Detection of Intentional Ambiguity in Informal Language

Lawrence J. Mazlack

1 Overview and Objectives

Knowledge is distilled into the informal or natural language used in ordinary discourse. Much of it is computationally stored in texts. Computational natural language understanding is critical to handling the large amounts of text that already exist, as well as the geometrically growing volume that will come to exist. Informal language may be ambiguous. Knowledge cannot be readily summarized or accessed unless inherent utterance ambiguity is detected and resolved. While there are many existing disambiguation tools, *intentional* ambiguity recognition has not received much attention. Recognizing *intentional* ambiguity is a vertical step to understanding natural language. Humor often depends on intentional ambiguity; it offers a clear, distinct research focus. Therefore, the *purpose* of this research is to detect intentional ambiguity that results in humor. Humor has been studied since the ancient Greeks. However, only small steps have been made in computational humor recognition. Necessary to achieving the goals is a flexible system for representing, storing and accessing data, information and knowledge. Also needed are automated methods of drawing conclusions from data and knowledge.

The *long term goal* of this research is to computationally detect, represent and handle ambiguities in text; the focus is on computational humor recognition in short, humorous texts. Considering short jokes that are dependent on lexical ambiguities and phonological similarities further focuses the proposed work. The *objective of this research*, which is a step toward our long-term goal, is to computationally detect short jokes intended for young children. The *central hypothesis* of this research is that a description logic ontology that is suitable for annotating knowledge in children's texts is also suitable for joke recognition. The hypothesis has been formulated on the basis of preliminary analysis of short children's texts and jokes, and our prior experience in using non-ontological methods. Restricting the domain to young children's jokes is expected to decrease complexity while retaining the core issues. The domain restriction reduces needed background knowledge, which leads to a reduction of knowledge to be captured. This, in turn, leads to a smaller, more manageable ontology, and fewer inferences drawn from it. It allows a concentration on humor recognition methods with a reduced emphasis on the problems of capturing and inferring knowledge. The development of such methods is a step closer to an understanding of the product of human cognitive processes and "more human" intelligent systems, which is a *rationale* for undertaking this research. In addition to our supportive preliminary data, we are particularly well prepared to undertake the proposed research, because we have already performed and published investigations of statistical techniques; and, we have published explorations that prepare the foundation for our proposed approach. Our working environment contains investigators who are working on parallel projects involving theoretical aspects of ontologies, as well as, the ontological reasoning tools that we propose to use. This creates a rich intellectual environment in which to perform the work (See the Facilities, Equipment and Other Resources section).

Significant interest in this research has been indicated by exceptional, extensive, worldwide popular media attention to our preliminary results in hundreds of media outlets (discussed in the Background section).

The central hypothesis will be tested and the objective of this research will be accomplished by pursuing the following two *specific aims*:

1. ***Build a description logic-based imprecise ontology, containing concepts and relationships between them for annotating texts for young children***.
   Based on the preliminary data, the *working hypothesis* is that a children's dictionary defines all nouns that are needed to create concepts for the ontology. The relationships between the concepts will be semi-automatically created with knowledge extracted from a children's dictionary and a collection of children's texts.
2. ***Determine a method for recognition of script overlap and opposition that create jokes and that are based on lexical ambiguity or phonological similarity***.
   A script is a framework for typical activity. Our *working hypothesis* is that salient scripts can be determined as a composition of concepts in the jokes and inferences drawn from these concepts.

The proposed research is *creative and original* because it is ontologically based. The few humor detectors that exist today are statistically, not ontologically based, and much more limited in the humor form. It is *expected* that the results will fundamentally advance the field of computational humor. It is expected that once an ontology-based children's joke detector is developed, the ontology can be expanded to include other knowledge. It is expected that knowledge contained in such ontology can be used to detect more sophisticated jokes, using similar, or slightly modified, algorithms. The *broad applications and positive impact* can range from aiding in achieving *sociable computing,* helping children and second-language learners master language, adding joke recognition to word processing software, as well as enhancing search engine results. Humor recognition may be useful in commercial natural language translation applications. Aim One will enable the recognition and representation of unambiguous text elements. Aim Two will enable the recognition of ambiguities. Collectively, the two aims will accomplish the overall objective of this application.

2 Expected Significance

Computational recognition of humor offers a focused domain for investigating intentional ambiguity. Computational humor recognition is a difficult task [Raskin, 1999] [Ritchie, 1999]; and very few computational detectors exist today [Mihalcea, 2005] [Taylor, 2004] [Yokogawa, 2002]. Yet humor is such a fundamental part of human interaction that no computational intelligent conversational agent is complete without an ability to understand humor [Binsted, 2006]. While humor has been studied since the ancient Greeks [Attardo, 1994], there is no formal theory that defines the basic terms formally, or that can make precise falsifiable predictions [Ritchie, 2004]. As an outcome of the proposed investigation, a formal method for recognizing short jokes that are intended for young children and are based on lexical ambiguity or phonological similarity will be developed and tested.

This contribution is significant because formal methods of humor recognition that are needed in order for computers to effectively detect and understand informal language will be developed for a restricted domain. Moreover, world knowledge in general [Zadeh, 2004], and humor in particular [Ruch, 2001] is perception based. The proposed ontology will be built taking advantage of methods developed for perception-based information [Zadeh, 2001]. The perception-based ontology and a formal method for recognizing short jokes will provide foundational elements to analytical disambiguation in whole language text. For example, multiple meanings of texts that may be undetected by a person could be discovered. This can lead to discovery of hidden messages in otherwise innocent texts. Such messages can range from simple innocent jokes that are undetected by a human due to a lack of background information, to possibly more serious information intentionally hidden by the sender. Beyond humor, Aim's One and Two might be used for computational understanding of children's texts. This in turn,

can be used for assessing language proficiency of young children; as well as, second language learners. Consequently, the outcomes are not only expected to fundamentally advance the field of computational humor, but also to have simultaneous broad and highly positive societal impact.

3 Relation to the Principal Investigator's Long-Term Goals

The work proposed in this application is a step along a continuum of research that the Principal Investigator is pursuing. The PI's longer-term goal is to systematically extract higher-level descriptions (summaries) from semantically tagged or untagged whole language (natural language). Extraction requires recognition of ambiguities and their disambiguation. The humor recognition in this proposal involves recognition of ambiguities, as such; it is directly relevant to the PI's goals. Semantically tagged data is an aspect of the Semantic Web that the proposed work will extend to natural language processing. Reference to ontologies may be necessary. Ontologies often need to incorporate imprecision. The proposed work involves semantically tagged text and ontologies. Specific longer-term goals include: autonomous knowledge extraction, autonomous semantic tagging, ontology development, management of imprecision within ontological frameworks, approximate and opportunistic reasoning. The PI is interested in imprecise data and often uses techniques from soft computing (including rough sets and fuzzy sets). Pursuit of this line of research is expected to yield a more detailed understanding of the application of soft computing methodologies to ontologies. Work performed to date has positioned the applicant to undertake what is proposed in this application, because that body of information suggested that the formal application of soft computing techniques would be useful in handling imprecise ontologies. Successful completion of what is proposed here will strategically position us to undertake the next steps that we anticipate will be needed in pursuit of our longer-term goal. Specifically, those steps are expected to include: general management of imprecision within ontological frameworks, autonomous knowledge extraction, autonomous semantic tagging, and disambiguation of figurative whole language dialogue.

4 Background

Often, computational natural language research has concentrated on tractable subsets of standard sentences that were intended to be unambiguous. Where there was ambiguity, the issue was how to reduce the ambiguity, not to embrace it. In order to computers to communicate more effectively, ambiguity has to be recognized and represented in the machine-readable structures. Humor often requires exploitation of intentionally retained ambiguities. For example, in

Judge: Why did you hit your husband with a chair?
Wife: I couldn't lift the table.

Consequently, there are excellent reasons to undertake detection and representation of intentional ambiguities. If computers are to communicate naturally and effectively with humans, they must be able to use humor [Binsted, 2006]. More fundamentally, "humor provides insight into how humans process language … By modeling humor generation and understanding on computers, we can gain a better picture of how the human brain handles not just humor but language and cognition in general" [Binsted, 2006]. Similarly, an intelligent machine must have the ability to laugh in the right places [Olsen, 2006]. Our central issue in computational humor is recognition or detection. However, very few computational recognizers have been attempted either in English [Purandare, 2006] [Mihalcea, 2005] [Taylor, 2004], or in other languages [Yokogawa, 2002]. The recognizers in English have used statistical methods. Additionally, there have only been a few humor generators attempted [Lessard, 1992] [Binsted, 1996] [Binsted, 1998] [McDonough, 2001] [McKay, 2002] [Stock, 2002] [Hempelmann, 2006]. Unlike humor generators, text generation in general has a long history.

Computational recognition of humor is not an easy task. It requires: the capability to interpret natural language, subtle and flexible inferences, and have access to a vast store of knowledge about the world. [Ritchie, 2004, p. 188] None of these components have been fully achieved.

The greatest strength of natural languages is that they have great expressive power [Sowa, 1999]. However, this strength is also one of the greatest obstacles to computational understanding. "Since the early 1950s, highly talented linguists, logicians, and computer scientists have been working on the problem of language understanding by computer. Yet no program today can read a newspaper at a level of high-school student." [Sowa, 1999]

One of the approaches to understanding natural language is the use of ontologies. There are many different kinds of ontologies with the aim of representing most human knowledge. They differ in formality, knowledge representation and reasoning methods. One of the earliest attempts was Cyc [Lenat, 1990]. Yet, none are completely successful in language interpretation and reasoning. The ontological burden was decreased by reducing the domain size to young children's jokes. While the ontology still has to contain and represent knowledge for everyday events, it does not have to encompass politics, science, professional terminology, etc. The domain reduction is expected to decrease the complexity and sophistication of language to be analyzed. This in turn decreases the knowledge that needs to be captured.

World knowledge in general [Zadeh, 2004], and humor in particular [Ruch, 2001] is perception based. Precisiated Natural Language (PNL) [Zadeh, 2004] deals with perception-based information, which "is intrinsically fuzzy" [Zadeh, 2004]. "PNL is based on fuzzy logic and has the capability to deal with partiality of certainty, partiality of possibility and partiality of truth. These are the capabilities that are needed to be able to draw on world knowledge for assessment of relevance, and for summarization, search and deduction." [Zadeh, 2004] Ontology based system that is designed for natural language interpretation and representation and takes advantage of methods developed for perception-based information [Zadeh, 2001] is expected to improve machine-based language interpretation and representation. In particular, this research will use the notion of precisiation and computational theory of perceptions [Zadeh, 1999; 2000] methods in the ontology.

Knowledge simply collected in the ontology is not enough. Inference is also needed. The meaning of a text is more than just its explicit meaning. On the basis of general linguistic, non-linguistic knowledge, and the situational context; inferences are drawn and the full meaning of the intended message is reconstructed. These inferences are needed to develop a full picture of the processed text that includes background knowledge. Logic-based systems provide methods to draw inferences from received information. This information has to be in a form that the system can deal with. In this research, description logics will be used.

Description logics are logical reconstructions of frame-based knowledge representation languages, with aim of providing well established declarative semantics to capture the meaning of the most popular features of structured representation of knowledge. Knowledge is represented as a tuple of three sets of axioms - terminological (concept) inclusions (TBox), role inclusions (RBox) and assertions of objects memberships to concepts and roles (ABox). The semantics of the language is provided via the interpretation function that maps symbolic definitions to the real world objects. Description logics have been used in natural language processing [Weischedel, 1989] [Allgayer, 1989] [Fehrer, 1994] [Herzog, 1991] [Stock, 1991; 1993] [Samek-Lodovici, 1990] [Lavelli, 1992] [Franconi, 1994] [Wahlster, 2000] [Rychtyckyj, 1999]. This project will use fuzzy description logics in stead of crisp description logics because natural language information is often imprecise.

Using fuzzy features in the description logics has been previously suggested [Straccia, 1998] [Tresp, 1998] [Holldobler, 2004] [Hajek, 2005] [Straccia, 2006]. The vagueness of concepts is captured by representing them as fuzzy sets. To accomplish this, TBox axioms are represented through fuzzy sets inclusion and ABox - as fuzzy membership assertions. Restricted inference services have been developed by extending classical crisp tableaux-based algorithms [Baader, 2000] to the fuzzy case [Stoilos, 2005]. In this proposal, the same approaches will be followed for the semantic component. The resulting knowledge is captured as a collection of fuzzy concepts, and semantic relationships between them.

Humans store and process words in mental lexicons. These lexicons contain lexical entries with information about a word's spelling, pronunciation, meaning and syntactic category. Based on the lexical decision task, Rubenstein [1970] showed that high frequency words are accessed faster than low frequency words. In our experiments we use Kučera-Francis word frequency [Kučera, 1967]. Experiments with semantic priming [Meyer, 1971] showed that words that are semantically related to a source are activated faster. Similarly, experiments with phonological priming [Evett, 1982] showed that words that sound similar to the source are activated faster. Parkin [1986] showed that words with unusual spelling (that have few spelling neighbors) are activated slower than words that have lots of spelling neighbors. Thus, at the very least, lexical access can be influenced by word frequency, word meaning, word phonology and spelling regularity [Simner, 2004]

Experiments were conducted to compare frequency of target and source in puns [Taylor, 2007a]. The goal was to discover a heuristic basis for a target recognition and selection methodology in joke recognizer. A target is a word or a word phrase that is similar to a source according to some metric. In a pun, this similarity is in pronunciation. A meaningful target, that is based on a source, has to be found for a text to be humorous. The results of the experiment showed that in 71% of the data the Kučera-Francis frequency (KF) value for a source was at least 10% lower than corresponding KF frequency value for a target. These numbers can serve as an input to guide a heuristic wordplay selection of target source pairs. Another experiment was conducted to compare the KF value of the source of the joke to the median KF value of the same joke [Taylor, 2007a]. The results showed that the source of the joke had a lower KF frequency than the median in 49 jokes out of 50.

For computational humor discovery, formal methods have to exist to determine when humor is found. However, there is no universally agreed theory of humor. This research will use the conceptually described Script-Based Semantic Theory of Humor [Raskin, 1985]. According to this theory, there are two necessary and sufficient conditions for a text to be humorous:

- A text has to be semantically compatible, fully or in part, with two different scripts.
- The two scripts with which the text is compatible oppose, and must overlap fully or partially.

A script variant that will be used in this research is defined as "a graph with lexical nodes and semantic links between the nodes." [Raskin, 1985]. This works well for this proposal's purposes as the knowledge is stored in an ontology, where concepts are lexical nodes with semantic relationships between them. As script overlap and script opposition are not fully formally defined in the Raskin's Humor theory, part of this research will be to define them for the class of jokes that will be considered.

Significant interest in this research has been indicated by exceptional, extensive, worldwide popular media attention to our preliminary results in hundreds of media outlets. Some of the outlets are:

- *New Scientist*
- *Wall Street Journal*
- *First Science News*

- *FreeRepublic*
- *Washington Post*
- *Eureka Alert*
- *British Computer Society News (United Kingdom)*
- *The Mail (United Kingdom)*
- *National Post (Canada)*
- *Andhra News (India)*
- *Daily India (India)*
- *New Kerala (India)*
- *Times of India (India)*
- *Iafrika Cooltech (South Africa)*
- *Australian Broadcasting Corporation (Australia)*
- *Daily Telegraph (United Kingdom)*
- *The Inquirer (United Kingdom)*
- *Tech Gadgets*
- *Mumbai Mirror (India)*
- *ITweek*
- *Telegraph (United Kingdom)*
- *Pakistan Uncut (Pakistan)*
- *Daily Times (Pakistan)*
- *ACM Tech News*
- *iiRobotics (United Kingdom)*
- *CBC News (Canada)*
- *Softpedia*
- *Columbus Dispatch*
- *Punjab Kesari (India)*
- *Australian IT (Australia)*
- *MC Press (Canada)*
- *One India (India)*
- *Austrian Broadcasting Corporation (Austria)*
- *Dr. Dobb's*
- *Inky Circus*
- *Ars Technica*
- *Scientific American*
- *Science Daily*
- *The Poorhouse (United Kingdom)*
- *CNET Reviews*
- *Los Angeles Times*
- *AAAS Science Update*
- *Discovery Channel*
- *Sawf News (India)*
- *Andhra News (India)*
- *The Economic Times (India)*
- *Malaysia Sun (Malaysia)*
- *Times of India (India)*
- *As It Happens (Canada)*
- *Radio New Zealand (New Zealand)*
- *ScienceTicker (Germany)*
- *Welt (Germany)*
- *Le Nouvel Observateur (France)*
- *Robot Impact (France)*
- *Pequei no Google (Brazil)*
- *Nuevo Excelsior (Mexico)*
- *O Popular On-Line, Tecnologia(Brazil)*
- *Filosofi (Belgium)*
- *Sostav (Ukraine)*
- *Ekapija (Serbia)*
- *Ny Teknik (Sweden)*
- *Pravda (Russia)*
- *ENI News (Bosnia)*
- *Lifestyle Bosnia (Bosnia)*
- *Vista (Vietnam)*
- *Robotyka (Poland)*
- *Coolstreaming (Italy)*
- *RHO Empreendedor (Portugal)*
- *RHO Empreendedor (Portugal)*
- *NeoTeo (Argentina)*
- *Tweakers (Netherlands)*
- *Silicon (Germany)*
- *KL!K (Croatia)*
- *Startpagina (Netherlands)*
- *Franklin Magazine*

A larger sample of the media interest can be seen at:
`http://www.ece.uc.edu/~mazlack/Applied.AI.Lab/InTheNews.html`

Lastly, as a humor recognizer needs a vast store of knowledge about the world, it is expected that limiting the domain of jokes to young children's jokes significantly reduced the amount of needed information. The success of natural language processing corpus-based techniques [Manning, 1999] points to the fact that knowledge can be learned from a collection of texts. Knowledge from children's texts can be collected through statistical and other approaches.

It is expected that the limiting the domain to children's jokes will also reduce the complexity of knowledge, inferences and sophistication of language of the analyzed text. Adding techniques for handling imprecise information, and experience with statistical-based humor recognition put us in a strong position for successful detection of jokes for young children. In a possible extension to this work, it is suspected that incrementally adding more knowledge to the developed system will result in recognition of intentional ambiguities in larger domains.

5 Research Plan

The central hypothesis will be tested and the objectives of this research will be accomplished by pursuing the following two specific aims.

Aim 1: Build a description logic based ontology capable of capturing imprecision and containing concepts and inter-concept relationships to support representation of knowledge contained in young children's texts; the annotation facilitates computational joke detection.

Introduction. The amount of background knowledge necessary to understand general natural language sentences is difficult to computationally represent. Reducing a discourse's domain size reduces the amount of computationally represented background knowledge needed. This work reduces the general discourse domain to a young children's discourse domain; this results in a substantial reduction of necessary background knowledge. For example, knowledge about political news and professional domains does not have to be captured. Typically any utterance can be represented in term of any of the three components: spelling, pronunciation or meaning. The approach that is followed in this proposal represents any texts using three parameters: they way the words are spelled (orthographic component), the way they are pronounced (phonological component) and what they mean (semantic component). Entities of each component are linked with their corresponding entities in the other two components, as shown in Figure 1. For example, the spelling of the word cat is in the orthographic component, the pronunciation is in the phonological component, and meaning is in the semantic component.

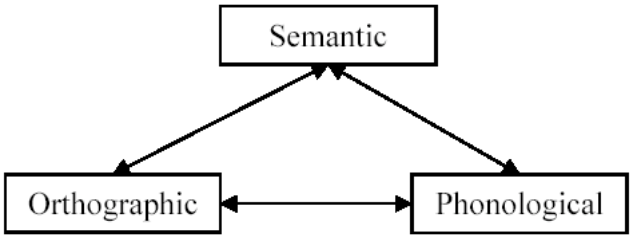

Figure 1: Components of text processing

Natural language often contains imprecise knowledge. A mechanism for dealing with imprecise knowledge is needed to fully handle natural language. Ontologies provide the capability to represent objects, concepts and other entities that exist in an area of interest as well as the relationships that hold among them. Background knowledge [Hassell, et al., 2006] and data meaning (semantics) [Beale, et al., 2004] can be provided by ontologies. Imprecise ontologies provide the capability to handle imprecise knowledge. The *objective* is to construct an imprecise ontology from a children's dictionary and a collection of children's texts for subsequent semi-autonomous annotation of children's books or texts. This will be done by: collecting the terms from a children's dictionary and from young children's texts; and then representing them in a description logic based knowledge base. The working hypotheses are: (1) A children's dictionary defines nouns that are needed to create concepts used for semantic tagging of young children's texts; (2) Knowledge extracted from a children's dictionary and a collection of children's text is sufficient to create needed relationships between the concepts; (3) The use of an imprecise ontology results in a higher accuracy of annotation compared to using crisp ontology.

A prototype for the ontology has been developed; it contains crisp concepts a from children's dictionary. It will be complemented by imprecise terms and relationships extracted from children's texts. The *rationale* for doing this is that the resulting ontology will formally represent the vocabulary that children use generally as well as in jokes. The ontology is expected to facilitate the process of annotating children's texts with accuracy superior to that provided by a crisp ontology. It is also anticipated that the resulting ontology will be extendable and larger vocabularies can be incorporated later as necessary. The jokes to be recognized are based on

lexical ambiguity and phonological similarity; consequently, most senses and pronunciations of words will be considered. The *expected outcome* of this aim is an ontology that can be used to annotate structurally simple texts, built from a relatively small vocabulary.

Experimental design. The following tasks will be accomplished to test the first hypothesis: *A children's dictionary defines most the nouns that are needed to create concepts used for semantic tagging of texts for young children.*

- Create a knowledge base with words containing in a children's dictionary. This will be referred as the orthographic component. Orthographic component is concerned with words' spelling and morphology. A children's dictionary, containing approximately 2500 words will be used. The words from the dictionary are entered into the knowledge base. Additionally, the words are rated with familiarity values [Wilson, 1987], Kučera-Francis frequency [Kučera & Francis, 1967] [Wilson, 1987], and possible syntactic categories. These additional values are also entered into the knowledge base. Given any two entries in the orthographic component, an orthographic similarity value can be calculated based on the words' spelling.
- Build a concept hierarchy from noun entries in the orthographic component. This will be referred as semantic component. A working prototype of such hierarchy has already been developed. The concepts hierarchy organization models the noun hierarchy of WordNet [Fellbaum, 1998]. Each noun in the orthographic component is linked to at least one concept in the concept hierarchy. The concept hierarchy is represented in description logic. The relationships in the hierarchy consist of is-a and part-of types. The "base layer" of a semantic component is created by combining similar entities (synonyms) in the orthographic component. The entries in the orthographic component are instances of the concepts in a hierarchy. Each concept describes common properties of a collection of instances. An instance belongs to each concept to a certain degree (membership value). The degree of membership assigned to the relationship is based on fuzzy logic [Zadeh, 1965]. When an instance does not belong to a concept, a membership value of 0 will be used. Initially, an instance *a* has a non-zero degree of membership to a concept β if a dictionary explains *a* in terms of β.
- Annotate a collection of young children's texts using the existing concept hierarchy. The first part of working hypothesis will be tested on a sample collection of young children's texts. The first five books that Amazon.com recommends as young children's texts will be used. The texts will be digitized. An existing tokenizer will be run on the collection of texts. The output from the tokenizer will be automatically annotated. The aim of the annotation is to find all possible meanings of nouns in texts, *not* to find the optimal meaning. For each noun received from the tokenizer's output, all meanings of the noun that are entered in the hierarchy will be found. A noun annotation will be considered unsuccessful only if none of the found meanings are correct. If at least one meaning is acceptable by a human expert, it will be assumed that the hierarchy contains the needed meaning of the word, regardless of the number of other meanings of the same word.

The following tasks will be accomplished to test the second hypothesis: *Knowledge extracted from a children's dictionary and a collection of young children's texts is sufficient to create needed relationships between the concepts*.

- Using a children's dictionary, build relationships between concepts and map them into the orthographic component. Relationships between concepts in the concept hierarchy will be built based on the dictionary terms. Adding relationships between concepts will transform the concept hierarchy into an ontology. A hierarchy of the relationships will be developed.

Similarly to the concept hierarchy, semantic component relationships are created by combining similar entities (synonyms) in the orthographic layer. The resulting ontology will then be enhanced with membership functions of the relationships. A membership value will be computed between instances of a set of all concepts. A relationship between concepts α and β exists if a dictionary definition of instance of α refers to an instance of β. If a relationship between two concepts exists, a non-zero membership value will be assigned. It is expected that most concept pairs will not have a relationship defined from the dictionary. The membership values will therefore be 0. The relationships will be represented as roles in description logic. The resulting ontology will contain all explicit conceptual knowledge recorded in the children's dictionary.

- Using a collection of young children's texts, build relationships between concepts and map them into the orthographic component. Relationships between concepts in the concept hierarchy will be added from a collection of young children's texts. The texts will contain new, non-repeating relationships. In other words, if text *A* contains relationships a_1, a_2, …, a_n, text *B* is only useful for finding new relationships if it contains relationships other than a_1, a_2, …, a_n. To keep track of which relationships were learned from which texts, ontology versioning [Klein, 2001] will be used. The membership values for the relationships of each text will be added as a newer version of the ontology. As all concept elements and hierarchy will remain the same, the difference between versions of the ontology will be in the added membership values.

  The choice of corpus is important as it will constrain the relationships that can be added. The corpus will be chosen so that the topics of the texts are similar to the topics of the jokes. To do so a candidate list of texts will be randomly selected. The texts will be examined for the selected criteria. After five satisfactory texts are discovered, the process stops, the corpus is selected.

  Because of the amount of work it involves, the texts will be computationally syntactically annotated. A Link Grammar Parser [Sleator, 1993] will be used to find syntactic relationships between words. It is expected that the relationships found by Link Grammar Parser can point to semantic relationships between unambiguous words: if there is a link between an unambiguous instance i_1 of a concept χ_1 and an unambiguous instance i_2 of a concept χ_2 then a relationship between χ_1 and χ_2 will be added. A difficulty arises when a word (an instance) can belong to more than one concept. In this case, a disambiguation will be done manually:

  if
  (i) an instance i_1 belongs to concepts χ_1 or χ_2;
  (ii) an instance i_2 belongs to concepts χ_3 or χ_4;
  (iii) a link between i_1 and i_2 is found by the parser;
  (iv) manual disambiguation results in an instance i_1 corresponding to χ_1 and i_2 corresponding to χ_3
  then
  a degree of membership of a relationship between χ_1 and χ_3 is high, and degree of membership of a relationship between χ_1 and χ_4 , χ_2 and χ_3 , χ_2 and χ_4 is low.

  It is expected that recording relationships with low membership values will assist in recognition of humor that is due to lexical ambiguity. The membership magnitude is less important than the relative order of membership values that will be taken into account.

- Annotate a different collection of young children's texts using the existing ontology. The aim of the annotation is *not* to disambiguate texts, but rather to find all interpretations that

are possible using the existing ontology. The words in the texts are entries of the orthographic component. These entries are instances of the concepts or roles of the ontology. Noun annotation results will be disregarded for evaluation of this sub-aim. The results will be considered to be successful if most relationships that exist in the test collection are recorded in the ontology, no matter how low the degree of membership is (except zero).

The following tasks will be accomplished to test the third hypothesis: *The use of imprecise ontology results in higher accuracy of annotation compared to crisp ontology*.

- Build a crisp ontology. A crisp ontology will be created using similar methods to creation of the fuzzy ontology. The only difference will be in the absence of membership values. Relationships from young children's texts will then be added to the crisp ontology. Similarly to a fuzzy ontology, relationships from non-ambiguous instances will be added. Unlike relationships in a fuzzy ontology, when ambiguous instances are manually disambiguated, only correct relationships will be added.
- Calculate the difference between imprecise and crisp ontologies for text annotation. Text annotation using crisp ontology and fuzzy ontology will be compared to annotation by a human. The metric will be based on the number of correct annotations. The significance of the difference can be determined using statistical analysis.

Expected Results: Upon completion of this aim, it is expected to have an orthographic component containing all words in the children's dictionary, and an ontology (semantic component) whose instances are taken from the orthographic component. The ontology can be used to annotate structurally simple texts with vocabulary contained in the dictionary. The vocabulary in the dictionary will correspond to instances of concepts in the ontology. The notions of precisiation [Zadeh, 2004b] and computational theory of perceptions [Zadeh, 2001] will be used in determining degrees of memberships to handle imprecision in natural language.

Potential problems and alternative strategies: Although the proposed method for building an imprecise ontology is supported by a preliminary investigation, the resulting ontology may not contain all necessary knowledge for correct annotation of texts. In this case, other knowledge that is required by determination of a human will be added to the domain. While the ontology may not represent all possible facts, it is expected that the new knowledge will not contradict the existing knowledge; and what it represented is expected to be true for the domain. In the event that a contradiction exists, an ontology will be manually modified to remove undesired elements. Lastly, the selection of corpus will affect the relationships that are created. In the event that the corpus does not contain necessary relationships, other texts will be added to the corpus.

Aim 2: ***Determine a method for the recognition of script overlap and opposition that create jokes and that are based on lexical ambiguity or phonological similarity***.

Introduction: The outcome of the previous aim is an imprecise ontology that can be used to support the interpretation of the meaning of the texts. This *section's objective* is to develop and test a method that can recognize short jokes, using possible interpretations of texts. The method is based on one of the existing humor theories, namely Script-Based Semantic Theory of Humor [Raskin, 1985]. The theory suggests that a short text is humorous if it contains two scripts that overlap and oppose. An algorithm will be developed to recognize phonological similarity or lexical ambiguity in potential jokes that *leads* to different interpretations. These interpretations will be used in choosing the appropriate scripts. For example, consider the joke "What did the beaver say to the tree? – It's been nice gnawing you." Phonological similarity of *gnawing* and *knowing* leads to two different interpretations of the sentence. The first one points to a farewell

statement; the second – to chewing on a tree. Notice that the second interpretation does not work without the presence of "beaver" in the first sentence. The beaver, gnawing on (or saying farewell to) the tree becomes salient concept of the joke. The *working hypothesis* is that salient scripts can be determined as a combination of concepts and relationships in the jokes (containing phonologically similar instances, which leads to ambiguity) and inferences drawn from these concepts.

Computational understanding of humor is important as humor "facilitates social interaction, ameliorates communication problems and affects attention and memory" [Binsted, 2006]. Drawn attention can improve learning curve of a subject. A humor recognizer can add a new and useful feature to a second language learning software. The *rationale* for computationally recognizing children's jokes is that jokes that result in wordplay can aid with the language learning. In addition, humor is not always intended. For example, a spelling mistake can result in humorous utterance, which existing word-processing applications do not catch. Recognizing humorous ambiguities in text can improve word-processing applications that so many people rely on today.

Experimental design: The experiment will consist of analyzing a sample of 100 question-answer jokes and 100 question-answer non-jokes. The question-answer jokes are typical children's jokes. They have a simple structure and are easy to follow. Fifty jokes will be based on lexical ambiguity; fifty will be based on phonological similarity. The working hypothesis will be tested by undertaking the following steps:

- Create a phonological component for entries in the orthographic component. The goal is to recognize jokes that are based on phonological similarity of words with different meaning. In order to do that, phonological word representations have to be recorded. Phonological representation of words in the orthographic component will be obtained from The CMU Pronouncing Dictionary [Weide, 1998]. The entries in the orthographic component are mapped into the phonological component. Thus, two entries of the orthographic component that have the same pronunciation (including stress) are mapped into one entry of phonological component. Additionally, the phonological component will contain a phoneme similarity table [Frisch, 1996], speech errors [Frisch, 1996], and the phoneme cost table of going to a target from a source in a pun [Hempelmann, 2003]. Using these factors, the phonological similarity of any two entries of the orthographic component can be calculated.
- Finding words in texts that lead to lexical ambiguity or phonological similarity. It will be assumed that lexical ambiguity exists among homographs (words with same spelling, different meaning). Words that have at least one character difference will not be considered at this point. Phonological ambiguity will be considered in words with similar, but not identical pronunciation as well as homophones (words that sound the same).
    - Lexical ambiguity. A tokenizer will be run on the text. The output of the tokenizer will be analyzed for lexical ambiguity. It will be assumed that a text contains lexical ambiguity if there is an entry of the orthographic component that can be an instance of several concepts of the imprecise ontology with non-zero degree of membership.

      Fifty lexical jokes and fifty non-jokes will be tested. The test will be considered successful if all ambiguities that lead to jokes are found. Other ambiguities that are found will give an idea of "overhead" computations, such as computation of ambiguities that do not lead to humor. A heuristic will be found to help concentrate on desired ambiguities and disregard the "overhead". One of the possible heuristics is to concentrate on the answer of the question, which is likely to be the punch line of the joke.

- Phonological similarity. Phonological similarity will be tested similarly to lexical ambiguity. It will be assumed that there is phonological similarity if two entries of the orthographic component result in a high phonological similarity value (of the phonological component) and have high membership values of different concepts of the semantic component (ontology). The high membership value of different concepts is important because in order for phonologically similar words result in a joke, they have to be conceptually different (mean different things). For example, *knowing* and *gnawing*. Initial forms of the words that have highly similar pronunciation but different spelling will be added as instances of concepts with low degree of membership. (It is assumed, again, that only instance relationship will have to be added to a concept). For example, if i_1 is an instance of concept α, and pronunciation (i_2) is very close to pronunciation (i_1), then i_2 will be added as an instance to concept α with low degree of membership. The degrees of membership of i_2 will be computed based on the degree of membership of i_1; and, degree of similarity of pronunciation between i_1 and i_2. These concept instances and their degrees of membership can be added as a new version of the ontology, similarly to the addition of relationships from different texts.

  To reduce the computational complexity, a heuristic function will be developed to guide the selection order of entries chosen for phonological comparison.

  Similarly to the previous experiment, 50 phonological jokes and 50 non-jokes will be tested. The success and the "overhead" will be measured similarly to the evaluation of the lexical ambiguity identification.

- Determine if the scripts containing concepts with lexically ambiguous instances, or phonologically similar instances lead to jokes. Not all lexical or phonological ambiguities result in jokes. For example, consider "Do you know what a dog will do to a bone? Gnaw on it." This contains phonologically similar words, *know* and *gnaw*, but the dialog is not a joke. In order for the two entries found in the previous step to trigger a joke, they have to belong to two *different* scripts that overlap and oppose.

  - Script creation and recognition. A scripted event should have a precondition and an effect [Triezenberg, 2003]. Scripts in a description logic ontology may be represented as a combination of concepts and relationships between them. A precondition and an effect of a script both are a set of concepts and roles. Script creation will be a manual process. The scripts will be selected from the existing concepts and relationships. Other relationships and concepts will be added only if needed.

    Lexical ambiguity or phonological similarity may result in a joke. Two algorithms will be developed: one to recognize lexical ambiguity, another to recognize phonological similarity that may result in a joke. Two concepts, χ_1 and χ_2, containing instances that are orthographically or phonologically similar, must be in the scripts. For a joke to occur, the concepts should be in different scripts; i.e., concept χ_1 should be part of $script_1$ and concept χ_2 should be a part of $script_2$.

    Inferential scripts [Hempelmann, 2001] can be activated by context of a text. If an inferential $script_3$ is activated by a text that has activated sentential $script_1$, $script_3$ and $script_1$ will be combined into $script_4$ [Clark, 2000]. Both $script_4$ and $script_1$ will be used separately in joke recognition.

To test script recognition, all previously used jokes and non-jokes will be considered. The script recognition task will be considered successful if the desired scripts are found in most of the previously used texts.

- Script opposition and overlap. Script overlap will be defined as a fuzzy intersection of concepts and relationships that are in the two scripts. This is similar to an approach proposed by Hempelmann [2001], who suggested treating script overlap as intersection of sets/scripts.

  Script opposition will be defined as a fuzzy intersection of scripts' effects. This approach is similar to the one reported in Taylor[2007b]. It is expected that *two different scripts create a joke when fuzzy intersection of two scripts is high (or at least exists); and, intersection of the scripts effects is low or does not exist*.

  Again, all previously used jokes and non-jokes will be considered. In the case of incorrect automatic script recognition, the scripts will be manually identified and used in another phase of investigation. Thus, the test of script overlap and opposition will not be influenced by the results of script definition. It is expected that concepts χ_1 and χ_2 that contain instances that lead to jokes, are parts of two different scripts.

  Both jokes and non-jokes are expected to have similar results in script overlap. Script opposition in non-jokes is expected to be similar to script overlap in non-jokes. Script opposition in jokes is expected to give significantly different results. A text will be considered to be a joke by the recognizer if script opposition results and script overall results are significantly different. A text will be considered a non-joke by the recognizer if the script overlap results and script opposition results are similar.

A test will be considered successful if the recognition of jokes as such and non-jokes as such is above random; and the difference is statistically significant.

Expected results: Completion of this aim will: (a) Add a phonological component to semantic and orthographic components, thus creating a triangle of pronunciation, spelling and meaning; (b) Result in a computationally based algorithm that will find lexical ambiguities in texts that result in humor; (c) Identify scripts that are necessary for the ambiguity to result in humor; (d) Identify words that have similar pronunciation that may lead to wordplay jokes, given the necessary scripts. In addition, the computationally based algorithm can be used in identifying ambiguities in text that do not lead to jokes. Consequently, texts that are either humorous or non-humorous will be identified; and the identification will be explained.

Potential problems and alternative strategies: The meaning of a sentence may not always be determined from syntactic analysis due to the difference between syntactic and semantic compositionality [Nirenburg, 2004]. In this event, the sentence will be manually semantically annotated to preserve its proper meaning. There is a chance, although unlikely, that a word does not exist in the ontology. If this is the case, a needed instance will either be manually added or substituted for a synonym. There is also a chance that a word exists as an instance, but is not an instance of a needed concept. This can lead to incorrect classification of jokes and non-jokes. The word will then be added as an instance to the needed concept with low degree of membership. For tracking purposes, such modifications can be added as a new version of the ontology. If a text is successfully classified after the additions, the result of Aim 2 for this text will be considered successful, while result of Aim 1 will not.

Future Directions

The long term goal of this research is to computationally recognize and handle intentional ambiguities in text; the focus is on computational humor detection in short, humorous texts. It is expected that the methods developed in this research can be applied to recognition of other ambiguities, intentional and unintentional. Once young children's jokes are recognized, the ontology will be incrementally updated with new knowledge. It is expected that ambiguity in more complex texts will be recognized based on the new knowledge added to the ontology.

Timeline

	Year 1	Year 2	Year 3
Build a DL-based imprecise ontology, containing minimal number of concepts and relationships between them for annotation of young children's texts	**xxxxxx**	**xxx**	
Create a knowledge base (orthographic component) with words containing in children's dictionary	x		
Build a concept hierarchy from noun entries in the orthographic component	xxx		
Annotate a collection of young children's texts using the existing concept hierarchy	x		
Using children's dictionary, build relationships between concepts and map them into the orthographic component	xxx		
Using a collection of children's texts, build relationships between concepts and map them into orthographic component	xx	xx	
Annotate a different collection of young children's texts using the existing ontology		x	
Determine a method for recognition of script overlap and opposition that create jokes and that are based on lexical or phonological ambiguity		**xxx**	**xxxxxx**
Create a phonological component for entries in the orthographic component		x	
Finding words in texts that lead to lexical ambiguity		x	
Finding words in texts that lead to phonological similarity using heuristic function		xx	
Determine if the scripts containing concepts with lexically ambiguous instances or phonologically similar instances lead to jokes			xxxxxx

6 Relation to Other Work in Progress

a. By the Principal Investigator: The field of computational humor has been a research area of the Primary Investigator. It is directly related to his overall research interests in natural language processing and the semantic web. While the proposed investigation is integral to the continuum of research that is being pursued by the PI, the specific investigation proposed do not overlap with other pending-support projects that are ongoing in the laboratory (see Current and Pending Support). They are highly complementary to that ongoing research, however, which makes possible extensive sharing of technology and equipment. It also assures that input and constructive criticism from the personnel working on two ongoing projects will be available to the investigators proposed here.

b. Elsewhere: The focus of the proposed investigation is on an under-explored area of computational humor. The existing research in computational humor is mainly in humor generation (as opposed to recognition and understanding). Some work has been done in statistical recognition of humor. To the best of our knowledge, we are alone in the investigating ontology-based recognition of children's humor. What is being proposed is not duplicative of what is being investigated elsewhere. Given the expected impact of the outcomes, it is anticipated that the interest of other investigators will be sparked by the published results, resulting in a subsequent amplification of productivity in this area of research. At the least, the results are expected to be complementary to what is being done elsewhere and, more likely, to be augmentative.

7 Broader Impacts

The proposed research has broad applications. It will *integrate research and teaching* and *advance discovery while promoting teaching and learning.* The PI heads his university's multi-disciplinary Data and Knowledge Management group. The Education Plan of this group focuses on undergraduate participation in research activities beginning early in their education; the PI's research is a magnet. It has been suggested that joke generation "could help children or second-language learners explore language." [Ritchie, 2004] [McKay, 2002] [O'Mara, 2002] Similarly, wordplay recognition may tackle the same issues using a complimentary approach: A human tries to produce jokes containing wordplay; if the word that the joke is based on is correctly used, the person understands at least two meanings of the word; gaining this knowledge is educationally useful. The deployment of an affinity group model in concert with targeted preparedness activities is expected to lead to *enhanced infrastructure* within the PI's department and college. The *educational infrastructure* will be enhanced by incorporating the proposed research in the Semantic Web courses and in computational intelligence courses in general. This will introduce young people to the excitement and utility of computing research. The intent is to increase their interest in developing a strong background and, for some, to motivate them to pursue careers as computer scientists and computer engineers. The research will also have an important impact on *underrepresented groups*, particularly women, as one of the group's members is a woman who acts as a mentor to undergraduate women who wish to become involved with Computing research. Similarly, the proposed research may become part of the University of Cincinnati Society of Women Engineers Middle School outreach program for girls grades 5 to 8. Likewise, this research will be part of the University of Cincinnati's summer Research Experience for Women Undergraduates. The widespread media attention already received indicates that the research will be a powerful attractor. This research has the potential to influence positively another underrepresented group: rural persons. Some of the students at this university come from rural backgrounds and have had limited exposure to the potential that exists in language-oriented fields. Advancement of such students is expected to have *broad societal impact,* which will be complemented by the positive effects that the research outcomes are expected to have on the society as a whole. The results of the research are expected to enable improvement of text based applications, such as: filtering outgoing emails in corporations that wish to restrict jokes, improving search engines, and word processing software recognition of unintentional jokes. Humor recognition may be used in commercial natural language translation software; often, these applications provide unintentionally humorous (and sometimes meaningless) translation due to incorrect interpretation of ambiguous utterances. Detection of such utterances and recognition of resulting humor will improve translation's accuracy and meaning. Outcomes of both the research and educational endeavors will be *broadly disseminated.* The results will be published in peer-reviewed journals and major conference proceedings; and, all code developed will be made available through open-source mechanisms.