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 inexplicit 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 proposed 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 event or states 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 events have or have not occurred, might be likely to occur or not, are desirable or not, and, for entities, how their qualities compare to those of other entities.

These questions point to a more prominent role for adjectives and adverbs. However, there is 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, because they ignore important semantic interactions between the various textual elements as well as usage of linguistically untrained ("naïve") speakers. This requires training machine learning (ML) algorithms to reconstruct the often inexplicit 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 proposed 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 event /states 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 events have or have not occurred, might be likely to occur or not, are desirable or not, and, for entities, how their qualities compare to those of other entities. These questions point to a more prominent role for adjectives and adverbs. However, there is 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, because they ignore important semantic interactions between the various textual elements as well as the linguistic use of "naive" (non-linguist) speakers.

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 specific aims of this proposal are:

- Establish an initial model for each adjective class, and identify templates of structure-to-inference mappings from corpora constructed using data mining techniques over the Web, using expert annotators.
- Have the same data interpreted "in the wild", with crowdsourced judgments and larger textual contexts, using Amazon Mechanical Turk (AMT) annotators.
- Determine whether the "non-expert" annotations diverge from the expected judgments. Where they do differ, if inter-annotator agreement is high, identify what previously untagged contextual parameters are contributing to the interpretation. Revise and enrich the model accordingly.

We concentrate on three diverse semantic types of adjectives, in order to both: (a) test the applicability of the methodology to different linguistic 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*. For each adjective type, the work will produce a small Gold standard inference corpus created by a standard linguistic annotation effort following explicit guidelines indicating the structure-to-inference mapping for each type of adjective. But, contrary to previous studies, 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 naïve 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 naïve judgments. We will use these differential measurements in judgment (trained linguist vs. naïve annotator) to classify the implicatures according to two dimensions: how strong an inference is, given that it arises; how stable the inference is to changes in linguistic context; and which contextual factors can contribute to weakening or eliminating the inference. On the basis of this study, we will develop an improved gold standard and test it again with naïve 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 more complex lexical relations between different adjectives as well as pragmatic factors contributing to the interpretation of lexical items in richer linguistic contexts.

Adjectives can be divided into different classes, depending on what dimensions of analysis are being used. Classic semantic field analysis (cf. [16, 36, 46]) 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.

As intuitive as these classes might be for organizing aspects of the lexis of a language, they fail to provide a coherent guide to the inferential patterns associated with adjectival modification. An alternative

[1] It should be noted that [46] , however, also discuss inferential patterns for distinct classes.

approach is to adopt a conceptually conservative but more formally descriptive and operational distinction, one which groups adjectives into inferential classes. [1] and [2], following [30] and [31], 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 ([1]). 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 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 ([11]) asks automatic systems to evaluate the truth or falsity of a statement (the Hypothesis, H) given a prior statement (the Text, T). A 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.

  Apart from recognizing that New Jersey is the Garden State (information available in WordNet, see [10],[9],[8]) and the relation between weather and temperature, knowing that *arctic* unilaterally entails *cool* would allow a more confident evaluation of the Hypothesis. If Text and Hypothesis were switched, the symmetric synonymy relation between the nouns would not facilitate a correct evaluation of H, whereas the downward entailing intensity relation might lead a system to evaluate a Hypothesis containing *arctic* to be false if the Text referred to *cool*. An RTE system with knowledge of intensity relations among its adjectives is thus potentially more powerful.

- 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 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.

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

Hence, this inference would be false. Now consider the pair below:

(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.

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; for verbs, lists of special inference patterns have been constructed starting from the work of [34] and [32] by [42], [48, 49] and [35]. 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 [53] 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 of current computational linguistics is based 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 machine learning algorithms that are being developed to process natural language texts.

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* [45], as illustrated in Figure (1).

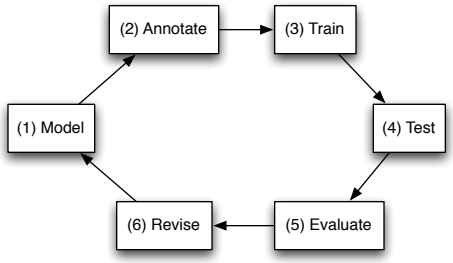

Figure 1: The MATTER Methodology

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.

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 annotated by trained annotators are, for practical reasons rather limited and rarely exhibit all the combinations of possibly relevant context factors. While sophisticated statistical techniques [refs by somebody who understands this better than i do] can fill in some gaps, experimental methods are necessary to supplement this shortcoming. With experiments we can achieve two goals: 1. enrich the spectrum of interpretations that the expert annotators have given with variants of English that will not be represented by the language use of the necessarily very limited number of culturally rather similar annotators 2. allow a systematic testing of the contextually factors that the expert annotators have isolated. it is well known that isolating the factors that contribute to the perception of an inference is extremely difficult. Testing contrastive contexts with a great 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 in FactBank, it is hypothesizing that 'NP be lucky to VP' has the meaning 'it is highly unlikely that NP VP'. We soon realized that this was too general and that this meaning is mainly found in the future. But even there more subtle factors need to isolated. 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.

Three key elements play a role in the proposed work:

(8) a. The interpretation, I, will focus on specific *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 constructed on the basis of the features isolated by the expert annotators but done by naive native speakers.

c. Iterative refinement and enrichment of the model is accomplished by comparing naive and expert judgments (interpretations of SIMs for adjectives), and then revising the model appropriately. This can be referred to as *context-based adjudication*, revealing unseen contextual features that can then be added as model-based primitives.
d. We reach a final model when a new set of naturally occurring data receives the same annotation by the experts and the naive users (modulo the dialect variation that has been detected.)

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 naive annotations create corresponding gold standards, from which we perform context-based adjudication.

2.2 Scalar adjectives

Linguistic fluency entails the ability to reconstruct implicit information that is not overtly expressed in the utterance. This is true in particular for information accessing scales referring to gradable expressions. [26] shows how speakers generate and interpret scalar implicatures in a number of specific linguistic contexts characterizing everyday, natural conversation. [27] establishes some intuitive scales among adjectives and notes that the affirmation of a weak(er) term implies the inappropriateness of a stronger term. For example, affirming *Mary is gorgeous* implies that *Mary is pretty*, an implication understood by the listener.

[5] distinguishes dimensional adjectives like *cold, long, high, large* from evaluative adjectives such as *lazy, diligent, beautiful, ugly, comfortable, gentle, tired*. Like dimensional adjectives, these are mostly gradable but a given scale of evaluative adjectives tends to be more richly lexicalized, perhaps reflecting a need to specify qualities of everyday objects as well as a human tendency to judge people's personal, physical and intellectual traits. Scales for such attributes include a good number of adjectives that each express a different degree of intensity of the shared attribute.

For a few points on a given scale, the adjectives' relative intensity is easy to judge intuitively. Speakers would probably agree that *excellent* entails *good*, and not vice versa. But where exactly do adjectives like *superb, stellar, outstanding, terrific, wonderful* fall on the scale? A related question is, does each adjective on a given scale express a different degree of a shared attribute or are some roughly synonymous for a given speaker or across speaker groups?

In prior work, the direction (positive or negative) of orientation has been widely studied, but the degree has not. [25] extract adjectives with "positive" and "negative" orientation from the Wall Street Journal corpus. They show that conjunctions such as *and* link predominantly adjectives of the same orientation, whereas *but* links mostly adjective pairs with opposed orientation. Relying on the observation that the most frequent member of an antonymic pair has positive orientation, they assign orientation labels to the adjective clusters they derived. [54] derive orientation through association (measured with LSA and PMI), based on the assumption that words with a given orientation co-occur with other words of the same orientation. Other related work such as [7] measured the relative strength of pairs only rather than that of all scale mates. [37] learn subtle sentiment analysis in an unsupervised manner using a vector space model.

We propose an empirical large-scale and in-depth investigation of gradability among adjectives in different semantic domains. We cover a representative sample of twelve different semantic classes ([5], [22], [15]) comprising 100 dimensional and evaluative adjectives. For each scale we manually identify the target lexemes using WordNet's adjective clusters and "similarity" relation as a point of departure. We exclude adjectives with low frequency (based on the BNC), as we do not expect to find enough corpus data nor reliable human judgments to allow us to place them on a given point on the scale. We build on, and extend, the work of [51], who developed a method for constructing full scales.

2.3 Inferences of Predicative adjectives with clausal complements

Another class of adjectives that is a rich source of inferences are predicative adjectives with clausal arguments (*that* S, *to* VP or *ing* complements). With these adjectives the inferences are about the (likelihood of the) occurrence or non-occurrence of the events or states described in the embedded clauses. There are hundreds of such adjectives but, as we will discuss, their inferential classification is not always easy.

These inferences will not depend on the adjective on its own but need to take at the very least the syntactic frame into account. 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 an opinion to the writer. 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 that imply that the author is committed to the factuality of the state of affairs described in the complement even if the matrix clause is negated or questioned. Take for example the following sentence.

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

From this sentence, the reader infers that the author believes the following proposition to be true: "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. But not all adjectives that fit into the [It is ADJ that S] pattern behave the same way, nor is this the only pattern that predicative adjectives with complements allow. With *annoying*, for example, neither negation nor questioning changes the veridicity of the *that* clause, as illustrated in (10).

(10) 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?

Both these sentences still carry the inference that *people post stuff that no one cares about on blogs.* This is not the case, however, with *certain*:

(11) a. It is certain that people post stuff that no one cares about on blogs.
b. Is it certain that people post stuff that no one cares about on blogs?

In (11b) it is the that-complement that is questioned, whereas with *annoying* it was the evaluation of the that complement as annoying that was the focus of the question. It would be of use to know which adjectives behave like *annoying* and which ones behave like *certain*.

Different subcategorization frames can also carry different implications. Consider the following two sentences.

(12) John is sure that Bill left.

(13) John is sure to have left.

Whereas (12) ascribes the belief that "Bill left" to John (the protagonist) and leaves open what the writer thinks, (13) ascribes an opinion to the writer.

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 we can distinguish two broad types of inference patterns:

(14) a. *factive* adjectives as illustrated above in (10). They imply that the author is committed to the factuality of the state of affairs described in their complement even if the matrix clause is negated or questioned.

Table 1: implicative and factive patterns

Implicative	example	illustration
two way	manage to	If Kim manage to do it, Kim did it; If Kim didn't manage to do it, Kim didn't do it.
	forget to	If Kim forgot to do it, Kim didn't do it; If Kim didn't forget to do it, Kim did it.
one way	force to	If Kim forced Sandy to do it Sandy did it.
	refuse to	If Kim refused to do it, Kim didn't do it.
	attempt to	If Kim didn't attempt to do it, Kim didn't do it.
	hesitate to	If Kim didn't hesitate to do it, Kim did it.
Presuppositional	example	notation
factives	forget that	If Kim forgot that the stove was on, the stove was on.
counterfactives	pretend that	If Kim pretended that the stove was on, the stove was not on.
Neutral	want to	no implication

b. *certainty* adjectives that directly assert the degree of certainty that the speaker, or a protagonist, ascribe to the complement, as illustrated in (11). When these adjectives express absolute certainty of absolute denial of the truth of the embedded clause, they give rise to logical entailments; they are *implicative*. Implicatives come in different varieties, as illustrated in table 1. When they do not express absolute certainty, they are part of the means an author uses 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 [48] 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).

A supplementary factor that plays a role in certain cases, as illustrated above (12 and 13), is whether the judgment is that of the writer or is attributed by the writer to a protagonist in the sentence or discourse, e.g., the subject of the sentence with a predicate of saying.

The most common type of inferential pattern for this class is the factive one (illustrated above in 10). It is also the most studied. [44] is to our knowledge the most extensive study of adjectives taking *that* and *to* clauses. [59] and [4] are less extensive studies with claims about factivity. Implicative patterns ([32]) and degree-of-certainty adjectives are only mentioned in passing in the literature. [40] looks at the syntax of 51 frequent adjectives taking *that* clauses in the BNC but without any attention to the semantics.[55] report on a corpus study of deontic-evaluative adjectives concentrating on *important, essential, crucial* and *appropriate, proper and fitting*. They attempt to distinguish between propositional and mandative uses (the former describes a state of affairs, whereas the latter intends to bring a state of affairs in existence or to prevent it from coming into existence) and show that the correlation between the mandative use and the subjunctive is not absolute. The syntax of *eager* and *easy* adjectives has been well-studied. But not much attention has been given to their semantics. We ignore the literature on these classes as it is irrelevant to our purpose. We are not aware of any computational work that focuses on any of these adjective classes.

The nature of the available data calls for several comments for the motivation of our research plan.

(a) The linguistic classification encodes veridicity or, if interpreted from the reader's point of view, reader inferred veridicity, but the judgments are made by a narrow set of linguists. There are no data reflecting the judgments of naive native readers. Specifically with respect to degree-of-certainty classes, linguists assume that the 'literal' meaning of the adjective reflects the degree of certainty. This needs to be tested.

(b) Linguists choose their examples so as to control for variations in e.g. tense or the presence of absence of the syntactically optional arguments, that might obscure the judgements they are interested in. The studies above are based on classic linguistic tests where the test sentences are typically in the past and the relevant arguments have all specific referents. But our preliminary experiments show that there is a big difference in the way people interpret a sentence such as *It was annoying of Bill to leave early* and a sentence like *It is annoying to leave early*. There is no explicit study of the factors that might override the factivity interpretation of a so-called factive adjective. It is important to understand these factors

(genericity in the case illustrated[2]) if one wants to predict the way a sentence will be interpreted in a running context. We will first check the linguistic judgments with their restrictions by submitting them to naive native speakers and then broaden the investigation by submitting variants that do not respect these constraints.

(c) Language variation. For some of the adjectives that are classified as factive, one finds on the web examples that clearly imply a non-factive interpretation. For instance, we find instance such as *This is my first trip to Italy, so I was not brave to venture out alone*, where the implication is clearly *I did not go out alone*. The adjective is not used as a factive but rather as a two-way implicative. A preliminary study indicates that these readings are accepted as unexceptional by a sizable minority (20%) of native American speakers. In some cases we can even distinguish two rather different meanings for a same construction. This is the case for *lucky* in the future tense, as discussed in [33].

We will ask naive native speakers about their judgments, both testing for their passive acceptance of the various patterns and for their willingness to produce such patterns.

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:

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

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

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

However, contrary to what is claimed in [1], 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, (17a) below, there are many contexts where the epistemic scope is reduced to a modification or additional attribution of the nominal head, as shown in (17b).

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

In Example (17a), 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 (17b), 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.

(18) 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. $[A_{int}\ A_2\ N] \models N$

[2]For treatment that dovetails with our preliminary findings, see [38]

[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.

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

(19) 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 (19) involves an additional aspect of the NP, beyond the unassailable characteristics of the entailed head. That is, the object is clearly a vase (in (19a)) and demonstrably a sonata (in (18b)). 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 (18a) as the "Wide-scope reading", and the inferences in (18b-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.

(20) 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 (21), where *alleged* is modifying an event nominal.

(21) 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 (21a-b) follow from the template in (18a). For sentence (21c), 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.

(22) Given the construction $[A_{int}\ N]$, where N is an event nominal, with certain feature, then:
a. $[A_{int}\ N] \not\models N$
$\models N'$ 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 (23b).

(23) 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 (23a) behaves according to the canonical template, while (23b) 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 (18), where the modification (argument specification, in fact) is postnominal.

(24) Given the construction $[A_{int}\ N\ XP_{mod}]$, where XP_{mod} is a modification or argument, then:
a. $[A_{int}\ N\ XP_{mod}] \not\models N\ XP_{mod}$
c. $[A_{int}\ N\ XP_{mod}] \models N$

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:

(25) Structure-to-Inference Mappings:
a. Wide-scope reading: $[A_{int}\ N] \not\models N$
b. Narrow-scope reading 1: $[A_{int}\ A_2\ N] \not\models A_2, \models N$
c. Narrow-scope reading 2: $[A_{int}\ N\ XP_{mod}] \models N$
d. Hypernym reading: $[A_{int}\ N] \models N'$ 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 naive 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 [52] and [41]. 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

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

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.

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 [51].

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 [50], 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

[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.

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 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 [51] and [39] 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.

(26) (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 naive 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 develop an initial categorization of the 200 adjectives that most often appear in clausal complements on the basis of the inferential signature(s) most suited to them. To do this we will follow the same method as proposed in the previous section; we explore the web with regular expressions, looking for patterns such as "It BE ADJ to". We extract 2,000 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. We will balance this corpus so that each adjective occurs at least 10 times. Linguist annotators will classify these snippets according to the initial categorization.

We take 1,000 of these snippets to construct the examples that fit linguistic test patterns. For that, the context following the pattern itself will be changed from the naturally occurring one. For instance, it is easier to judge the factivity of an adjective when it is negated and followed by a negation of the embedded clause. If in "It wasn't silly for John to make a trip around the world but he didn't go," the adjective is taken to be factive, the result is an incoherent discourse.

We will submit these stimuli to MTurkers with set of instructions similar to the following:

- In each task, you will be shown a statement and you will be given possible interpretations of what the author seems to believe assuming that she is truthful. Select the one that you think represents the author's belief based solely on the statement without making use of any information you might have independently. If the statement does not does not make sense, or if you for some other reason cannot decide, choose the "Cannot decide" option.

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 "don't know" answer, in the other, they will chose on the 7-point scale, developed in [48, 47] and validated in [12].

For the patterns that are not recognized by linguists but are found on the Web, we will add follow-up questions, asking: "Could you imagine yourself or a native English speaker saying the Test Sentence? If not, how would you change the Test Sentence?"

On the basis of these judgments we will be able to calculate whether the different adjective classes we have constructed on the basis of our linguistic criteria correspond to classes that occur in the real world. This step will allow us to ascertain whether there are substantial differences in the judgments about veridicity as conceived by linguists and those made by naive native readers.

The same set of 1,000 snippets will be used without changes in the contexts and submitted to the subjects under the similar conditions as sketched above. This second set will help us isolate the factors that influence judgments in real life.

On the basis of these crowdsourced judgments, we will be able to construct a model of the structural factors that influence readers judgments and revise the initial categorization. We will then develop annotations guidelines for trained annotators. These annotations will capture the factors that play a role in the categorization. The annotators will annotate 1,000 new, naturally occurring snippets. These 1,000 snippets will be annotated by the MTurkers according to the protocol sketched above using the scale (the 7-point or the 3-point one) which gave the most information. We will use these data to estimate how well the factors we have isolated capture the MTurkers data by building statical 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 (27) and (28) below. In these examples, the inference in (27) is legitimate, while that in (28) is false.

(27) 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.

(28) 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 [50], 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 [50]. 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 will use the same 1,000 snippets and submit them to the Brandeis TARSQI system to mark up the events. We consider all the events as being factual. We develop a set of rules based on our classifications and evaluate how well the resulting system identifies the events using the scale used in the last MTurker experiment. MORE

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 and Princeton as well as the consultants from Stanford will maintain regular contact via weekly skype conferences. One annual meeting is planned, alternating between Princeton and Brandeis, as well as a meeting at a domestic or international conference or workshop focusing on topics of shared interest.

3.6 Milestones and Deliverables

Year One of the project is dedicated to:

Q1	Identify target adjectives; collect relevant syntactic patterns.
Q2	Perform corpus mining, derive initial semantic classifications and structure-to-inference mappings.
Q3	Preliminary annotation schema. Pilot MTurking experiments.
Q4	Evaluate corpus data; linguists annotate first sets of snippets. Update classifications and mappings. Prepare articles for publication.

Year Two is dedicated to:

Q1	Begin MTurking work; first sets of stimuli for MTurkers.
Q2	Start creation of gold standard; run first experiments with MTurkers.
Q3	Analyze results of MTurker data with gold standard.
Q4	Continue MTurking work. Update classifications and mappings. Prepare articles for publication.

Year Three is focused on:

Q1	Analyze results of second MTurker experiments.
Q2	Compare with gold standard; revise annotation schema.
Q3	Construct WordNet scales; run next experiment and build revised model.
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 University.; (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:** [29]; [28]; [**?**].

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:** [58]; [56]; [57].

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:** [50].

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:** [21],[18],[17],[6],[43].

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:** [19]; [20] [3] [14].

References

[1] 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.

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

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

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

[5] M. Bierwisch. The semantics of gradation. *Dimensional adjectives*, 71:261, 1989.

[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] T. Chklovski and P. Pantel. Verbocean: Mining the web for fine-grained semantic verb relations. In *Proceedings of EMNLP*, volume 4, pages 33–40, 2004.

[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. *Where have all the adjectives gone?: and other essays in semantics and syntax*, volume 107. De Gruyter Mouton, 1982.

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

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

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

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

[20] 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.

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

[22] W. Frawley. *Linguistic semantics.* Lawrence Erlbaum Associates, Inc, 1992.

[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] V. Hatzivassiloglou and K.R. McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 174–181. Association for Computational Linguistics, 1997.

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

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

[28] 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.

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

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

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

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

[33] 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.

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

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

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

[37] A.L. Maas, A.Y. Ng, and C. Potts. Multi-dimensional sentiment analysis with learned representations.

[38] Thomas Mathew and Graham Katz. Supervised categorization of habitual and episodic sentences. In *Sixth Midwest Computational Linguistics Colloquium*. Indiana University, Bloomington, Indiana, 2009.

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

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

[41] 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.

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

[43] 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.

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

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

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

[47] 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.

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

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

[50] 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.

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

[52] 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.

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

[54] P. Turney and M.L. Littman. Measuring praise and criticism: Inference of semantic orientation from association. 2003.

[55] 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.

[56] 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.

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

[58] 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.

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