My primary research is in the representation of semantic information and its use in natural language processing applications. The meaning of a sentence is a central aspect of natural language understanding, yet an elusive one, since there is no accepted methodology for determining meaning. There is not even a consensus on criteria for distinguishing word senses, as can easily be seen by comparing entries for the same word in any two dictionaries. Linguistics offers potentially useful insights into semantic representations, and initially I used Jackendoff’s Lexical Conceptual Structures as a basis for computational lexical semantics [1]. This was implemented in the Pundit/Kernel text processing system at Unisys [2]. These representations proved to be effective for driving reference resolution, temporal analysis and recovery of implicit information, and led to this system being internationally recognized as providing path breaking coverage of semantics and pragmatics [3]. However, this experience also revealed the domain-specific limitations of the approach, and the difficulty of extending hand-crafted lexical entries to new vocabulary. The desire to develop a more robust technology for semantic processing focused my attention on more data driven techniques, with new insights from linguistics. Beth Levin has correlated syntactic alternations with a verb’s semantic content to create classes of verbs with similar syntactic and semantic behavior [4]. The accessibility of syntactic structure gives rise to the potential for using a distributional analysis of text as a methodology for determining semantic components. My students and I have been developing VerbNet, a class based computational English verb lexicon that contains explicit syntactic frames and semantic components for individual verbs [5, 6]. As a validity check on the semantic components we have used them to drive animations of the actions the verbs describe [7]. VerbNet has recently been incorporated in the PARC parser used by PowerSet [8], as the basis for lexical acquisition at Rochester [9], and as a lexical component for discourse analysis at the University of Illinois/Chicago [10].

Data-driven techniques rely on supervised machine learning, so my current goal is the definition of a level of shallow semantics (sense tags, semantic role labels and more recently event relations and discourse relations) that can be consistently annotated on a large scale. While chair of SIGLEX[icon], I initiated a series of international workshops that culminated in the SENSEVAL (now SEMEVAL) Word Sense Disambiguation (WSD) evaluations based on publicly available sense-tagged data in several languages [11, 12]. (The call for task proposals for SENSEVAL-2010 has just gone out). In addition to supplying English sense-tagged training data based on WordNet, we are implementing automatic WSD systems. The performance of Hoa Dang’s Maximum Entropy WSD system [13] for highly polysemous English verbs, 62.5%, was surpassed only by Jinying Chen’s [14], which in turn is currently matched by Dmitry Dligach’s Support Vector Machine system [15]. This high performance is achieved through the use of rich linguistic features derived from syntactic arguments of the verb and corresponding semantic class constraints. A similar approach has been applied to Chinese, although with less benefit from syntax [16].

A pervasive problem with sense tagging is finding a sense inventory with clear criteria for sense distinctions. WordNet is the most widely used public domain on-line English lexicon, but it is often criticized for its extremely subtle and fine-grained sense distinctions. Would more consistent and coarse-grained sense distinctions be more suitable for natural language processing applications? By grouping the highly polysemous verb senses in WordNet (on average reducing the >16 senses per verb to 8) we have taken an important first step in providing a more flexible granularity for WordNet senses [13]. This approach to grouping has been applied to the large-scale sense tagging effort of English, Chinese and Arabic nouns and verbs that has been funded by DARPA-GALE OntoNotes [17]. The Frameset sense tags
associated with the PropBank, as discussed below, provide an even more coarse-grained and replicable level of sense distinctions, forming the top level of a sense hierarchy.

Based on a consensus of colleagues participating in the NIST ACE (Automatic Content Extraction) program, my students and I developed a Proposition Bank, or PropBank, which provides semantic role labels for verbs, nominalizations and participial modifiers for the 1M word Penn TreeBank II [18]. Our work with VerbNet proved invaluable for defining the appropriate semantic roles in this endeavor. For example, *John* is the Agent or Arg0 of *John broke the window*, and the *window* is the Patient or Arg1; *IBM* is the Theme or Arg1 of *IBM rose 1.2 points*. In addition, for just over 700 of the most polysemous verbs in the Penn TreeBank, we have defined two or more Framesets – major sense distinctions based on differing sets of semantic roles. These Framesets overlap closely (95%) with our manual groupings of the SENSEVAL2 verb senses, and thus they can be combined to provide an hierarchical set of sense distinctions. The PropBank was made publicly available through the Linguistic Data Consortium in February 2004, and formed the basis of the Conference for Natural Language Learning (CoNLL) 2004 and 2005 shared tasks. It was used again in conjunction with NomBank for CoNLL-08. Semantic role labeling based on PropBank figured prominently in the ACE and ARDA-AQUAINT programs. We are currently funded through an NSF-ITR to map PropBank to FrameNet, Chuck Fillmore’s complementary lexicography project at Berkeley, and to train automatic labelers for English and Chinese. We have already demonstrated that mappings from PropBank roles to VerbNet roles can improve accuracy on Arg2 labeling for a new genre by over 10% [19]. The parallel Chinese Treebank and PropBank data developed under the same grant [20, 21] have been included in the annotated parallel Treebanks/PropBanks for DARPA-GALE OntoNotes. The automatic semantic role labelers we are building use features that are very similar to our WSD system features, and we find that semantic role label features improve WSD and vice versa [16]. Dan Gildea, Nianwen Xue and I have committed to writing an SRL textbook [27].

Multilingual TreeBanks and PropBanks are facilitating the training of automatic parsers, sense taggers and semantic role labelers for languages other than English. The levels of accuracy that we are now achieving for English allow for powerful information extraction and text mining applications [Surdeanu, et. al., ACL2003], and has sparked strong interest in porting to new domains such as biomedical journal articles. I am currently working closely with biomedical colleagues on this task [22]. There is also a compelling hypothesis that these corpora will inform statistical machine translation systems, improving their ability to preserve essential elements of dependency structure in the generation of translations [23], the lack of which is currently having a major negative impact on translation quality. They will also allow us to investigate cross-linguistic semantic generalizations, [24], moving us one step closer to the all too elusive goal of rich, independently validated semantic representations. With colleagues from Hyderabad, India, Columbia and the Universities of Washington and Massachusetts Bhuvana Narasimhan (CU) and I have just begun a new NSF project to develop a Hindi/Urdu Dependency Treebank/PropBank which we will be automatically converting to an additional Phrase Structure Treebank/PropBank. We are also beginning a new project with Northrop-Grumman on social role analysis for English and Hindi, under IARPA funding.

In addition, my move to Colorado opened up new avenues for collaborations with psychologists on human sentence processing models. The first fruit is a study by Susan Brown, my Linguistics PhD student, on the validity of an hierarchical approach to sense distinctions [25]. The success of this study has encouraged us to set more ambitious goals, and funded by a summer ICS Science of Learning catalyst grant, a psychologist colleague (AI Kim) and I are now investigating the cognitive processing of figurative language using ERP studies, corpus linguistics studies and response time studies [26].


