). Factivity is defined in terms of its discourse effects. Very roughly, a verb is factive if upon uttering a sentence containing a factive verb with a subordinate clause, a speaker takes the content of the subordinate clause for granted regardless of propositional operators placed around the propositional attitude verb: in particular, negation (13b)/(12b) or questioning (13c)/(12c). For instance, each sentence in (12) commits the speaker to (14) being true, but modulo the context, the sentences in (13) do not. That is, in uttering the sentences in (12), the speaker presupposes (14) (Stalnaker, 1973). This suggests that *know*, *love*, and *hate* are factive, while *think*, *believe*, and *say* are not.

(12) a. Bo {knew, loved, hated} that Jo went to the store.
b. Bo didn't {know, love, hate} that Jo went to the store.
c. Did Bo {know, love, hate} that Jo went to the store?

(13) a. Bo {thought, believed, said} that Jo went to the store.
b. Bo didn't {think, believe, say} that Jo went to the store.
c. Did Bo {think, believe, say} that Jo went to the store?

(14) Jo went to the store.

Factivity truly cross-cuts the representationality distinction in that there are verbs representing all four possible combinations: (i) representational (cognitive) factives, like *know*, *realize*, and

[10]There is a further distinction in the literature made between S-lifts involving first person and third person propositional attitude verb subjects (Reinhart, 1983; Asher, 2000; Rooryck, 2001). We incorporate this first-third distinction into our experiment, but the data regarding this syntactic distinction are murky at best.

understand, (ii) preferential (emotive) factives, like *love* and *hate*, (iii) representational nonfactives, like *think* and *say*, and (iv) preferential nonfactives, like *want* and *prefer*.[11]

The factivity distinction appears to be tracked most closely by whether the verb allows both question and nonquestion subordinate clauses (Hintikka, 1975; Ginzburg, 1995; Lahiri, 2002; Sæbø, 2007; Egré, 2008; Uegaki, 2012; Anand and Hacquard, 2014; Spector and Egré, 2015). For instance, the factive *know* can occur with both nonquestion (15a) and question (15b) subordinate clauses, while the nonfactive *think* can occur with nonquestion subordinate clauses (16a) but not question subordinate clauses (16b).[12]

(15) a. Jo knows that Bo went to the store.
b. Jo knows {if, why} Bo went to the store.

(16) a. Jo thinks that Bo went to the store.
b. *Jo thinks {if, why} Bo went to the store.

This generalization has two well-known types of exceptions. First, many nonfactive communication predicates, such as *tell* and *say*, allow both question and nonquestion subordinate clauses; second, some mental predicates, such as *decide*, *assess*, and *evaluate*, also allow both question and nonquestion subordinate clauses.

(17) a. Jo hasn't {told me, said} whether Bo went to the store.
b. Jo hasn't yet {decided, assessed, evaluated} whether to go to the store.

1.3.3 Assertivity

Further cross-cutting representationality and factivity is the "assertivity" distinction (Hooper, 1975).[13] Like factivity, assertivity is defined in terms of its effects on discourse. Again very roughly, a verb is assertive if it can be used in situations where its subordinate clause is relevant to the main point of the utterance (see Urmson 1952; Simons 2007; Anand and Hacquard 2014 for discussion). For instance, *think* and *say* seem to allow this (18a), but *hate* does not (18b).

(18) a. **A:** Where is Jo?
B: Bo {thinks, said} that she's in Florida.
b. **A:** Where is Jo?
B: # Bo hates that she's in Florida.

Assertivity correlates with the availability of S-lifting and the propositional anaphor object *so*. Assertives, like *think* and *say*, can occur with S-lifted subordinate clauses (19a) and *so* (20a), but *hate* cannot occur with either S-lifting (19b) or *so* (20b).

(19) a. She's in Florida, Bo {thought, said}.
b. *She's in Florida, Bo hated.

[11] One question that arises here is whether, given the existence of representational+preferential verbs like *hope*, there could also be such representational+preferential factives. In a certain sense, this may be the case for the emotive factives, since it seems like sentences containing them imply that the holder of the emotion also believes the subordinate clause to be true. If all preferential factives are emotive (and show this behavior), this might suggest that there are no non-representational factives. One must tread carefully here, however, since not all entailments need be encoded in the meaning of the verb—i.e. this belief entailment could plausibly arise via the same sorts of pragmatic processes that give rise to the factive presupposition in the first place. In the remainder of this paper, we treat all emotives—factive (e.g. *love*, *hate*) or non-factive (e.g. *hope*, *worry*)—as both representational and preferential, since we believe it to be the most consistent treatment for our purposes, but we recognize the subtlety.

[12] This paradigm is filled out by what Lahiri (2002) calls rogatives, like *wonder* and (for some speakers) *ask*. *Wonder*, at least, takes only subordinate questions and not nonquestions.

[13] Whether assertivity fully cross-cuts representationality is unclear, since the only verbs that have both a preferential component and are plausibly assertive are the emotive doxastics (e.g. *worry*, *hope*, *fear*) which also arguably have a representational component.

(20) a. Bo {thinks, said} so.
b. *Bo hates so.

Hooper (1975) claims that the assertivity distiction cross-cuts the factivity distinction to give rise to a further split between semi-factives (assertive factives), like *know*, and true factives (nonassertive factives), like *love* and *hate* (see Karttunen 1971 for an early description of this distinction).[14] Important for our purposes is that the semi-factive v. true factive distinction appears to correlate (i) with the (semantic) representationality distinction—semi-factives also tend to be cognitive factives and true factives, emotive factives—and (ii) at least two sorts of syntactic distinctions.

First, semi-factives tend to allow both polar (21a) and WH (21b) questions, but true factives tend to allow only WH questions (22b), not polar questions (22a) (Karttunen, 1977).[15]

(21) a. Jo knows if/whether Bo sliced the bread.
b. Jo knows if/whether Bo sliced the bread.

(22) a. *Jo {loves, hates} if/whether Bo sliced the bread.
b. Jo {loves, hates} how Bo sliced the bread.

Guerzoni (2007) notes, however, that this correlation is not perfect, since some canonical semi-factives like *realize* resist polar questions in many contexts.

Second, semi-factives tend to allow complementizer ommission (23a), but true factives tend not to (23b). This second correlation is less strong and is likely modulated by syntax: expletive subject emotive factives appear to be better with complementizer omission, particularly when they passivize (see Grimshaw 2009 for further recent discussion of complementizer ommission).

(23) a. I {know, realize} (that) Jo already went to the store.
b. I {hate, love} *(that) Jo already went to the store.

(24) a. It {amazed, bothered} me ???(that) Jo already went to the store.
b. I was {amazed, bothered} ?(that) Jo already went to the store.

1.3.4 Communicativity

Communicativity, transparent from its name, roughly corresponds to whether a verb refers to a communicative act, or perhaps more generally, (manner of) externalization of linguistic form. This distinction cross-cuts at least the representationality distinction—there are both representational communicatives, like *say* and *tell*, and preferential communicatives, like *demand*—and perhaps other distinctions as well, such as the factive-nonfactive distinction (see Anand and Hacquard 2014 for extensive discussion of whether communicativity truly cross-cuts factivity or not).[16]

The syntactic correlates of communicativity seem quite apparent on the surface. Communicative verbs, along with a subordinate clause, tend to take noun phrase (25a) or prepositional phrase (25b) arguments representing their communicatee (Zwicky, 1971).

[14]The pragmatic effects that distinguish semi-factivity from true factivity are beyond the scope of this paper. Much ink has been spilled regarding the nature of semi-factivity in recent years, however, so the interested reader is encouraged to see, e.g., Simons 2001; Abusch 2002; Abbott 2006; Romoli 2011.

[15]Kyle Rawlins (p.c.) points out that the traditional description of this correlation concerns cognitive v. emotive factives and not necessarily semi- v. true factives. At least in the verbal domain (and possibly outside it), these two descriptions seem to be extensionally equivalent, though we recognize that which distinction is relevant does matter (cf. Abels, 2004).

[16]Whether *say* and *tell* are only representational is a question. Both can be used to talk about commands conditional on their taking a nonfinite subordinate clause. In any case, they plausibly have something like a representational use with finite subordinate clause.

(25) a. Bo told me that Jo went to the store.
b. Bo said to me that Jo went to the store.

But though this is often treated as a clearly marked distinction, there are various reasons to be cautious about it. For instance, note that *demand* and *tell* can occur in string-identical contexts with *want* and *believe*. These string-identical contexts appear to be be distinguished only given some parse of the string. Wanting and believing don't seem to involve anything besides a wanter/believer and a thing wanted/believed. In contrast, telling and demanding seem to require a tellee/demandee.

(26) Bo {told, demanded, wanted, believed} Jo to be happy.

This is plausibly syntactically encoded. Note that the pleonastic element *there*, which is plausibly an overt cue to the particular syntactic configuration in question, is only allowed with *want* and *believe*, but not *tell* and *demand*. This has been used to suggest that *tell* and *demand* in (26) involve an underlying object while *want* and *believe* do not.

(27) Bo {*told, *demanded, wanted, believed} there to be a raucous party.

Further, there are some string-identical contexts that both communicative and noncommunicative verbs can appear in which plausibly have no syntactic (or perhaps even selectional) distinctions. For instance, the communicative verb *promise* and the verb *deny*, which is plausibly noncommunicative in this syntactic context, both allow constructions with two noun phrases.

(28) Bo {promised, denied} Bo a meal.

This is not to say that the semantic distinction has no syntactic correlates, of course; it is just to say that they may not be apparent from the string context.

1.3.5 Perception

The final class we consider is the perception predicates, like *see*, *hear*, and *feel*. These predicates form a somewhat small class, and all allow bare verb phrase subordinate clauses (29a) and gerundive verb phrase subordinate clauses (29b).

(29) a. Jo {saw, heard, felt} Bo leave.
b. Jo {saw, heard, felt} Bo leaving.

While bare verb phrase subordinate clauses are relatively distinctive of perception predicates, gerundive verb phrase subordinate clauses can occur with predicates that do not clearly involve perception, such as *remember*.

(30) Jo remembered Bo leaving.

1.3.6 Discussion

In this section, we briefly reviewed the literature on the syntax and semantics of propositional attitude verbs. One thing we note throughout the section is that, though there appear to be some possibly useful correlations between attitude verb syntax and semantics, in many cases, it remains unclear how robust these correlations actually are. This raises two questions: in cases where the correlations are only partial, (i) is this because the syntax really only partially correlates with that distinction or have some other correlations with that distinction been missed?; and (ii) is the distinction in question even the best one to generalize from or is there some other distinction correlated with it that is better?

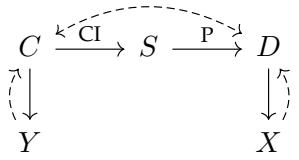


Figure 3: Schematization of Fisher et al.'s methodology in relation to the model presented above.

To some extent these questions can be answered using traditional distributional analysis of the sort reviewed above. To address the first question, one holds constant the semantic properties in question—e.g. representationality, factivity, etc.—and then investigates more and more fine-grained hypotheses about the syntactic distributions that correlate with them. To address the second question, one holds the hypotheses about the syntactic distributions constant—e.g. by investigating verbs that take both question and nonquestion subordinate clauses—and then tests more fine-grained hypotheses about the semantic properties of those verbs.

This seems to be exactly the right tack. The worry, however, is that these finer-grained hypotheses will be very difficult to test using informal (syntactic or semantic) judgments alone due to the inherent uncertainty that lies in these judgments when given by only a single researcher. In the next section, we discuss a method for remedying this worry by augmenting traditional distributional analysis with quantitative tools.

1.4 Augmenting traditional distributional analysis

In their seminal paper, Fisher et al. (1991) present a similar worry about traditional distributional analysis. They argue that

> ...only those semantic generalizations that can be readily labeled by the investigator are likely to be discerned. It may well be that there are semantic abstractions which, while correlated with the syntax, are not so easy to puzzle out and name. (p. 342)

This methodological problem arises, they argue, as a consequence of confounding isolation of a property and labeling of that property, since "...disagreements over labels for semantic features can get in the way of deciding whether those features are marked in the syntax" (ibid, p. 342). Note that this is analogous to the labeling problem (discussed above): even assuming syntactic distribution is attended to, how does a learner link the appropriate features of that distribution (syntactic contexts) to the appropriate meaning components?

Their methodological solution has two components: (i) some method for independently measuring (a) the array of syntactic contexts a verb i can occur in (its syntactic distribution) and (b) that verb's meaning; and (ii) some way of comparing these two measures. Within the above abstraction, (i) involves measuring the syntactic distributions $\mathbf{D}$ and the concepts $\mathbf{C}$—likely indirectly—while (ii) involves specifying a method for (a) reconstructing $\mathbf{C}$ and $\mathbf{D}$ and (b) comparing these reconstructions. This methodology is schematized in Figure 3.

To implement these components, Fisher et al. begin by obtaining, for a set of verbs spanning the lexicon, semantic similarity judgments for those verbs—call the resulting data $\mathbf{y}$. The idea here is that such quantitative representations allow one to bypass the sort of explicit labeling inherent to the traditional method, since distinctions among features salient to the participants are not explicitly invoked.[17]

[17]There is a question here to what extent the semantic proxy $\mathbf{y}$ is solely dependent on $\mathbf{C}$ and not, e.g., $\mathbf{D}$ itself. Fisher et al. give various arguments that $\mathbf{y}$ is plausibly the product of participants utilizing some aspects of the meaning of the words in the task independently of the correspond syntactic distributions $\mathbf{D}$. As far as we can

Their goal is to then compare this proxy $\mathbf{y}$ with a quantitative representation of those verbs' syntactic distributions gathered using an acceptability judgment task.[18] We refer to these sorts of quantitative representations as $\mathbf{x}$, an approximation of $\mathbf{D}$. Fisher et al.'s question is then how well the semantic similarity judgments $\mathbf{y}$ and acceptability judgments $\mathbf{x}$ match up and in what ways do they match up.

Fisher et al. use this methodology to study the high-level correlations between semantics and syntax, using a fairly coarse-grained notion of syntactic frame that does not distinguish among the various syntactic features that are potentially relevant to attitude verb semantics. In the remainder of this paper, we deploy Fisher et al.'s methodology to investigate these correlations at a more fine-grained level.

1.5 Discussion

We began this section by reviewing the problem of hard words in word-learning, focusing on a particular subset of such words: the propositional attitude verbs. We discussed two problems for attitude verb learning: the observability problem and the multi-faceted meaning problem. We then reviewed the now-standard solution to this problem proposed in the syntactic bootstrapping literature. We noted that, though this solution is fairly standard, little is known about how effective it could be for attitude verb learning, since it is still unclear how strong the correlations between attitude verb syntax and semantics are.

In this paper, we fill this lacuna by employing a methodology originally used by Fisher et al. to study high-level correlations between verb syntax and semantics. In Section 2, we present an experiment aimed at measuring the acceptability of a variety of propositional attitude verbs in different syntactic contexts. There, we ask how well the classification laid out in Section 1.3 can be predicted using the syntax. In Section 3, we present two experiments aimed at getting a measure of how similar in meaning naïve speakers take the propositional attitude verbs from the first experiment to be. There, we ask (i) how well the classification laid out in Section 1.3 can be used to predict the similarity judgments and (ii) how well the data from the first experiments can be used to predict these same judgments. In Section 4, we conclude.

2 Experiment 1: verb-frame acceptability

In this section, we present an experiment aimed at measuring the acceptability of a variety of propositional attitude verbs in different syntactic contexts. Our goal here is to assess the extent to which claims from traditional distributional analysis regarding correlations between syntax and semantics hold up. To do this, we compare the results of the acceptability judgment experiment against a previous classification of attitude verbs that synthesizes classifications from much prior theoretical literature. This allows us to quantitatively assess how closely these previous classifications are tracked in the syntax.

2.1 Design

Thirty propositional attitude verbs were selected in such a way that they evenly spanned the classes in Hacquard and Wellwood's (2012) semantic classification. This classification is essentially a more elaborated version of the classification presented in Section 1.3, synthesizing

discern, it would be nearly impossible to tell whether the similarity judgments $\mathbf{y}$ are a product (to some extent) of comparing of verbs' syntactic distributions $\mathbf{D}$ or whether they are a product of conceptual feature correlated with those distributions.

[18] Lederer et al. (1995) took a similar tack, using the same sort of semantic similarity judgment task but replacing acceptability judgments with syntactic distributions extracted from a corpus.

much of the previous theoretical literature on the propositional attitude verb classes.[19]

We then selected 19 syntactic features whose distribution has been claimed to be sensitive to attitude verb lexical semantics (see Section 1.3). These features consist in five broad types: clausal complement features, noun phrase (NP) complements, prepositional phrase (PP) complements, expletive arguments, and anaphoric arguments.[20] (Note that we break these features into types for expository purposes only. No special status is afforded to these groupings in the analysis.)

2.1.1 Features of interest

Six types of clausal complement features were selected: finiteness, complementizer overtness, subordinate subject overtness, subordinate question type, S-lifting, and small clause type. Finiteness had two values: finite (31a) and nonfinite (31b).

(31) a. Jo thought that Bo went to the store.
b. Jo wanted Bo to go to the store.

Complementizer presence had two values: present (32a) and absent (32b).

(32) a. Jo thought that Bo went to the store.
b. Jo thought Bo went to the store.

Embedded subject presence had two values: present (33a) and absent (33b) and is relevant only when the clause is finite and has no overt complementizer.

(33) a. Jo wanted Bo to go to the store.
b. Jo wanted to go to the store.

Embedded question type had three values: nonquestion (34a), polar question (34b), and WH question (34c). Only adjunct questions were used, since constituent questions are ambiguous on the surface between a question and a free relative reading.

(34) a. Jo knows that Bo went to he store.
b. Jo knows if Bo went to he store.
c. Jo knows why Bo went to he store.

S-lifting had two values: first person (35a) and third person (35b).

(35) a. Bo went to the store, I think.
b. Bo went to the store, Jo said.

Small clause type had two values: bare small clause (36a) and gerundive small clause (36b).

(36) a. Jo saw Bo go to the store.
b. Jo remembered going to the store.

Two NP structures were selected: single (37a) and double objects (37b). NPs were chosen so as not to have an interpretation in which they could be interpreted to have propositional content (Moulton, 2009a,b; Uegaki, 2012; Rawlins, 2013; Anand and Hacquard, 2014).

[19] Other large scale attitude verb classifications exist—see, for instance, the extensions to VerbNet (Kipper-Schuler, 2005) proposed in Korhonen and Briscoe (2004); Kipper et al. (2006) and the classifications given in FrameNet (Baker et al., 1998). This classification was chosen because it hews most closely to classes discussed above and in the theoretical literature more generally.

[20] A sixth feature—degree modification—was also selected for investigation. We exclude this from our analyses since the information degree modification carries is likely purely—or at least mostly—semantic in nature.

(37) a. Jo wanted a meal.
b. Jo promised Bo a meal.

A third feature relevant to NP complements—passivization—was also included (38). The availability of structures like (38) and the unavailability of structures like (33a), appears to correlate with whether a predicate's eventivity and/or its encoding of manner (Postal, 1974, 1993; Pesetsky, 1991; Moulton, 2009a,b, see also Zwicky 1971 for other syntactic and semantic features that track manner of speech).

(38) Bo was said to be intelligent.

Two types of PP complement were selected: PPs headed by *about* (39a) and PPs headed by *to* (39b).

(39) a. Jo thought about Bo.
b. Jo said to Bo that she was happy.

Three types of expletive arguments were selected: expletive *it* matrix subject, expletive *it* matrix object, and expletive *there* matrix object/embedded subject.

(40) a. It amazed Bo that Jo was so intelligent.[21]
b. Bo believed it that Jo was top of her class.
c. Bo wanted there to be food on the table.

Three types of anaphoric complement features were selected: *so* (41a), null complement/intransitive (41b), and nonfinite ellipsis (41c). (Note that we cannot be sure that the structure in (41b) involves null complements in either Williams's (2015, Ch. 5) broad or narrow sense for all verbs. See Hooper 1975; Hankamer and Sag 1976; Grimshaw 1979; Depiante 2000; Williams 2012 for further discussion of these structures.)

(41) a. Jo knew so.
b. Jo remembered.
c. Jo wanted to.

2.1.2 Stimulus construction

These 19 features were then combined into 30 distinct abstract frames. (Again, note that the features are mentioned for expository purposes only. They do not enter into the analysis in any formal sense.) These abstract frames are listed along the y-axis in Figure 5. Each categorial symbol in the frame should be interpreted as follows:

NP NP constituent (e.g. *Jo*)
WH (Adjunct) WH word (e.g. *why*)
V Bare form of verb (e.g. *think*)
VP Verb phrase with verb in bare form (e.g. *fit the part*)
S Tensed clause without complementizer (e.g. *Bo fit the part*)

For each abstract frame, three instantiations were generated by inserting lexical items, resulting in 102 frame instantiations. These 102 frame instantiations were then crossed with the 30 verbs to create 3060 total items.

Thirty lists of 102 items each were then constructed subject to the restriction that the list should contain exactly 3 instances of each verb and exactly 3 instances of each frame and that

[21] It is difficult to force the subject in a sentence like (40a) to be interpreted nonreferentially. As we see in Figure 5, this likely affected the judgments for verbs like *tell*, which are fine in this frame if the subject is interpreted referentially.

the same verb should never be paired with the same frame twice in the list. (That is, no verb showed up with more than one instantiation of the same frame in a single list.)

These lists were then inserted into an Ibex (version 0.3-beta17) experiment script with each sentence displayed using an unmodified `AcceptabilityJudgment` controller (Drummond, 2014). This controller displays the sentence above a discrete scale. Participants can use this scale either by typing the associated number on their keyboard or by clicking the number on the scale. A 1-to-7 scale was used with endpoints labeled *awful* (1) and *perfect* (7). All materials, including the instructions participants received, are available on the first author's github.

2.2 Participants

Ninety participants (48 females; age: 34.2 [mean], 30.5 [median], 18–68 [range]) were recruited through Amazon Mechanical Turk (AMT) using a standard Human Intelligence Task (HIT) template designed for externally hosted experiments and modified for the specific task. Prior to viewing the HIT, participants were required to score seven or better on a nine question qualification test assessing whether they were a native speaker of American English. Along with this qualification test, participants' IP addresses were required to be associated with a location within the United States, and their HIT acceptance rates were required to be 95% or better. After finishing the experiment, participants received a 15-digit hex code, which they were instructed to enter into the HIT. Once this submission was received, participants were paid $3.50.

2.3 Data validation

Even with the stringent requirements listed above—a qualification test, IP restriction, and high HIT acceptance rate—some participants attempt to game the system. There are two main ways that participants do this: (i) submitting multiple HITs despite being instructed not to and (ii) not actually doing the task—e.g. choosing responses randomly.

The first is easy to detect. When data are submitted in Ibex, the submitting participant's IP address is converted into an MD5 hash, which is in turn associated with the responses they submit. This hash can then be used to check whether participants followed instructions in only submitting a single HIT. Two participants submitted multiple HITs: one participant submitted three and another submitted two. In both of these cases, only the first submission was used.[22]

The second requires more care to detect. Here, we use the fact that multiple participants did the same list. The idea is to compare each participant's responses against those of all other participants that saw the same list. If a participant has low agreement with the other participants that saw the same list and the other participants show high agreement with each other, then we conclude that the disagreeing participant was providing lower quality data and remove them from the analysis.

To implement this, we calculated Spearman rank correlations between each participant's responses and those of every other participant that did the same list. For instance, if participants x, y, and z all did list 1, we would compute the correlation between x's and y's responses, x's and z's, and y's and z's. We then inspected the distribution of these correlations for outliers.

The median Spearman rank correlation between participant responses is 0.64 (mean=0.63, IQR=0.69-0.58). To find outliers, we use Tukey's method. In Tukey's method, an observation—here, a correlation between two subjects' judgments—is deemed an outlier if that observation falls below the first quartile (Q1) minus 1.5 times the interquartile range (IQR)—i.e. the difference between the third quartile and the second—or above the third quartile (Q3) plus 1.5 times the interquartile range. Four comparisons fall below Q1-1.5*IQR and none fall above

[22]Note that this method does not distinguish between one participant attempting to submit multiple HITs from the same IP and two participants each submitting a single HIT from the same IP. We err on the side of caution in filtering all the but the first HIT from the same IP.

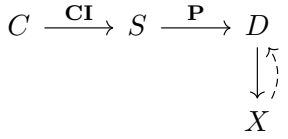


Figure 4: Schematization of data normalization in relation to the model presented in Section 1.

Q3+1.5*IQR. The four that fall below are due to two participants, each from a different list. Perhaps not coincidentally, those participants were also the ones that submitted multiple HITs. The remainder of the analyses exclude responses from these two participants, resulting in 86 unique participants.

After excluding these participants, the median remains the same (to two significant figures) and the mean shifts upward slightly, from 0.63 to 0.64. (This is to be expected since the mean is more sensitive to outliers.) The IQR becomes slightly smaller, and Q1 shifts slightly upward (IQR=0.69-0.59). These correlations are comparable to those reported by Fisher et al. (1991).

2.4 Results

In this section, we investigate the extent to which attitude verb classifications based on traditional distributional analysis are predictable from our acceptability judgment data. This investigation has three parts: (i) normalizing the acceptability judgment data across participants and items to remove noise in the judgments arising from those two sources (Section 2.4.1); (ii) preparing the acceptability judgment data to act as predictors by decorrelating the judgments using various matrix factorization methods—e.g. PCA (Section 2.4.2); and (iii) predicting verb classes from traditional distributional analysis using those decorrelated judgments (Section 2.4.3).

These three steps straightforwardly relate to the model presented in Section 1—a relationship we discuss in detail through the section. The first step can be viewed as inferring the true syntactic distributions $\mathbf{D}$ from acceptability judgments $\mathbf{x}$; the second can be viewed as inferring grammatically relevant semantic features $\mathbf{S}$ from the syntactic distributions $\mathbf{D}$; and the third can be viewed as mapping from these grammatically relevant semantic feature $\mathbf{S}$ into a representation of $\mathbf{S}$ derived from traditional distributional analysis. These steps pave the way for our comparison of the grammatically relevant features $\mathbf{S}$ inferred below with the representation of $\mathbf{C}$ we infer from the semantic similarity judgments presented in Section 3.

2.4.1 Data normalization

In this section, we address how to obtain an estimate of the actual acceptability of a verb-frame pair from our ordinal scale judgments. This can be seen as a way of taking our acceptability judgments $\mathbf{x}$ and inferring the underlying acceptabilities $\mathbf{D}$ that gave rise to them. Within the model laid out in Section 1, this can be schematized as in Figure 4.

One option for doing this is to just take the mean over the ordinal scale judgments for a particular verb-frame pair as though they were in fact interval-valued—i.e. as though the distance in acceptability space between a 1 and a 2 rating were the same as that between a 2 and a 3 rating, a 3 and a 4 rating, etc.

This method ignores two issues. First, it is well-known that participants use the same ordinal scale in different ways: some use the endpoints (here, 1 and 7) almost exclusively while others might not use them at all, preferring the middle values (here, 2 through 6).[23] This

[23]This phenomenon is well-known in the psychometrics literature, but see Schütze and Sprouse 2014 and citations therein for discussion of acceptability judgments in particular.

CUTPOINTS	ADDITIVE	MULTIPLICATIVE	LOG-LIKELIHOOD	AIC
Equidistant	True	False	-8078	24254
Equidistant	False	True	-8509	25116
Equidistant	True	True	-7268	22808
Varying	True	False	-7989	24084
Varying	False	True	-8061	24228
Varying	**True**	**True**	**-7084**	**22448**

Table 2: Comparison of normalization models for acceptability judgments.

means, for instance, that one participant's 3 rating might be equivalent to another participant's 1 rating, depending on their scale use preferences. Second, the ratings for a particular verb-frame pair come from 3 distinct items, introducing the possibility that certain items may draw an acceptability rating up or down at random.

The standard way of addressing the first issue (in the psycholinguistics literature) is to normalize the data by-participant—e.g. using a methods such as z-scoring or ridit scoring. The problem, here, is that neither such method can simultaneously address the second issue—item-based variability—without modification. Insofar as the assumptions of these transformations are satisfied, this is not an issue in other studies, since acceptability is generally treated as a dependent variable, not a predictor, and thus item variability can be taken into account in whatever confirmatory analysis follows the transformation—generally, using random intercepts for item in a linear mixed model.

To address these two issues here, we employ an ordinal mixed model similar in form to the polytomous Rasch model (Rasch, 1960; Andersen, 1977; Andrich, 1978; Masters, 1982). A review of ordinal mixed models—specifically, proportional-odds cumulative logit models (see Agresti 2014 for a more extensive review)—can be found in Section A.1 of the Appendix.

2.4.1.1 Comparison of mixed effects ordinal regression normalization models In this section, we fit six different kinds of ordinal mixed model to our data to obtain an estimate of the acceptability of a verb v in a frame f, filtering out participants' variable scale use and item effects. Following the discussion in Section 1, we denote the latent acceptability of verb v in a frame f as d_{vf}. Supposing that x_{vfmn} is the response given by participant n to item m of verb-frame pair (v, f), the item random intercepts are given by $u_{1m} \sim \mathcal{N}(0, \sigma^2_{item})$ and the subject random intercepts are given by $u_{2n} \sim \mathcal{N}(0, \sigma^2_{subj-add})$ and $u_{3n} \sim \mathcal{N}(0, \sigma^2_{subj-mult})$, we consider the following six models.

Equi add $\mathbb{P}(x_{vfmn} \leq r \mid d_{vf}, i, \mathbf{U}) = \text{logit}^{-1}((r-1)i - (d + u_{1m} + u_{2n}))$

Equi mult $\mathbb{P}(x_{vfmn} \leq r \mid d_{vf}, i, \mathbf{U}) = \text{logit}^{-1}(|u_{3n}|(r-1)i - (d + u_{1m}))$

Equi add-mult $\mathbb{P}(x_{vfmn} \leq r \mid d_{vf}, i, \mathbf{U}) = \text{logit}^{-1}(|u_{3n}|(r-1)i - (d + u_{1m} + u_{2n}))$

Varying add $\mathbb{P}(x_{vfmn} \leq r \mid d_{vf}, \mathbf{i}, \mathbf{U}) = \text{logit}^{-1}\left(\left[\sum_{j=1}^{r} i_j\right] - (d + u_{1m} + u_{2n})\right)$

Varying mult $\mathbb{P}(x_{vfmn} \leq r \mid d_{vf}, \mathbf{i}, \mathbf{U}) = \text{logit}^{-1}\left(\left[|u_{3n}| \sum_{j=1}^{r} i_j\right] - (d + u_{1m})\right)$

Varying add-mult $\mathbb{P}(x_{vfmn} \leq r \mid d_{vf}, \mathbf{i}, \mathbf{U}) = \text{logit}^{-1}\left(\left[|u_{3n}| \sum_{j=1}^{r} i_j\right] - (d + u_{1m} + u_{2n})\right)$

For each model, we use gradient descent with momentum (Rumelhart and McClelland, 1986; McClelland and Rumelhart, 1986) and learning rate annealing with a *search-then-converge* schedule (Darken and Moody, 1990) to obtain the Maximum Likelihood Estimate (MLE) of $\mathbf{D}$, i or $\mathbf{i}$,

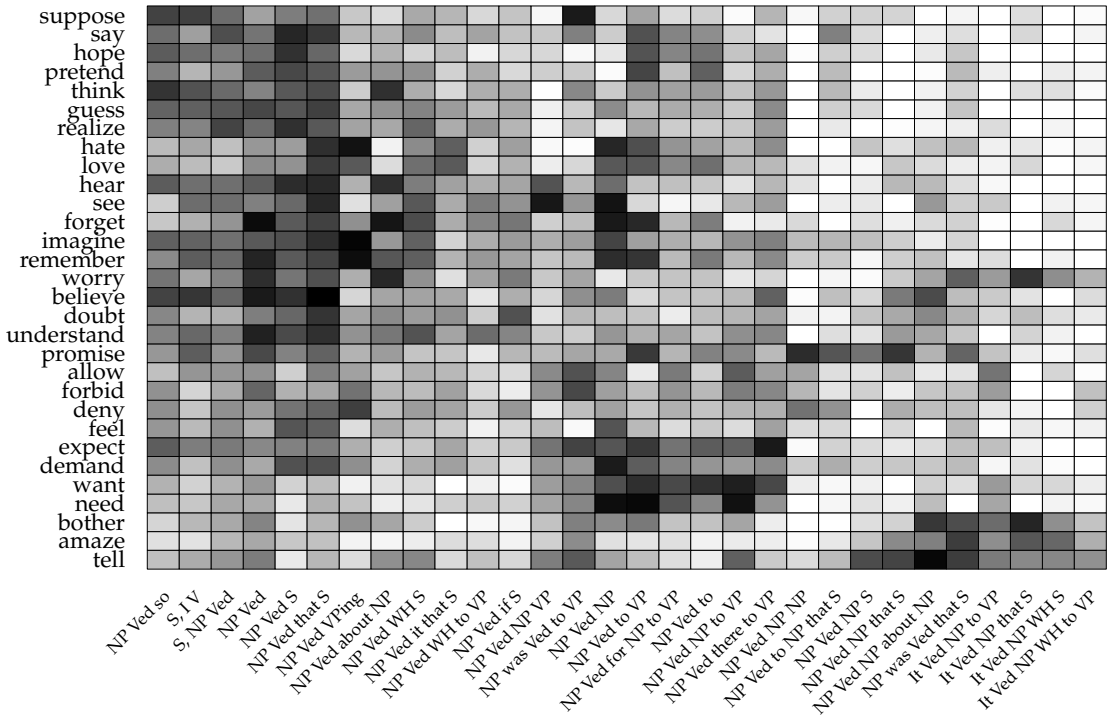


Figure 5: Normalized verb-frame acceptability judgments. Darker shades mean higher normalized ratings.

$\mathbf{U}$, σ^2_{item}, and σ^2_{subj}—call this estimate $\boldsymbol{\theta}_{\text{MLE}}$.[24] Each model was implemented in version 0.7 of the python package `theano` (Bergstra et al., 2010; Bastien et al., 2012); this implementation is available on the first author's github.

The log-likelihood of the MLE, $\ln \mathcal{L}(\boldsymbol{\theta}_{\text{MLE}} \mid \mathbf{x}; \mathcal{M})$, for each model $\mathcal{M}$ was then computed. To conduct model comparison, we use the Akaike Information Criterion (AIC; Akaike 1974). The AIC for a model $\mathcal{M}$ with parameters $\boldsymbol{\theta}$ is given by

$$\text{AIC}(\mathcal{M}) = 2k - 2 \ln \mathcal{L}(\boldsymbol{\theta}_{\text{MLE}} \mid \mathbf{x}; \mathcal{M})$$

where k is the number of parameters in $\boldsymbol{\theta}_{\text{MLE}}$. This metric rewards a model for associating the data with a lower likelihood but penalizes it proportional to the number of parameters it has, since the more parameters a model has, the more likely it is to overfit the data—i.e. fit to noise.

Table 2 shows the log-likelihood and AIC for each of the six normalization models. The best fitting model, after penalizing for complexity using AIC, is the varying cutpoint additive-multiplicative model.[25] We extract the MLE of $\mathbf{D}$—the matrix of verb-frame acceptabilities, given in Figure 5—from this model and feed this matrix into various matrix factorization algorithms to decorrelate.

2.4.2 Data decorrelation

When the predictors entered into a model are correlated, fitting procedures often have a hard time deciding which of the correlated variables to use. This causes instability in the model fits that can affect downstream inference. This is a real worry for our data, since we have

[24] Indeed, the estimates of $\mathbf{U}$, σ^2_{item}, and σ^2_{subj} will be Maximum A Posterior (MAP) estimates.

[25] This suggests that, at least for this dataset, a standard normalization such as z-scoring would have been inappropriate, even controlling for item effects. We urge other psycholinguistics researchers to take heed of this possibility when analyzing ordinal scale data, since using possibly inappropriate normalization methods like z-scoring could well affect downstream inferences.

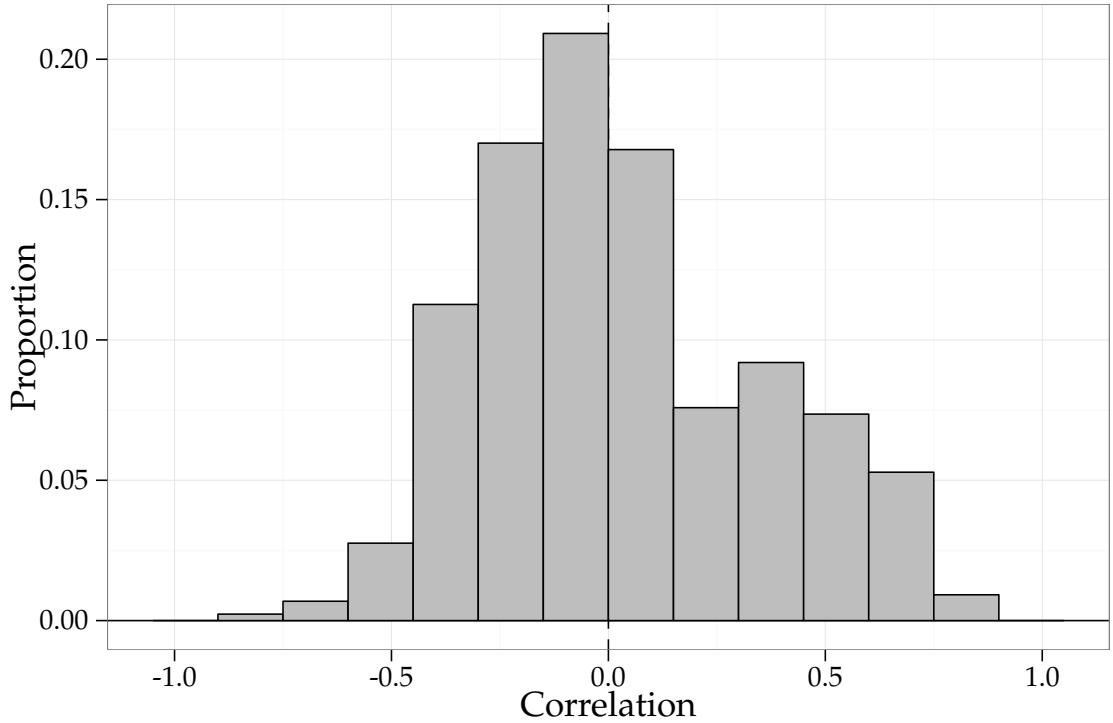


Figure 6: Correlation between frame judgments (columns of Figure 5).

a fairly large number of highly positively correlated frame judgments. This can be seen in Figure 6, which shows the distribution of these correlations. To remedy this instability, it is common to apply an information preserving (reversible/invertible) transformation to the data that produces uncorrelated (orthogonal) predictors. The most popular such transformation is Principal Component Analysis (PCA; Jolliffe 2002).

Beyond its importance as a technique for decorrelating predictors, we believe that PCA (and its ilk) might be useful as a theoretical tool (cf. Landauer and Dumais, 1997). Specifically, we would like to suggest that PCA and similar methods might fruitfully be viewed as discovering grammatically relevant semantic features $\mathbf{S}$ given syntactic distributions $\mathbf{D}$, schematized in Figure 7. This view highlights PCA's use as a method for conducting matrix factorization, wherein one representation is factored into two: one that encodes relationships between verbs in the original dataset on new dimensions and another that maps from this encoding to the original dataset.

Different sorts of matrix factorization techniques are more or less apt at discovering certain kinds of patterns in data, raising the possibility that different techniques will give rise to verb semantic representations that are better or worse approximations to the ones humans actually have. For instance, Landauer and Dumais (1997) propose that PCA—or more specifically, an intimately related technique known as Singular Value Decomposition (SVD)—produces good approximations of human semantic representations. More recently, another family of techniques known under the heading of Nonnegative Matrix Factorization (NMF; Lee and Seung 1999) has been claimed to produce good lexical semantic representations (Murphy et al., 2012; Fyshe et al., 2014, 2015). In the latter case, it has been claimed that this method is best coupled with *sparsity* constraints on the semantic representations. (See Section A.2 in the Appendix for a high-level review of PCA, NMF, and their sparse variants.)

In the remainder of this section, we apply each of these matrix factorization techniques to our normalized acceptability judgment data with the mutually supportive goals of (i) reducing correlations among predictors and (ii) constructing a representation of each verb's grammatically relevant semantic features. We then use the resulting representations to predict the

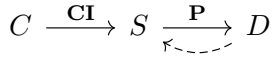


Figure 7: Schematization of matrix factorization in relation to the model presented in Section 1.

attitude verb classification reviewed in Section 1. Our ultimate goal is to assess whether the classification presented in Section 1 is indeed encoded in the syntactic distributions, but this comparison can also be taken as a contribution to the growing literature on methods for extracting semantic features from linguistic context.

2.4.3 Predicting attitude verb classes

In this section, we ask how well the acceptability judgments can be used to predict the classification of attitude verbs presented in Section 1. This has two parts: (i) a transformation of the normalized acceptability judgment data $\mathbf{D}$ to (a) reduce correlation between variables and (b) extract verbs' grammatically relevant semantic features; and (ii) use of that transformed data to predict each class discussed in Section 1, represented as a binary variable. The transformations we consider are PCA, sparse PCA, NMF, and sparse NMF, and the classes we consider are REPRESENTATIONAL, PREFERENTIAL, PERCEPTION, FACTIVE, COMMUNICATIVE, and ASSERTIVE. The classification we use is given in Table 3.

2.4.3.1 Methodology Each transformation was applied to the normalized acceptability judgment data $\mathbf{D}$ using the implementation available in the `decomposition` module in version 0.16.1 of the python `sklearn` package (Pedregosa et al., 2011). Prior to applying either PCA transformation to the normalized acceptability judgment data, the frame judgments (each column of Figure 5) were first mean-centered and standardized, and prior to applying either NMF transformation to the normalized acceptability judgment data, the frame judgments were first shifted, so that the minimum value was zero, and then divided by the resulting maximum, so that all values lie on the unit interval and the highest value in each column was equal to one. For PCA, all 30 principal components—equal to the number of frames—were retained for later prediction; for sparse PCA, NMF, and sparse NMF, 30 components were learned. For sparse PCA and sparse NMF, grid search was used to select the regularization parameter from the values 0.1, 0.2, 0.5, 1.0, 2.0, 5.0, and 10.0 based on the cross-validation described below. For sparse NMF, we induce sparsity on the verb representation.

Each transformed dataset was entered into logistic regressions with L1 regularization (cf. LASSO regression; Tibshirani 1996).[26] As for the transformation sparsity, grid search was used to select the regularization parameter from the values 0.1, 0.2, 0.5, 1.0, 2.0, 5.0, and 10.0 based on the cross-validation described below. This regularization helps to force the model to select only a few features to predict from. To validate these regressions, leave-one-out crossvalidation was used, wherein the transformed data for one verb was removed from the training set, the model fit, and that verb's class predicted. This was carried out for all verbs in all transformations for all classes. The accuracy over verbs for each tuple of transformation, class, sparsity parameter, and regularization parameter was then computed by averaging over the cross-validation accuracies. The maximum of the averages over the transformation sparsity parameter (for sparse PCA and sparse NMF) and regression regularization parameter was taken. Table 4 shows these accuracies—with an asterisk denoting an accuracy that is better than always guessing

[26] See Rooth 1995; Stevenson and Merlo 1999; Schulte im Walde 2000; Merlo and Stevenson 2001; Korhonen 2002; Schulte im Walde and Brew 2002; Schulte im Walde 2003, 2006; Vlachos et al. 2008, 2009; Alishahi and Stevenson 2008 for alternative methods of evaluating existing classifications with respect to syntactic distributions.

VERB	REPR	PREF	FACT	ASSERT	COMM	PERCEPT
allow		✓				
amaze	✓	✓	✓			
believe	✓			✓		
bother	✓	✓	✓			
demand		✓			✓	
deny	✓				✓	
doubt	✓					
expect	✓			✓		
feel	✓			✓		✓
forbid		✓			✓	
forget	✓		✓	✓		
guess	✓			✓		
hate	✓	✓	✓			
hear	✓		✓	✓		✓
hope	✓	✓		✓		
imagine	✓					
love	✓	✓	✓			
need		✓				
pretend	✓					
promise	✓			✓	✓	
realize	✓		✓	✓		
remember	✓		✓	✓		
say	✓			✓	✓	
see	✓		✓	✓		✓
suppose	✓			✓		
tell	✓			✓	✓	
think	✓			✓		
understand	✓		✓	✓		
want		✓				
worry	✓	✓		✓		

Table 3: Classification of 30 verbs in experiment based on literature reviewed in Section 1.

TRANSFORM	REPR	PREF	PERCEPT	FACT	COMM	ASSERT	MEAN
PCA	**0.933***	0.833*	**0.967***	0.767*	**0.867***	0.733*	0.850
sparse PCA	0.900*	0.833*	0.933*	0.800*	0.833*	**0.933***	0.878
NMF	0.900*	0.700*	0.900	0.667	0.767	0.600	0.761
sparse NMF	0.900*	**0.933***	**0.967***	**0.900***	0.833*	0.833*	**0.894**

Table 4: Cross-validation accuracy for each transformation and attitude verb class. Asterisks represent accuracies that are better than a model that always chooses the most numerous class. Bolding shows the transformation with the highest accuracy on that class.

the most likely category—a standard baseline model—and bolding denoting the best accuracy across models for that class.

2.4.3.2 Results We see in Table 4 that all classes are predictable above the baseline accuracy by all transformations of the data besides standard NMF, which only predicts the REPRESENTATIONAL and PREFERENTIAL classes above baseline. This suggests that all six of these classes—or at least some distinction correlated with each—are tracked in the syntax. It may also suggest that, even if the syntactic distributions are poorly encoded, as they appear to be in standard NMF, the REPRESENTATIONAL and PREFERENTIAL classes are encoded so robustly in the syntactic distributions as to shine through even the poor encoding.

With respect to the performance of particular transformations, we see that sparse NMF shows the best results overall. This is further corroboration of Murphy et al.'s (2012) finding that sparse NMF, which they refer to as Nonnegative Sparse Embedding (NNSE), is superior to other factorization methods in various tasks. To our knowledge, this is the first direct comparison of factorization methods on acceptability judgment data.

Besides knowing that a class is predictable, it is also useful to know which frames predict it best. To assess this we analyze the logistic regression coefficients—which weight the dimensions that each transformation produces—in conjunction with the mapping $\mathbf{P}$ to the original transformation—which represents the projection rules. Thus, we can extract the relationship between each frame and each class by weighting the mapping $\mathbf{P}$ (averaged over the cross-validation) by the logistic regression coefficients and summing across the latent dimensions.[27]

Figure 8 shows the resulting weights for each class and each frame after standardization by class (row). Darker shades show more positive values. To a large extent, these weights corroborate the syntax-semantics relationships laid out in Section 1. The five highest weighted frames for the representational class—*NP Ved that S*, *S, I V*, *NP Ved if S*, *NP Ved*, and *NP Ved there to* —tend to be finite. The five highest weighted frames for the preferential class—*It Ved NP to VP*, *NP Ved for NP to VP*, *NP Ved NP to VP*, *NP Ved to VP*, *It Ved NP that S*, and *NP was Ved to VP*—tend to be nonfinite. The five highest weighted frames for the perception class—*NP Ved NP*, *NP Ved NP VP*, *NP Ved S*, *NP Ved WH S*, *NP Ved that S*—include the bare VP subordinate clause and NP complements. The five highest weighted frames for the factive class—*NP Ved WH S*, *NP Ved WH to VP*, *NP Ved about NP*, *NP Ved*, *NP Ved S*—include both question and nonquestion complements. And the five highest weighted frames for the communicative class—*NP Ved to NP that S*, *NP Ved NP S*, *NP was Ved to VP*, *NP Ved NP NP*, *NP Ved NP that S*—are all frames that involve NP and PP objects. The assertive class has only four nonzero frames—*S, NP Ved*, *S, I V*, *NP Ved*, and *NP Ved about NP*—with *S, NP Ved* being much stronger than the others. The S-lifting frames (*S, NP Ved*, *S, I V*) are two of the ones that Hooper (1975) suggests are associated with assertivity.

[27] That is, we take the dot product of the regression coefficients, averaged over the cross-validation sets, for each class and the mapping $\mathbf{P}$.

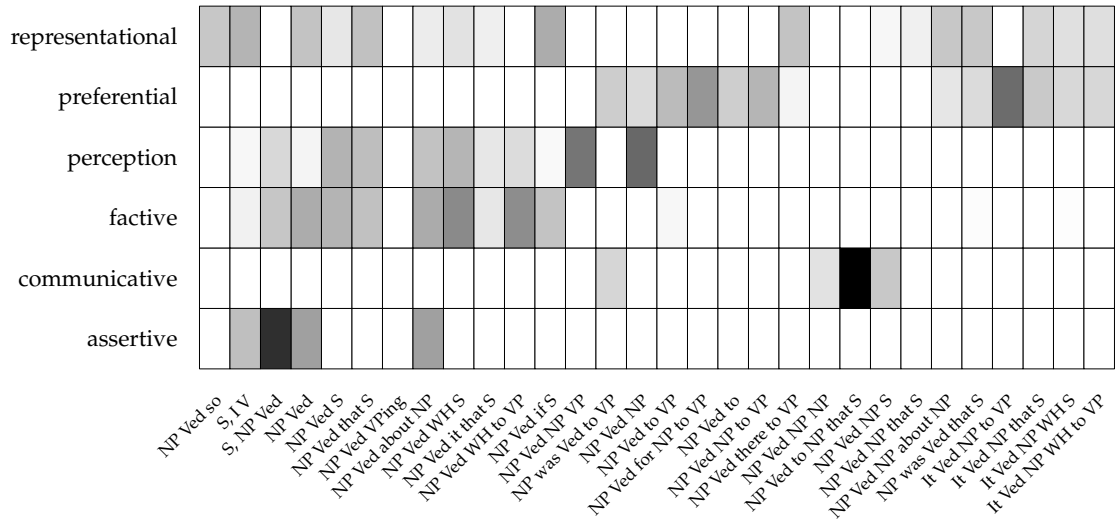


Figure 8: Frame weights (standardized) for each class derived from logistic regression coefficients and representation of projection rules **P**.

2.5 Discussion

In this section, we presented an experiment aimed at getting a measure of how acceptable a variety of propositional attitude verbs are in different syntactic contexts. After normalizing this acceptability measure with respect to subject and item effects, we employed four different kinds of transformations—PCA, sparse PCA, NMF, and sparse NMF—to reduce correlations among the frames. We suggested that this procedure has a further theoretical benefit in the sense that it can be viewed as a way of extracting both grammatically relevant semantic features **S** and projection rules **P** from the normalized acceptability judgments **D**.

For each of the above transformations, we compared the transformed data to the attitude verb classification laid out in Section 1. We found that all of these classes could be predicted well by most of the transformed datasets, suggesting that the classifications from traditional distributional analysis are at least on track. We then showed that, beyond predicting the classes well, the frames that traditional distributional analyses have suggested are associated with each class are indeed associated with that class.

The worry that we raised in Section 1 still looms, however. Our predictive models in this section did better than baseline, but their accuracies were not perfect. This imperfection could be due to the fact that the classifications given in traditional distributional analysis literature are merely correlated with the true distinction, and/or it could be that these distinctions are not discrete at all. These possibilities could in turn arise (i) due to entrenchment of traditional distinctions that work well enough and (ii) due to the precision of the sort of qualitative methods inherent to traditional distributional analysis.

In the next section, we present two experiments aimed at constructing a classification that is less biased in these respects. To do this, we employ two sorts of semantic similarity judgment tasks: a generalized semantic discrimination task and an ordinal scale similarity task. The results of both tasks are compared to both the classification used above as well as the acceptability judgment data presented in this section.

3 Experiments 2 & 3: verb similarity

In this section, we present two experiments aimed at getting a measure of how similar in meaning naïve speakers take the propositional attitude verbs from Experiment 1 to be. The first

experiment (Experiment 2) employs a generalized semantic discrimination task—also known as a triad or "odd man out" task—in which participants are given lists of three words and asked to choose the one least like the others in meaning (Wexler, 1970; Fisher et al., 1991). The second experiment (Experiment 3) employs an ordinal (likert) scale similarity task, in which participants are asked to rate the similarity in meaning of a word pair on a 1-7 scale. The reason we employ two similarity judgment tasks instead of one is so that we can obtain an upper bound on the amount of variability we can reasonably expect a model that uses syntactic distribution as a predictor of semantic similarity to explain. To this end, we compare the similarity judgments against (i) each other, (ii) the classification presented in Section 1, and (iii) the acceptability judgments.

In Sections 3.1 and 3.2, the two experiments are described, along with the normalization models we employ to filter out participant-based variability in the judgments. In Section 3.3, the three comparison listed above are conducted. In that last section, we also sketch out a method for using these comparisons to derive a quantitatively based attitude verb classification.

3.1 Experiment 2: generalized semantic discrimination task

3.1.1 Design

In this experiment, we aim to get a measure of how similar in meaning naïve speakers take the propositional attitude verbs from Experiment 1 to be. To do this, we constructed a list containing every three-combination of the 30 verbs from Experiment 1 (4060 three-combinations total). Twenty lists of 203 items each were then constructed by randomly sampling these three-combinations, which we refer to as triads, without replacement.

These lists were then inserted into an Ibex (version 0.3-beta15) experiment script with each triad displayed using an unmodified `Question` controller (Drummond, 2014). This controller displays an optional question above a list of answers. In this case, the question was omitted and the verbs making up each triad constituted the possible answers. Participants could select an answer either by typing the number associated with each answer or clicking on the answer. As for the acceptability judgment experiment presented in Section 2, all materials, including the instructions participants received, are available on the first author's github.

3.1.2 Participants

Sixty participants (28 females; age: 34.5 [mean], 31 [median], 18–68 [range]) were recruited through AMT using a standard HIT template designed for externally hosted experiments and modified for the specific task. All qualification requirements were the same as those described in Section 2. After finishing the experiment, participants received a 15-digit hex code, which they were instructed to enter into the HIT. Once this submission was received, participants were paid $3.

3.1.3 Data validation

The data validation procedure is the same one described in Section 2 with the exception that we calculate Cohen's κ instead of Spearman's ρ.[28] The median Cohen's κ between participant

[28] Both Fisher et al. and Lederer et al. compute Spearman rank correlations over count matrices constructed from judgments across participants. The method they use is not available to us without significant alteration since we collected data from more than two participants per list. Instead, we opt for a more standard measure of interrater agreement here. This measure is preferable in any case since (i) it allows us to assess each participant's reliability at the same time as we assess overall agreement and (ii) it can be applied to the raw data instead of a statistic of the data, as in the cases of Fisher et al. and Lederer et al..

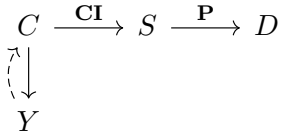


Figure 9: Schematization of data normalization (Sections 3.1.4 and 3.2.4) in relation to the model presented in Section 1.

responses is 0.45 (mean=0.45, IQR=0.52-0.37).[29] To find outliers, we use Tukey's method. No comparisons fall below Q1-1.5*IQR and none fall above Q3+1.5*IQR. Thus, we exclude no participants.

The median agreement here is quite a bit lower than the interrater agreement found by either Fisher et al. or Lederer et al.[30] This is likely driven by the fact that we are investigating a much smaller portion of the lexicon and thus are bound to find that participants have less certainty about which verbs are more semantically similar.[31] Another possible contributor to this lower correlation is that Cohen's κ is more conservative than Spearman's ρ. As we see in Section 3.2.3, however, the conservativeness of Cohen's κ is not likely to be the culprit here, since even Spearman's ρ shows roughly the same amount of agreement among participants using a different measure.

3.1.4 Data normalization

The fact that verbs are displayed in a list raises the worry that effects of position may arise, either as an overall preference for a particular position and/or as a participant-specific preference. We see both such preferences. Across participants, there is a bias for earlier positions—proportion for position 1: 0.36, position 2: 0.34, position 3: 0.30—but substantial variability among participants—interquartile range of participant bias for position 1: [0.33, 0.39], position 2: [0.31, 0.36], position 3: [0.27, 0.34].[32] Thus, as in Section 2, we normalize the data prior to analysis to control for biases a particular participant may have to choose a verb in a particular position.

To carry out this normalization, we use a mixed effects multinomial logistic (maximum entropy) model. This model predicts which verb position in a triad is chosen based on (i) the (latent) similarities between each pair in the triad and (ii) the (latent) bias each participant has to choose a verb in a particular position. This results in the following model

$$\mathbb{P}(y_{v_1v_2v_3n} \mid \mathbf{C}, \mathbf{U}) = \mathrm{softmax}(c_{v_2v_3} + u_{n1}, c_{v_1v_3} + u_{n2}, c_{v_1v_2} + u_{n3})$$

where v_1, v_2, v_3 are the verbs in positions one through three, respectively; $y_{v_1v_2v_3n}$ represents the response of participant n to the triad v_1, v_2, v_3; $\mathbf{C}$ is a verb-by-verb matrix representing verb similarity; $\mathbf{u}_n$ represents the bias of participant n to choose particular positions; and $\mathrm{softmax}$ is the standard multinomial generalization of the logistic function.[33]

[29] An analysis of the distribution of Fleiss' κ (the multi-rater generalization of Scott's d) by list corroborates this analysis (median=0.45, mean=0.45, IQR=0.48-0.40).

[30] Fisher et al. report Spearman's ρ=0.81 (Exp. 1); 0.78 (Exp. 2); 0.76 (Exp. 3), 0.79 (Exp. 4), 0.72 (Exp. 5). Lederer et al. report Spearman's ρ=0.81.

[31] If this is indeed true, interrater agreement on this and other similarity judgment tasks could be a way of investigating the "semantic density" of a lexical neighborhood. Modeling reaction time, as a proxy for uncertainty, might also be fruitful in future research.

[32] Some of this variability may be attributable to the list each participant received. Since the lists were between-subjects and we are only concerned with controlling for the variability to normalize the data, this can safely be explained by the random effects components.

[33] We furthermore impose a symmetry constraint on the $\mathbf{C}$, wherein $c_{ij} = c_{ji}$. A model without this constraint

CUTPOINTS	ADDITIVE	MULTIPLICATIVE	LOG-LIKELIHOOD	AIC
Equidistant	True	False	-4397	9816
Equidistant	False	True	-4652	10326
Equidistant	True	True	-4125	9392
Varying	True	False	-4361	9752
Varying	False	True	-4392	9814
Varying	**True**	**True**	**-4070**	**9290**

Table 5: Comparison of normalization models for ordinal similarity judgments.

We find the Maximum Likelihood Estimate (MLE) of **C** and **U** using gradient descent implemented in version 0.7 of the python package `theano`. This implementation can be found on the first author's github. In Section 3.3, we use the MLE of **C** as our normalized representation of the generalized semantic discrimination judgments.[34]

3.2 Experiment 3: ordinal similarity

3.2.1 Design

As in Experiment 2, we aim to get a measure of how similar in meaning naïve speakers take the propositional attitude verbs from Experiment 1 to be. To do this, we constructed a list containing every pair of the 30 verbs from Experiment 1 along with the verb *know* (460 unordered pairs, 920 ordered pairs). Twenty lists of 62 ordered pairs were then constructed such that every verb was seen an equal number of times and no pair—either unordered or ordered—was seen twice.

These lists were then inserted into an Ibex (version 0.3.7) experiment script with each pair displayed using an unmodified `AcceptabilityJudgment` controller (Drummond, 2014). This controller displays the verb pair separated by a pipe character—e.g. *think* | *want*—above a discrete scale. Participants could use this scale either by typing the associated number on their keyboard or by clicking the number on the scale. A 1-to-7 scale was used with endpoints labeled *very dissimilar* (1) to *very similar* (7). To encourage them to make a symmetric similarity judgment, participants were instructed to rate "the similarity between the meanings of the two verbs" as opposed to rating how similar the first verb was to the second (or vice versa). All materials, including the instructions participants received, are available on the first author's github.

3.2.2 Participants

Sixty participants were recruited through AMT. All qualification requirements were the same as those described in Section 2.2. After finishing the experiment, participants received a 15-digit hex code, which they were instructed to enter into the HIT. Once this submission was received, participants were paid $1.

3.2.3 Data validation

The data validation procedure is the same one described in Section 2. The median Spearman rank correlation between participant responses is 0.40 (mean=0.41, IQR=0.52-0.32). To find

was also fit but had a worse AIC than the model with the symmetry constraint. This suggests that participants are in fact making judgments based on a symmetric understanding of similarity (or at least that we do not have enough evidence to show that they are not).

[34] In fact, it may be more precise to consider this estimate to be of some intermediary representation that is not itself **C**, but rather relationships among representations in **C**.

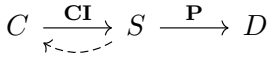


Figure 10: Schematization of similarity prediction (Sections 3.3.2 and 3.3.3) in relation to the model presented in Section 1.

outliers, we use Tukey's method. No comparisons fall below Q1-1.5*IQR and none fall above Q3+1.5*IQR. Thus, we exclude no participants.

3.2.4 Data normalization

Since these data are ordinal, we use the data normalization procedure described in Section 2.[35] Table 5 gives the log-likelihood and AIC for each model. As in Section 2, the best fitting model, penalizing for complexity, is the model with varying cutpoints and both additive and multiplicative subject random effects. We use the MLE of the similarities inferred by this model in the remainder of this section.

3.3 Predicting similarity judgments

In this section, we construct three models for predicting the normalized similarity judgments. The first model, employed for comparing the two similarity datasets, uses the normalized similarities from one similarity dataset to predict the other. This model is used throughout the rest of the section as a baseline with which to assess the results of the other two models, which predict similarities based on the classification given in Section 1 and the transformed acceptability judgments described in Section 2. This use as a baseline is licensed by the fact that predicting one set of similarities from another represents a reasonable upper bound on the expected performance of models not using similarity as a predictor. At the end of this section, we return to the idea, laid out in Section 1, that these analyses can be seen as a quantitative augmentation of traditional distributional analysis, and we sketch out a general method for inferring a quantitatively derived classification from these datasets.

3.3.1 Comparsion of similarity datasets

Figure 11 plots the generalized semantic discrimination judgments against the ordinal scale similarity judgments, both after normalization and standardization. Overall, the correlation between responses on the generalized semantic discrimination task and those on the ordinal scale task are at about the same level as the correlations among participants within each experiment (Spearman's ρ=0.437, $p < 0.001$). This suggests not only that these two tasks are tapping similar aspects of participants' semantic knowledge but that they do so at the limit of what we would expect given inter-annotator agreement within each experiment.

Fitting a linear model to these data, with the (normalized and standardized) ordinal scale similarity ratings as the dependent variable and the (normalized and standardized) generalized semantic discrimination similarity as the independent variable, we obtain an r^2 of 0.172.[36] In the next two sections, this r^2 is used as a baseline against which to compare (i) models that predict the similarity judgments given the classification presented in Section 1 and (ii) models that predict the similarity judgments given the normalized and transformed acceptability judgments.

[35] As for the generalized semantic discrimination normalization model, we constrain the similarities inferred to be symmetric. Each model was also fit without this constraint, and the AICs were found to be worse in comparison to the corresponding model with the constraint.

[36] Note that the r^2 is the same regardless of which similarities are taken as the dependent variable.

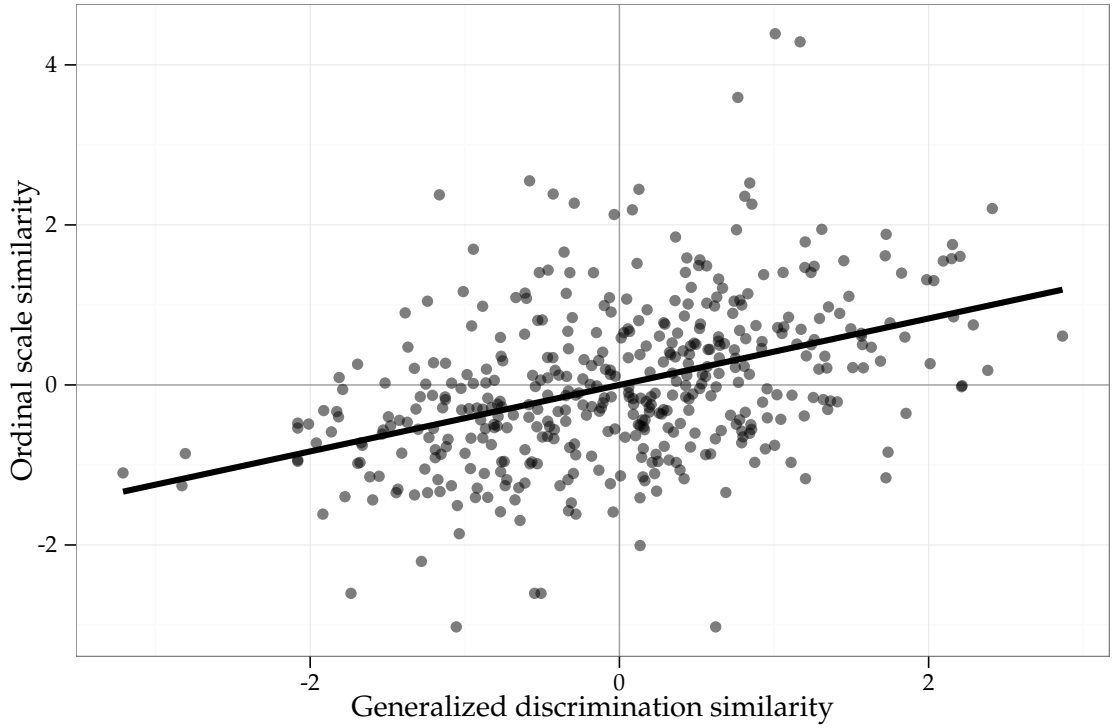


Figure 11: Generalized semantic discrimination semantic similarity ratings (normalized and standardized) plotted against ordinal scale similarity ratings (normalized and standardized).

3.3.2 Predicting similarities given attitude verb classes

Each normalized similarity dataset was entered into a linear regression with SIMILARITY as the dependent variable and the value of REPRESENTATIONAL, PREFERENTIAL, PERCEPTION, FACTIVE, COMMUNICATIVE, and ASSERTIVE for the two verbs whose similarity is being predicted as the dependent variables.[37] All main effects were included as well as interactions between each verb's value on same classes. That is, the interaction between the first verb's value for REPRESENTATIONAL and the second verb's value for REPRESENTATIONAL was included but not, e.g., the interaction between the first verb's value for REPRESENTATIONAL and the second verb's value for PREFERENTIAL.

To validate these results, we use 30-fold cross-validation. In each fold of this cross-validation, all normalized similarities for one of the 30 verbs tested were removed, the model trained, the normalized similarities predicted, and the r^2 for these predicted similarities calculated.[38] The average predicted r^2 over each fold of the cross-validation was 0.046 for the normalized generalized semantic discrimination similarities and 0.072 for the normalized ordinal similarities. These r^2 appear quite low, but it should be noted that, in comparison to the r^2 that results when comparing the two normalized similarity datasets to each other, they are quite good, with about 25% and 40% as much variance being explain by each model as by the model from the previous section, which explicitly gets access to similarity judgments for prediction.

The performance of these models appears to be drawn down by the existence of intuitively distinct semantic features not included in the classification. This can be seen in Figure 12, which shows the predicted r^2 broken out by verb and dataset.[39] Here, we see that the simi-

[37] Adding regularization to these regressions only worsened the cross-validation r^2.

[38] It is well-known that, just as for log-likelihood, r^2 can be a problematic measure for assessing model fit, especially when the number of parameters in the models being compared differ. Note, however, that using cross-validation and predicted r^2 alleviates this worry, since it naturally controls for the overfitting that gives rise to this problem.

[39] Note that predicted r^2 can be negative for a particular verb because the model is not trained on that verb's

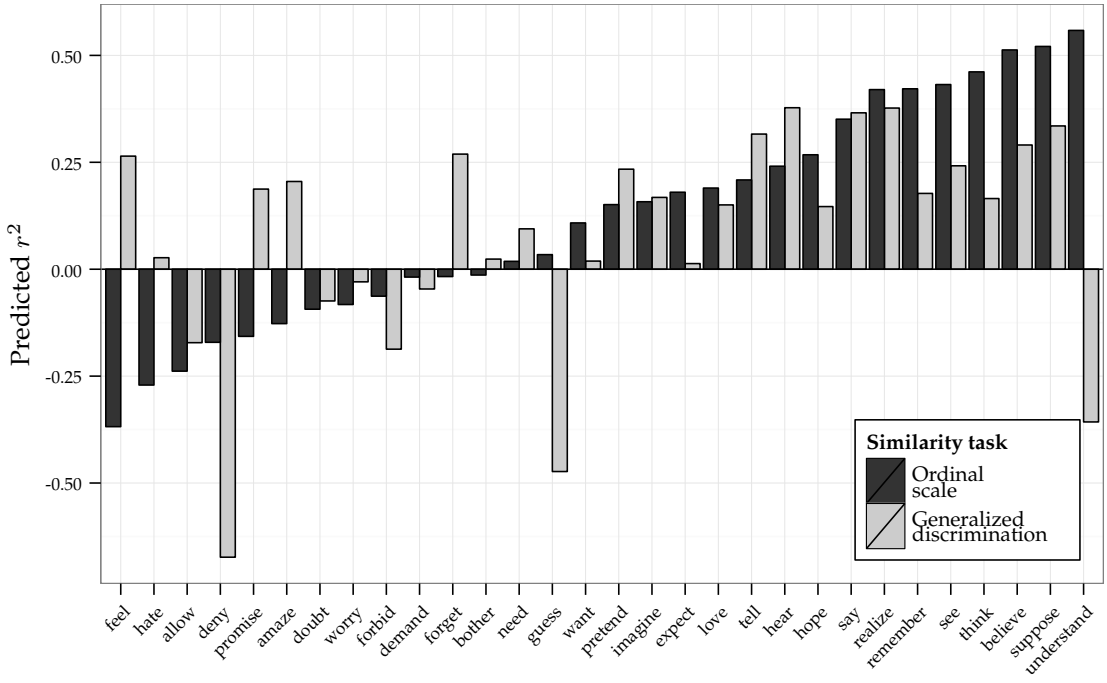


Figure 12: By-verb predicted r^2 for model based on classification from Section 1.

larities of verbs involving negation—*deny*, *forbid*, *doubt*, *demand*, and *worry*—and verbs involving permission, obligation, or commitment—*allow*, *forbid*, *deny*, *promise*—tend to be predicted poorly by the classification. This is not surprising. As observed in Section 1, there are semantic properties which are not relevant to syntactic distribution. Participants were simply told to find similarities, which does not tell them what dimensions to compare along. So, we expect there to be some similarities that are based on features that have no grammatical relevance and some based on features that do. So, since our regression is only about how the similarity relates to the linguistically relevant classes we identified, we do not expect to account for all of the variance.

One the other hand, similarities among many of the representationals (*think*, *believe*, *suppose*, etc.) are predicted quite well, the predicted r^2 for the classification-based model reaching far above the r^2 for the similarity-based model from the last section. Indeed, on the whole, the mean predicted r^2 for representationals (generalized discrimination r^2=0.105, ordinal scale r^2=0.164) is far better than that of preferentials (generalized discrimination r^2=-0.021, ordinal scale r^2=0.021) which shows essentially no improvement from always guessing the mean. This suggests that the classification given in Section 1 is not fine-grained enough from the point of view of predicting participants' similarity judgments, though it remains an open question whether it is sufficiently fine-grained to cover the grammatically relevant features. We address this in the next section.

3.3.3 Predicting similarities given syntactic distributions

In this section, we assess how well the normalized similarities presented above can be predicted by the syntactic distributions $\mathbf{D}$ (inferred from acceptability judgments $\mathbf{x}$) presented in Section 2. As in Section 2, the syntactic distributions $\mathbf{D}$ are first transformed in order to (i) lower correlations among the frame variables and (ii) extract a potential representation of

similarities.

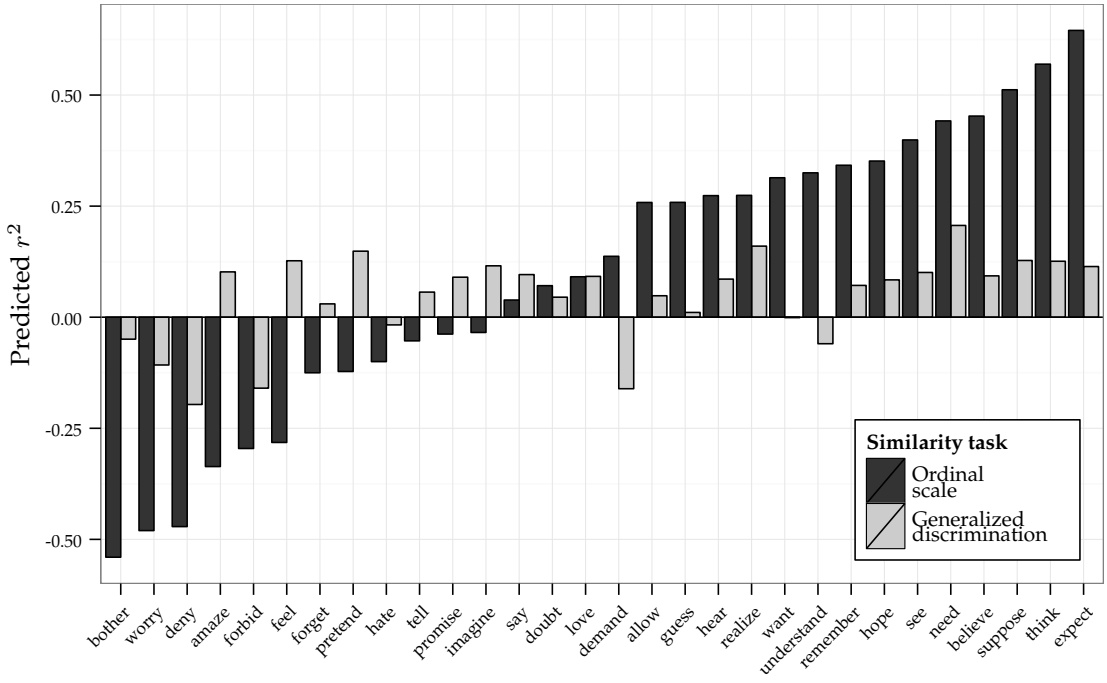


Figure 13: By-verb predicted r^2 for model based on acceptability judgments presented in Section 2.

the grammatically relevant semantic features **S**. The same transformation considered in that section—PCA, sparse PCA, NMF, and sparse NMF—are used here as well. Again, the ultimate goal here is to assess how robustly verb semantics (as proxied by participants' similarity judgments) are encoded in the syntactic distributions, but since PCA and NMF have been proposed as useful methods for extracting words' semantic features, we utilize both here.

3.3.3.1 Methodology Each transformation was applied to the normalized acceptability judgment data **D** using the implementation available in the `decomposition` module in version 0.16.1 of the python `sklearn` package (Pedregosa et al., 2011). Prior to applying either PCA transformation to the normalized acceptability judgment data, the frame judgments (each column of Figure 5) were first mean-centered and standardized. Prior to applying either NMF transformation to the normalized acceptability judgment data, the frame judgments were first shifted, so that the minimum value was zero, and then divided by the resulting maximum, so that all values lie on the unit interval and the highest value in each column was equal to one. For PCA, all 30 principal components—equal to the number of frames—were retained for later prediction; for sparse PCA, NMF, and sparse NMF, 30 components were learned. For sparse PCA and sparse NMF, grid search was used to select the regularization parameter from the values 0.1, 0.2, 0.5, 1.0, 2.0, 5.0, and 10.0 based on the cross-validation described below. For sparse NMF, we induce sparsity on the verb representation.

Each transformed dataset was entered into both a standard linear regression and linear regressions with L1 regularization (LASSO regression; Tibshirani 1996). A fixed effects structure analogous to that used in the last section was used here. As for the transformation sparsity, grid search was used to select the regularization parameter from the values 0.0 (standard linear regression), 0.1, 0.2, 0.5, 1.0, 2.0, 5.0, and 10.0 based on the cross-validation described below. To validate these regressions, 30-fold cross-validation was used, wherein the transformed data and similarities for one verb was removed from the training set, the model fit, and the similarities for that verb predicted based on the transformed data. This was carried out for all verbs

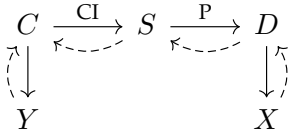


Figure 14: Schematization of our own methodology in relation to the model presented above.

in all transformations. The average predicted r^2 over verbs for each pairing of transformation and class was then computed by averaging over the cross-validation predicted r^2.

3.3.3.2 Results The model-transformation combination with the best such mean predicted r^2 for the generalized discrimination similarity dataset was sparse PCA (α=2.0) with no regularization (mean predicted r^2=0.046) and the best for the ordinal scale similarity dataset was PCA with a small amount (λ=0.1) of regularization (mean predicted r^2=0.096).

Overall, the mean predicted r^2 for the ordinal scale similarities is slightly higher than that seen for the classification-based model presented in the last section, explaining more than half (0.558) as much variance as the similarity-based model. In contrast, the mean predicted r^2 for the generalized discrimination similarities is quite a bit lower than that found for the classification, explaining only about a quarter as much variance (0.267) as the similarity-based model. This suggests that we have improved upon the classification derived from traditional distributional analysis as far as the ordinal scale judgments are concerned, though the generalized discrimination judgments are predicted equally well.

Figure 13 shows the same kind of by-verb breakdown of predicted r^2 seen in the last section. As in that section, similarities for representational verbs are predicted well on average, and similarities for verbs involving negation—*bother*, *worry*, *deny*, *forbid*, *forget*, and *hate*—are predicted poorly. This suggests two things: (i) whatever component the negative verbs share is not tracked in the syntactic contexts we tested; and (ii) the original classification was likely right to exclude this component as a grammatically relevant semantic feature, at least in English.

3.3.4 A sketch of a quantitative method for discovering attitude classifications

The above results are useful as an omnibus measure of how much semantic information lies in attitude verb distributions, but they do not tell us how that semantic information is structured. To investigate this, we need some way of parceling out which aspects of the syntactic distributions are predictive of the similarity judgments. Doing this constitutes an augmentation of traditional distributional analysis in the sense that, like the linguist, we employ pairings of concepts and syntactic distributions $(\mathbf{c}_i, \mathbf{d}_i)$ to construct a representation of the grammatically relevant semantic features $\mathbf{s}_i$.

Luckily, this augmentation requires only a conceptual step; we have already described all the quantitative methods that are needed. (Figure 14 shows a combined schematization of our analyses.) Note that, if the combination of normalization and transformation carried out on the acceptability judgments presented in Section 2 can be viewed as extracting grammatically relevant semantic features $\mathbf{S}$ and if the normalization of the similarity judgments described in this section can viewed as extracting an estimate of $\mathbf{C}$ (or at least something sufficiently related to $\mathbf{C}$), then we need merely assess which aspects of the grammatically relevant semantic features $\mathbf{S}$ are most predictive of the our estimate of $\mathbf{C}$. This can be done quite straightforwardly by analyzing the regression coefficients for the model from the last section. In the remainder of this section, we focus in on the model-transformation combinations from the last section with

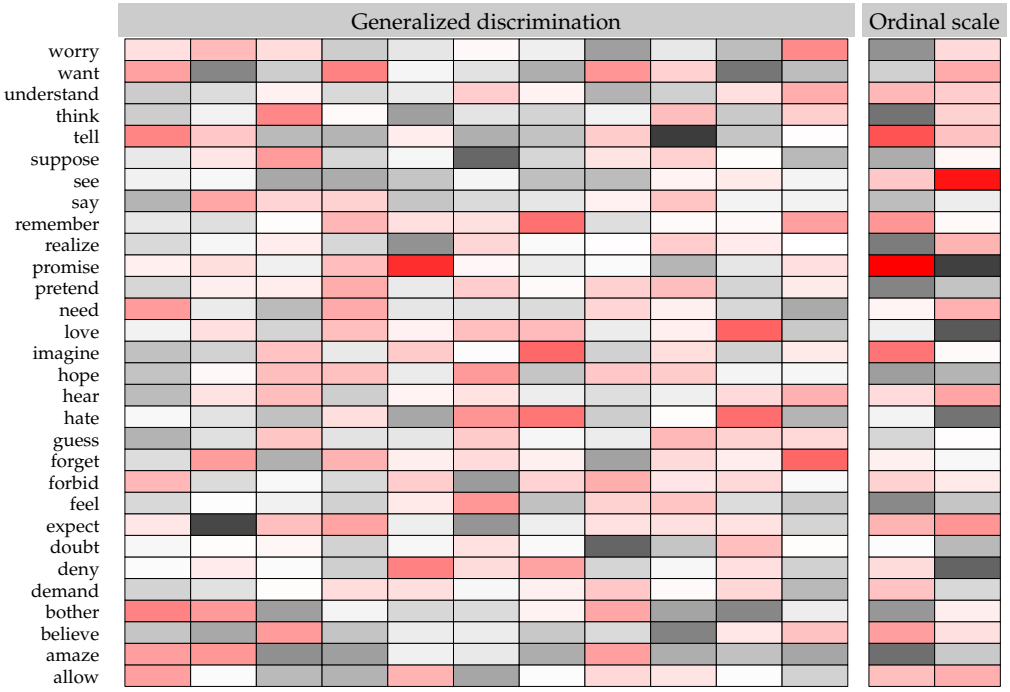


Figure 15: Features that contribute positively to predicting similarity on average (estimate of $\mathbf{S}$).

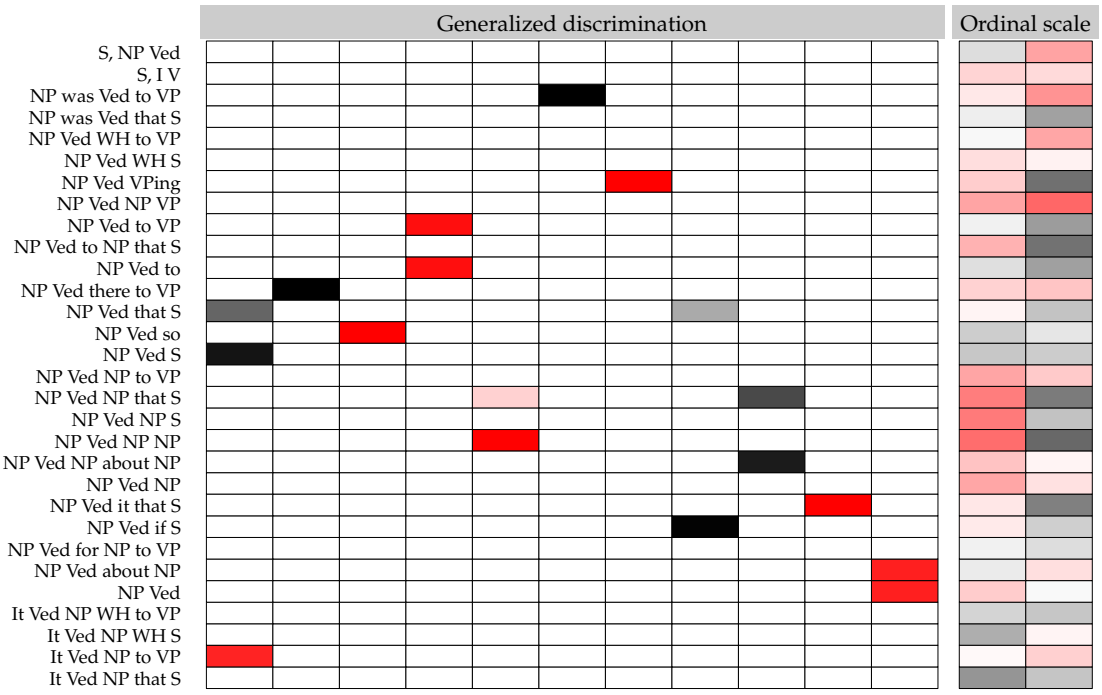


Figure 16: Mapping from positively contributing features to syntactic distributions (estimate of $\mathbf{P}$).

the best mean predicted r^2: sparse PCA (α=2.0) with no regularization for the generalized discrimination similarity dataset and PCA with a small amount (λ=0.1) of regularization for the ordinal scale similarity dataset.

Ideally, we would be able to analyze the coefficients directly; the more positive the coefficient for a dimensions of $\mathbf{S}$ extracted using, e.g., PCA, the more important that feature for predicting similarity. But since our models necessarily included interactions, this is not possible. To circumvent this issue, we instead ask how positively a particular feature influences predictions across datapoints. To do this, we construct the design matrix for the transformation with the best cross-validation predicted r^2 for each dataset. We then scale each column in the design matrix by its corresponding coefficient but do not sum as one would do to get the expected similarity. Note that this results in three columns relevant to each feature in $\mathbf{S}$: one for the main effect of each verb being compared and one for the interaction. We take the sum for each of these grouping of three to get a weight for each feature in $\mathbf{S}$ for each datapoint, and then average these weights. A positive mean for a particular feature means that, on the whole, this feature contributes to two verbs being similar.

Figure 15 shows, for each similarity dataset, all the features in $\mathbf{S}$ that contribute positively to similarity in the sense described above. Figure 16 shows, for each similarity dataset, the corresponding mappings back into the original syntactic distributions, which we noted in Section 2 can be thought of as estimates of the projection rules $\mathbf{P}$. In both cases, red marks negative weights and black marks positive weights. We do not analyze these results here, as we would merely like to sketch the method, and thus we leave there analysis for future investigation.

3.4 Discussion

In this section, we presented two experiments aimed at getting a measure of how similar in meaning naïve speakers take the propositional attitude verbs from Experiment 1 to be. We analyzed the data from these experiments in three ways. First, we compared the similarity judgments against each other, finding that there is about as much agreement between the two similarity judgment datasets as there is between participants' response within each dataset. Second, we compared both datasets against the classification presented in Section 1, finding that that classification explains between a quarter and half as much variance in the generalized discrimination and ordinal scale datasets as that explained by the other similarity dataset, respectively. Finally, we compared both datasets against the acceptability judgments from Section 2, finding a similar pattern of results to that seen for the classification-based model, with slightly higher accuracy for the ordinal scale similarities. This higher accuracy suggests that the classification from Section 1 likely misses grammatically relevant semantic features. We then sketched a method for finding out what those features are, leaving deeper investigation for future research.

4 General discussion

We began this paper by reviewing the problem of hard words in word-learning, focusing in particular on the propositional attitude verbs. We discussed two problems for attitude verb learning: the observability problem and the multi-faceted meaning problem. We then reviewed the now-standard solution proposed in the syntactic bootstrapping literature—noting that, though this solution is fairly standard, little is known about how effective it could be for attitude verb learning, since it is still unclear how strong the correlations between attitude verb syntax and semantics are.

We addressed this by employing a methodology originally used by Fisher et al. to study high-level correlations between verb syntax and semantics. In Section 2, we presented an experiment aimed at measuring the acceptability of a variety of propositional attitude verbs in

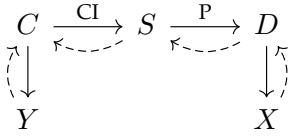


Figure 17: Schematization of our own methodology in relation to the model presented above.

different syntactic contexts and asked how well it predicted previous attitude verb classifications. We found that all classes within this classification were predictable given the syntax.

In Section 3, we presented two experiments aimed at getting a measure of how similar in meaning naïve speakers take the propositional attitude verbs from the first experiment to be. In that section, we asked (i) how well the classification referenced above can be used to predict the similarity judgments and (ii) how well the data from the first experiment can be used to predict these same judgments. We found that both predict the similarity judgments surprisingly well but for cases that involve negativity in some way—a feature that was not included in the original classification and which does not appear to be tracked in the syntax. We further found that the acceptability judgment-based model was better able to predict the ordinal scale similarities, suggesting that the classification derived from traditional distributional analysis has missed some grammatically relevant semantic feature(s). We then gave a sketch of a method that extracts those features, leaving deeper analysis of these features for future work. We take these results to be indicative that a quite large amount about an attitude verb's fine-grained semantics could in principle be learned from syntactic distributions.

In the remainder of the paper, we discuss these results relative to this paper's motivating themes, rounding back to the relationship between the methods we describe above and traditional distributional analysis as well as drawing explicit links to future cross-linguistic investigations.

4.1 Augmenting traditional distributional analysis, redux

In Section 1, we introduced a schematization of both the architectural assumptions made in traditional distributional analysis as well as the methodology that Fisher et al. (1991) used to quantitatively investigate high-level correlations between verb syntax and semantics. We then used this schematization to structure the analyses we deployed throughout the paper, which can be found repeated in Figure 17. This schema admits of three primary types of cognitive objects—a space of concepts C, a space of grammatically relevant semantic features S, and a space of syntactic distributions D—linked by two mappings—a mapping **CI** from the concept space to the space of grammatically relevant semantic features and a mapping **P** from the space of grammatically relevant semantic features to the syntactic distribution space (the projection rules).

Assuming this cognitive architecture, the problem that the verb-learner faces, given a word's syntactic distribution, is to reverse these mappings to discover the concept (or subspace of concepts) that is associated with that word. Embodied in a cognitive mechanism, this reversal is one way of viewing syntactic bootstrapping (Landau and Gleitman, 1985; Pinker, 1989; Gleitman, 1990; Kako, 1997; Lidz et al., 2004). The question that arises here is how precisely this reversal could determine the concept a word is associated with. This question is particularly pressing in the case of propositional attitude verb semantics, since (i) it seems likely that attitude verb learning is highly dependent on information from syntactic distribution and (ii) this specific domain remains quantitatively understudied, likely because these words' abstract nature makes their semantics hard to experimentally measure. It is worth reflecting on this measurement issue, both on the meaning side and on the syntactic distribution side.

4.1.1 Measuring meaning

Following Fisher et al., we utilized a quite general form of measurement, semantic similarity judgments embodied in a dataset $\mathbf{y}$, to approach this problem. This form of measurement is useful since it is possible to gain a bird's eye view of the landscape, and there are ready methods for working with these data, which we utilized above. One issue with this approach is that, though it is possible to extract quite fine-grained information from these judgments, it is not necessarily easy to label some of the aspects of the semantics that are discovered. One need only look to Figures 15 and 16 to see that this labeling can be quite hard.

A remedy for this is to replace semantic similarity judgments with a more targeted measure. For instance, if one were interested in factivity specifically, it would be straightforward to replace the model linking the transformed dataset $\mathbf{S}$ to the normalized similarity similarity judgments $\mathbf{C}$ with one that links $\mathbf{S}$ to force choice judgments of the kind employed in, e.g., Karttunen et al. 2014; Križ and Chemla 2015 or reaction times, as in Schwarz 2014. We are actively pursuing this direction.

4.1.2 Altering representational assumptions

In this paper, we utilized off-the-shelf data transformations for deriving grammatically relevant semantic features $\mathbf{S}$ from the normalized acceptability judgments $\mathbf{D}$. This is useful since it makes our results comparable to others', but it also forces us into representational assumptions we may or may not wish to make when thinking about how these findings relate to psychological representations of syntax-semantics. For instance, in using PCA or NMF, we commit ourselves to a real-valued semantic feature space S. Is this the correct sort of representation for grammatically relevant semantic features?

In Section 3's investigation of the differing capabilities of PCA and NMF to extract semantic features $\mathbf{S}$ that explain the similarity judgments, we have given preliminary results in the outcome of altering representational assumptions. But we believe that further investigating what tweaking the assumptions a matrix factorization makes regarding the formal properties of semantic feature will be a fruitful future direction. For more discussion of this, see White 2015, which gives a method for discovering discrete (binary) semantic features from both normalized acceptability judgments and corpus frequency-based syntactic distributions.

This second foray—into corpus frequency-based syntactic distributions—is important. If any further use is to be put to the methods and models described in this paper, it is necessary to show how the sorts of methods we deploy above can be adapted to such data. And while in-roads have been made in investigating the information carried in attitude verb syntactic distributions in corpora (see Buttery and Korhonen, 2005; Buttery, 2006; Barak et al., 2013, 2014; White, 2015), there is still much work to be done. One area in particular that we hope this paper opens up in earnest is the use of vector-based models, such as the ones deployed above, for doing work in areas currently served only by traditional distributional analysis. Recent work on composing such vector-based representations—such as that found in Murphy et al. 2012; Fyshe et al. 2013, 2014, 2015; Socher et al. 2012, 2013; Socher 2014—and those based explicitly on syntactic representations of a word's context, such as that found in Levy and Goldberg 2014 (following Mikolov et al. 2013)[40]—may be particularly fruitful if coupled with a strong theoretical foundation.[41]

[40] See also Boyd-Graber and Blei (2009) for a very similar idea couched in a topic modeling framework.

[41] There are many other models of verb representation that may turn out to be useful here. See Footnote 26 for some references, and see White 2015 for a review.

4.2 Cross-linguistic instability

Throughout this paper we have seen that the representational-preferential distinction is quite robustly tracked by the syntax. One of the best indicators of this distinction in English is tense, corroborating claims in the attitude verbs literature. One question that arises here is how cross-linguistically stable this correlation is. The answer is that it appears not to be very stable, yet learners still learn these words at similar points in development (Perner et al., 2003).

This instability arises in at least two ways: languages where the distinction is roughly tracked by mood—in the Romance languages, representationals tend to take indicative mood and preferentials tend to take subjunctive mood (Bolinger, 1968; Hooper, 1975; Farkas, 1985; Portner, 1992; Giorgi and Pianesi, 1997; Giannakidou, 1997; Quer, 1998; Villalta, 2000, 2008, a.o.)—and languages where the distinction is tracked by the availability of verb second (V2) syntax (Truckenbrodt, 2006; Scheffler, 2009).

An instance of the correlation with mood can be seen in Spanish. In Spanish both the representational (belief) verb *creer* (*think/believe*) and the preferential (desire) verb *querer* (*want*) take finite subordinate clauses. The difference between these subordinate clauses is that, whereas verbs like *creer* (*think*) take subordinate clauses with verbs inflected for indicative mood (42a), verbs like *querer* (*want*) take subordinate clauses with verbs inflected for subjunctive mood (42b).

(42) a. Creo que Peter va a la casa.
think.1S.PRES that Peter go.PRES.IND to the house.
b. Quiero que Peter vaya a la casa.
want.1S.PRES that Peter go.PRES.SBJ to the house.

This makes the subordinate clause under *creer* look more like the declarative main clause in Spanish, whose tensed verb is inflected for indicative mood.

(43) Peter va a la casa.
Peter go.PRES.IND to the house.

An instance of the correlation with V2 can be seen in German and other Germanic languages—e.g. Dutch. V2, which is generally found in main clauses, is a phenomenon in which a clause's tensed verb appears as the second word in a sentence. For instance, (44) shows a German main clause with the tensed form of the auxiliary verb *sein* (*be*) occurring as the second word of the sentence (in second position).

(44) Peter ist nach Hausen gegangen
Peter is to home gone

In subordinate clauses headed by the complementizer *dass* (*that*), this verb occurs clause-finally, which evidences the fact that German is underlyingly a subject-object-verb (SOV) language. Both the verb *glauben* (*think*) and the verb *wollen* (*want*) can take such clauses, in which the main verb is tensed.

(45) a. Ich glaube, dass Peter nach Hausen gegangen ist.
I think that Peter to home gone is.
b. Ich will, dass Peter nach Hausen geht.
I want that Peter to home goes.

Only *glauben* (*think*), however, allows a second sort of structure more akin to the main clause in the position of the tensed verb (Scheffler, 2009). If the complementizer *dass* (*that)* is not present, *glauben* (*think*) can take a subordinate clause with syntax that looks exactly like that of the main

clause—compare the main clause in (44) with the subordinate clause in (46a). *Wollen* does not allow this (46b).

(46) a. Ich glaube, Peter ist nach Hausen gegangen.
I think Peter is to home gone.
b. *Ich will, Peter geht nach Hausen.
I want Peter goes to home.

Thus, though both Spanish and German take tensed complements, militating against a hard-coded link between tense and representationality, they still show language-internal correlations between representationality and some more abstract aspect of the clausal syntax. Further, the aspect of the clausal syntax that occurs with only the representational verbs—indicative mood in Spanish and V2 in German—also tends to show up in declarative main clauses.

4.3 Main clause syntax

This apparent language-internal correlation has led some authors to conclude that, rather than there being a relationship directly between representationality and tense, as is evidenced in English, the relationship needs to be specified more abstractly. One idea is that this more abstract mapping between semantics and syntax should be specified in terms of *(declarative) main clause syntax* (Dayal and Grimshaw, 2009; Hacquard, 2014). One reason that such a correlation might exist is that (declarative) main clauses are often used to assert content and many representationals are assertive.

Under this view, the apparent relationship between tense in English, mood in Spanish (and the rest of Romance), and V2 in German (and other Germanic languages besides English) is really the outgrowth of a more abstract relationship between some cluster of syntactic features—call them MAIN CLAUSE features—that are language-specific but likely highly constrained. The way in which they are constrained is that they tend to be associated with properties of the subordinate clause's that are "close" to the attitude verb. For instance, both complementizers and mood tend to be assumed to be quite high within the clausal structure (cf. Cinque, 1999; Speas, 2004), which in turn seems to make them amenable to selection by particular semantic classes of verbs—e.g. representationals or preferentials. Indeed, ideally, one could pin the relevant feature to some particular type of head which carries the relevant selection information—e.g. the complementizer— and is "as high as possible" within the subordinate clause so as to make selection maximally local.

Suggestive of this possibility is that the standard analysis of German V2, which has that V2 is a particular kind of complementizer-driven movement akin to that seen in English WH-movement (Den Besten, 1983). English may be amenable to such an analysis in the sense that complementizer drop with finite subordinate clauses tends to only occur with representationals (Dayal and Grimshaw, 2009).

(47) a. Bo {thinks, believes, knows} (that) Jo is out of town.
b. Bo {loves, hates} *(that) Jo is out of town.

This latter fact is furthermore suggestive, since of course English main clauses do not have complementizers, bolstering the relationship between main clause syntax and representationality, at least in English. This, however, also raises a potential problem for languages like Spanish, which lack complementizer drop in any subordinate clauses but whose declarative main clauses do not have complementizers.

But regardless of whether main clause syntax information can be carried solely in the complementizers themselves—thus allowing for an extremely local form of selection giving rise to the relationship between representationality and main clause syntax—or whether somewhat longer distance relationships need to be posited, there is nonetheless a potential rela-

tionship between the representational-preferential distinction and this language-specific-yet-highly-constrained MAIN CLAUSE feature.

The importance of this for current purposes is that, if such a correlation between representational and main clause syntax exists, it may signal a possible candidate for a hard-coded-yet-flexible projection rule that allows for a solution to the labeling problem in this particular case. Further, since the main clause syntax itself is presumably observable to the same extent that subordinate clause syntax is, the language specific instantiation of the MAIN CLAUSE feature may well itself be learnable. We are currently pursuing this idea. See White 2015 for a preliminary investigation of the efficacy of such an approach.

References

Abbott, Barbara. 2006. Where have some of the presuppositions gone. *Drawing the boundaries of meaning: Neo-Gricean studies in pragmatics and semantics in honor of Laurence R. Horn* 1–20.

Abels, Klaus. 2004. Why surprise-predicates do not embed polar interrogatives. In *Linguistische Arbeitsberichte*. Leipzig: Universität Leipzig.

Abusch, Dorit. 2002. Lexical alternatives as a source of pragmatic presuppositions. In *Proceedings of SALT*, volume 12, 1–19.

Agresti, Alan. 2014. *Categorical data analysis*. John Wiley & Sons.

Akaike, Hirotugu. 1974. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on* 19:716–723.

Alishahi, Afra, and Suzanne Stevenson. 2008. A computational model of early argument structure acquisition. *Cognitive science* 32:789–834.

Anand, Pranav, and Valentine Hacquard. 2013. Epistemics and attitudes. *Semantics and Pragmatics* 6:1–59.

Anand, Pranav, and Valentine Hacquard. 2014. Factivity, Belief and Discourse. In *The Art and Craft of Semantics: A Festschrift for Irene Heim*, ed. Luka Crnič and Uli Sauerland, volume 1, 69–90. Cambride, MA: MIT Working Papers in Linguistics.

Andersen, Erling B. 1977. Sufficient statistics and latent trait models. *Psychometrika* 42:69–81.

Andrich, David. 1978. A rating formulation for ordered response categories. *Psychometrika* 43:561–573.

Asher, Nicholas. 2000. Truth conditional discourse semantics for parentheticals. *Journal of Semantics* 17:31–50.

Baker, Carl L. 1979. Syntactic theory and the projection problem. *Linguistic Inquiry* 533–581.

Baker, Collin F., Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, 86–90. Association for Computational Linguistics.

Baker, Mark C. 1988. *Incorporation: A theory of grammatical function changing*. University of Chicago Press Chicago.

Barak, Libby, Afsaneh Fazly, and Suzanne Stevenson. 2014. Gradual Acquisition of Mental State Meaning: A Computational Investigation. In *Proceedings of the Thirty-Sixth Annual Conference of the Cognitive Science Society*.

Barak, Libby, Afsaneh Fazly, and Suzanne Stevenson. 2013. Acquisition of Desires before Beliefs: A Computational Investigation. In *Proceedings of the 17th Conference on Computational Natural Language Learning (CoNLL)*.

Bastien, Frédéric, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian Goodfellow, Arnaud Bergeron, Nicolas Bouchard, David Warde-Farley, and Yoshua Bengio. 2012. Theano: new features and speed improvements. *arXiv preprint arXiv:1211.5590* .

Bergstra, James, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio. 2010. Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for scientific computing conference (SciPy)*, volume 4, 3. Austin, TX.

Berwick, Robert C. 1985. *The Acquisition of Syntactic Knowledge*. Cambridge, MA: MIT Press.

Bolinger, Dwight. 1968. Postposed main phrases: an English rule for the Romance subjunctive. *Canadian Journal of Linguistics* 14:3–30.

Boyd-Graber, Jordan L., and David M. Blei. 2009. Syntactic topic models. In *Advances in neural information processing systems*, 185–192.

Buttery, Paula, and Anna Korhonen. 2005. Large-scale analysis of verb subcategorization differences between child directed speech and adult speech. In *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*.

Buttery, Paula J. 2006. Computational models for first language acquisition. Doctoral Dissertation, University of Cambridge.

Carey, Susan, and Elsa Bartlett. 1978. Acquiring a single new word. *Papers and Reports on Child Language Development* 15:17–29.

Carter, Richard. 1976. Some linking regularities. *On Linking: Papers by Richard Carter Cambridge MA: Center for Cognitive Science, MIT (Lexicon Project Working Papers No. 25)* .

Chomsky, Noam. 1981. *Lectures on Government and Binding: The Pisa Lectures*. Walter de Gruyter.

Cinque, Guglielmo. 1999. *Adverbs and functional heads: A cross-linguistic perspective*. Oxford University Press.

Connor, Michael, Cynthia Fisher, and Dan Roth. 2013. Starting from scratch in semantic role labeling: Early indirect supervision. In *Cognitive aspects of computational language acquisition*, 257–296. Springer.

Darken, Christian, and John Moody. 1990. Note on Learning Rate Schedules for Stochastic Optimization. In *Proceedings of the 1990 Conference on Advances in Neural Information Processing Systems 3*, NIPS-3, 832–838. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. URL `http://dl.acm.org/citation.cfm?id=118850.119956`.

Dayal, Veneeta, and Jane Grimshaw. 2009. Subordination at the interface: the Quasi-Subordination Hypothesis.

Den Besten, Hans. 1983. On the interaction of root transformations and lexical deletive rules. *On the formal syntax of the Westgermania* 47–131.

Depiante, Marcela Andrea. 2000. The syntax of deep and surface anaphora: a study of null complement anaphora and stripping/bare argument ellipsis. Doctoral Dissertation, University of Connecticut.

Dowty, David. 1991. Thematic proto-roles and argument selection. *Language* 67:547–619.

Drummond, Alex. 2014. Ibex. URL `https://github.com/addrummond/ibex`.

Egré, Paul. 2008. Question-embedding and factivity. *Grazer Philosophische Studien* 77:85–125.

Farkas, Donka. 1985. *Intensional descriptions and the Romance subjunctive mood*. Taylor & Francis.

Fillmore, Charles John. 1970. The Grammar of Hitting and Breaking. In *Readings in English Transformational Grammar*, ed. R.A. Jacobs and P.S. Rosenbaum, 120–133. Waltham, MA,: Ginn.

Fisher, Cynthia, Henry Gleitman, and Lila R. Gleitman. 1991. On the semantic content of subcategorization frames. *Cognitive psychology* 23:331–392.

Fodor, Jerry, and Ernie Lepore. 1999. Impossible Words? *Linguistic Inquiry* 30:445–453. URL `http://www.jstor.org/stable/4179071`.

Fodor, Jerry A., and Ernie Lepore. 1998. The emptiness of the lexicon: reflections on James Pustejovsky's The Generative Lexicon. *Linguistic Inquiry* 29:269–288.

Frank, Michael C., Noah D. Goodman, and Joshua B. Tenenbaum. 2009. Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science* 20:578–585.

Fyshe, Alona, Partha Talukdar, Brian Murphy, and Tom Mitchell. 2013. Documents and Dependencies: an Exploration of Vector Space Models for Semantic Composition. *CoNLL-2013* 84. URL `http://www.aclweb.org/anthology/W/W13/W13-35.pdf#page=96`.

Fyshe, Alona, Partha P. Talukdar, Brian Murphy, and Tom M. Mitchell. 2014. Interpretable semantic vectors from a joint model of brain-and text-based meaning. In *Proceedings of ACL*.

Fyshe, Alona, Leila Wehbe, Partha P. Talukdar, Brian Murphy, and Tom M. Mitchell. 2015. A Compositional and Interpretable Semantic Space. In *Proceedings of the NAACL-HLT*. Denver.

Giannakidou, Anastasia. 1997. The landscape of polarity items. Doctoral Dissertation, University of Groningen.

Gillette, Jane, Henry Gleitman, Lila Gleitman, and Anne Lederer. 1999. Human simulations of vocabulary learning. *Cognition* 73:135–176.

Ginzburg, Jonathan. 1995. Resolving questions, II. *Linguistics and Philosophy* 18:567–609.

Giorgi, Alessandra, and Fabio Pianesi. 1997. *Tense and Aspect: Form Semantics to Morphosyntax*. Oxford: Oxford University Press.

Gleitman, Lila. 1990. The structural sources of verb meanings. *Language acquisition* 1:3–55.

Gleitman, Lila R., Kimberly Cassidy, Rebecca Nappa, Anna Papafragou, and John C. Trueswell. 2005. Hard words. *Language Learning and Development* 1:23–64.

Goodman, Nelson. 1955. *Fact, fiction, and forecast*. Harvard University Press.

Grimshaw, Jane. 1979. Complement selection and the lexicon. *Linguistic inquiry* 10:279–326.

Grimshaw, Jane. 1990. *Argument structure*. Cambridge, MA: MIT Press.

Grimshaw, Jane. 1994. Lexical reconciliation. *Lingua* 92:411–430.

Grimshaw, Jane. 2009. That's nothing: the grammar of complementizer omission.

Gruber, Jeffrey Steven. 1965. Studies in lexical relations. Doctoral Dissertation, Massachusetts Institute of Technology.

Guerzoni, Elena. 2007. Weak Exhaustivity and 'Whether': A Pragmatic Approach. In *Semantics and Linguistic Theory 17*, 112–129.

Hacquard, Valentine. 2014. Bootstrapping attitudes. In *Semantics and Linguistic Theory*, volume 24, 330–352.

Hacquard, Valentine, and Alexis Wellwood. 2012. Embedding epistemic modals in English: A corpus-based study. *Semantics and Pragmatics* 5:1–29.

Hale, Ken, and Samuel Jay Keyser. 2002. *Prolegomena to a Theory of Argument Structure*. Cambridge, MA: MIT Press.

Hankamer, Jorge, and Ivan Sag. 1976. Deep and surface anaphora. *Linguistic inquiry* 391–428.

Harrigan, Kaitlyn. 2015. Syntactic bootstrapping in the acquisition of attitude verbs. Doctoral Dissertation, University of Maryland.

Heim, Irene. 1992. Presupposition projection and the semantics of attitude verbs. *Journal of semantics* 9:183–221.

Hintikka, Jaakko. 1975. Different Constructions in Terms of the Basic Epistemological Verbs: A Survey of Some Problems and Proposals. In *The Intentions of Intentionality and Other New Models for Modalities*, 1–25. Dordrecht: D. Reidel.

Hooper, Joan B. 1975. On Assertive Predicates. In *Syntax and Semantics*, ed. John P. Kimball, volume 4, 91–124. New York: Academy Press.

Horn, Laurence Robert. 1972. On the semantic properties of logical operators in English. Doctoral Dissertation, UCLA.

Jackendoff, Ray. 1972. *Semantic interpretation in generative grammar*. MIT press Cambridge, MA.

Jolliffe, Ian. 2002. *Principal component analysis*. Wiley Online Library.

Kako, Edward. 1997. Subcategorization Semantics and the Naturalness of Verb-Frame Pairings. *University of Pennsylvania Working Papers in Linguistics* 4:11.

Karttunen, Lauri. 1971. Some observations on factivity. *Paper in Linguistics* 4:55–69. URL `http://www.tandfonline.com/doi/abs/10.1080/08351817109370248`.

Karttunen, Lauri. 1977. Syntax and semantics of questions. *Linguistics and philosophy* 1:3–44.

Karttunen, Lauri, Stanley Peters, Annie Zaenen, and Cleo Condoravdi. 2014. The Chameleon-like Nature of Evaluative Adjectives. In *Empirical Issues in Syntax and Semantics 10*, ed. Christopher Piñón, 233–250. CSSP-CNRS.

Kilgarriff, Adam. 1997. I don't believe in word senses. *Computers and the Humanities* 31:91–113.

Kiparsky, Paul, and Carol Kiparsky. 1970. Fact. In *Progress in Linguistics: A Collection of Papers*, ed. Manfred Bierwisch and Karl Erich Heidolph, 143–173. The Hague: Mouton.

Kipper, Karin, Anna Korhonen, Neville Ryant, and Martha Palmer. 2006. Extending VerbNet with novel verb classes. In *Proceedings of LREC*, volume 2006, 1. Citeseer.

Kipper-Schuler, Karin. 2005. VerbNet: A broad-coverage, comprehensive verb lexicon. Doctoral Dissertation, University of Pennsylvania.

Korhonen, Anna. 2002. Subcategorization Acquisition. Doctoral Dissertation, University of Cambridge.

Korhonen, Anna, and Ted Briscoe. 2004. Extended lexical-semantic classification of English verbs. In *Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics*, 38–45. Association for Computational Linguistics.

Kripke, Saul A. 1982. *Wittgenstein on rules and private language: An elementary exposition*. Harvard University Press.

Križ, Manuel, and Emmanuel Chemla. 2015. Two methods to find truth-value gaps and their application to the projection problem of homogeneity. *Natural Language Semantics* 23:205–248. URL `http://link.springer.com/article/10.1007/s11050-015-9114-z`.

Lahiri, Utpal. 2002. *Questions and answers in embedded contexts*. Oxford University Press.

Landau, Barbara, and Lila R. Gleitman. 1985. *Language and experience: Evidence from the blind child*, volume 8. Harvard University Press.

Landauer, Thomas K., and Susan T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review* 104:211.

Lasnik, Howard. 1989. On certain substitutes for negative data. In *Learnability and linguistic theory*, 89–105. Springer.

Lederer, Anne, Henry Gleitman, and Lila Gleitman. 1995. Verbs of a feather flock together: Semantic information in the structure of maternal speech. *Beyond names for things: Young children's acquisition of verbs* 277.

Lee, Daniel D., and H. Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401:788–791.

Levin, Beth. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago Press.

Levin, Beth, and Malka Rappaport Hovav. 2005. *Argument realization*. Cambridge University Press.

Levy, Omer, and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, 302–308.

Lidz, Jeffrey, Henry Gleitman, and Lila Gleitman. 2004. Kidz in the 'hood: Syntactic bootstrapping and the mental lexicon. In *Weaving a Lexicon*, ed. D.G. Hall and S.R. Waxman, 603–636. Cambridge, MA: MIT Press.

Markman, Ellen M. 1990. Constraints children place on word meanings. *Cognitive Science* 14:57–77.

Markman, Ellen M., and Jean E. Hutchinson. 1984. Children's sensitivity to constraints on word meaning: Taxonomic versus thematic relations. *Cognitive psychology* 16:1–27.

Markman, Ellen M., and Gwyn F. Wachtel. 1988. Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive psychology* 20:121–157.

Marr, David. 1982. Vision: a computational investigation into the human representation and processing of visual information. *Henry Holt and Co.* .

Masters, Geoff N. 1982. A Rasch model for partial credit scoring. *Psychometrika* 47:149–174.

McClelland, James L., and David E. Rumelhart. 1986. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Psychological and Biological Models*, volume 2. Cambridge, MA: MIT Press.

Merlo, Paola, and Suzanne Stevenson. 2001. Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics* 27:373–408.

Merriman, William E., and Laura L. Bowman. 1989. *The mutual exclusivity bias in children's word learning*. Number 220 in Monographs of the Society for Research in Child Development. Hoboken, NJ: John Wiley & Sons.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

Mintz, Mike, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, 1003–1011. Association for Computational Linguistics.

Moulton, Keir. 2009a. Clausal Complementation and the Wager-class. In *Proceedings of the 38th annual meeting of the North East Linguistic Society*, ed. Anisa Schardl, Martin Walkow, and Muhammad Abdurrahman, 165–178. Amherst, MA: GLSA.

Moulton, Keir. 2009b. Natural selection and the syntax of clausal complementation. Doctoral Dissertation, University of Massachusetts, Amherst.

Murphy, Brian, Partha Pratim Talukdar, and Tom M. Mitchell. 2012. Learning Effective and Interpretable Semantic Models using Non-Negative Sparse Embedding. In *COLING*, 1933–1950.

Papafragou, Anna, Kimberly Cassidy, and Lila Gleitman. 2007. When we think about thinking: The acquisition of belief verbs. *Cognition* 105:125–165. URL `http://www.sciencedirect.com/science/article/pii/S0010027706002009`.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 12:2825–2830. URL `http://dl.acm.org/citation.cfm?id=1953048.2078195`.

Perner, Josef, Manuel Sprung, Petra Zauner, and Hubert Haider. 2003. Want That is Understood Well before Say that, Think That, and False Belief: A Test of de Villiers's Linguistic Determinism on German–Speaking Children. *Child development* 74:179–188.

Pesetsky, David. 1991. Zero syntax: vol. 2: Infinitives.

Pesetsky, David Michael. 1982. Paths and categories. Doctoral Dissertation, Massachusetts Institute of Technology.

Pinker, Steven. 1989. *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA: MIT Press.

Pinker, Steven. 1994. How could a child use verb syntax to learn verb semantics? *Lingua* 92:377–410.

Portner, Paul, and Aynat Rubinstein. 2013. Mood and contextual commitment. In *Proceedings of SALT*, volume 22, 461–487.

Portner, Paul Howard. 1992. Situation theory and the semantics of propositional expressions. Doctoral Dissertation, University of Massachusetts, Amherst.

Postal, Paul M. 1993. Some defective paradigms. *Linguistic Inquiry* 347–364.

Postal, Paul Martin. 1974. *On raising: one rule of English grammar and its theoretical implications*. Current Studies in Linguistics. Cambridge, MA: MIT Press.

Quer, Josep. 1998. Mood at the Interface. Doctoral Dissertation, Utrecht Institute of Linguistics, OTS.

Quine, Willard Van Orman. 1960. *Word and object*. MIT press.

Rasch, Georg. 1960. *Probabilistic models for some intelligence and attainment tests.*. Chicago: The University of Chicago Press.

Rawlins, Kyle. 2013. About 'about'. In *Proceedings of the 23rd Semantics and Linguistic Theory Conference*, 336–357.

Reinhart, Tanya. 1983. Point of view in language–The use of parentheticals. In *Essays on Deixis*, ed. Gisa Rauh, volume 188, 169–194. Tübingen: Narr.

Resnik, Philip. 1996. Selectional constraints: An information-theoretic model and its computational realization. *Cognition* 61:127–159.

Romoli, Jacopo. 2011. The presuppositions of soft triggers aren't presuppositions. In *Proceedings of SALT*, volume 21, 236–256.

Rooryck, Johan. 2001. Evidentiality, part I. *Glot International* 5:125–133.

Rooth, Mats. 1995. Two-dimensional clusters in grammatical relations. In *AAAI Symposium on representation and acquisition of lexical knowledge*.

Ross, John Robert. 1973. Slifting. In *The formal analysis of natural languages*, ed. Maurice Gross, Morris Halle, and Marcel-Paul Schützenberger, 133–169. The Hague: Mouton de Gruyter.

Rumelhart, David E., and James L. McClelland. 1986. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*, volume 1. Cambridge, MA: MIT Press.

Scheffler, Tatjana. 2009. Evidentiality and German attitude verbs. *University of Pennsylvania Working Papers in Linguistics* 15.

Schwarz, Florian. 2014. Presuppositions are Fast, whether Hard or Soft-Evidence from the visual world. In *Semantics and Linguistic Theory*, volume 24, 1–22.

Schütze, Carson T., and Jon Sprouse. 2014. Judgment data. In *Research Methods in Linguistics*, ed. Robert J. Podesva and Devyani Sharma, 27–50. Cambridge University Press.

Simons, Mandy. 2001. On the conversational basis of some presuppositions. In *Proceedings of Semantics and Linguistic Theory 11*, ed. R. Hasting, B. Jackson, and Z. Zvolensky, 431–448. Ithaca, NY: Cornell University.

Simons, Mandy. 2007. Observations on embedding verbs, evidentiality, and presupposition. *Lingua* 117:1034–1056.

Snedeker, Jesse, and Lila Gleitman. 2004. Why it is hard to label our concepts. *Weaving a lexicon* 257–294.

Snow, Rion, Daniel Jurafsky, and Andrew Y. Ng. 2004. Learning syntactic patterns for automatic hypernym discovery. *Advances in Neural Information Processing Systems 17* .

Socher, Richard. 2014. Recursive Deep Learning for Natural Language Processing and Computer Vision. Doctoral Dissertation, Stanford University.

Socher, Richard, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 1201–1211. Association for Computational Linguistics.

Socher, Richard, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, 1642. Citeseer.

Speas, Margaret. 2004. Evidentiality, logophoricity and the syntactic representation of pragmatic features. *Lingua* 114:255–276.

Spector, Benjamin, and Paul Egré. 2015. A uniform semantics for embedded interrogatives: An answer, not necessarily the answer. *Synthese* 192:1729–1784.

Stalnaker, Robert. 1973. Presuppositions. *Journal of philosophical logic* 2:447–457.

Stalnaker, Robert. 1984. *Inquiry*. Cambridge University Press.

Stevenson, Suzanne, and Paola Merlo. 1999. Automatic verb classification using distributions of grammatical features. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, 45–52. Association for Computational Linguistics.

Sæbø, Kjell Johan. 2007. A whether forecast. In *Logic, Language, and Computation*, ed. B.D. ten Cate and H.W. Zeevat, 189–199. Verlag, Berlin, Heidelberg: Springer.

Tenenbaum, Joshua B., and Thomas L. Griffiths. 2001. Generalization, similarity, and Bayesian inference. *Behavioral and brain sciences* 24:629–640.

Tibshirani, Robert. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288.

Truckenbrodt, Hubert. 2006. On the semantic motivation of syntactic verb movement to C in German. *Theoretical Linguistics* 32:257–306.

Uegaki, Wataru. 2012. Content nouns and the semantics of question-embedding predicates. *Proceedings of Sinn und Bedeutung 16* .

Urmson, James O. 1952. Parenthetical verbs. *Mind* 61:480–496.

Villalta, Elisabeth. 2000. Spanish subjunctive clauses require ordered alternatives. In *Proceedings of SALT*, volume 10, 239–256.

Villalta, Elisabeth. 2008. Mood and gradability: an investigation of the subjunctive mood in Spanish. *Linguistics and Philosophy* 31:467–522.

Vlachos, Andreas, Zoubin Ghahramani, and Anna Korhonen. 2008. Dirichlet process mixture models for verb clustering. In *Proceedings of the ICML workshop on Prior Knowledge for Text and Language*.

Vlachos, Andreas, Anna Korhonen, and Zoubin Ghahramani. 2009. Unsupervised and constrained Dirichlet process mixture models for verb clustering. In *Proceedings of the workshop on geometrical models of natural language semantics*, 74–82. Association for Computational Linguistics.

Schulte im Walde, Sabine. 2000. Clustering verbs semantically according to their alternation behaviour. In *Proceedings of the 18th conference on Computational linguistics-Volume 2*, 747–753. Association for Computational Linguistics.

Schulte im Walde, Sabine. 2003. Experiments on the Automatic Induction of German Semantic Verb Classes. Doctoral Dissertation, Universität Stuttgart.

Schulte im Walde, Sabine. 2006. Experiments on the automatic induction of German semantic verb classes. *Computational Linguistics* 32:159–194.

Schulte im Walde, Sabine, and Chris Brew. 2002. Inducing German semantic verb classes from purely syntactic subcategorisation information. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 223–230. Association for Computational Linguistics.

Wexler, Kenneth N. 1970. Embedding structures for semantics. University of California, Irvine.

Wexler, Kenneth N., and Henry Hamburger. 1973. On the insufficiency of surface data for the learning of transformational languages. In *Approaches to natural language*, 167–179. Springer.

White, Aaron Steven. 2015. Information and incrementality in syntactic bootstrapping. Doctoral Dissertation, University of Maryland.

Williams, Alexander. 2012. Null Complement Anaphors as definite descriptions. In *Proceedings of SALT*, volume 22, 125–145.

Williams, Alexander. 2015. *Arguments in Syntax and Semantics*. Cambridge University Press.

Xu, Fei, and Joshua B. Tenenbaum. 2007. Word learning as Bayesian inference. *Psychological review* 114:245.

Zwicky, Arnold M. 1971. In a manner of speaking. *Linguistic Inquiry* 2:223–233.

A Appendix

A.1 Review of (cumulative logit) ordinal mixed models

In an ordinal model, one assumes that ordinal scale judgments are made on the basis of some continuous latent value (for instance, acceptability) and that these continuous latent values are associated with a discrete probability distribution over ordinal scale points (acceptability judgments). Intuitively, the larger the latent value, the higher the probability of higher scale points should be; and conversely, the smaller the latent value, the higher the probability of lower scale points should be.

In a fixed effects ordinal model, one assumes some fixed latent parameter d. In our setting, this parameter is the acceptability of a verb-frame pair.[42] To obtain a (discrete) probability

[42] In the more general regression setting, this parameter may be a linear combination of some fixed predictors.

distribution over the ordinal scale points (1 through 7) given such a parameter, one or more auxiliary variables are assumed. In the simplest such model, one assumes a single value i (for interval), which gives the distance between what are termed cutpoints $\mathbf{c}$. Cutpoints are generally assumed to be ordered from least to greatest, and in this simplest model, $c_j - c_{j-1} = i$ for all j. For convenience, one usually sets $c_0 \equiv -\infty$, $c_1 \equiv 0$, and $c_R \equiv \infty$, where R is the number of possible ordinal responses (here, 7). The response distribution associated with the latent parameter π is then given by

$$\mathbb{P}(x \leq r \mid d, i) = \text{logit}^{-1}((r-1)i - d) = \text{logit}^{-1}(c_j - d)$$

where x is the response associated with d. This model (up to isomorphism) is essentially the one assumed when "raw" acceptability judgments are averaged. And like that method, it does not take into account preferences for particular types of scale use—e.g. preferring the endpoints of midpoints of the scale.

The probability of a particular response in this model is then given in the standard way for discrete distributions.

$$\mathbb{P}(x = r \mid d, i) = \mathbb{P}(x \leq r \mid d, i) - \mathbb{P}(x < r \mid d, i) = \mathbb{P}(x \leq r \mid d, i) - \mathbb{P}(x \leq r - 1 \mid d, i)$$

A slightly more complex model allows the cutpoints to have variable distances between them. Where the previous model requires one auxiliary parameter i, this model requires $R - 2$ parameters $i_j \equiv c_j - c_{j-1}$, for $j = 2, \ldots, R-1$. The response distribution is then given only a slight modification from the above.

$$\mathbb{P}(x \leq r \mid d, \mathbf{i}) = \text{logit}^{-1}\left(\left[\sum_{j=2}^{r} i_j\right] - d\right) = \text{logit}^{-1}(c_j - d)$$

Either of these models can be augmented with a random effects component in the standard way to make a mixed effects model. For instance, we can add item random intercepts $u_{item} \sim \mathcal{N}(0, \sigma^2_{item})$ and participant random intercepts $u_{subj} \sim \mathcal{N}(0, \sigma^2_{subj})$ in the standard way.

$$\mathbb{P}(x \leq r \mid d, \mathbf{i}) = \text{logit}^{-1}(c_j - (d + u_{item} + u_{subj}))$$

This augmentation allows us to take into account the possibility that participants might differ in how often they use particular parts of the scale and that particular items might make a particular sentence type—e.g. verb-frame pair—sound better or worse.

But while this addition of random effects addresses the issues that motivated this section in the first place, it is unclear that it appropriately addresses possible participant differences in scale use, unless scale use differences are of a very particular type. By invoking an additive term to account for subject variability, we can account for participants who use the higher end of the scale more than the lower end (or vice versa) but not necessarily participants who use the endpoints more often than the midpoints or the midpoints more often than the endpoints.

To account for this, we need a slightly different type of random effect: one that scales the cutpoint distances instead of shifting the latent parameter. Assuming the same distributions on the u_{item} and u_{subj}, this model can be specified as follows.

$$\mathbb{P}(x \leq r \mid d, \mathbf{i}) = \text{logit}^{-1}(|u_{subj}|c_j - (d + u_{item}))$$

Thus, as σ^2_{subj}, the variance on u_{subj}, gets larger, we expect more participants to (tend to) use the middle of the scale; and as it gets smaller, we expect more participants to (tend to) use the endpoints. We refer to the former model, with the additive participant term, as the additive model, and the latter, with the multiplicative participant term, as the multiplicative model.

As for the additive model, the multiplicative model can be combined with either the single parameter model (the equidistant cutpoint model) or the $R-2$ parameter model (the varying cutpoints model). Thus, crossing the equidistance-varying distinction with the additive-multiplicative distinction, we generate four possible normalization models. If we furthermore allow for models with both additive effects $u_{subj-add}$ and multiplicative effects $u_{subj-mult}$, two further models result, for a total of six. (With equidistant cutpoints and without item effects, this model corresponds to standard z-scoring.)

$$\mathbb{P}(x \leq r \mid d, \mathbf{i}) = \text{logit}^{-1}(|u_{subj-mult}|c_j - (d + u_{item} + u_{subj-add}))$$

A.2 Review of matrix factorization techniques

A.2.1 A high-level overview of PCA

PCA's effect can be viewed from two perspectives. On the one hand, it can be seen as mapping the original data (the normalized verb-frame acceptability judgment matrix $\mathbf{D}$) to a new representation that preserves all the relationships (between verbs) that existed in the original data but which encodes those relationships on new, uncorrelated dimensions (the principal components). On the other hand, PCA can be seen as a form of matrix factorization, wherein the original data (the normalized verb-frame acceptability judgment matrix $\mathbf{D}$) are pulled apart (factored) into two distinct representations. As under the previous view, one of these representations encodes the relationships between verbs in the original dataset on new, uncorrelated dimensions (the principal components). But distinct from the previous view, the other representation is viewed as a mapping from the new representation to the original data $\mathbf{D}$, instead of from the original to the new.

Thus, the only difference from the previous view is the direction that the mapping is specified. The reason these two views of PCA are possible is because the mapping from the original dataset to the new one is reversible (invertible). And when this mapping from the original dataset to the new one is reversed (inverted), the mapping from the new to the old results.

We would like to point out that the second view of (the result of) PCA—the one that views PCA as factoring the original dataset, a representation of verbs' syntactic distributions, into two distinct representations, a representation of verb relationships and a mapping from that representation to the original data—is isomorphic to the model of projection we set out in Section 1. In that section, we specified the projection rules $\mathbf{P}$ as a mapping from some representation of a verb's grammatically relevant semantic features $\mathbf{s}_i$ to that verb's syntactic distribution $\mathbf{d}_i$. This isomorphism raises the possibility that PCA might serve as a model of how semantic features are discovered given syntactic distributions $\mathbf{D}$. That is, PCA might be viewed as a way of discovering grammatically relevant semantic features $\mathbf{S}$. In fact, since this method also produces a representation of the mapping from the $\mathbf{S}$ to $\mathbf{D}$, it might furthermore be viewed as discovering projection rules $\mathbf{P}$.

Indeed, this suggestion can be extended to any method that has the general form of a matrix factorization—wherein one representation is factored into two: one that encodes verb relationships in the original dataset on new dimensions and another that maps from this new encoding to the original dataset. There are many such methods, but one that has recently become popular within the computational linguistics and natural language processing literature (cf. Murphy et al., 2012; Fyshe et al., 2014, 2015) is known as Nonnegative Matrix Factorization (NMF; Lee and Seung 1999).

A.2.2 A high-level overview of NMF

Like PCA, NMF attempts to factor the syntactic distributions $\mathbf{D}$ into two representations: one that encodes verb relationships in the original dataset on new dimensions and another that

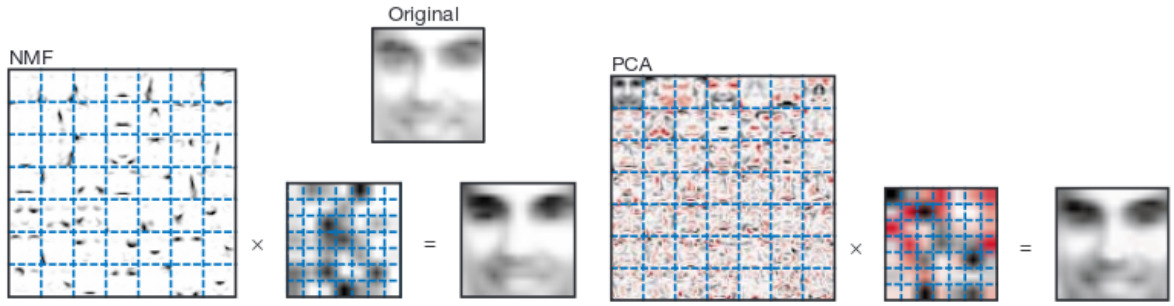


Figure 18: Example of NMF and PCA from Lee and Seung 1999.

maps from this new encoding to the original dataset. Unlike PCA, however, NMF requires the resulting representations to be non-negative (hence its name).[43] This results in what Lee and Seung (1999) refer to as a *parts-based representation* in contrast to PCA's *holistic representation*. The distinction between these two representations is roughly that, in a parts-based representation, the new dimensions created in the matrix factorization target intuitively distinct parts of the factored dataset; in contrast, in a holistic representation, intuitively distinct parts of the observed distribution are spread across all of the new dimensions, with as much information as possible packed into as few dimensions as possible.

An example of these differing representations comes from Lee and Seung, who employ image compression as an example. Using images of faces, they show that NMF discovers new dimensions that correspond to parts of faces—like noses, mouths, and eyes. In contrast, PCA discovers representations that involve each part of a face to differing degrees. This is shown in Figure 18 from Lee and Seung 1999, where black shows positive values and red shows negative. We see here that the representation that NMF learns (left matrix, left side) focuses on particular areas of an image, whereas the representation that PCA learns (left matrix, right side), is spread across the image.

The reason that NMF results in parts-based representations has to do with the fact that it can only encode the original data in terms of adding pieces together. This contrasts with PCA, which also allows subtracting pieces out. Thus, though both NMF and PCA transform the data in such a way that variables become less correlated—necessary for our analysis, as a practical matter—they do so in very different ways. Returning to our suggestions that matrix factorization methods like PCA and NMF are a way of modeling the discovery of grammatically relevant semantic features $\mathbf{S}$ given syntactic distributions $\mathbf{D}$, PCA and NMF represent very distinct models of how such grammatically relevant semantic features might be discovered, while both serving the practical analytical purpose of lowering the correlations between variables.

A.2.3 Parts-based representations and sparsity

Another way of inducing the sorts of parts-based representations inherent to NMF is to enforce *sparsity* on the representations that result from matrix factorization. Sparsity refers to the overall distribution of values in a representation. In sparse representations, there are many values close to zero and very few that are far from zero. This contrasts with a method, like standard PCA or standard NMF, which does not enforce sparsity, tending to produce *dense* matrices with many values not close to zero. A sparse version of NMF, which Murphy et al. (2012)

[43] Before moving forward, it is worth noting another, less important difference between PCA and NMF. Unlike PCA, NMF is not guaranteed to yield transformed representations with uncorrelated dimensions; thus, it does not in fact carry out true decorrelation. However, the dimensions that it produces do tend to show less correlation than the original matrix, which is sufficient for our purposes.

refer to as Nonnegative Sparse Embedding (NNSE), has grown in popularity as a method for representing word meanings and modeling compositionality (see Fyshe et al., 2014, 2015)

Sparsity encourages parts-based representations in a slightly different way from the one that the nonnegative constraints on NMF do. Where NMF produces parts-based representations by only allowing the addition of components, sparsity yields parts-based representations by ensuring that each underlying dimension is associated with only a few pieces.

In the next section, we apply PCA and NMF—both standard and sparse—to our normalized data and then attempt to predict classifications of attitude verbs based on traditional distributional analysis. We then compare the accuracy of these predictions using both PCA- and NMF-based predictors.