Project Description

IIS-RI MEDIUM: Collaborative Research
Discovering Inference Patterns for Adjectives

1 Introduction

Linguistic communication relies on the ability of language users to identify the often implicit semantic dependencies in an utterance. It is on the basis of interpreting the lexical and structural patterns encoding these implicit dependencies that we make "semantic inferences". An understanding of how speakers identify and exploit such lexical and structural inference patterns has the potential to significantly enrich models in theoretical linguistics. At the same time, it furthers a major goal of current NLP research: allowing natural language understanding systems to recover more of the rich semantic information encoded in the text, e.g.: Did the event mentioned in the text actually occur? When did it happen? What sentiment was conveyed regarding the event? What qualities are being attributed to the participants in the event?

Identifying such inferences in a computational setting requires the reconstruction of lexical, syntactic, and contextual information in texts, for the same reason that effective communication requires language users to recover information not expressed explicitly. For this reason, existing systems typically rely on training or development corpora annotated manually or automatically, using information derived from lexical classes or from syntax-semantics correspondences. To date, most research in the field has concentrated on identifying events and their participants; this has led to a focus on identifying verbs and verb senses, and nouns within named entities. While many of these semantic relations follow directly from the syntax accompanying a verb and its associated semantic roles, there are many other crucial semantic relations that cannot be read off the syntactic structure as readily: for example, comparative information about graded properties of the individuals and events described, and information about the stance of the author of a text towards the events described. Such information that is more *evaluative* in nature is frequently conveyed by the use of adjectives. For instance, in scalar adjectives, the entity characterized is evaluated against others of the same class; in intensional adjectives, the epistemic uncertainty determines how the description of an entity or event is evaluated; and veridicity-related adjectives (evaluative, emotive and epistemic predicates) express both the author's stance with respect to the truth of the event described in the complement and, in the case of evaluative and emotive ones, an evaluation of the character of the protagonist of that event.

Adjectives make up a significant portion of typical English texts, usually between 5%, and 10%. They influence the interpretation of texts in ways that make their study particularly interesting, now that information extraction is going beyond simple entity and event identification, in order to take into account both modal and subjective factors in the text. For instance, Wiebe [43] notes that the presence of an adjective is a good indicator of the subjective stance of an utterance. The most common treatment of adjectives has been to assume that they are, by default, INTERSECTIVE – that is, an object described by [A N] is both "an A" and "an N". Hence, an "American woman" is both an American and a woman; a "tall building" is both tall and a building, and so on. It is usually assumed that such uses are objective, but in fact most adjectives are rarely interpreted intersectively [19]. Nevertheless, most computational treatments within the field have ignored the non-intersective behavior of most adjectives. When exceptions are noted, they are usually handled by simply blocking the inference: e.g., an "alleged thief" cannot be inferred to be a thief, or to be alleged (whatever this would mean).

As we demonstrate here, however, the inferential reach of adjectives extends far beyond the treatment of the adjective on its own, even if we extend the classification to account for subsective, privative, and other non-subsective adjectives.[1] Each of the three classes of adjectives that we investigate—veridicity-related, intensional, and scalar—will typically influence the interpretation of the surrounding linguistic context, and generate robust and predictable inferences. They do this by taking semantic scope over significant portions of text in a context-sensitive way, or by generating entailments and implicatures involving comparisons

[1] In the construction [A N], an adjective A is: SUBSECTIVE if the object described is A relative to N, but not necessarily in general; PRIVATIVE if the object is not an N, by virtue of being A; and NON-SUBSECTIVE if the object may or may be N.

to non-textual material [40]. Even two of the examples given above—*tall* and *alleged*—are associated with inferences of this type, and are not well-treated on the standard model. Our preliminary investigations suggest that this is a common profile, and that simple intersective adjectives are rare in natural corpora.

In the case of nouns, the WordNet hierarchies have proved useful in numerous studies (e.g., [59]); for verbs, there is VerbNet and FrameNet as well as lists of special inference patterns have been constructed starting from the work of [38, 33] by [48, 55, 56, 41]. Information about the inferential properties of adjectives is, however, much less easy to come by. For the categories of adjectives that we are interested in, the existing resources have severe shortcomings or are non-existent, in part because the contribution of adjectives tends to be more subtle and more dependent on the rest of the linguistic context. A few resources have been developed in the area of sentiment analysis (e.g. [23]) but even in those, as in the few proposals to annotate adjectives (cf.[54]), there is no attempt to model adjective senses, and no consistent methodology has been developed for characterizing and encoding their inferences or their relations. While WordNet does contain a certain amount of adjective information, the 'dumbbell' model is fairly impoverished for these purposes, as discussed below. As a consequence most of the shared tasks in the SemEval challenges in the field have had to largely ignore the semantic role and inferential impact contributed by adjectives in the language.

There is thus a serious gap between current practice and available resources, on the one hand, and the needs of robust inference-enabled systems on the other. The research proposed here addresses this issue with an in-depth investigation of the lexical semantics of three well-attested adjective classes, using a new and coherent methodology for identifying and annotating the semantic inferences associated with adjectives in texts. The methodology proposed goes beyond what is the practice in the NLP annotation community by integrating corpus research and experimental data and by relying on the judgments of non-expert native speakers rather than on those of students in linguistics.

To capture the manner in which adjectives impact the interpretation of text in what appear to be non-compositional ways, it is necessary to identify those contextual factors at work when making such inferences. In this research, we propose both a methodology that identifies these contextual interpretive factors, and an encoding scheme that associates an adjective with specific inferences and implicatures, as interpreted from a context. We call these *Context-dependent Inferences (*CDIs*)*.

The availability of digital corpora (the biggest one being the Web itself) and crowd-sourcing techniques to elicit the judgments of a larger and more diverse group of native speakers allow us to go beyond the narrow evidential base of traditional lexical semantic studies. Our methodology aims to discover systematic linguistic inferences of specific lexical classes in natural language through an integration of lexical resources and corpus annotation and to construct inferential models for these classes of lexical items. It differs from the current annotation methodology by

- using many linguistically untrained speakers instead of a few experts;
- treating inference as graded rather than categorical;
- taking into consideration a rich set of semantic interactions of the textual elements involved instead of artificially constraining them.

In this proposal, we concentrate on three diverse semantic classes of adjectives, with the aim of testing the validity of this methodology and articulating precise conceptual models for each lexical class. The adjective types studied are: (i) veridicity-related adjectives, showing varying implications of veridicity over a clausal complement, e.g., *rude, annoying, likely*, etc.; (ii) intensional adjectives, introducing implications of modal subordination, e.g., *alleged, supposed, so-called*; and (iii) dimensional and evaluative adjectives with scalar values and associated scalar implicatures, e.g., *pretty, beautiful, large, huge*.

Together these classes cover attributive as well as predicative uses of adjectives and the major non-intersective classes that have rich and unexpected inferential properties[2]. Quantitatively, these classes contain hundreds of the most frequently used adjectives. A list of the 2,000 most used words in the BNC contains 274 adjectives. Among these, 13 are epistemic adjectives and 94 can have either emotive or evaluative propositional complements; 81 are scalar adjectives.

[2]Classic semantic field analysis (cf. [12, 42, 53]) categorizes attributes denoted by adjectives according to a thematic organization lexically encoded in the language. Only [53] discusses inferential patterns for distinct classes.

The work will start out with developing initial inferential models for our classes of adjectives based on existing annotation efforts, focusing in particular on FactBank [56]. We will then test these models against large corpus data and analyze sources of deviation from the models. Next, we will use this information to design test sets to elicit inferential judgments from non-expert native speakers, in particular Mechanical Turk workers (MTurkers). Our preliminary studies have lead us to expect that there will be considerable deviation from the linguistically-derived models. We hypothesize that an important part of this variance is caused by textual factors that are abstracted away in linguistic studies, but are important to explain the non-expert judgments. On the basis of this study, we will develop a 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 end product will be, for each class of adjectives:

- annotation guidelines for crowd source annotation;
- annotated corpora in the RTE style (T-H pairs);
- conceptual models that capture the behavior of the adjective classes studied, including the CDIs for each adjective;
- statistical models that evaluate how well the hypothesized factors explain the data of the experiments;

This research is significant in two major respects. First, it develops inference models for adjectives that are richer than the ones currently envisioned. Second, it develops a new method for textual annotation, combining crowdsourcing and corpus exploitation and producing multi-level annotations representing graded strength of inference.

2 Theoretical Background

2.1 Veridicity inferences of adjectives with clausal complements

One common and inferentially rich class of adjectives are predicative adjectives with clausal arguments (*that* S, *to* VP, or *ing* complements). These adjectives are typically used to communicate an agent's epistemic stance on the likelihood that an eventuality will occur, and some convey in addition an emotive or evaluative attitude of the agent. 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 author thinks, whereas *John is sure to have left* ascribes a belief to the author. This minimal pair shows that the adjective is not sufficient on its own to determine the type of inference. Nor does the syntactic frame determine the type of inference alone: not all adjectives that fit into this frame behave in the same way, as discussed below.

Three broad classes have been distinguished based on their epistemic inference patterns.

1. Certainty adjectives. These adjectives directly encode an agent's degree of certainty in the complement.

(1) 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?

(1b) has the opposite inference from that of (1a), and in (1c) it is the *that*-complement itself that is questioned. CDIs for these adjectives, therefore, would need to take into account the syntactic frame and would also minimally need to distinguish between positive and negative polarity contexts and declaratives and interrogatives.

Some of the adjectives in this class express absolute certainty in or absolute denial of the truth of the embedded clause, and hence give rise to logical entailments; they are *implicative* ([33]). Others, such as

possible, probable, impossible, improbable, constitute a means for the author to indicate the probability that (s)he attaches to the factuality of the state of affairs expressed in the embedded clause. In this study, we follow [55] and approximate this probability by the following scale: CT+ (certain), PR+ (probable), PS+(possible), U (none), PR- (improbable), PR- (impossible) and CT- (certainly not).

Apart from the adjectives that express an epistemic stance directly, there are adjectives expressing other modalities that can lead to epistemic inferences. 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.

2. 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, (2).

(2) It is annoying that people post stuff that no one cares about in blogs.

From (2), the reader infers that the author presents as true the proposition that people post stuff that nobody cares about in blogs. This inference is derived 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 (3). The focus of the question in (3b) is the evaluation of the *that*-complement as annoying or not.

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

[50] proposed two subclasses: emotive (e.g. *sad*) and evaluative (e.g. *stupid*) adjectives. But the empirical picture is more complicated: the status of the factuality inference varies among speakers ([36] and below).

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. Pilot studies we have conducted on various subclasses have revealed that their inferential behavior is dependent on fine-grained structural and contextual factors.

In addition to the extensive study of factive adjectives in [50], there are the more limited studies in [65] and [3]. Implicative ([33]) and degree-of-certainty adjectives are only mentioned in passing in the literature. [45] looks at the syntax of 51 frequent adjectives taking *that* clauses in the BNC but without any attention to the semantics. [63] report on a corpus study of deontic-evaluative adjectives concentrating on *important, essential, crucial*, *appropriate, proper*, and *fitting*.

For this project we start from CDIs based on the syntactic characterization of the adjectives and extend them further with the contextual factors our corpus and experimental studies reveal. For instance, there are more than a thousand adjectives that can occur in the [it be ADJ (of NP) to VP] frame. Several hundreds have been listed as factive in [50]. But our investigations show that 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* rarely. The factive interpretation reliably arises only in the past tense with a specific *of NP*. The CDI needs at the very least to specify the tense as in (4):

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

Moreover, adjectives without epistemic entailments may still be interpretated as assigning high probability to the truth of the complement. For instance, *It was essential for researchers to collect accurate information* is judged by MTurk workers to be factual for more than 50% of them and probable for another 35%. This suggests that a simple CDI assigning a 'don't know' inference to those is insufficient.

With personal constructions of the type [NP be ADJ to VP], many hundreds of adjectives are classified as giving rise to a factive interpretation. But a corpus search in enTenTen[3] shows in about half of the cases

[3] http://www.sketchengine.co.uk/documentation/wiki/Corpora/enTenTen

the interpretation is implicative. The preferred interpretation of *Kim wasn't stupid to send money* is that no money was sent, while that of *Kim wasn't stupid to save money* is the expected factive interpretation. We hypothesized that the crucial determinant is whether there is a *harmonic* (to waste money is stupid) or a *disharmonic* (to save money is not stupid) relationship between the evaluative attitude expressed by the adjective and a general evaluative assessment of the activity described by the complement clause. This was corroborated by an experimental study ([36]). An initial CDI based on the accepted wisdom, such as

(5) $[NP_{pers}$ was ADJ_{eval} to $VP\ ] \models NP_{pers}\ VP$

will certainly need to be revised and split based on the polarity of the matrix clause and the 'harmonicity' of the ADJ-VP pair.

Our preliminary investigation of so-called factive constructions with *that* complements suggests that they indeed 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 ([69]). The initial CDI corresponding to an impersonal syntactic frame is given in (6).

(6) [it be ADJ that $S\ ] \models S$

So far, we have identified 43 CDIs of the form given above that need further investigation.

2.2 Identifying Epistemic Uncertainty: Intensional Adjectives

The second adjective class we examine for their inferential properties is the set of non-subsective adjectives. This class is further divided into two classes: privatives such as *fake* or *pretend*, and intensional adjectives such as *alleged*. Privative adjectives can be analyzed as follows:

(7) $\|A\ N\| \cap \|N\| = \varnothing$

Intensional adjectives introduce epistemic uncertainty for the elements within their scope; examples include *alleged*, *supposed*, and *presumed*. These adjectives call into question some property of the nouns they modify, and traditionally it has been claimed that no informative inference can be associated with this construction [31]:

(8) 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[4], several distinct patterns of inference emerge. While the typical resulting composition entails uncertainty of whether the nominal head belongs to the mentioned sortal, (9a) below, there are many contexts where the epistemic scope is reduced to a modification or additional attribution of the nominal head, as shown in (9b).

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

In Example (9a), 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 (9b), 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 by the CDIs below, where we distinguish between the relations of entailment ($\models$) and *implication* ($\rightsquigarrow$).

[4]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.

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

The CDI in (10c) is quite common, as illustrated by further examples in (11).

(11) 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 (11) involves an additional aspect of the NP, beyond the characteristics of the head noun. That is, the object is clearly a vase (in (11a)) and a sonata (in (10b)). Such evidence, however, will not always be available within the sentence, but will need to be derived from the context. We will refer to the canonical CDI in (10a) as the "Wide-scope reading", and the inferences in (10b-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 occurring in the corpus are shown below, along with apparent scoping behavior.

(12) 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 (13), where *alleged* is modifying an event nominal.

(13) 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 (13a-b) follow from the template in (10a). For sentence (13c), however, there is an implicature that there was, in fact, a killing, although it is uncertain whether it was a murder. This requires the CDI below, where the hypernym of the event nominal is implicated from the context.

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

We refer to this inference rule as the "Hypernym reading". Similarly, the scope of an intensional adjective modifying an undergoer can be lowered to a modification of the event description, as in (15b).

(15) 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 (15a) behaves according to the canonical template, while (15b) 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 (10), where the modification (argument specification, in fact) is postnominal.

(16) 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. $[A_{int}\ N\ XP_{mod}] \rightsquigarrow N$

Summarizing the semantic behavior for this class, we have identified at least three distinct CDIs associated with intensional adjectives. They are:

(17) Context-dependent Inferences (CDIs):
a. Wide-scope reading: $[A_{int}\ N] \not\models N$
b. Narrow-scope reading 1: $[A_{int}\ A_2\ N] \not\models A_2, \rightsquigarrow N$
c. Narrow-scope reading 2: $[A_{int}\ N\ XP_{mod}] \rightsquigarrow N$
d. Hypernym reading: $[A_{int}\ N] \rightsquigarrow N'$ where $N \subseteq N'$

2.3 Scalar adjectives

As discussed in the introduction, language understanding requires active reconstruction of information which is merely implicit in an utterance, or which can be reconstructed on the basis of an utterance together with its linguistic, social, and worldly context [21, 26, 5, 22, 51]. This is true in particular for scalar adjectives, which have highly context-sensitive meanings (e.g., compare the inferred size of a *big baby*, *big tree*, or *big planet*). Many scalar adjectives are organized into entailment scales, where items high on the scale asymmetrically entail lower items [27, 28]. The use of items lower on the scale, in turn, can lead to a defeasible inference that the sentence would be false if the lower item were replaced by the higher [21, 27, 39].

(18) a. HEAT_SCALE: warm < hot < scorching
b. *Dallas is scorching* **entails** *Dallas is hot*, **which entails** *Dallas is warm*
c. *Dallas is warm* **implicates** *Dallas is not hot* and *Dallas is not scorching*
d. *Dallas is hot* **implicates** *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 [25, 18, 20]. 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:

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

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

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

(19) illustrates a standard scalar implicature, where the choice to use "hot" when "scorching" is relevant leads to an inference that Dallas is not scorching. In (20), this inference is much less robust, presumably because it is not clear whether "scorching" is a relevant alternative. In (21) no implicature regarding "scorching" arises, but (perhaps surprisingly) there is a robust inference that Dallas is not beautiful. These examples illustrate the importance of taking context into account when drawing inferences from scalar adjectives.

For each scalar adjective, we define the appropriate context-dependent inferences (CDIs) for a specific usage. For example, using the generic scalar CDI in (22), we can recover the entailments given in (18b) above (where '$\models$' stands for "entails" and '$\rightsquigarrow$' for "implicates").

(22) a. [NP be A_i]
b. $\models$ [NP be A_j], given a scale, where $A_j < A_i$
c. warm < hot < scorching

Similarly, many scalar implicatures can be identified as CDIs, as illustrated in (23), corresponding to that given in (19b).

(23) a. $\langle$[be NP A_i?], [NP be A_j]$\rangle$
b. $\rightsquigarrow$ [NP be $\neg A_i$], given a scale, where $A_j < A_i$

A second complication in modeling inferences from scalar adjectives involves their DIMENSIONALITY. While *tall* and *heavy* pick out a single scalar dimension (height and weight, respectively), multidimensional adjectives such as *big* and *huge* are more complex, placing requirements on multiple scalar dimensions (e.g., 3-D size and perhaps weight, in the case of *big*). Emotive adjectives such as *happy* are even more complex, since the dimensions that they rely on are not easy to identify or quantify. These relationships among adjectives thus raise difficult questions for an inferential model of adjective semantics. Since it is possible for something to be *big* without being *tall* (e.g., a *big city*), these expressions do not share all of their dimensions and thus are not in an asymmetric entailment relationship. 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*?

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

1. determine which adjective pairs share scalar dimension and polarity, differing only in strength;[5]
2. determine which adjective pairs differ only in polarity (*cool* vs. *warm*), or display partial overlap of dimensions (*big* vs. *tall*);
3. identify factors influencing the contextual weighting of dimensions for multidimensional adjectives.

2.4 Methodology

Much work in modern computational linguistics relies on the creation of annotated datasets focused on one or on several 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 model, which is a triple, $M = \langle T, R, I \rangle$, consisting of a vocabulary of terms, T, the relations between these terms, R, and their interpretation, I. This is often a partial characterization of quite extensive theoretical research in an area, encoded as specification elements for subsequent annotation. These annotations provide the features that are then used for training and testing classification or labeling algorithms over the dataset. Depending on a system's performance, various aspects of the model or related specification will be revised, retrained, and then retested. For this reason, we can refer to this methodology as the MATTER cycle: *Model-Annotate-Train-Test-Evaluate-Revise* [52], as illustrated in Figure (1).

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

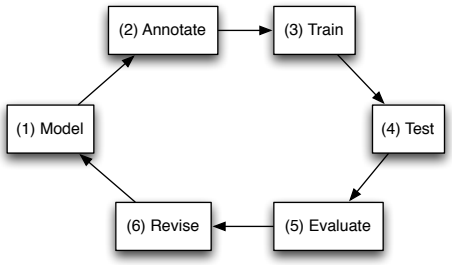

Figure 1: The MATTER Methodology

Isolating the factors that contribute to the perception of an inference is notoriously difficult. Testing contrastive contexts with a large number of native speakers is one way to discover the precise factors at work. For instance, from FactBank we learn that 'NP be lucky to VP' has the meaning 'it is highly unlikely that NP VP'. In previous work project consultants Karttunen and Zaenen [34, 68] observe that this is far too general, and that the "unlikely" meaning is mainly found in future-tense sentences, whereas past tense sentences are overwhelmingly factive. Experimental work has confirmed this while also pointing to further factors that influence the interpretation [35], [32].

[5] This is the logical notion of "polarity" used in formal semantics [37], not the emotive concept from sentiment analysis [67, 66].

In light of this, we propose in the present work a significant enrichment to the standard methodology. Traditionally, annotation is done by two more or less trained annotators and a trained adjudicator. This is sufficient as long as the phenomena to be annotated fall in reasonable clearcut categories; one does not need many judgments to conclude that native speakers of English interpret *table* as a piece of furniture in a sentence such as *Jane put the food on the table.* But, once we get to linguistic phenomena such as the adjective classes studied here, reliance on a few judgments might not lead to reliable conclusions because contextual factors and pragmatic effects are critical. Moreover, the corpora that can be created using linguistically trained annotators are rather limited and rarely exhibit all the combinations of relevant context factors, resulting in lack of data capturing what a rich, human-like understanding of texts should be. This limits the usefulness of many machine learning methods that are widely used in natural language understanding [44, 64, 47]: sophisticated statistical models cannot produce rich understanding without a linguistically informed understanding of what is being modeled. Our project seeks to augment standard annotation practices with both new experimental methods based on crowdsourcing and corpus studies, in order to address this lacuna for the important, understudied category of adjectives.

Crowdsourced annotations tasks, when carefully constructed and analyzed, have been shown in previous work to have reliability comparable to traditional expert annotations in some domains including certain RTE tasks [61, 46]. Adopting this method as part of our approach will help us achieve three related goals. First, it makes possible systematic testing of the 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 reliable but not categorical: for example, the defeasible inference that someone described as *attractive* is probably not *stunning*. Third, systematic use of untrained annotators will make available data on the interpretations of people with backgrounds that go beyond those typically involved in standard annotation efforts. We conjecture they will bring in contextual factors that linguists have been trained to ignore.

We take it for granted that in the adjective classes under study the inferential potential is in general influenced in a non-deterministic way by linguistic and non-linguistic context.

Five key elements play a role in our methodology:

(24) a. The interpretation, I, focuses on determining CDIs, indicating how a given adjective type associated with its environment, which we will call a CONTEXT, contributes to or enables inferences. These CDIs will be formulated by experts. The initial CDIs will by and large be categorical.

b. The CDIs, encoded in the initial interpretation, are checked against large corpus data. This step is heuristic: we analyze large corpora (en-Ten-Ten, CoCa, BNC, the web itself) by means of finite-state templates to find instances of the patterns of our CDIs. The templates will depend on the corpora being syntactically analyzed. The data collected in this way will inform an updated formulation of the CDIs, which are then used to construct experimental material.

c. This experimental material is presented for annotation to Amazon Mechanical Turk workers (AMT). The material will cover the combination of features isolated by experts. Annotation instructions will be adapted for use by linguistically untrained native speakers, borrowing methods from experimental psychology, where appropriate. Given the complexity of the factors, the test items will often be larger than one sentence.

d. The heart of the project is the iterative refinement and enrichment of CDI models importing new contextual factors. This will be done through the loop of hypothesis construction by the experts and experimental evaluation by native annotators.

e. The product of the iterative refinement is a new annotation standard. In contrast to traditionally annotations which give a value that is based on a few judgments, our methods provide graded strength information derived from statistical patterns in large scale annotations. To make this information more easily interpretable while maintaining a reasonably high degree of granularity, we map it on a seven point scale.

3 Project Plan

As described above, we develop Context-dependent Inferences (CDIs) for each of the three adjective classes. Based on corpus research, we adapt and enrich the existing inferential models for all three types of adjectives. We then select an initial set of target adjectives in each class and construct experimental items based on the CDIs. These are submitted to non-expert annotators (e.g. via Mechanical Turk). The results are analyzed, new hypotheses are incorporated in revised CDIs, leading to a new set of experiment items and ultimately to a new annotation scheme.

Corpus data. We will use a variety of corpus resources, including in some cases the Web, for the extraction of patterns identified as inferentially relevant in the initial model and in subsequent corpus investigations. The advantages and disadvantages of using web data vs. smaller, more carefully controlled corpora are well-known. In many cases — especially when dealing with short patterns whose diagnostic usefulness has already been established — the size and diversity of styles and genres on the web, as well as its access to diverse speaker communities, gives us vital information which compensates for the increase in noise and the possibility of multiple counts. In other cases, it will be necessary to use resources such as BNC, COCA, and Gigaword, or to develop methods to use these resources to complement each other. This is especially true in portions of the project which attempt to use semi- and unsupervised methods to identify patterns of interest, where POS-tagged and parsed data is needed. [6] Note further that our iterative methodology will eliminate many potential false hits from web data and other sources alike, since patterns identified as non-diagnostic by human annotators will be removed or down-weighted subsequently.

Experimental data. Several studies in the last years have shown that crowd-sourced annotation tasks can deliver reliable results when carefully constructed and analyzed (cf. [61, 46] and section 2.4 above). In constructing tasks on AMT, we will present test questions to ascertain that the workers are native speakers of English, and then ask them to make judgments about inferences (both potential entailments and implicatures) with varying amounts of linguistic context. Best practices for AMT annotation are not yet firmly established, and we expect that achieving our goals will require us to explore a variety of approaches. Our approach to inferring quantitative patterns of inference as well as inferences associated with specific contexts will rely primarily on this method of data collection, while standard annotations will be used as a tool in building AMT tasks and as a sanity check for the results.

3.1 Veridicity of Adjectival Complement Clauses

Ignoring patterns with *-ing* complements, we currently have 43 patterns. We turn those into CDIs based on existing linguistic claims about their inferential properties and test them by searching the web and other large corpora. We select an initial set of examples of the pattern and submit them for initial judgments to the research team. On the basis of their judgments, we formulate a hypothesis and consequently a new CDI, e.g. contrary to existing linguistic literature, the pattern 'NP be evaluative ADJ to VP' is used implicatively as well as factively and the judgment is influenced by harmonicity (see section 2.1). We then construct experimental data to test that conjecture on the 20 adjectives that are most used with each pattern in the enten-ten corpus. These data will be in the RTE format of TH pairs, in most cases to be judged 'True'/'Don't know'/'False'. We submit these experimentally constructed pairs for judgment to AMT workers, ensuring we get at least 30 judgments per pair. We analyze the data using statistical models to evaluate how well the hypothesized factors account for the experimental results. If necessary, we revise our conjectures and do a new round of experiments. In these revisions we will take into account not only structural and contextual factors but also frequency data in order to more finely characterize the adjective and the patterns. This might lead to further splits in several patterns. These new experiments deliver an annotated corpus, where judgments of the 30 MTurkers are distilled into a single probabilistic value. Based on these results, we build revised conceptual models that use a -3 to +3 point scale indicating for each revised CDI how likely they are to be interpreted as implying the truth of their complement (e.g. for evaluative adjectives with negative polarity +3 indicates a close to 100% 'True' interpretation, whereas -3 would indicate that the complement

[6] Where necessary, we will process the corpora with off-the-shelf part-of-speech taggers and dependency and syntactic analyzers.

is inferred to be false by close to 100% of our subjects). If there is substantial reliability variation among the adjectives, we will also give a second matrix with confidence ratings for each judgment. These results should be amenable to implementation into systems such as Biutee [62].

3.2 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. We have already idenfied 3,000 snippets for intensional adjectives.

The methodology is very similar to the one used for the veridicity adjectives in the previous section. By studying the 3,000 snippets we adapt and expand the 20 patterns we have so far (some of them presented in section 2.2) and create an initial set of CDIs. We will first do this on a set of 1,000 and then extend the analysis to all 3,000 snippets. Using the patterns in the CDIs, we will extract snippets from the en-ten-ten corpus and create TH pairs to test the inferences. That is, we construct examples that fit the identified patterns in the test CDIs, as shown in (25) and (26) below. In these examples, we hypothesize that the inference in (25) is legitimate, while that in (26) is false.

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

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

We submit these stimuli to AMT workers (30 MTurkers per TH pair) and examine the differences in judgments. For those cases that do not accord with the pre-assigned classification, we try to isolate the factors contributing to the divergence. 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. If necessary, we will revise our CDIs and conduct a next round of analysis and AMT experimentation. The end result will be a corpus of TH pairs, where the TEXT part matches the pattern(s) in a CDI and where the HYPOTHESIS is not simply annotated as valid or invalid, but given a probabilistic value.

3.3 Scalar Adjectives

[57, 58] used a methodology similar to Hearst's [24] to demonstrate the usefulness of hand-selected syntactic patterns in identifying co-scalarity and relative intensity of adjective pairs. [57] showed that when adjectives X and Y are "semantically similar" according to WordNet — and so likely to be co-scalar — it is possible to learn which is stronger by examining the frequency of patterns "X, even Y" and "if not Y, at least X". If these and other carefully chosen patterns are frequent, Y is likely to entail X asymmetrically. A recent pilot experiment revealed that this method has high precision but low recall as applied to 40 scalar adjectives in the Google Web1T corpus. We propose to extend it by learning the relevant patterns from dependency-parsed corpora, following Snow et al. [59, 60]. This approach reduces noise inherent in the use of raw counts, and should result in an improvement on WordNet's baseline in both precision and recall. When using the web or other corpora, we will also explore methods for reducing noise by iteratively refining the patterns employed by careful error analysis and weighting based on results from parsed corpora. Building on the results of these investigations, we will design AMT-based annotation tasks which allow us to collect inferential judgments for the 50 most common adjectives in English, a set which includes hundreds of scalar adjectives. Finally, we will iterate by evaluating the success of the patterns learned from text in predicting human judgments.

Given the theoretical background on graded and context-dependent inference, it is necessary to return not just categorical judgments about entailment and implicature, but also graded information about the **strength** of an inference on a multi-level scale (-3 to 3, where a negative number indicates that the truth of the first supports the judgment that the second is false). Strength of inference will be estimated from quantitative patterns in annotators' judgments. With these results in hand, we will revisit the original,

context-independent judgments used to build the model, exploring in what ways the inclusion of richer context modulates entailment and implicature judgments. We will also explore methods for predicting context-dependent judgments from tagged and parsed corpora, considering at all available linguistic features of the context. The product of this work will be similar to the adjective-context ratings of inference strength which will be employed in the veridicity and intensional components of the project, but with an additional dimension since we are examining *pairs* of adjectives in various contexts. For every context and pair of adjectives in the set investigated, the triple (context, adjective1, adjective2) will be annotated with a number from -3 to 3 indicating the strength of the inference from the first to the second, with negative numbers representing likely falsehood. If there is significant variability among adjectives in annotators' reliability, we will also include a second set of ratings giving a confidence rating for each judgment, and use this information to set penalties in the RTE task that we now describe.

3.4 Evaluation

Our evaluations adopt the format of RTE tasks as developed in [10]. This format has been shown to lead to improvements of NLP inference capabilities and will allow for the further use of our results by teams building NLP systems.

For the evaluation of the veridicity inferences, we perform a web search for naturally occurring data exhibiting the factors we have discovered as being relevant for the inference profiles of our CDIs. On the basis of the web texts, we also construct an RTE-like task (using the snippets as TEXT and constructing HYPOTHESES based on our conjectures) for the 20 most used adjectives of each pattern, developing at least 4 pairs per pattern. We associate with these pairs predictions based on our conceptual models encoded in the CDIs. We submit the pairs to AMT workers, ensuring we get at least 30 judgments per TH pair. For each item, we map the results on our -3 to +3 scale and calculate the discrepancy between our predicted values and the actually obtained values. We also map our initial CDIs on our 7 point scale and calculate their discrepancy with our predictions. A measure of our success is the comparative closeness of the final experimental results to our predictions.

The evaluation for intensional adjectives is similar in approach to that above for veridicity. Using the 1,000 snippets of naturally occurring data containing intensional adjectives, we will construct an RTE-like test (with TEXT-HYPOTHESIS pairs), exploiting the inference profiles from the CDIs. We develop at least 4 pairs/pattern for the most frequent 30 adjectives in the class. The TEXT consists of the full snippet, and the HYPOTHESIS is derived from the CDIs associated with the adjective set. As above, we use AMT workers to judge each TH-pair, making sure we collect at least 30 judgments/pair, the output of which is mapped to the -3 to +3 point scale mentioned earlier. The difference is calculated against the initial CDIs, where accuracy is measured as relative proximity of the results to those predicted.

For scalar adjectives, we propose to evaluate the scales constructed in an RTE task using methods for measuring the contribution of specific WordNet relations developed by [6, 7, 8]. We will similarly quantify the contribution of scalar orderings among adjectives in WordNet to the RTE task using a new test set involving adjectives that we have analyzed. To perform the evaluation, will encode new scales in WordNet following the model described in [57], where WordNet's "dumbbells" are augmented with arcs connecting some adjectives on each half of the dumbbells to specific points on the scale. This preserves the original WordNet representation for one central adjective (e.g., *rich*) and a set of "semantically similar" adjectives (*wealthy, comfortable*, etc.) while also indicating their intensity relative to the central adjective and one another. This representation is amenable to external evaluation with systems like [8].

3.5 Coordination Plan

The PIs at Brandeis, Princeton, and Stanford will maintain regular contact via biweekly conference calls. All data, corpora, and tools will be maintained through a shared GitHub resource. One annual meeting is planned, alternating between Princeton, Brandeis, and Stanford, as well as regular meetings at both national and international conferences or workshops focusing on topics of relevance to the proposed research.

James Pustejovsky will oversee the experiments at Brandeis, Christiane Fellbaum the experiments at Princeton, and Daniel Lassiter and Lauri Karttunen will be responsible for running the experiments at Stanford. Lassiter will further help in coordination of the AMT experiments for all sites. Karttunen will consult Brandeis in the construction of experimental materials for the intensional adjective experiments. Annie Zaenen, Cleo Condoravdi and Karttunen will design the materials for the clausal complement adjectives, and Zaenen will be responsible for the analysis of the data. Lassiter will be involved in the data analysis phase on all components of the project.

3.6 Milestones and Deliverables

Year One of the project is dedicated to:

Q1	Collection of target adjectives; Perform corpus mining; Collect relevant patterns for clause-selecting, intensional, and scalar adjectives.
Q2	Derive initial semantic classifications and CDIs; MTurk hit design; coordination of annotation specs; preliminary inferential hypotheses.
Q3	Pilot MTurking experiments; Evaluate corpus data by research teams.
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	Run experiments with MTurkers.
Q2	Analyze/Evaluate results of MTurker data with/against initial CDIs.
Q3	Continue MTurking work; Update classifications and mappings.
Q4	Identify detailed contextual parameters accounting for judgment divergence; revise CDIs 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 CDIs.
Q3	Design a way to represent different adjective classes in lexicons, e.g. WordNet (for scalars, model developed in [57] can be further developed).
Q4	Evaluation; Data collection protocols; Prepare articles for publication. Make sure that all the data is publicly available. Final report.

4 Outreach and Education Plan

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

Adjective Inference Challenge. To actively engage the community in the adoption and use of the annotation methodology and the resources developed therewith, we will organize an NLP shared task in the third year of the project, focused on three specific tasks involving a relatively straightforward challenge, identifying inferences in textual data associated with the adjective classes being studied. The challenge will be run in a way similar to the Shared Tasks of the Conference on Natural Language Learning (CONLL) or SemEval, where colleagues are invited to compete to obtain best results on a specified task and data set. Our challenges will require use of the adjective inference datasets (CDIs) developed for training the competing algorithms. We plan to host a workshop at the Language Resources and Evaluation Conference (LREC) in May, 2018, where we will engage the community in further refining the scope and nature of deep textual inferences.

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

Tutorials and Training. We will design a tutorial on how the annotation methodology can be applied and deployed to other annotation tasks and CL challenges. This will be submitted for inclusion at the major conferences in the field (ACL, NAACL, EACL, AFNLP-sponsored conferences, ICGL, LREC, COLING), beginning in spring, 2016 and continuing to the end of the project. We will also propose tutorials at summer schools such as NASSLLI, ESSLLI, and LSA.

5 Broader Impact

The proposed work makes several significant contributions to a broader community of computational linguists, AI researchers, and psychologists. Our work lays theoretical groundwork for large-scale annotation of adjectives in order to support automatic systems in inferencing tasks. A second contribution is a more sophisticated theory of the role of lexical information in human inferential behavior, a topic of considerable psychological interest. Third, the work holds out the promise of developing new methodologies for large-scale annotation and combining experimental and corpus investigation that could benefit the development of more human-like systems for natural language understanding.

6 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) 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:** [30]; [29].

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

workshop participants are incorporated into the present proposal. **Publication:** [57].

CI-ADDO-EN: A Second-Generation Architecture for WordNet*CNS 0855157*(PI: Christiane Fellbaum) 07/29/2009 - 07/31/2012 $396,231.00. This grant supports the design and creation of a relational database for WordNet as well as numerous lexicographic improvements and community support. **Publications:** [17],[14],[13],[4],[49].

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

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

Three supplements supported grant IIS -0705199, 08/17/2007 - 07/16/2011: RI: Collaborative Proposal: Complementary Lexical Resources: Towards an Alignment of WordNet and FrameNet, PIs C.Fellbaum and C. Baker (ICSI). **CNS 0835139**, awarded 06/12/2008, $6,000.00; **RI: 1007133**, awarded 12/29/2009, $6,000.00; **IIS 0903358**, awarded 10/31/2008 $6,000.00. The original grant and the three supplements supported the manual alignment of FrameNet and WordNet. An important by-product was the manual annotation of all senses of the targeted word forms in the American National Corpus. **Publications:** [15]; [16] [2] [11].

Workshop on Semantics for Textual Inference *NSF 1064068*, (PI Cleo Condoravdi, Co-PI Annie Zaenen). The PIs organized two workshops, one at the LSA Institute 09-10/07/2011, the other at CSLI, Stanford, 09-10/03/2012. **Publications:** [9].