Project Description

IIS-RI MEDIUM: Collaborative Research
Inference Discovery through Paired Annotation: A Case Study with Adjectives

1 Introduction

Effective communication relies on the ability of language users to recover information that is not explicitly expressed in utterances. Much of this covert information can be identified as "semantic inferences" associated with identifiable structural or lexical patterns. An understanding of how speakers identify and exploit structural and lexical inference patterns has the potential to enrich models in theoretical linguistics. At the same time, it can support a major goal of current NLP research: allowing natural language understanding systems to recover more of the rich semantic information encoded in the text. Did events mentioned in the text happen? Are they deemed desirable? Which qualities are being ascribed to entities in the text? To answer subtle questions such as these, careful analysis of the lexical semantics of adjectives is needed.

Identifying such inferences requires training machine learning algorithms to reconstruct the often subtle lexical, constructional, and contextual cues in texts, just as effective human communication relies on the ability of speakers to recover information that is not explicitly expressed in utterances. Existing systems typically rely on training or development corpora that have been annotated automatically or by hand, in most cases derived from lexical classes or from syntax-semantics correspondences. Previous research has also concentrated on coarse-grained semantic tasks such as identifying events and their participants; this has led to a focus on properties of verbs and nouns.

The work proposed here focuses on the subtle information conveyed by the important but neglected class of adjectives. However, there is very little readily exploitable linguistic information available for these lexical categories, and it is not clear how their relations should be modeled. Moreover, the studies reported in the linguistic literature do not address the questions that motivate this proposal, as they ignore important semantic interactions between the various textual elements as well as the usage of linguistically untrained speakers. Finally, existing systems generally treat inference as an all-or-nothing matter. We propose to develop a methodology that addresses these shortcomings and to validate it on three subsets of adjectives. This approach will result in a more complete characterization of how these adjectives are interpreted, through a better integration of lexical resources and corpus annotation.

We develop a methodology for the discovery and exploitation of systematic linguistic inferences identified with specific lexical classes in natural language by constructing an inferential model for adjectival semantics and using **paired annotation**. Specifically, we propose to:

- Establish an initial model for each adjective class, combining existing background from linguistic theory with data mining over large corpora to identify structure-to-inference mappings, i.e., syntactic and distributional correlates of judgments by trained annotators.
- Create larger labeled data sets using linguistically untrained annotators recruited on crowdsourcing platforms, notably Amazon Mechanical Turk (AMT).
- Draw systematic comparisons between the judgments of trained and untrained annotators, iteratively refining experimental techniques with the goal of establishing practical methods to create larger-scale annotated corpora than is achievable using small groups of trained annotators;
- Revise and enrich models in light of results, with the goal of enriching lexical resources, e.g., WordNet.

We concentrate on three diverse semantic types of adjectives, in order to: (a) test the applicability of the methodology to different semantic classes; and (b) to articulate just how the structure-to-inference mapping can be modeled within each lexical class. The adjective types studied are: (i) dimensional and evaluative adjectives with scalar values and associated scalar implicatures, e.g., *pretty, beautiful, large, huge*; (ii) veridicity-related adjectives, showing varying implications of veridicity over a clausal complement, e.g., *rude, annoying, likely*, etc.; and (iii) intensional adjectives, introducing implications of modal subordination, e.g., *alleged, supposed, so-called*. The work will result in a small Gold standard inference corpus created by using a standard linguistic annotation effort following explicit guidelines indicating the structure-to-inference mapping for each type of adjective. Unlike most previous efforts, this standard is not the end product to be used in learning: we will construct paired annotations by comparing the baseline structure-to-inference mappings to inferential judgments made by non-expert native speakers, in particular Mechanical Turk workers (MTurkers). Our preliminary studies lead us to expect that there will be variance from the baseline. We hypothesize that an important part of this variance is caused by textual factors that are abstracted away in linguistic studies, but are important to explain the non-expert judgments. We will use these differential measurements in judgment (trained linguist vs. non-expert annotator) to classify inferences according to how stable they are regardless of linguistic context and which (if any) contextual factors contribute to blocking the inference. On the basis of this study, we will develop an improved gold standard and test it again with non-expert native speakers. We will then build a model to gauge how well our distinctions explain the behavior of these speakers. Our approach will allow us to account for the interactions of different structural and lexical factors instead of seeing them as independent from each other.

This research is significant in two major respects. First, it lays theoretical and methodological groundwork for a large-scale annotation of adjectives in order to support automatic systems in inferencing tasks. Second, it leads to a more sophisticated theory of the contribution of lexical information to inferencing. By studying inferencing "in the wild" and how it differs from the baseline established from gold standard corpora, we can begin to identify the pragmatic factors contributing to the interpretation of lexical items in richer linguistic contexts.

Classic semantic field analysis (cf. [17, 43, 53]) categorizes the attributes denoted by adjectives according to a thematic organization, as lexically encoded in the language.[1] An alternative approach is to adopt a conceptually conservative but more formally descriptive and operational distinction which groups adjectives into inferential classes. [2, 3] make just such a move, adopting a four class distinction based on inferential properties noted by [34, 35]:

(1) In the construction, [A N], A can be classed as:
  a. INTERSECTIVE: the object described is both A and N.
  b. SUBSECTIVE: the object described is A relative to the set of N, but not independent of N.
  c. PRIVATIVE: the object described is not an N, by virtue of A.
  d. NON-SUBSECTIVE: there is epistemic uncertainty whether the object is N.

These constructions constitute patterns that license specific inferences associated with classes of adjectives, and can be exploited in the context of text-based inference systems, such as the RTE ([2]). This classification, however, is both too broadly defined to model the finer inferential distinctions within each class, and too narrow to include the behavior of other adjective classes, in particular, those taking clausal complements.

For these reasons, we have chosen to study the three different classes of adjectives mentioned above, which require refinements and additions to the recognized inference patterns. Examples of the types of inferences we intend to capture are the following:

Scalar Adjectives. The PASCAL Recognizing Textual Entailment task ([1, 12]) requires automatic systems to evaluate the truth or falsity of a statement (the Hypothesis, H), given a prior statement (the Text, T). The system must decide whether or not H is true or false given T, as in:

(2) T: **Arctic** weather swept across New Jersey.
  H: The Garden State experienced **cool** temperatures.

A system which hypothesizes a symmetric synonymy relation between *cool* and *Arctic* would incorrectly infer an entailment relation also if T and H were switched: an awareness of the asymmetry of entailment encoded in our model is crucial to making the correct judgment here. In addition, scalar adjectives license inferences based on complex contextual and probabilistic considerations, as in (3).

(3) T: The Empire State Building is **huge**.
  H: New York City's most famous building is **tall**.

[1] It should be noted that [53] , however, also discuss inferential patterns for distinct classes.

Even an RTE system that could manage the difficult coreference task here would generally fail to infer that this is a valid entailment in context, because *huge* does not entail *tall* in a context-independent way. However, these adjectives overlap in the scalar dimensions that they refer to (size, including height as a special case), and a system which is able to recognize this fact as well as the fact that height is the most relevant form of size for a typical building could capture this very common type of inference.

Adjectives with clausal complements. In order to recognize that the Text does not entail the Hypothesis in the following example, it is not enough to recognize events and their participants; one has additionally to understand the stance the Text takes with respect to the described event:

(4) T: It is **unlikely** that the attack on the consulate in Benghazi was the work of Al Qaeda.
H: The attack on the consulate in Benghazi was the work of Al Qaeda.

Intensional adjectives. The effect of modifying the nominal head is frequently the introduction of "epistemic uncertainty" regarding the description.

(5) T: The police arrested the **alleged** criminal.
H: A criminal was arrested.

The inference from T from H would not be justified here, since it is not known whether the allegation is true. However, consider the pair in (6):

(6) T: Archeologists discovered an **alleged** paleolithic stone tool.
H: A stone tool was discovered.

This inference is legitimate because the epistemic scope of the adjective *alleged* is the adjective *paleolithic*, and not the nominal head itself (tool).

Supervised or unsupervised annotation is typically used to improve automatic inference based on the lexical items in texts. These annotations reflect the inferential potential associated with lexical items. Some aspects of this potential have been studied in the linguistic literature; computational approaches tend to take these results for granted. In the case of nouns, the WordNet hierarchies have proved useful in numerous studies (e.g., [59]); for verbs, lists of special inference patterns have been constructed starting from the work of [41, 36] by [48, 55, 56, 42]. Information about the inferential properties of adjectives is, however, much less easy to come by. For the categories of adjectives that we are interested in, the existing resources have severe shortcomings or are non-existent, in part because the contribution of adjectives tends to be more subtle and more dependent on the rest of the linguistic context. This difficulty requires adopting a more careful methodology for syntactic and semantic lexical categorization tasks. The availability of digital corpora (the biggest one being the Web itself) and crowd-sourcing techniques to elicit the judgments of a larger and more diverse group of native speakers allow us to go beyond the narrow base that traditionally lexical studies were based on. Moreover, the development of statistical modeling techniques allow us to test theoretical hypotheses with large datasets. This should allow us to obtain better data to feed into automatic inferencing systems (such as BiuTee [62] or [10]).

2 Theoretical Background

2.1 Methodological Preliminaries

Much work in modern computational linguistics relies on the creation of annotated datasets focused on one or more related linguistic phenomena. Such gold standard corpora are essential for training and tuning the statistical models on which natural language processing tasks largely rely.

In the development of a gold standard corpus using rich linguistic annotation, it is typical to establish an initial model for the phenomena being studied. This includes a triple, $M = \langle T, R, I \rangle$, consisting of a vocabulary of terms, T, the relations between these terms, R, and their interpretation, I. This is often a partial characterization of quite extensive theoretical research in an area, encoded as specification elements for subsequent annotation. These annotations provide the features that are then used for training and testing classification or labeling algorithms over the dataset. Depending on a system's performance, various aspects of the model or related specification will be revised, retrained, and then retested. For this reason, we can refer to this methodology as the MATTER cycle: *Model-Annotate-Train-Test-Evaluate-Revise* [52], as illustrated in Figure (1).

The "Model Testing" phase of this cycle involves iterating over model development followed by subsequent testing by annotation. This (Model-Annotate)* technique assumes a classic iterative software development cycle, as applied to the creation of a rich specification language to be used for linguistic annotation. That is, as issues are encountered with the model when instantiated in a specification and applied to data through the annotation process, the model is revised to accommodate these observations.

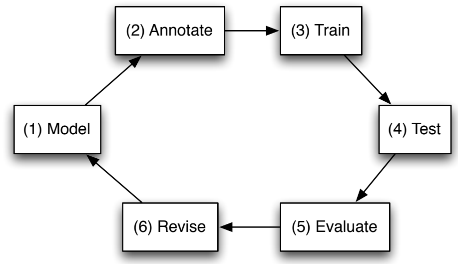

Figure 1: The MATTER Methodology

In the present work, we propose a significant enrichment to this methodology, in order to better model contextual and pragmatic factors that are often ignored or down-played in this strategy. These involve linguistic phenomena (such as the adjective classes studied here) for which contextual factors and pragmatic effects are critical in how the annotations are interpreted.

The corpora that can be created using linguistically trained annotators are rather limited and rarely exhibit all the combinations of relevant context factors, resulting in lack of data capturing what a rich, human-like understanding of texts should be. This limits the usefulness of many machine learning methods that are widely used in natural language understanding [44, 68, 47]: sophisticated statistical models cannot produce rich understanding without a linguistically informed understanding of what is being modeled. Our project seeks to use standard annotation, new experimental methods based on crowdsourcing, and corpus studies together in order to address this lacuna for the important, understudied category of adjectives.

Crowdsourced annotations tasks, when carefully constructed and analyzed, have been shown in previous work to have reliability comparable to traditional expert annotations in some domains including certain RTE tasks [61, 46]. Adopting this method as part of our approach will help us achieve three related goals. First, it makes possible systematic testing of contextual factors that have been claimed in the linguistic literature to be relevant to the inferences licensed by the use of an adjective, as well as patterns isolated on the basis of existing annotations. Second, statistical analysis of the results of larger-scale annotations gathered by experimental means make it possible to reliably identify inferences which are probabilistic, rather than deterministic, in nature; for example, the defeasible inference that someone described as *attractive* is probably not *stunning*. Third, systematic use of untrained annotators will make available data on the interpretations of people with linguistic backgrounds that go beyond those associated with the narrow socioeconomic groups typically involved in standard annotation efforts.

Isolating the factors that contribute to the perception of an inference is notoriously difficult. Testing contrastive contexts with a large number of native speakers is one way to discover the precise factors at work. For instance, from FactBank we learn that 'NP be lucky to VP' has the meaning 'it is highly unlikely that NP VP'. In previous work project consultants Karttunen and Zaenen [37, 72] observe that this definition is far too general, and that the "unlikely" meaning is mainly found in future-tense sentences. However, they show that the facts are more subtle even when tense is taken into account: in (7), the (a) example has the highly unlikely meaning, whereas the (b) example does not (note that replacing 'at least' with 'in any case' makes the highly unlikely reading again more prominent).

(7) a. Your son will be lucky to escape a jail term.
b. At least your son will be lucky to escape a jail term.

Here and in many other cases, linguistic and non-linguistic context influences the inferences associated with the use of an adjective, though in general non-deterministically. The discovery and treatment of such data involving evaluative adjectives in [37, 72, 39] relied crucially on a mixture of experimental investigation, standard annotation, corpus work, and linguistic analysis of the type proposed here, and serves as an example of the promise of the methods.

Four key elements play a role in our methodology:

(8) a. The interpretation, I, focuses on *structure-to-inference mappings* (SIMs), indicating how a given adjective type contributes to or enables inferences associated with its embedding syntactic contexts. This is achieved by the conventional expert annotation.

b. The conventional "expert" annotation cycle is followed by a separate annotation of data by Amazon Mechanical Turk workers, using instructions developed on the basis of features isolated by expert annotators but adapted for use by linguistically untrained native speakers, borrowing methods from experimental psychology where appropriate and using larger linguistic contexts as warranted.

c. Iterative refinement and enrichment of SIM models to include contextual and probabilistic factors is accomplished on the basis of data from both trained and untrained annotators. This can be referred to as *context-based adjudication*, revealing unseen contextual features that can then be added as model-based primitives.

d. In several key places, machine learning techniques applied to large corpora will be used to extend conclusions beyond what it is feasible for even crowdsourced human annotators to accomplish. These studies will make crucial use of probabilistic information from large-scale annotation.

We proceed by establishing for each adjective class being studied (Scalar, Veridicity-related, Intensional), an initial model, incorporating the appropriate SIM as the interpretation function. Expert and non-expert annotations create corresponding gold standards, from which we perform context-based adjudication.

2.2 Scalar adjectives

As discussed in the introduction, language understanding requires active reconstruction of information which is merely implicit in an utterance, or which can be reconstructed on the basis of an utterance together with its linguistic, social, and worldly context [25, 29, 7, 26, 51]. This is true in particular for scalar adjectives, which have highly context-sensitive meanings (e.g., compare *big baby*, *big tree*, and *big planet*). Many scalar adjectives are organized into entailment scales, where items high on the scale asymmetrically entail items lower on the scale [30, 31]. The use of items lower on the scale, in turn, can lead to a defeasible inference that the sentence would be false if the lower item were replaced by the higher [25, 30].

(9) a. warm < hot < scorching

b. *Dallas is scorching* **entails** *Dallas is hot*, **which entails** *Dallas is warm*

c. *Dallas is warm* **may implicate** *Dallas is not hot* and *Dallas is not scorching*

d. *Dallas is hot* **may implicate** *Dallas is not scorching*

The entailments and pragmatic inferences illustrated here are both important parts of the total understood context of sentences involving scalar adjectives. However, whether the inference arises and how strong it is are highly context-sensitive matters, depending on defeasible assumptions about the speaker's information and the alternative possible utterances that are relevant in context [28, 23, 24]. Already in the example above, we see implicatures of varying strength: the speaker's choice to use *warm* will often lead to an inference that the sentence would have been false if *hot* were used, and will likely lead to an even stronger inference that the sentence would have been false if *scorching* were used.

As a further example of importance of rich linguistic context, compare these three dialogues:

(10) a. Is Dallas scorching?
b. It is hot.

(11) a. Is Dallas hot?
b. It is hot.

(12) a. Is Dallas beautiful?
b. It is hot.

(10) illustrates a standard scalar implicature, where the choice to use "hot" when "scorching" is relevant leads to an inference that Dallas is not scorching. In (11), this inference is much less robust, presumably because it is not clear whether "scorching" is a relevant alternative. In (12) no implicature regarding "scorching" arises, but (perhaps surprisingly) there is a robust inference that Dallas is not beautiful. These examples illustrate the importance of taking context into account when drawing inferences from scalar adjectives.

A second complication in modeling inferences from scalar arises due to DIMENSIONALITY. While adjectives such as *tall* and *heavy* pick out a single scalar dimension (height and weight, respectively), multidimensional adjectives such as *big* and *huge* are more complex, placing requirements on multiple scalar dimensions (e.g., 3-D size and perhaps weight, in the case of *big*). Emotive adjectives such as *happy* are even more complex, since the dimensions that they rely on are not easy to identify or quantify. These relationships among adjectives thus raise difficult questions for an inferential model of adjective semantics. Since it is possible for something to be *big* without being *tall*, these expressions do not share all of their dimensions and thus are not in an asymmetric entailment relationship as *warm* and *hot* are. But then how do we explain contextual entailments such as the one illustrated in (3) above? In addition, there are many cases in which it is simply not clear whether two adjectives are in a scalar relationship: does *happy* asymmetrically entail *content*? Subtle human judgments, gathered under tightly controlled conditions, will be crucial in building and refining a model which will allow NLU systems to approach human performance for such adjectives.

In constructing a formal model of the inferences associated with scalar adjectives, it will be necessary to

1. determine which adjective pairs are "co-scalar", sharing polarity and all scalar dimensions, and differing only in strength;[2]

2. quantify the difference in strength between co-scalar adjective pairs, in order to predict the strength of implicature (cf. *warm/hot/scorching*);

3. determine which adjective pairs overlap partially, differing e.g. only in polarity (*cool* vs. *warm*) or display partial overlap of dimensions (*big* vs. *tall*);

4. Identify factors which influence the weighting of different dimensions in the contextual meaning of adjectives, e.g., the factors which allow people to infer that height is the relevant size dimension in interpreting *big* when buildings are under discussion, but not when cities are.

Adjectival polarity for sentiment analysis [70] correlates only weakly with polarity in our formal sense (see fn. 2). More directly related are methods for identifying total entailment relations between adjective pairs developed in [58, 57]. These methods can be improved by using semi- and un-supervised methods for learning lexical relations from parsed corpora rather than hand-specified patterns on raw text (cf. [59, 13, 63]). Moreover, there has been virtually no investigation of the other issues that we will consider: the semantic and pragmatic effects of partial scalar overlap, or of the effects of differing scalar distance on the strength of pragmatic inferences. Notably, the use of crowdsourced annotation will be a crucial in enabling us to pursue these fine-grained aspects of lexical structure. This is because quantitative information is needed to discover scalar distance, partial overlap relations, and strength of pragmatic inference in context, and gathering this information requires statistical analyses on the responses of many annotators.

2.3 Veridicity: Inferences of adjectives with clausal complements

Another class of adjectives that are a rich source of inferences are predicative adjectives with clausal arguments (*that* S, *to* VP, or *ing* complements). The class comprises hundreds of adjectives whose use communicates an agent's epistemic stance on the likelihood of the (non-)occurrence of the eventualities described by their clausal argument. Some of the adjectives convey, in addition, an emotive or evaluative attitude of the agent. The inferential classification of this class poses interesting challenges.

The relevant agent as well as the type of inferences that arise depend on both the adjective and its syntactic frame. The agent with the implied epistemic stance is sometimes the speaker/writer and sometimes

[2] This is the logical notion of "polarity" used in formal semantics [40], not the emotive concept from sentiment analysis [71].

the referent of the subject of the predicative construction (the 'protagonist'). For instance, *John is sure that Bill left* ascribes the belief that Bill left to John (the protagonist) and leaves open what the writer thinks, whereas *John is sure to have left* ascribes a belief to the writer. This minimal pair shows that the adjective is not sufficient on its own to determine the type of inference. Similarly, the syntactic frame alone does not suffice to determine the type of inference. For instance, not all adjectives that fit into the [It is ADJ that S] pattern behave the same way.

There is no generally accepted classification of predicative adjectives taking clausal complements on either syntactic or semantic grounds, but three broad classes have been distinguished based on their epistemic inference patterns.

1. Factive adjectives These are adjectives implying that the author is committed to the factuality of the state of affairs described in the complement even when the matrix clause is negated or questioned. They are traditionally analyzed as presupposing the truth of their complement. Take, for example, (13).

(13) It is annoying that people post stuff that no one cares about on the web.

From (13), the reader infers that the author presents as true the proposition that people post stuff that nobody cares about on blogs. This inference is derived directly from the semantics of the adjective *annoying*, when used in such a construction. Neither negation nor questioning changes the veridicity of the *that* clause, as illustrated in (14). The focus of the question in (14b) is the evaluation of the *that*-complement as annoying or not.

(14) a. It isn't annoying that people post stuff that no one cares about on blogs.
b. Is it annoying that people post stuff that no one cares about on blogs?

[50] proposed two large subclasses: emotive (e.g. *sad*) and evaluative (e.g. *stupid*) adjectives. However, the empirical picture is more complicated: the status of the factuality inference varies among speakers (cf. [39] and below).

2. Certainty adjectives These adjectives directly assert the degree of certainty that the writer, or a protagonist, ascribes to the complement, as illustrated in (15).

(15) a. It is certain that people post stuff that no one cares about on blogs.
b. It is not certain that people post stuff that no one cares about on blogs.
c. Is it certain that people post stuff that no one cares about on blogs?

(15b) has the opposite inference from that of (15a), and in (15c) it is the *that*-complement itself that is questioned. Structure-inference patterns for these adjectives then would minimally need to distinguish between positive contexts, negative contexts and questions.

Some of the adjectives in this class express absolute certainty or absolute denial of the truth of the embedded clause, and hence give rise to logical entailments; they are *implicative* ([36]). Others, such as *possible, probable, impossible, improbable*, constitute a means for the author to indicate the probability that (s)he attaches to the factuality of the state of affairs expressed in the embedded clause. In this study, we follow [55] and approximate this probability by the following scale: CT+ (certain), PR+ (probable), PS+(possible), U (none), PR- (improbable), sc pr- (impossible) and CT- (certainly not).

Apart from the adjectives that express an epistemic stance directly, there are adjectives expressing other modalities that have epistemic consequences. These include *able (to), unable (to), willing (to), not willing (to)*, also *unthinkable (that), unbelievable (that)*. Their negative versions may carry negative entailments (*unable to VP* implies that the situation described by the VP complement did not come about), while their positive versions lead to the expectation that the situation described by the complement has occurred or will occur but without warranting an entailment relation.

3. Adjectives with no epistemic implications These adjectives fall into several subclasses, e.g. *easy* adjectives, dispositional adjectives such as *afraid (to), keen (to)*, mandative adjectives such as *important (to), essential (to)*, etc. While these do not lead to logical entailments, some of them *invite* the inference that the writer thinks that their complement is factual or at least very likely to have happened. The factors that trigger these invited inferences need further study.

In addition to the extensive study of factive adjectives in [50], there are the more limited studies in [69] and [5]. Implicative ([36]) and degree-of-certainty adjectives are only mentioned in passing in the literature. [45] looks at the syntax of 51 frequent adjectives taking *that* clauses in the BNC but without any attention to the semantics. [64] report on a corpus study of deontic-evaluative adjectives concentrating on *important, essential, crucial*, *appropriate, proper*, and *fitting*.

Pilot studies we have conducted on various subclasses have revealed that their inferential behavior is dependent on fine-grained structural and contextual factors. We discuss some of our findings to illustrate that getting a proper inferential classification needs further, systematic study.

a. Impersonal constructions of the type [it be ADJ (of NP) to VP] Adjectives in this syntactic pattern can belong to any of three inferential classes described above. [50] lists several hundred as factive. But even among those the situation is more complicated. A sentence like *It was audacious of John to make a trip around the world* readily gets a factive interpretation but one like *It is audacious of anyone to make a trip around the world* very rarely. Our preliminary investigation of the evaluative adjectives among these shows that a factive interpretation reliably arises only in the past tense with a specific *of NP*. For this case we can have the structure-to-inference mapping in (16).

(16) $[\text{ it was } ADJ_{eval} \text{ of } NP_{spec} \text{ to } VP\ ] \models\ NP_{spec}\ past\ \ VP$

Adjectives without epistemic entailments may nevertheless give rise to interpretations where a high degree of probability is ascribed to the truth of the complement. For instance, *It was essential for researchers to collect accurate information* is judged by MTurk workers to be factual for more than 50% of them and probable for another 35%. Preliminary results thus suggest that for this syntactic pattern there are several subclasses of the three broad, traditionally recognized classes, for which the exact conditioning factors have yet to be identified.

b. Personal constructions of the type [NP be ADJ to VP] We have discovered that factive adjectives in this frame can be implicative under certain circumstances ([39]). The preferred interpretation of *Kim wasn't stupid to send money* is that no money was sent, while that of *Kim wasn't stupid to save money* is the expected factive interpretation. We looked at 60 occurrences of *is/was stupid to* in the enTenTen English corpus, one of the only curated corpora that includes blogs, and found that 25 were clearly factive, 23 clearly implicative and 12 either unclear or part of a different construction. Previous theoretical assumptions would have predicted only the factive interpretations in both cases. We hypothesized that the crucial determinant is whether there is a harmonic or a disharmonic relationship between the evaluative attitude expressed by the adjective and a general evaluative assessment of the activity described by the complement clause. This was corroborated by a pilot experimental study. The pattern is clearly dependent on extra-linguistic factors, since there is no situation-independent metric of stupid or clever actions.

c. Personal and impersonal constructions with a *that*-complements Our preliminary investigation suggests that *that*-complements of factive adjectives give rise to rather solid factive interpretations but a more detailed study needs to be done. A preliminary classification of these adjectives is available on-line ([73]). A structure-to-inference mapping corresponding to an impersonal syntactic frame is given in (17).

(17) $[\text{ it be } ADJ \text{ that } S\ ] \models\ S$

2.4 Identifying Epistemic Uncertainty: Intensional Adjectives

The third adjective class we examine for their inferential properties is the set of non-subsective intensional adjectives. The intensional adjectives can be split into privatives and non-subsective. Privatives, such as *fake* or *pretend*, can be analyzed as follows:

(18) $\|A\ N\| \cap \|N\|\ =\ \emptyset$

Intensional non-subsective adjectives introduce an epistemic uncertainty for the elements within their scope. Examples of this class include *alleged*, *supposed*, and *presumed*, and they call into question some predicative property of the nouns they modify. Following [35], no informative inference is associated with this construction:

(19) a. $[A\ N]$ (alleged criminal)
b. $\not\models\ N$

However, contrary to what is claimed in [2], non-subsective adjectives do appear to license specific inferences when examined in a broader context than the [A N] construction usually studied. From preliminary corpus studies of this class[3], several distinct patterns of inference emerge. While the typical resulting composition entails uncertainty of whether the nominal head belongs to the mentioned sortal, (20a) below, there are many contexts where the epistemic scope is reduced to a modification or additional attribution of the nominal head, as shown in (20b).

(20) a. The **alleged criminal** fled the country.
b. Archeologists discovered an **alleged paleolithic tool**.

In Example (20a), the adjective *alleged* calls into question the predicative property of 'criminality' of the *criminal*. When a predicative property is called into question by adjectives of this class, are there any systematic inferences to be made about the semantic field? E.g., is the semantic field still guaranteed to be some hypernym of *criminal*? Even if the individual does not belong to the set of "criminals", it does still seem to belong to the set of "persons". In example (20b), contrastively, at least under one interpretation, it is whether the *tool* is *paleolithic* or not that is called into question: i.e., the object belongs to the set of "tools" regardless if it is truly *paleolithic* or not. This inference is schematically represented below.

(21) Given the construction $[A_{int}\ N]$, where A_{int} is *alleged, ...*, then:
a. $[A_{int}\ N] \not\models\ N$
b. $[A_{int}\ A_2\ N] \not\models\ A_2$
c. $[A_{int}\ A_2\ N] \models\ N$

Such an inference pattern is subject to contextual variables, many of which are not available to sentential compositional mechanisms, but some constraints can be identified. For example, the closer the head noun is to a sortal base level category, such as *bird*, *table*, or *tool*, the more likely the inference in (21) will go through. Consider the examples below:

(22) a. The store bought an alleged antique vase.
b. The researcher found an alleged Mozart sonata.

These cases make it clear that the epistemic uncertainty in (22) involves an additional aspect of the NP, beyond the unassailable characteristics of the entailed head. That is, the object is clearly a vase (in (22a)) and demonstrably a sonata (in (21b)). Such evidence, however, will not always be available within the composition of a sentence, but will be derivable from context (if at all). We will refer to the canonical inference in (21a) as the "Wide-scope reading", and the inferences in (21b-c) as the "Narrow-scope reading".

Another interesting distinction emerging in the basic [A N] construction with intensional adjectives is one based on the type of the nominal head. The most common semantic types occuring in the corpus are shown below, along with apparent scoping behavior.

(23) a. EVENT NOMINAL: *violation*, *misconduct*, *murder*, *assault*. The more specific nominal descriptions carry greater inferential force for the hypernym. That is, *murder* suggests inference of a death.
b. AGENTIVE NOUN: *collaborator*, *perpetrator*, *murderer*, *criminal*. Epistemic scope is over the entire sortal. The canonical form, "the alleged criminal".
c. UNDERGOER NOUN: *victim*. While not always the case, the scope is narrowed to a modification of the event: For example, "the alleged victims of Whitey Bulger".

Consider the sentences in (24), where *alleged* is modifying an event nominal.

[3]The initial corpus has been collected from directed CQL queries over two Sketch Engine corpora, Ententen12 and BNC. Three sentence "snippets" have been compiled from this source.

(24) a. He denies the alleged assault on the police.
b. The greatest number of alleged violations occurred in California.
c. He's been charged in connection with the alleged murder of John Smith, whose mutilated body ...

The inferences associated with (24a-b) follow from the template in (21a). For sentence (24c), however, we need to infer that there was, in fact, a killing, although it is uncertain whether it was a murder. This requires the inference rule below, where the hypernym of the event nominal is infererable from the context.

(25) Given the construction $[A_{int}\ N]$, where N is an event nominal, with certain feature, then:
a. $[A_{int}\ N] \nvDash N$
$\vDash N'$ where $N \subseteq N'$

We refer to this inference rule as the "Hypernym reading". Similarly, the scope of an intensional adjective modifying an undergoer can be lowered to a modification of the event description, as in (26b).

(26) a. Testimony will be heard from the alleged victim in court.
b. The families of two alleged victims of James "Whitey" Bulger have received compensation.

Sentence (26a) behaves according to the canonical template, while (26b) involves a narrower scope of the epistemic uncertainty. That is, the inference should be made that there are victims, but the cause (or etiology) of this designation is uncertain. This rule is formally related to that presented above in (21), where the modification (argument specification, in fact) is postnominal.

(27) Given the construction $[A_{int}\ N\ XP_{mod}]$, where XP_{mod} is a modification or argument, then:
a. $[A_{int}\ N\ XP_{mod}] \nvDash N\ XP_{mod}$
c. $[A_{int}\ N\ XP_{mod}] \vDash N$

Summarizing the semantic behavior for this class, we have identified at least three distinct structure-to-inference mappings associated with intensional (non-subsective) adjectives. These are:

(28) Structure-to-Inference Mappings:
a. Wide-scope reading: $[A_{int}\ N] \nvDash N$
b. Narrow-scope reading 1: $[A_{int}\ A_2\ N] \nvDash A_2, \vDash N$
c. Narrow-scope reading 2: $[A_{int}\ N\ XP_{mod}] \vDash N$
d. Hypernym reading: $[A_{int}\ N] \vDash N'$ where $N \subseteq N'$

3 Project Plan

For each of the three adjective classes, we develop structure-to-inference mappings, which are templates associating textual constructions with allowable inferences from the linguistic content. We adapt and enrich the existing inferential models for all three types of adjectives. We then (a) select an initial set of target adjectives, (b) extract text snippets from corpora containing the target adjectives, and (c) construct on this basis small corpora in the format of RTE to be annotated by both linguistically trained and non-expert annotators.

Corpus data. We will use a variety of corpus resources, including in some cases the Web, for the extraction of patterns identified as inferentially relevant in the initial model and in subsequent corpus investigations. The advantages and disadvantages of using web data vs. smaller, more carefully controlled corpora are well-known. In many cases — especially when dealing with short patterns whose diagnostic usefulness has already been established — the size and diversity of styles and genres on the web, as and its access to diverse speaker communities, gives us vital information which compensates for the increase in noise and the possibility of multiple counts. In other cases, it will be necessary to use resources such as BNC, COCA, and Gigaword, or to develop methods to use these resources to complement each other. This is especially true in portions of the project which attempt to use semi- and un-supervised methods to identify patterns of interest, where POS-tagged and parsed data is needed, which can be more difficult to obtain starting with web data (see section 3.1 below). Note further that, where web data is appopriate, our iterative methodology will eliminate many potential false hits from web data and other sources alike, since patterns will be identified as non-diagnostic by human annotators will be removed or downweighted subsequently.

Human data. Several studies in the last years have shown that crowd-sourced annotation tasks can deliver reliable results when carefully constructed and analyzed (cf. [61, 46] and section 2.1 above). The PIs have considerable experience with traditional annotation tasks, and will proceed as usual in this respect. In constructing analogous tasks on AMT, we will present test questions to ascertain that the workers are native speakers of English, and then ask them to make judgments about inferences (both potential entailments and implicatures) with varying amounts of linguistic context. Best practices for AMT annotation are not yet firmly established, and we expect that achieving our goals will require us to explore a variety of approaches. To do so, we will draw on previous NLP research cited above as well as methods developed in recent experimental cognitive science (one of the Stanford co-PI's areas of research, and a field in which the use of AMT for data collection is firmly established). Our approach to inferring quantitative patterns of inference as well as inferences associated with specific contexts will rely primarily on this method of data collection, while standard annotations will be used as a tool in building AMT tasks and as a sanity check for the results.

3.1 Scalar Adjectives

[57, 58] used a methodology similar to Hearst's [27] to demonstrate the usefulness of hand-selected syntactic patterns in identifying co-scalarity and relative intensity of adjective pairs. [57] showed that when adjectives X and Y are "semantically similar" according to WordNet — and so likely to be co-scalar — it is possible to learn which is stronger by examining the frequency of patterns "X, even Y" and "if not Y, at least X". If these and other carefully chosen patterns are frequent, Y is likely to entail X asymmetrically.

In a recent pilot experiment, we used the Google Web1T corpus with a slightly expanded set of patterns to classify the 300 adjectives for which we have pairwise intensity ratings from AMT. The results indicate both the promise and the need for expansion of this method. If we simply threshold at 40 (the lowest count contained in the corpus), precision is high but recall is low. Many pairs judged by AMT workers to be co-scalar and differing in intensity did not appear, but most pairs returned were intuitively co-scalar and had the expected intensity relation: "**indecent** but not **obscene**"; "**sad**, almost **tragic**"; "**unfriendly**, even **hostile**"; "**satisfactory**, and sometimes even **good**" — but this method also returned "**good** but not **easy**". (Of course, it is not clear that the last result is a false hit rather than an indication that *good* and *easy* are in a context-dependent probabilistic inference relationship: cf. discussion in section 2.2 above.)

We propose to extend [57] by learning the relevant patterns rather than specifying them by hand. This approach follows [59, 60], who generalize Hearst's [27] method of hypernym discovery using WordNet together with a novel learning algorithm applied to a large corpus of dependency-parsed sentences [15]. This approach allowed them to make use of information contained in many patterns that Hearst had not considered, as well as eliminating noise inherent in the use of raw counts; this resulted in a considerable improvement on WordNet's baseline in both precision and recall.

We will combine and generalize these methods in several ways. First, we will collect a small gold-standard corpus of judgments of co-scalarity, intuitive strength, entailment, and implicatures among adjective pairs using linguistically trained annotators. Second, we will use adapt these methods to design AMT tasks which allow us to collect a larger corpus of judgments for the 500 most common adjectives in English, using gold-standard judgments also as a sanity check. Third, we will extract all n-grams from the Web1T and (1900-) Google Books corpora which contain any two relevant adjectives. Fourth, we will create a large corpus of dependency-parsed text and perform a similar analysis, looking at patterns in dependency relations rather than raw text. Finally, we will use statistical methods to identify, on the basis of these two data sets, which patterns are predictive of the human judgments and to what extent. We will evaluate the resulting model on AMT annotations and corpus data for held-out adjective pairs.

In addition to the new and linguistically important subject matter, our work contains an important methodological innovation: due to the of the probabilistic nature of many inferences involving scalar adjectives, we will analyze not only **binary** judgments about entailment and implicature but also the **probability** that an inference is appropriate, as estimated from quantitative patterns in AMT workers' judgments. This approach offers the hope of capturing the graded nature of many inferences involving scalar adjectives.

With these results in hand, we will revisit the original, context-independent judgments used to build the model, exploring in what ways the inclusion of richer context modulates entailment and implicature judgments. We will also explore methods for predicting context-dependent judgments from tagged and parsed corpora, considering at all available linguistic features of the context. This aspect of the project is likely to be challenging, but we expect that the results of the first section will aid us in identifying relevant features and appropriate learning methods.

3.2 Veridicity of Adjectival Complement Clauses

We will select the 100 most frequent adjectives with clausal complements in the en-Ten-Ten corpus. We will extract 1,000 corpus snippets from the Web based on these adjectives with their frame. (A corpus snippet is a text that contains the sentence in which the target adjective with its pattern occurs and the sentence before and the sentence after it. The en-Ten-Ten corpus, unfortunately, doesnt give enough context). We will balance this corpus so that each adjective occurs at least 10 times. Following the MATTER methodology, linguist annotators will annotate these snippets with their epistemic inference pattern and the factors that are judged to be relevant for this interpretation.

On the basis of these snippets we will construct examples that exhibit variation in the proposed features. These stimuli will be submitted to MTurkers. The possible interpretations will be presented according to two different conditions: in one case the subjects will have to chosen between a positive, a negative and a "dont know" answer; in the other, they will choose on the 7-point scale, developed in [55, 54] and validated in [14]. For inference patterns that are not recognized by linguists we will add follow-up tests, to find out whether the MTurker considers the expression as part of his/her language or not.

On the basis of these judgments we will estimate whether the different features based on the linguistic annotations correspond to those used in the real world. In the cases there is no fit, the linguist annotators will propose new features that will be tested as described above. The annotators will annotate 1,000 new, naturally occurring snippets. These 1,000 snippets will again be annotated by the MTurkers using both the 7-point and the 3-point scale. We will use these data to estimate how well the factors we have isolated capture the MTurkers data by building statistical models.

3.3 Intensional Adjectives

There are approximately 50 intensional (sub-selective) adjectives that we have identified, from which we will select the most frequent 30 for our investigation. Fewer than 10 of these are root adjectives (*superficial*, *putative*), and most are participial adjectival derivations, such as *alleged*, *supposed*, and *believed*. For each adjective, we have extracted 100 snippets from the corpus, where snippets are three-sentence fragments from the text. This gives us a corpus of 3,000 snippets for intensional adjectives.

We will develop an initial classification of 1,000 of these adjectives based on the inferential patterns discussed in the previous section; i.e., wide-scope, narrow-scope, and hypernym readings. These are the initial structure-to-inference templates which will constitute the small gold standard. This annotation is performed by undergraduate linguistics majors, with three annotations per snippet. That is, we construct the examples that fit the identified test patterns, as shown in (29) and (30) below. In these examples, the inference in (29) is legitimate, while that in (30) is false.

(29) Hypernym Reading:
(T): A teenage girl has been arrested over the **alleged murder** of a mourner at a funeral in London.
(H): A mourner died.

(30) Wide-Scope Reading:
(T): She was soon tried and executed in June by South Korea as an **alleged spy**.
(H): She was a spy.

We submit these stimuli to AMT workers with the same guidelines as those given to the linguists. We then submit the remaining 2,000 snippets to both linguists and MTurkers, and examine the differences in judgments. That is, for those cases that do not accord with the pre-assigned classification, we try to isolate the factors contributing to when the judgment goes against the expected inference. To this end, we perform a statistical analysis of the contexts of the adjective for both the cases that are in accordance with the classification and the cases that are not.

3.4 Evaluation

For scalar adjectives, we propose to evaluate the scales constructed in an RTE task using methods for measuring the contribution of specific WordNet relations developed by [8, 9, 10]. We will similarly quantify the contribution of scalar orderings among adjectives in WordNet to the RTE task using a new test set involving adjectives that we have analyzed. To perform the evaluation, will encode new scales in WordNet following the model described in [57], where WordNet's "dumbbells" are augmented with arcs connecting some adjectives on each half of the dumbbells to specific points on the scale. This preserves the original WordNet representation for one central adjective (e.g., *rich*) and a set of "semantically similar" adjectives (*wealthy, comfortable*, etc.) while also indicating their intensity relative to the central adjective and one another. This representation is amenable to external evaluation with systems like [10].

Concerning the evaluation of predicative adjectives with clausal complements, we base a first, controlled evaluation on the Brandeis TARSQI system. We submit our final 1,000 snippets to the system and consider a baseline markup where all the events are factual. We then develop a set of rules based on our findings and evaluate how well the resulting TARSQI system identifies the events. Larger-scale evaluation proceeds by incorporating our rules and lexical patterns to BiuTee and evaluating the resulting improvement.

The evaluation for intensional adjectives is similar in approach to that above. We will use the 2,000 snippet corpus to train both a Naive Bayes and a MaxEnt classifier, where we take all mentions of the adjective to be invoking the wide-scope reading rule. We take this as our baseline and compare the same two classifiers trained on the differentiated structure-to-inference mappings that were discovered, first by the linguists, and then, as they were enriched by the inferences in the wild.

Finally, the structure-to-inference mappings for all three adjective classes are evaluated by applying the mappings to a held out evaluation set of snippets. We compare the mappings as generated after the corpus mining phase to the revised mappings that were created after analysis of the crowdsourcing results. Additional annotated snippets may be generated for this evaluation if needed.

3.5 Coordination Plan

The PIs at Brandeis, Princeton, and Stanford will maintain regular contact via biweekly Skype conferences. One annual meeting is planned, alternating between Princeton, Brandeis, and Stanford, as well as regular meetings at both national and international conference or workshops focusing on topics of shared interest.

3.6 Milestones and Deliverables

Year One of the project is dedicated to:

Q1	Complete collection of target adjectives; Perform corpus mining; Collect relevant syntactic patterns for clause-selecting, intensional, and scalar adjectives.
Q2	Derive initial semantic classifications and structure-to-inference mappings; MTurk hit design; coordination of annotation specs; preliminary annotation schema.
Q3	Pilot MTurking experiments; Evaluate corpus data; Linguists annotate first sets of snippets.
Q4	Update classifications and mappings; Begin MTurking work; First sets of HIT stimuli for MTurkers; Prepare articles for publication.

Year Two is dedicated to:

Q1	Complete gold standard for expert annotators; Run experiments with MTurkers.
Q2	Analyze/Evaluate results of MTurker data with/against gold standard.
Q3	Continue MTurking work; Update classifications and mappings.
Q4	Identify detailed contextual parameters accounting for judgment divergence; revise structure-to-inference mappings accordingly; Prepare articles for publication; Organize workshop.

Year Three is focused on:

Q1	Revise the annotation specs based on analysis in Y2Q4; develop semantic interpretation of effect of contextual parameters.
Q2	Develop enhanced, layered gold standard.
Q3	Design a way to represent different adjective classes in WordNet (for scalars, model developed in [57] can be developed).
Q4	Evaluation; Data collection protocols; Prepare articles for publication. Final report.

4 Outreach and Education Plan

In the early stages of the project we will disseminate information about the paired annotation methodology and the gold standard being developed, by means of presentations at conferences, workshops, and other meeting venues. We will exploit the relations we have built up through work in ISO groups for language resources to reach those in our field and in related fields such as ontology, linked data, and terminology.

Adjective Inference Challenge. To actively engage the community in the adoption and use of the paired annotation methodology and the resources developed therewith, we will organize an NLP shared task in the third year of the project, focused on three specific tasks involving a relatively straightforward challenge, identifying inferences in textual data associated with the adjective classes being studied. The challenge will be run in a way similar to the Shared Tasks of the Conference on Natural Language Learning (CONLL), where colleagues are invited to compete to obtain best results on a specified task and data set. Our challenges will require use of the adjective inference datasets developed for training the competing algorithms. We plan to host a workshop at the Language Resources and Evaluation Conference (LREC) in May, 2016, where we will engage the community in further refining the scope and nature of deep textual inferences.

Education. New graduate courses will be developed within the Computer Science Department at Brandeis and the Linguistics Department at Stanford, associated with the project research. The courses, envisioned as ”Semantic Annotation and Text-based Inference”, taught by the P.I.s, will have students engage in the methodology developed from the proposal, over new and diverse textual inference phenomena (e.g., bridging, accommodation, shared beliefs). Starting from initial models with expert annotators, students will learn how to deploy the data over a crowdsourced annotation environment, and examine how to resolve the potential variance or deviation from the initial model. Princeton would contribute materials and develop a local version of the course after it has been offered once. Syllabi and materials from these courses will be made available to the community through mechanisms such as the ACL wiki.

Tutorials and Training. We will design a tutorial on how the paired annotation methodology can be applied and deployed to other annotation tasks and CL challenges. This will be submitted for inclusion at the major conferences in the field (ACL, NAACL, EACL, AFNLP-sponsored conferences, ICGL, LREC, COLING), beginning in spring, 2015 and continuing to the end of the project. We will also propose tutorials at summer schools such as NASSLLI, ESSLLI, and LSA.

5 Results from Prior NSF Support

SI2-SSI: The Language Application Grid: A Framework for Rapid Adaptation and Reuse *NSF 1147912* (PI: James Pustejovsky) 7/2012-6/2015; $1,962,526. The goal of this project is to build a comprehensive network of web services and resources within the NLP community. This involves: (1) the design, development and promotion of a *service-oriented architecture* for NLP development that defines atomic and composite web services for NLP, along with support for service discovery, testing and reuse; (2) the construction of a *Language Application Grid* (LAPPS Grid) based on Service Grid Software developed at NICT and Kyoto Unversity.; (3) deployment of an open advancement (OA) framework for component- and application-based evaluation; and (4) community involvement with the LAPPS Grid.

RI: Small: Interpreting Linguistic Spatiotemporal Relations in Static and Dynamic Contexts *NSF 1017765* (PI: James Pustejovsky) 8/01/10-7/31/13; $493,862.00. This grant focuses on developing spatial processing algorithms to automatically capture locations, paths, and motion constructs in text. Results of this work include the working draft specification of ISO-Space, the implementation of a place identifier, and the mapping of DITL ouput, a dynamic temporal logic, to ISO-Space representations, for subsequent use by extraction and inferencing algorithms.

INTEROP: Sustainable Interoperability for Language Technology *NSF 0753069* (PI: Nancy Ide; co-PI: James Pustejovsky) 9/2008-8/2013; $503,620. This collaborative effort with the EU-funded FLaReNet project is aimed at establishing standards and principles of interoperability within the corpus construction and natural language technology fields, and implementing state-of-the-art formalisms that support interoperability of language processing components and frameworks. **Publications:** [33]; [32].

CRI: Towards a Comprehensive Linguistic Annotation of Language *CNS 0551615 CRI* (PI James Pustejovsky), awarded 08/22/2005, $1,935,867.00. This work explored how to merge annotations from different layers of semantic annotation, working from the assumption that it is the combination of these layers that proves useful for applications. This grant spawned two supplementals: (i) CNS 0832940 CRI, awarded 04/03/2008, $6,000.00, for annotation support, and (ii) CNS 083670 CRI, awarded 05/15/2008, $10,000.00, to support organization of the North American Computational Linguistic Olympiad (NACLO). **Publications:** [67]; [65]; [66].

Workshop on Scalar Adjectives *NSF 1139844*, (PI Christiane Fellbaum). The PI organized a community workshop on "Extracting, Constructing, Modeling and Applying Scales for Gradable Adjectives" at the NSF in Virginia, 09/ 30 - 10/011, 2011. Participants agreed that a number of applications, including Word Sense Disambiguation, reasoning and inferencing would benefit from the study of scalar adjectives and the encoding of scales in WordNet. The unidirectional entailments that can be derived from scales and that allow implicatures are likely to boost deep language understanding. Specific recommendation from workshop participants are incorporated into the present proposal. **Publication:** [57].

CI-ADDO-EN: A Second-Generation Architecture for WordNet*CNS 0855157*(PI: Christiane Fellbaum) 07/29/2009 - 07/31/2012 $396,231.00. This grant supports the design and creation of a relational database for WordNet as well as numerous lexicographic improvements and community support. **Publications:** [22],[19],[18],[6],[49].

CNS: 1204573 CI-P: Collaborative Research: LexLink: Aligning WordNet, FrameNet, PropBank and VerbNet PI Christiane Fellbaum, awarded 06/01/2002, §45,000.00. This grant funded a community workshop at LREC 2012 to explore the linking of four lexical resources, WordNet, FrameNet, PropBank, VerbNet. Participants agreed that the transitive closures among the current partial links would result in numerous benefits for the NLP community.

CCF 0937139: Interactive Discovery and Semantic Labeling of Patterns in Spatial Data PI: T. Funkhauser, co-PIs: D. Blei, A. Finkelstein, C. Fellbaum, awarded 08/25/2009. $499,934.00. This work explored the use of WordNet for labeling spatial data.

Three supplements supported grant IIS -0705199, 08/17/2007 - 07/16/2011: RI: Collaborative Proposal: Complementary Lexical Resources: Towards an Alignment of WordNet and FrameNet, PIs C.Fellbaum and C. Baker (ICSI). **CNS 0835139**, awarded 06/12/2008, $6,000.00; **RI: 1007133**, awarded 12/29/2009, $6,000.00; **IIS 0903358**, awarded 10/31/2008 §6,000.00. The original grant and the three supplements supported the manual alignment of FrameNet and WordNet. An important by-product was the manual annotation of all senses of the targeted word forms in the American National Corpus. **Publications:** [20]; [21] [4] [16].

Workshop on Semantics for Textual Inference *NSF 1064068*, (PI Cleo Condoravdi, Co-PI Annie Zaenen). The PIs organized two workshops, one at the LSA Institute 09-10/07/2011, the other at CSLI, Stanford, 09-10/03/2012. **Publications:** [11].

References

[1] Recognizing textual entailment (rte) corpus. `http://www.nist.gov/tac/2010/RTE/`.

[2] M. Amoia and C. Gardent. Adjective based inference. In *Proceedings of the Workshop KRAQ'06 on Knowledge and Reasoning for Language Processing*, pages 20–27. Association for Computational Linguistics, 2006.

[3] M. Amoia, C. Gardent, et al. A test suite for inference involving adjectives. *Proceedings of LREC'08*, pages 19–27, 2008.

[4] Collin F. Baker and Christiane Fellbaum. Wordnet and framenet as complementary resources for annotation, 2009.

[5] Chris Barker. The dynamics of vagueness. *Linguistics and Philosophy*, 25(1):1–36, 2002.

[6] C. Chiarcos, J. McCrae, P. Cimiano, and C. Fellbaum. Towards open data for linguistics: Linguistic linked data. In *New Trends of Research in Ontologies and Lexical Resources*. Springer, New York, in press.

[7] Herbert H Clark. *Using language*, volume 4. Cambridge University Press Cambridge, 1996.

[8] P. Clark, C. Fellbaum, and J. Hobbs. Using and extending WordNet to support question-answering. *Proc. 4th GWC*, 2008.

[9] P. Clark, C. Fellbaum, J.R. Hobbs, P. Harrison, W.R. Murray, and J. Thompson. Augmenting WordNet for deep understanding of text. *Proceedings of STEP*, pages 45–57, 2008.

[10] P. Clark, W.R. Murray, J. Thompson, P. Harrison, J. Hobbs, and C. Fellbaum. On the role of lexical and world knowledge in RTE3. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 54–59. Association for Computational Linguistics, 2007.

[11] C. Condoravdi, V. de Paiva, and A. Zaenen (eds.). *Linguistic Issues in Language Technology: Special Issue on Perspectives on Semantic Representations for Textual Inference*. Forthcoming.

[12] Ido Dagan, Oren Glickman, and Bernardo Magnini. The PASCAL recognising textual entailment challenge. In *MLCW*, pages 177–190, 2005.

[13] Dmitry Davidov and Ari Rappoport. Unsupervised discovery of generic relationships using pattern clusters and its evaluation by automatically generated sat analogy questions. In *Proceedings of ACL-08: HCL*, pages 692–700, 2008.

[14] Marie-Catharine de Marneffe, Christopher D. Manning, and Christopher Potts. Did it happen? The pragmatic complexity of veridicality assessment. *Computational Linguistics*, 38:301–333, 2012.

[15] Marie-Catherine De Marneffe, Bill MacCartney, Christopher D Manning, et al. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454, 2006.

[16] G. de Melo, C. Baker, N. Ide, R. Passonneau, and C. Fellbaum. Empirical comparisons of MASC word sense annotations. In *Proceedings of LREC, Istanbul, Turkey*, 2012.

[17] R.M.W. Dixon. *A new approach to English grammar on semantic principles*. Oxford University Press, 1991.

[18] Christiane Fellbaum. Wordnet. In *Theory and Application of Ontology: Computer Applications*, pages 231 – 243. Springer New York, 2012.

[19] Christiane Fellbaum. Wordnet. In *The Encyclopedia of Applied Linguistics*. Wiley/Blackwell, to appear 2013.

[20] Christiane Fellbaum and Collin Baker. Representing verb meaning in complementary resources. *Linguistics*, in press.

[21] Christiane Fellbaum and Collin F. Baker. Can WordNet and FrameNet be made interoperable? In *Proceedings of The First International Conference on Global Interoperability for Language Resources*, page 6774, 2008.

[22] Christiane Fellbaum and Piek Vossen. Challenges for a multilingual WordNet. *Language Resources and Evaluation*, 46(2):313–326, 2012.

[23] Michael C Frank and Noah D Goodman. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998, 2012.

[24] Noah D Goodman and Andreas Stuhlmüller. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, 5(1):173–184, 2013.

[25] H.P. Grice. Logic and conversation. pages 64–75, 1975.

[26] Joy E Hanna, Michael K Tanenhaus, and John C Trueswell. The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory and Language*, 49(1):43–61, 2003.

[27] Marti Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*. 1992.

[28] Julia Hirschberg. *A Theory of Scalar Implicature*. Garland Press, 1991.

[29] Jerry R Hobbs, Mark E Stickel, Douglas E Appelt, and Paul Martin. Interpretation as abduction. *Artificial Intelligence*, 63(1):69–142, 1993.

[30] Laurence Horn. *A natural history of negation*. The University of Chicago Press, 1989.

[31] L.R. Horn. Pick a theory, not just any theory. *Negation and Polarity. Syntactic and Semantic Perspectives*, pages 147–192, 2000.

[32] Nancy Ide and Harry Bunt. Anatomy of annotation schemes: Mapping to GrAF. In *Proceedings of the Fourth Linguistic Annotation Workshop (LAW IV)*. Association for Computational Linguistics, 2010.

[33] Nancy Ide and Keith Suderman. Bridging the gaps: Interoperability for GrAF, GATE, and UIMA. In *Linguistic Annotation Workshop*, pages 27–34, 2009.

[34] H. Kamp. Two theories about adjectives. In *Formal Semantics of Natural Language*, pages 123–155. University Press, 1975.

[35] H. Kamp and B. Partee. Prototype theory and compositionality. *Cognition*, pages 57–129, 1995.

[36] Lauri Karttunen. Implicative verbs. *Language*, 47:340–358, 1971.

[37] Lauri Karttunen. You will be lucky to break even. In Tracy Holloway King and Valeria dePaiva, editors, *From Quirky Case to Representing Space: Papers in Honor of Annie Zaenen*, pages 167–180. CSLI Publications, Stanford, CA, 2012.

[38] Lauri Karttunen, Cleo Condoravdi, Miriam Connor, Stuart Melton, Kenny Moran, Marianne Naval, Stanley Peters, Tania Rojas-Esponda, and Annie Zaenen. Double meaning: A systematic empirical study. Paper presented at the 20th International Congress of Linguists, July 2013.

[39] Lauri Karttunen, Annie Zaenen, Cleo Condoravdi, and Stanley Peters. When you are not stupid, you do not do stupid things: Evaluative uses of factive adjectives. Paper presented at the Colloque de Syntaxe et Sémantique à Paris (CSSP), September 2013.

[40] Christopher Kennedy. Polar opposition and the ontology of degrees. *Linguistics and Philosophy*, 24(1):33–70, 2001.

[41] Paul Kiparsky and Carol Kiparsky. Fact. In M. Bierwisch and K. E. Heidolph, editors, *Progress in Linguistics*, pages 143–173. Mouton, Hague, 1970.

[42] Amnon Lotan. A syntax-based rule-base for textual entailment and a semantic truth value annotator. Master's thesis, Tel Aviv University, 2012.

[43] John Lyons. *Semantics*. Cambridge University Press, 1977.

[44] Christopher D Manning and Hinrich Schütze. *Foundations of statistical natural language processing*, volume 999. MIT Press, 1999.

[45] Ilka Mindt. *Adjective Complementation: An Empirical Analysis of Adjectives Followed by that clauses*, volume 42 of *Studies in Corpus Linguistics*. John Benjamins, 2011.

[46] Robert Munro, Steven Bethard, Victor Kuperman, Vicky Tzuyin Lai, Robin Melnick, Christopher Potts, Tyler Schnoebelen, and Harry Tily. Crowdsourcing and language studies: the new generation of linguistic data. 2010.

[47] Kevin P Murphy. *Machine learning: a probabilistic perspective*. The MIT Press, 2012.

[48] Rowan Nairn, Cleo Condoravdi, and Lauri Karttunen. Computing relative polarity for textual inference. In *ICoS-5*, pages 67–76, 2006.

[49] Sonya Nikolova, Jordan Boyd-Graber, and Christiane Fellbaum. Collecting semantic similarity ratings to connect concepts in assistive communication tools. In *Modeling, Learning and Processing of Text Technological Data Structures*, pages 81–93. Springer, 2012.

[50] Neal R. Norrick. *Factive Adjectives and the Theory of Factivity*. Niemeyer, 1978.

[51] Steven T Piantadosi, Harry Tily, and Edward Gibson. The communicative function of ambiguity in language. *Cognition*, 122(3):280–291, 2012.

[52] James Pustejovsky and Amber Stubbs. *Natural language annotation for machine learning*. O'Reilly, 2012.

[53] V. Raskin and S. Nirenburg. Lexical semantics of adjectives. *New Mexico State University, Computing Research Laboratory Technical Report, MCCS-95-288*, 1995.

[54] R. Saurí and J. Pustejovsky. Are you sure that this happened? Assessing the factuality degree of events in text. *Computational Linguistics*, 38(2):261–299, 2012.

[55] Roser Saurí. *A Factuality Profiler for Eventualities in Text.* PhD thesis, Brandeis University, 2008.

[56] Roser Saurí and James Pustejovsky. Factbank 1.0. `http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2009T23`, September 2009.

[57] V. Sheinman, C. Fellbaum, I. Julien, P. Schulam, and T. Tokunaga. Large, huge or gigantic? identifying and encoding intensity relations among adjectives in wordnet. *Lexical Resources and Evaluation*, 2013.

[58] V. Sheinman and T. Tokunaga. Adjscales: Visualizing differences between adjectives for language learners. *IEICE TRANSACTIONS on Information and Systems*, 92(8):1542–1550, 2009.

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

[60] Rion Snow, Daniel Jurafsky, and Andrew Y Ng. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 801–808. Association for Computational Linguistics, 2006.

[61] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. Cheap and fast - but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of EMNLP*, 2008.

[62] Asher Stern and Ido Dagan. A confidence model for syntactically-motivated entailment proofs. In *Proceedings of RANLP 2011*. 2011.

[63] Peter Turney. A uniform approach to analogies, synonyms, antonyms, and associations. *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 905–912, 2008.

[64] An Van linden and Kristin Davidse. The clausal complementation of deontic-evaluative adjectives in extraposition constructions: a synchronic-diachronic approach. *Folia Linguistica: Acta Societatis Linguisticae Europaeae*, 43:171–211, 2009.

[65] Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. Semeval-2007 task 15: Tempeval temporal relation identification. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval- 2007)*, page 7580, Prague, Czech Republic, 2007. Association for Computational Linguistics.

[66] Marc Verhagen and James Pustejovsky. Interoperability of syntactic and semantic annotation schemes. In *Interoperability of Language Resources*, 2007.

[67] Marc Verhagen, Amber Stubbs, and James Pustejovsky. Combining independent syntactic and semantic annotation schemes. In *Proceedings of the Linguistic Annotation Workshop*, pages 109–112, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[68] Larry Wasserman. *All of statistics: a concise course in statistical inference*. Springer, 2004.

[69] Robert Wilkinson. Factive complements and action complements. In *CLS 6*, pages 425–444, 1970.

[70] Gbolahan K Williams and Sarabjot Singh Anand. Predicting the polarity strength of adjectives using wordnet. In *ICWSM*, 2009.

[71] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational linguistics*, 35(3):399–433, 2009.

[72] Annie Zaenen and Lauri Karttunen. Veridicity annotation in the lexicon? A look at factive adjectives. In *The Ninth Joint ACL – ISO Workshop on Interoperable Semantic Annotation*, 2013.

[73] Annie Zaenen, Lauri Karttunen, Cleo Condoravdi, and Stanley Peters. A polarity lexicon of adjectives. `http://www.stanford.edu/group/csli_faust/Lexical_Resources/Polarity-Lexicon-of-Adjectives/`. Lexical resource compiled at the Center for the Study of Language and Information, Stanford University, 2012.