Project Description

IIS-RI MEDIUM:
Lexical Inference Patterns for Adjectives in Natural Language

1 Introduction

Effective human communication relies on the ability of speakers to recover information that is not explicitly expressed in an utterance. According to [23]'s estimate, 90% of what is communicated is implicit and must be inferred. [24], in his investigation of the pragmatics of human communication, postulated principles of cooperative conversation, by which context-dependent meanings can be recognized and interpreted without being expressed literally or even at all. While some of this is due to the pragmatics of the situation and the conversation, much of such covert information can be identified as "semantic inferences", and as such, can be associated with identifiable structural or lexical patterns in natural language. An understanding of how speakers identify and exploit systematic covert inferences in language can enrich our models of compositionally derived inferences. At the same time, it can enhance the capabilities of natural language understanding systems to read beyond the surface forms of the text.

Performing such text-based inferences is one of the major goals of current NLP research. This requires training machine learning (ML) algorithms to reconstruct the often subtle lexical and constructional cues in texts, just as effective human communication relies on the ability of speakers to recover information that is not explicitly expressed in utterances. Systems usually rely on training or development corpora that have been annotated automatically or by hand. These annotations are in most cases derived from lexical classes or from syntax-semantics correspondences found in the linguistic literature. The work proposed here focuses on lexically based distinctions. Previous research has mainly concentrated on the properties of verbs and nouns because the main task was information extraction, defined as identifying "events" and their participants. More recently, however, attention has shifted to exploiting information about the stance of the participants in the textual information exchange. Our work goes beyond the mere identification of entities and events and addresses the questions as to whether or not events have occurred, might be likely to occur, are desirable, and for entities, how their qualities compare to those of other entities.

These questions point naturally to a more prominent role for how adjectives and adverbs are semantically interpreted. There is, however, much less readily exploitable linguistic information available for these lexical categories, and it is less 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, ignoring inferences that arise reliably but are not yet entailments. 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 in the wild, through a better integration of lexical resources and corpus annotation.

Our proposed research then, consists in developing a methodology for the discovery and exploitation of systematic linguistic inferences identified with specific lexical classes in natural language. We develop this methodology in the construction of an inferential model for adjectival semantics in natural language. To this end, the three specific aims of this proposal are:

- 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 between 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 inclusion in lexical resources such as WordNet.

We concentrate on three diverse semantic types of adjectives, in order to both: (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) evidentiality adjectives, showing varying implicatures of veridicity over a clausal complement, e.g., *rude, annoying, likely*, etc.; and (iii) intensional adjectives, introducing implicatures 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. However, contrary to most previous corpus annotation efforts, this standard is not the end product to be used in ML: we will compare these baseline structure-to-inference mappings to inferential judgments made by non-expert native speakers, such as 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 that are important to explain the non-expert judgments. We will use these differential measurements in judgment (trained linguist vs. non-expert annotator) to classify the implicatures according to two dimensions: how stable an inference is regardless of linguistic context; and which contextual factors contribute to blocking the expected 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.

The contributions of this research are significant to the computational linguistics community in two major respects. First, they lay the theoretical and methodological groundwork for a large-scale annotation of adjectives in order to support automatic systems in inferencing tasks. More importantly, , they contribute to a more sophisticated theory of the contribution of lexical information to inferencing. By studying the way inferences are made "in the wild" and how these differ from baseline expectations 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. [15, 41, 52]) categorizes the attributes denoted by adjectives according to a thematic organization, centered around a human frame-of-reference, as lexically encoded in the language, such the following classes:[1] DIMENSION, PHYSICAL PROPERTY, COLOR, EMOTIONS, TEMPORAL SPATIAL VALUE, MANNER.

An alternative approach is to adopt a conceptually conservative but more formally descriptive and operational distinction, one which groups adjectives into inferential classes. [2] and [3], following [32] and [33], make just such a move, adopting a four class distinction based on inferential properties of the adjective, as illustrated below:

(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,

[1] It should be noted that [52] , however, also discuss inferential patterns for distinct classes.

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 three different classes of adjectives that require refinements and additions to the inference patterns given above. These classes are:

(2) a. Scalar adjectives: both dimensional (*big, small*) and evaluative (*happy, pretty*) scalars have been categorized as subsective adjectives;

b. Adjectives with clausal complements: adjectives such as *annoying* and *nice*, when governing clausal complements, do not fit nicely into any of the above classes;

c. Intensional adjectives: adjectives such as *alleged* and *supposed* are non-subsective, but in complex ways that are dependent on the semantics of the nominal head.

Examples of the types of inferences we intend to capture are the following:

- The PASCAL Recognizing Textual Entailment task ([1, 11]) 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:

  (3) 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 (4).

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

  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.

- 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:

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

- Concerning the third adjective class, the intensional adjectives, the effect of modifying the nominal head is the introduction of "epistemic uncertainty" regarding the description.

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

  Making the inference of H from T here would not be justified. However, consider the pair below:

  (7) 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 (stone).

One of the resources typically relied upon to improve automatic inferencing based on texts in natural language is supervised or unsupervised annotation of the lexical items occurring in the text. These annotations reflect directly or indirectly the inferential potential that is associated with lexical items. Some aspects of this inferential potential have been studied in detail in the linguistic literature and the computational approaches tend to take the results of these studies for granted. In the case of nouns, the WordNet hierarchies have proved useful in numerous studies (e.g., [58]); for verbs, lists of special inference patterns have been constructed starting from the work of [39, 34] by [47, 54, 55, 40]. Information about the inferential properties of adjectives is, however, much less easy to come by. Our preliminary studies show that, for the categories of adjectives that we are interested in, the existing resources have severe shortcomings or are non-existent. One of the reasons is that the contribution of adjectives tends to be more subtle and more dependent to the rest of the linguistic context. This difficulty requires, in our view, a more careful methodology than the one that has been used up to now for syntactic and semantic lexical categorization tasks. The availability of, on the one hand, digital corpora (the biggest one being the Web itself) and, on the other, 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 [61] or [10]). Although we will start from the known linguistic literature, our main effort will be focused on corpus-based analysis with the Web itself as our main corpus and on crowd-sourcing experiments. The combination of these approaches will insure that our results are representative for a larger community of users of English but they will also help us notice textual interactions that tend to be ignored in studies based on linguists' armchair intuitions. It is important to have a solid understanding of these interactions before embarking upon a large scale annotation task: it is only if we can characterize the contribution of the targeted lexical items in its various linguistic contexts securely enough that corpus or lexicon annotation is truly useful.

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* [51], 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.

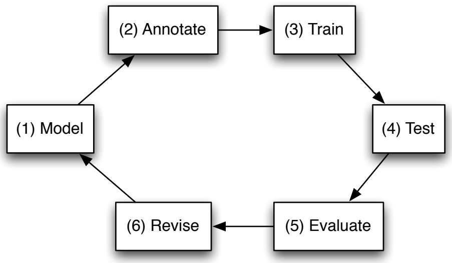

Figure 1: The MATTER Methodology

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.

For practical reasons, the corpora that can be created using linguistically trained annotators are rather limited and rarely exhibit all the combinations of possibly relevant context factors. The result, in many domains, is a 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 [42, 66, 46]: 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 and understudied linguistic 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 basic RTE tasks [59]. Adopting this method as part of 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 defesaible 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.

It is well known that isolating the factors that contribute to the perception of an inference is extremely difficult. Testing contrastive contexts with a large number of native speakers is one way to check whether the hypothesized factors are indeed the ones that are at work and whether they have been circumscribed sufficiently. 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 [35, 70] 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 (8), 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).

(8) 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 [35, 70, 37] 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:

(9) 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 as follows. For each adjective class being studied (Scalar, Factive, Intensional), we establish an initial model, incorporating the appropriate SIM as the interpretation function. Expert and naïve annotations create corresponding gold standards, from which we perform context-based adjudication.

2.2 Scalar adjectives

Human language understanding involves 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 [24, 27, 7, 25, 50]. 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 [28, 29]. The use of items lower on the scale, in turn, can lead to an defeasible inference that the sentence would be false if the lower item were replaced by the higher [24, 28].

(10) 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 [26, 21, 22]. 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:

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

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

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

(11) illustrates a standard scalar implicature, where the choice to use "hot" when "scorching" would clearly be relevant leads to an inference that Dallas is not, in fact, scorching. In (12), however, this inference is much less robust, presumably because it is not clear whether "scorching" would be a relevant alternative. Finally, in (13) no implicature regarding "scorching" arises, but (perhaps surprisingly) the inference that Dallas is not beautiful is very robust. These examples illustrate the importance of taking into account contextual information when drawing inferences from scalar adjectives.

A second complication in modeling inferences from scalar arises due to their DIMENSIONALITY and partial overlap of dimensions. 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, typically placing requirements simultaneously on multiple scalar dimensions (e.g., three-dimensional size and perhaps weight, in the case of *big*). Emotive adjectives such as *happy* are even more complex, since the scalar 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* (by being large but flat, as in a large city), we know that 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 then do we explain contextual entailments such as the one illustrated in (4) 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 when such adjectives are involved.

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*);

- 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.

Progress has been made in recent years on the question of determining adjectival polarity [68] and total entailment relations between adjective pairs [57, 56]. However, there is room for improvement in these methods, in particular by using unsupervised methods for learning lexical relations from parsed corpora rather than hand-specified patterns on raw text (cf. [58]). 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 crowd-sourced 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 us to perform statistical analyses on the responses of many annotations of the same text.

[2] Note that the concept of "polarity" intended here is the one used in theoretical linguistics rather than the somewhat different concept from sentiment analysis; it refers not to the typical emotional content attached to an expression [69] but to the "directionality" of a scale in a logical sense [38], as diagnosed by the meaning of the comparative form of the adjective.

2.3 Inferences of predicative 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. As we discuss below, 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 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. 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, as shown below. Thus, if one is interested in inferences that can be drawn on the basis of linguistic form, one has to consider lexical items together with their syntactic patterns.

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, (14).

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

From (14), 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 (15). The focus of the question in (15b) is the evaluation of the *that*-complement as annoying or not.

(15) 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?

The only comprehensive study of factive adjectives was done by Norrick in [49]. Norrick proposed two big subclasses : emotive (e.g. *sad*) and evaluative (e.g. *stupid*) adjectives. The empirical picture , however, is more complicated and the status of the factuality inference varies among speakers, as reported in [37] and discussed below.

2. Certainty adjectives These adjectives directly assert the degree of certainty that the write, or a protagonist, ascribes to the complement, as illustrated in (16).

(16) 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?

In (16c) it is the *that*-complement itself that is questioned, and (16b) has the opposite inference from that of (16a). Structure-inference patterns for these adjectives then would 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* ([34]). Others, such as *possible, probable, impossible, improbable*, do not express absolute certainty but 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 [54] 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 [49], there are the more limited studies in [67] and [5]. Implicative patterns ([34]) and degree-of-certainty adjectives are only mentioned in passing in the literature. [44] looks at the syntax of 51 frequent adjectives taking *that* clauses in the BNC but without any attention to the semantics.[62] report on a corpus study of deontic-evaluative adjectives concentrating on *important, essential, crucial* and *appropriate, proper and fitting*.

Pilot studies we have conducted on various subclasses of adjectives with clausal complements have revealed that their inferential behavior is dependent on fine-grained structural and contextual factors. An exhaustive list of all patterns cannot be given in the confines of this proposal. We discuss some of our observations 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. [49] 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 (17).

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

The other variants of the pattern (present tense, non-specific *NP* or no *of NP*) allow for much variation in interpretation.

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 Mechanical Turk workers to be factual for more than 50% of them and probable for another 35%.

Our 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 are implicative under certain circunstances for many people ([37]). The preferred interpretation of *I wasn't stupid to send money* is that no money was sent. 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. This result was corroborated by a pilot experimental study, which showed that for a sentence such as *Kim was not stupid to waste money* 66.7% of the subjects give an implicative interpretation (when given the choice between implicative, factive and don't know). Similar results obtain with *clever*. This pattern is clearly dependent on extra-linguistic factors: for *Kim was not stupid to save money*, we get 78.6% in favor of a factive interpretation. Previous theoretical assumptions would have predicted only the factive interpretations in both cases. Our studies show that the inference patterns need to be further specified with respect to the content of the VP. The exact formulation of these specifications is part of the proposed project.

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 ([71]). A structure-to-inference mapping corresponding to an impersonal syntactic frame is given in (18).

(18) $[\text{ it be } ADJ \text{ that } S\ ] \vDash S$

2.4 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:

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

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 [33], no informative inference is associated with this construction:

(20) a. $[A\ N]$ (alleged criminal)
b. $\nvDash N$

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, (21a) below, there are many contexts where the epistemic scope is reduced to a modification or additional attribution of the nominal head, as shown in (21b).

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

In Example (21a), 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 (21b), 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.

(22) Given the construction $[A_{int}\ N]$, where A_{int} is *alleged, ...*, then:
a. $[A_{int}\ N] \nvDash N$
b. $[A_{int}\ A_2\ N] \nvDash A_2$
c. $[A_{int}\ A_2\ N] \vDash N$

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 (22) will go through. Consider the examples below:

[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.

(23) 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 (23) involves an additional aspect of the NP, beyond the unassailable characteristics of the entailed head. That is, the object is clearly a vase (in (23a)) and demonstrably a sonata (in (22b)). 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 (22a) as the "Wide-scope reading", and the inferences in (22b-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.

(24) 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 (25), where *alleged* is modifying an event nominal.

(25) 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 (25a-b) follow from the template in (22a). For sentence (25c), 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.

(26) 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". Similar remarks hold for undergoer nominals in some contexts, where the scope of the intensional adjective can be lowered to a modification of the event description. This is illustrated below, in (27b).

(27) 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 (27a) behaves according to the canonical template, while (27b) 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 (22), where the modification (argument specification, in fact) is postnominal.

(28) 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:

(29) 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

The project consists of three specific aims: (1) developing an inferential model for adjectival semantics in natural language; (2) connecting this model to data by formulating templates of structure-to-inference mappings using data mining techniques over Web corpora; and (3) revising and enriching the theoretical model and inference templates by examining the same data "in the wild", that is, crowdsourced judgments using larger textual contexts.

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) manually select target adjectives, (b) apply regular expressions to the Web so as to extract text snippets containing the target adjectives and (c) construct, on the basis of the data culled from the web, small corpora in the format of RTE to be annotated by both linguistically trained and non-expert annotators. The former will be the "Gold" judgments and the latter the "Wild" judgments. On the basis of these judgments, we revise our models and test them again "in the wild."

The Web as corpus. We propose to use the Web as a corpus for the extraction of filled patterns. The principal advantages over corpora like the BNC or COCA are the size of the Web and access to broad and diverse speaker communities. Frequently cited disadvantages will most likely not affect the proposed work. Non-native or non-standard language, which is characterized by idiosyncratic lexical choices, non-standard morphology and ungrammatical constructions do not apply to the short and fixed patterns of interest. In the case of scalar adjectives, false positives like *makes you rich; if not wealthy enough to build such a house* when the aim is to find the pattern *rich if not wealthy* will be identified as such in a subsequent filtering process using the freely available Stanford parser ([13]). We expect the Web to yield sufficient data so as to allow us to detect outliers based on non-native or idiosyncratic intuitions. We represent the patterns as regular expression (REs) and apply them as search queries to the corpus.

To avoid missing tokens that exhibit variations of the target patterns, we formulate the pre-defined, "strict" patterns in a way that allows flexibility and apply these in addition to the strict patterns. "Flexibly" formulated REs will likely result in a higher number of false positives. A solution for maximizing the number of true positives while keeping noise to a minimum is to process the results and perform POS tagging and parsing. But rather than processing the entire corpus or all search results returned for a given query, we minimize the computational cost as follows. Search results generated by the strictly formulated REs will be considered as valid examples of the pattern and will be directly included in the dataset for semantic analysis. These false positives will be removed from the results returned by the flexibly formulated REs for the same pattern, and only the remaining results will be passed to the parser.

Crowdsourcing. Several studies in the last years have shown that crowd-sourcing experiments can deliver reliable results [60] and [45]. As with our preliminary experiments, we will use boto[4], the Python interface to AWS (Amazon Web Services), to build an application capable of creating and publishing HITs, collecting results and evaluating the judgments from MTurkers. Our approach to soliciting inferences associated with the contextualized interpretation of adjectives relies on the linguistic intuitions of non-expert native English speakers, who will provide the core judgments used to enrich the initial gold annotations, created by linguists. We will present test questions to ascertain that the workers are native speakers of English. Multiple MTurkers will be asked to make direct inferences, given the text presented to them.

The crowd sourcing techniques for the scalar adjectives serve mainly to measure the relative value of each lexical-semantic pattern as a discriminator for adjectives that express different degrees of a given attribute.

[4] http:boto.s3.amazonaws.com.

3.1 Scalar Adjectives

We select 100 frequent adjectives with scalar properties expressing different values of twelve different attributes. For adjectives on a given scale, their relative intensity becomes apparent in the context of lexical-semantic patterns. Thus, the patterns ″X even Y″ and ″if not Y, at least X″ indicate that *gorgeous* expresses a greater degree of the attribute ″beauty″ than *pretty*: *pretty, even gorgeous*; *if not gorgeous, at least pretty*. We represent the patterns as regular expression (REs) and apply them as search queries to the corpus, following [57].

For each query, the adjectives are pre-classifed as members of one half of a scale (or a part of a half scale) based on WordNet's "dumbbells" (clusters of adjectives containing two polar antonymic adjectives (such as *large* and *small*) and a number of adjectives that are "semantically similar" to one of the polar adjectives). A given query uses adjectives from oen half of the dumbbell only, including for example *large* and *enormous* but not *small* and *enormous*. Importantly, the patterns are such that they apply to adjectives with shared selectional restrictions (the nouns they modify), so that adjectives that were misclassified in WordNet and irrelevant senses of polysemous adjectives are not returned by the searches.

We extend the method as described in [56], applying additional patterns such as the Negative Polarity Items ("She's certainly not very bright, let alone brilliant."). A given lexically filled pattern will be applied with the lexemes in both orders to identify lexemes that are semantically similar enough to be considered synonyms rather than different in strength or variation inter-speaker variation. Given the pairwise orderings, we construct scales as follows. We process each half of a WordNet "dumbbell" (a central adjective like *rich* plus its ″similar″ adjectives *wealthy, comfortable, loaded* etc.) separately. For each pair (centroid, similar-adjective), we instantiate each pattern p in patterns that were extracted in the preprocessing stage to obtain phrases $s_1 = p(\text{head-word, similar-word})$ and $s_2 = p(\text{similar-word, centroid})$. We send $s1$ and $s2$ to a search engine as two separate queries and check whether $df^5(s_1) > weight \times df(s_2)$ and whether $df(s_1) > threshold$. The higher the values are for the $threshold^6$ and $weight^7$ parameters, the more reliable are the results. If p is of the type *intense*, then a positive value is added to the similar-word's score, otherwise if p is of the type *mild* a negative value is added. When all the patterns are tested, similar-words with positive values are classified as intense, while the similar-words with negative values are classified as mild. Words that score 0 are classified as *unconfirmed*. For each pair of words in each one of the subsets (mild and intense), the same procedure is repeated, creating further subsets of *mildest* words that have the most negative values within the mild subset, and *most intense* words for the words with the highest positive values within the intense subset. Adjectives of similar intensity are grouped together.

The Gold standard for scalar adjectives We plan to collect human judgments from linguistically sophisticated native speakers of the relative intensity of a subset of the adjectives. Our specific goals are (a) to determine whether the ordering derived from the Web data are consistent with human judgments, and (b) which lexical-semantic patterns are more reliable in revealing partial orderings among adjective scalemates. We build on informal pilot work.

A recent small experiment presented twenty Princeton University students with ten adjectives denoting size (*big, large, enormous, gigantic, huge, tremendous, colossal, gargantuan, monumental, humongous*). The adjectives were arranged two sets of five pairs, one set presenting them in the inverse order from the other set. Two groups of students were each given one set and asked to indicate which member of a given pair expressed a greater value of the attribute ″size;″ the option of equal value was allowed for as well. Across both groups, there was complete agreement on most pairs (e.g., all students agreed that *gigantic* is greater than *large* and that *huge* is smaller than *tremendous*). We found that for the groups agreed well with each other and the judgments for all ratings yielded a coherent picture, as follows: There was complete (100%) agreement on the weakest and the strongest member of each scale. For other adjectives, there some some disagreement with respect to their pairwise ordering (e.g., whether *colossal* was stronger than *gigantic* or

[5] df represents *document frequency*.

[6] *threshold* regulates the number of pages returned by the search engine that is considered sufficient to trust the result.

[7] *weight* regulates the gap between s_1 over s_2 that is required to prefer one over the other.

vise versa). The lower frequency of these adjectives may in part account for the lower agreement.

A subsequent task presenting the same adjectives in random order asked the students to place them on a single scale, again allowing for more than one adjective to occupy the same point on the scale. The scales that the raters constructed were consistent with the pairwise judgments, though the students were not able to consult their pairwise ratings when constructing the scales. All judges rated both *big* and *large* as being the least intensive adjective on the scale. There was also strong agreement that *colossal, gargantuan* and *monumental* expressed the most intense values of "size." *Gigantic* and *enormous* were placed towards the end of the scale; some disagreement was found towards the center of the scale, where *humongous* and *tremendous* were not clearly discriminated. In sum, the small experiment confirmed previous findings by [57] and [43] and showed that (a) human judges could perform the task, given clear, explicit instructions and illustrative examples; (b) there was good agreement among the judges with respect to the pairwise ordering as well as the scalar orderings; (c) the scales clearly reflected the pairwise orderings; (d) the scales showed the highest agreement at either end and less (though still good) agreement towards the center. These pilot studies confirm that speakers have scalar values as part of their representation of adjectives, that they can access these representations and that there is significant agreement across speakers.

As it is not practicable to collect judgments for the 100 target adjectives with the methodology applied in the pilot work, we collect the Gold standard using the RTE format. In the absence of existing Text-Hypothesis pairs that crucially involve scalar adjectives, we construct pairs such as exemplified below and ask for an evaluation of the Hypothesis, covering fifty frequent adjectives from six scales.

(30) (T): Pat's daughter is gorgeous.
(H): Pat's daughter is nice-looking.

We then compare the judgments with the data mining results and evaluate their agreement with the different lexical-semantic patterns. We expect some patterns to discriminate more clearly than others among adjectives expressing different intensities of the shared underlying attribute.

The "Wild" standard for scalar adjectives We submit the same T-H pairs to the Turkers and compare their judgments to the Gold standard. Given the large sample of speakers, we do not expect significant differences with respect to the orderings of specific adjective pairs. However, the results will indicate, first, whether specific lexical-semantic patterns better capture speakers' representation of the adjectives' meaning differences and second, whether the patterns that are better discriminators are the same for linguistically sophisticated and linguistically non-expert speakers. If such differences are found, they are likely to guide the future encoding of scalar adjectives in WordNet and the application of WordNet's data to automatic reasoning tasks, where the aim is to train systems to emulate human reasoning.

3.2 Clause-selecting Adjectives

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 [54, 53] and validated in [12]. 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 (31) and (32) below. In these examples, the inference in (31) is legitimate, while that in (32) is false.

(31) 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.

(32) 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 MTurkers 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 we construct in the context of an RTE task. Prior work [10], [9], [8] measured the contribution of specific WordNet relations to this task. It was found that a lexical matching strategy between Text and Hypothesis improved measurably when super- and subordinate terms, synonyms, meronyms, and various relations among verbs were considered. The next logical step is to quantify the contribution of scalar orderings among adjectives in WordNet to the RTE task. We manually inspected several datasets used in the RTE tasks and found that none included T-H pairs that critically involved scales. We will construct a new test set with some of the adjectives we focus on, including slightly modified data from existing sets as much as possible. Applying a system to such T-H pairs with and without including the scalar information in WordNet, as described below, will allow a straightforward evaluation. To perform the evaluation, we propose to encode three scales in WordNet. We follow the model described in [56], where WordNet’s ”dumbbells” are maintained and augmented with a linear scale where some – but not necessarily all – of the adjectives of each half of the dumbbells are linked with arcs to specific points on the scale. This dual representation preserves the original WordNet representation in terms of one central adjective (e.g., *rich*) and a set of undifferentiated ”semantically similar” adjectives like *wealthy, comfortable, prosperous, well-heeled, flush, loaded* etc. while also indicating their intensity relative to the central adjective and to one another, as described of [56]. This representation is amenable to an external evaluation with systems like [10], which allow the measurement of the contribution of the encoding of scales in WordNet.

Concerning the evaluation of the 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. We 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. For a larger scale evaluation we incorporate our rule system and lexical patterns to BiuTee and evaluate how much it improves the performance of that system.

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 weekly 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 [56] can be developed).
Q4	Evaluation; Data collection protocols; Prepare articles for publication. Final report.

4 Outreach and Education Plan

The construction of the Gold standard as proposed here involves graduate and undergraduate students, who will learn about the semantic, lexical and syntactic concepts driving our work. The PIs are teaching undergraduate and graduate level courses and will include discussions of adjective semantics and inferencing in future lectures. One PI (Fellbaum) serves as an adviser on a number of Undergraduate Independent Work research projects that focus on scalar adjectives. The other PI (Pustejovsky) serves on the ISO TC 37 /SC 4 committee, and will involve annotator students in the intial specification and development of a markup language for adjectivally induced modality.

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) promotion of adoption, use, and 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:** [31]; [30]; [**?**].

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:** [65]; [63]; [64].

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:** [56].

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:** [20],[17],[16],[6],[48].

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:** [18]; [19] [4] [14].

Workshop on Semantics for Textual Inference *NSF 1064068*, (PI Cleo Condoravdi, Co-PI Annie Zaenen).

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

[12] 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.

[13] Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. Generating typed dependency parses from phrase structure parses. In *LREC 2006*. 2006.

[14] 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.

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

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

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

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

[19] 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.

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

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

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

[23] A.C. Graesser and S.M. Goodman. Implicit knowledge, question answering, and the representation of expository text. *Understanding expository text*, pages 109–171, 1985.

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

[25] 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.

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

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

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

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

[30] 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.

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

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

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

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

[35] 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.

[36] 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.

[37] 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.

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

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

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

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

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

[43] Y.Y. Mathieu and C. Felbaum. Verbs of emotion in French and English. *Proceedings of GWC-2010, Mumbai, India*, 2010.

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

[45] 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.

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

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

[48] 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.

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

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

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

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

[53] 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.

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

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

[56] 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.

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

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

[59] 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 the conference on empirical methods in natural language processing*, pages 254–263. Association for Computational Linguistics, 2008.

[60] 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.

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

[62] 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.

[63] 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.

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

[65] 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.

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

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

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

[69] 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.

[70] 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.

[71] Annie Zaenen, Lauri Karttunen, Cleo Condoravdi, and Stanley Peters. A polarity lexicon of adjectives. Lexical resource compiled at the Center for the Study of Language and Information, Stanford University, 2012.