

Building a Human Information Discourse Interface to Uncover Scenario Content

A. Sanfilippo, B. Baddeley, A. J. Cowell, M. L. Gregory, R. Hohimer and S. Tratz

Pacific Northwest National Laboratory

902 Battelle Blvd

Richland, WA 99354

{antonio.sanfilippo; robert.baddeley; andrew.cowell; michelle.gregory; ryan.hohimer; stephen.tratz}@pnl.gov

Keywords: Analysis of Competing Hypotheses, Information Sharing and Collaboration, HUMINT, OSINT, Terrorism

Abstract

Software environments for intelligence analysis need to leverage scenario-based analytical processes to help analysts achieve a more comprehensive view of competing hypotheses in their risk assessments. While current solutions offer some of the tools and functionality required for such advancement, there still is no integrated set of capabilities that addresses the extraction and manipulation of scenario content from unstructured intelligence data sources. We propose to fill this gap through the development of a visual interactive environment for event analytics that integrates text mining, discourse analysis and visualization capabilities to support scenario manipulation and generation processes.

1. Introduction

The ability to support scenario¹ manipulation and generation processes is perhaps the greatest single challenge for today's intelligence analysis systems. Since the end of the Cold War, the concerns of the intelligence community have shifted focus from a single known enemy to an evolving set of networked, hard-to-track hostile groups. This transformation is directing the theory and practice of intelligence analysis toward more complex analytic models, where backtracking loops that help identify gaps, reframe problems, and gather additional evidence are necessary steps to generate actionable intelligence. Because of this shift, there has been increased emphasis on intelligence analysis techniques such as the Analysis of Competing Hypotheses (ACH) (Heuer, 1999). ACH requires identification, investigation, and simultaneous evaluation of multiple, alternative hypotheses and is highly instrumental in avoiding premature commitment

¹ Following Heuer (1999, p. 156), we regard a scenario as a series of events linked together in a narrative description leading to an anticipated outcome.

to a single expected outcome with consequent neglect of relevant evidence relative to other plausible outcomes.

Analysis techniques such as ACH are difficult to perform without computational aid. Because of memory limitations on human cognition (Miller, 1956), most people are simply unable to retain several hypotheses and relevant supporting facts in working memory. Moreover, supporting information for ACH needs to be distilled from potentially huge repositories of classified and unclassified documents, and properly vetted to ensure the privacy of individuals is respected. Such a task would require extravagant expenditure of human resources without the help of machine-aided information management and extraction processes.

Current intelligence analysis tools provide some of the pieces needed to support scenario-based processes such as ACH. For example, link analysis tools (e.g., Analysts's Notebook²) allow users to build scenarios manually (or automatically from structured data) and display graphical views of the timeline and network information encoded. Chappell *et al.* (2004) describe a system that allows users to extract facts from document sets automatically, provide timeline and network visualization of the information extracted and makes available an environment that supports hypothesis construction. Fikes *et al.* (2005) discuss how this system can be extended to handle alternative hypotheses through (1) the use of "contexts" in which each hypothesis is developed independently and (2) the determination of relationships

This manuscript has been authored by Pacific Northwest National Laboratory under Contract No. DE-AC06-76RL01830 Modification M375 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting this article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

² http://www.i2inc.com/Products/Analysts_Notebook.

across contexts to detect incompatibilities and common features among alternative hypotheses. However, no analytic tool is available yet that enables the automatic retrieval of facts supporting, contradicting, or originating a given event (Marcu & Echiabi, 2002) to sustain analysis techniques such as ACH. In addition, large event network displays tend to be unmanageable unless effective ways of condensing and compacting nodes that encode compatible information are available. Graph summarization techniques that are semantically driven are therefore necessary to distill the most significant patterns from large event networks. Such functionality is not yet available in link analysis tools or similar toolkits.

The goal of this paper is to describe a system designed to address these gaps via a visual interactive environment for event analytics that integrates text mining, ontological annotation, discourse analysis, and visualization capabilities to support scenario generation, exploration, and manipulation processes.

2. Characterizing Scenario Content

In developing an environment that supports scenario generation, exploration, and manipulation, our first objective is to construct an ontology that enables the specification of a scenario typology with associated parameterizations of content and activity, in terms of event structure, temporal relations and rhetorical structure. More specifically, we are developing a Scenario Content Ontology (SCO) that provides the building blocks and basic templates that analysts can use to build their own scenarios interactively. An example of scenario content characterization with SCO is shown in Figure 1, where ovals indicate ontology classes and boxes class instantiations; straight lines denote inheritance; dotted lines point to scenario content components; and dotted arrows designate properties of or relations between scenario subcomponents.

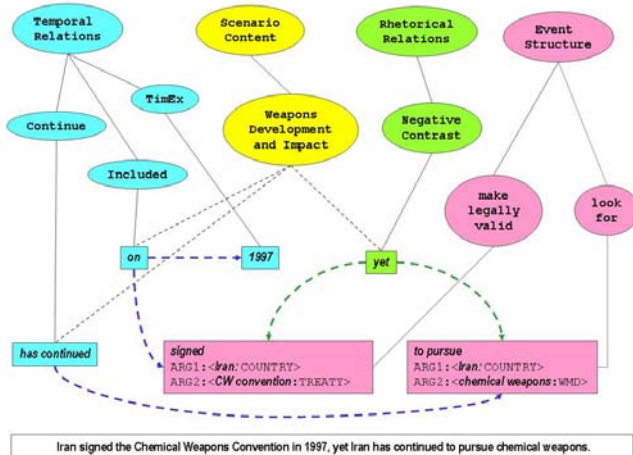


Figure 1: Sample scenario content characterization with SCO.

SCO is being developed as an OWL ontology³ using Protégé as the ontology editor environments⁴ and Jena⁵ as the semantic web framework in which to implement the ontology, handle reification, issue queries, and perform logical inference. Event structures are verbal frames, such as *signed* and *pursue* in Figure 1, which include verbs plus their arguments (the event participants). In our system, we use the verb concepts in WordNet⁶ to define events and make use of argument structure frames, such as expressed in VerbNet (Kipper *et al.*, 2000) and FrameNet (Ruppenhofer *et al.*, 2002) to define event structures.

One of the innovations of our system is the use of an event ontology to reduce a set of specific event structures (e.g., gesticulate, grimace, talk, write, telecommunicate) to a single and more general event structure (e.g., communicate). The event ontology was constructed by defining a selection of verb synonym sets in the WordNet database as event classes in SCO. Verb synonym sets that were less specific in meaning (e.g., *communicate* and *intercommunicate* vs. *gesticulate* and *gesture*) were chosen as event classes. In doing so, we chose the more frequent member of the synonym set (e.g., *communicate* for the *intercommunicate* synonym set) to name the class. The verbs in the synonym sets chosen as event classes (e.g., *communicate*, *intercommunicate*) as well as their troponyms (e.g., *gesticulate*, *gesture*, *motion*; *grimace*, *make a face*, *pull a face*) were declared as instances. An example of the SCO event ontology is shown in Figure 2, where verb senses associated with the folder icon indicate event classes while those associated with a bullet point are instances.

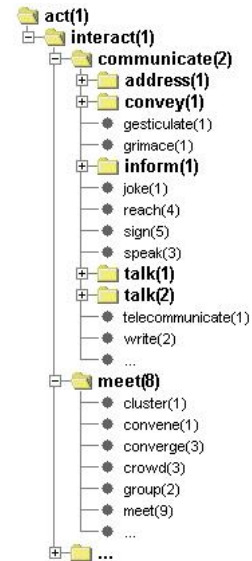


Figure 2: SCO event ontology fragment.

³ <http://www.w3.org/TR/owl-ref>.

⁴ <http://protege.stanford.edu>.

⁵ <http://jena.sourceforge.net>.

⁶ <http://www.cogsci.princeton.edu/~wn>.

To assess the specificity level of synonym sets, we used the notion of information content, as described in Resnik (1995). Synonym sets whose information content was below a given threshold were chosen as event classes because they designated more general event concepts. Following this method, we created 690 event classes out of a total of 24,632 verb synonym sets. These event classes are used by the system we have been developing to give users the option of contracting event structures into a more manageable space, as shown in Figure 6.

A second innovation in our system is the use of discourse connectives as an additional organizational layer for events. Discourse connectives, such as *and*, *if*, *but*, etc., are identified in the documents. Each connective represents a rhetorical relation between two events. For example, in the following utterance, two events (underlined) are connected by *however*.

*Iraq used its ballistic missiles as weapons of terror against Saudi Arabia and Israel. **However**, Iraq did not use its SCUDS with chemical or biological ware fare.*

The use of *however* indicates that the two events are in contrast to one another. Discourse connectives and the associated events were identified according the conventions of the Penn Discourse Treebank guidelines (Miltsakaki *et al.*, 2004). As for events, discourse connectives can be arranged into an inheritance hierarchy in which more general rhetorical relations (e.g., *contrast*) subsume sets of semantically similar rhetorical relations (e.g., *but*, *however*, *yet*) to offer users the option of summarizing relationships between event structures via generalization. To implement this capability, we created an ontology of rhetorical relations, based on the scheme proposed by Marcu & Echiabi (2002). The classes for this ontology include rhetorical relation concepts such as *contrast*, *concession*, and *elaboration*; each class (e.g., *contrast*) is linked to a set of connectives (e.g., *but*, *however*, *yet*). Using the SCO ontology of rhetorical relations, we can, for example, find all events that are in contrast with a particular event of interest in a document set.

We also plan to include Temporal Relations in SCO, using TimeML⁷ as modeling guidelines.

3. Conceptual Design

The underlying design criteria for the environment we are developing is based on the analytical processes used by individuals in the intelligence community. We utilized previous work, funded under ARDA's NIMD program, to refine our development approach (Chappell *et al.*, 2004). This prior effort used a three-phase approach to investigate how best to introduce novel analytical tools into the intelligence analysts' workflow. First, in order to refine our understanding of the intelligence analysis domain, we looked to the literature. Access to working intelligence

analysts early in a development lifecycle is challenging, so it is essential to have an understanding of the domain instead of wasting valuable time having them define it. Based on our review, we were able to identify a number of essential core references (Heuer, 1997; Dearth & Goodden, 1995; Krizan, 1999). The literature provided a foundation for understanding the domain and the challenges faced, in addition to giving us some indication of where a scenario-based solution might best fit.

Second, we performed an investigation into the formal task-based processes performed by various intelligence agencies in preparing their intelligence products (Thurman *et al.*, 2003). Our collapsed hierarchical "taskonomy," built from reducing replication across the reviewed models, contained over a hundred steps, detailing to a low level the individual tasks performed. This provided a means for us to determine at what stages of the analytical workflow our suite of technologies may provide the most value.

Finally, we were able to use the Advanced Research and Development Activity (ARDA) Novel Intelligence from Massive Data (NIMD)⁸ funded Glass Box Analytical Environment (GBAE)⁹ to see exactly how the analytical tradecraft is performed by working Battelle analysts. The GBAE recorded and time stamped all activities that were performed for a number of stereotypical taskings, including web browser and query events, active applications and window locations onscreen, files saved, text copied/pasted into documents, detailed keyboard/mouse data, and screens captured at a one-second rate. Based on observing the analytical process first hand through the Glass Box, we were able to identify elements of the analytical workflow that was receptive to the scenario discourse process and that could provide utility to the analyst.

From this work, we designed an analytical environment that attempts to mimic certain aspects of the analytic workflow. Our first screen, the "Document Space" gives the analyst the opportunity to interact with a document corpus of interest. This may be a collection of documents saved locally or potentially an interface to remote databases or even the internet. This screen is meant to replicate the collection process in the analytical workflow. The purpose of our second (middle) screen, the "Evidence Marshalling Space," is to allow analysts to perform their tradecraft in a highly flexible environment, where they can look at the entities, relationships, and events expressed within a selected set of documents and are able to manipulate them, move them around, and view them from alternative angles to really understand what is happening. It is the sandbox within which the analyst performs the analytical tradecraft. The final screen gives the analysts a natural interface to structure their analytical product. In the "Hypothesis Screen," analysts build up their arguments, supported by the docu-

⁷ www.cs.brandeis.edu/~jamesp/arda/time.

⁸ http://www.ic-arda.org/Novel_Intelligence.

⁹ http://www.ic-arda.org/Novel_Intelligence/glass_box.htm.

ments and analysis performed in the previous two screens. It is important to note that this workflow is not necessarily linear: movement back and forth across each of the screens is expected during a tasking. Users can investigate an information path, retreat and follow a new path, while constantly updating and reorganizing their hypothesis graph to create a final appreciation of the analytical results.

3.1 Support tools

Analysts can interface directly with the information contained in a large number of documents without having to read the individual documents or rely on summaries (although access to such information is always available) because our system is driven by the notion that information has underlying structure, independent of the documents in which they are found. As such, events and entities can be extracted from documents and organized in a way that adheres to the analyst’s cognitive expectations.

We extracted event structures (verbs and their arguments), discourse information on how the events were related, named entities, and other key concepts from a set of documents collected for the ARDA Metrics Workshop.¹⁰ These structures serve as the basis of the organization of information in our system, which enables analysts to visualize the relationships between event structures in multiple documents, independent of their origin.

4. Implementation

The conceptual design is realized through three “spaces”—each corresponding to a step in the analytical cycle—represented on individual monitors, as shown in Figure 3.



Figure 3: The Analytical Environment

4.1 The Document Space

The first screen in Figure 3 represents the “Document Space,” which corresponds to the collection phase of the analytic cycle. This is where analysts can investigate their own personal document collection and decide which documents they wish to use. In IN-SpireTM,¹¹ documents

are clustered based on their content and word frequency, as shown in Figure 4. A two-dimensional Galaxy view of all the documents is created, with points representing each document and clusters showing how those documents are related. The analyst may interact with the document collection at this level, zooming in and out to get a better appreciation of what documents are included and their main themes. Additional mechanisms can be used to select documents, including full-text search and “query by example.” When ready, the analyst may select a subset of the documents for further analysis in the “Evidence Marshalling Space.”

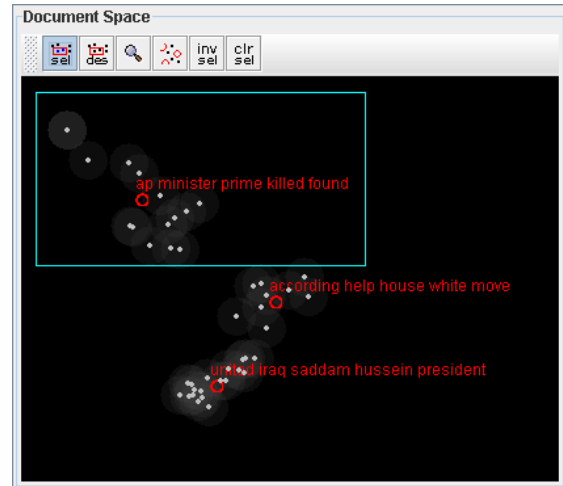


Figure 4: The Document Space

4.2 The Evidence Marshalling Space

As a set of documents is selected, the analyst uses the Evidence Marshalling Space, represented on the second (middle) screen, to interact directly with the information contained in the documents. This space is a sandbox where the analyst can look at the events, connectives, and entities to understand causal patterns, opposing views, and occurrences of named entities such as people and places. Using this window, users define the level of granularity desired for investigating the information.

Using the toolbar, the user can select documents and display salient entities and events contained in them. Entities and events are represented in the form of a link-node graph, as shown in Figure 5. At this stage, the user can abstract away from individual events by generalizing over subsets of events using the inheritance relationships and class-instance links encoded in the event ontology, as shown in Figure 6. Each event class (e.g., *communicate*, *react*, *interact* in Figure 6) subsumes a number of events that occur as instances or subclasses of the event class in the ontology. Users can choose to cluster events by class or to display only event classes to have a bird’s-eye view of the entire Evidence Marshalling Space, and then narrow their investigation by focusing on the entities.

¹⁰ <http://www.ic-arda.org/index.html>.

¹¹ <http://in-spire.pnl.gov>.

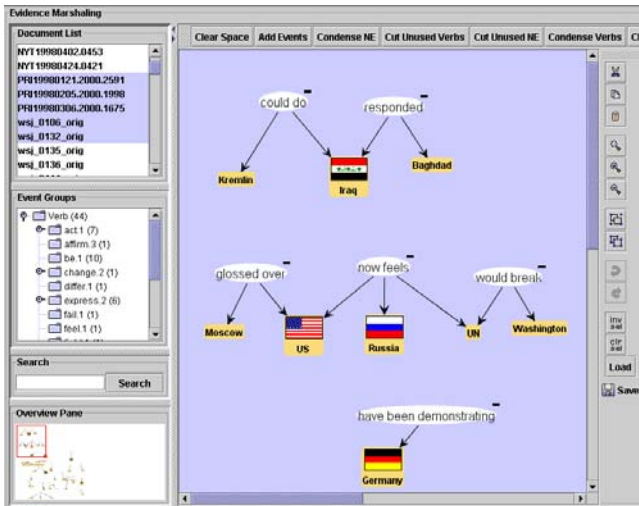


Figure 5: The Evidence Marshalling Space

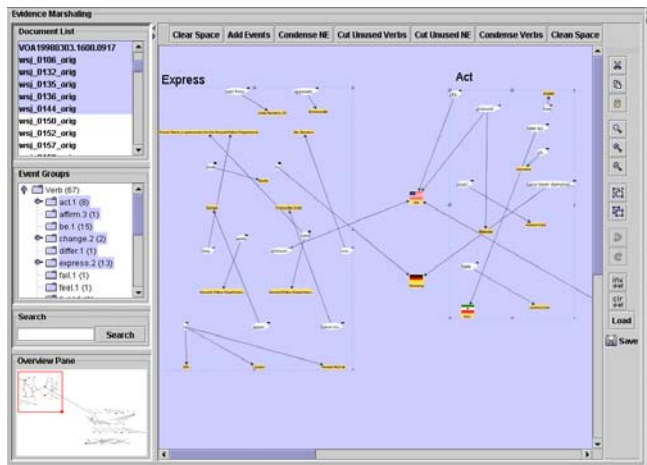


Figure 6: The verb ontology used to cluster event types into more general classes.

Relationships between individual events can be elucidated through the use of rhetorical relations. As for events, we used rhetorical relation classes (e.g., contrast) to generalize over sets of relationships between events (e.g., *but*, *however*, *yet*) exploiting the inheritance relationships and class-instance links encoded in the rhetorical relation ontology. Figure 7 provides an abstract representation of how the entity, event, and rhetorical relation layers are interleaved in terms of class inheritance and class-to-instance relationships in the SCO ontology.

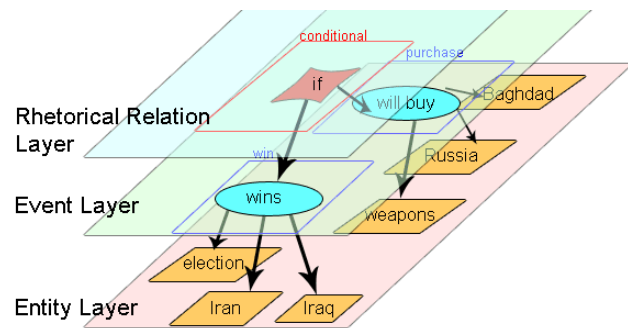


Figure 7: Conceptual overlays

4.3 The Hypothesis Space

As analysts perform their analytical tradecraft, investigating the myriad of entities, relationships, and events in the level of granularity appropriate for their task, they can begin to formulate their intelligence product using the “Hypothesis Space.” As schematically shown in Figure 8, the Hypothesis Space offers an environment in which analysts can build a graph of their findings and attach material (documents, paragraphs, or individual sentences) that support or refute the assertions under investigation. These actions underline the dynamic nature of the environment. Items from the other screens can be dragged and dropped from other screens to make these linkages. The Hypothesis Space serves as the basis from which analysts construct their intelligence product.

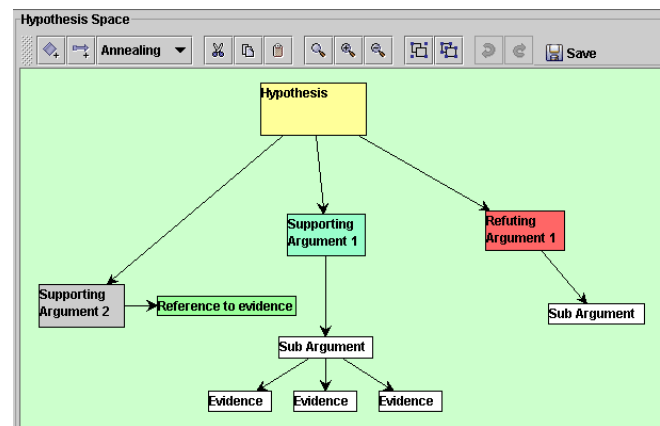


Figure 8: The Hypothesis Space

5. Discussion

We have created an interactive environment that supports scenario-based analysis of competing hypotheses by distilling coherent discourse structures with selectable levels of specificity from the glut of information contained in the hundreds of document resulting from a single query. The application presented here is not intended to supplant the tools that analysts (and other users) use to gather, synthesize, and present information. What we have provided is a tool that allows analysts to interact directly with the information found in multiple documents by

choosing the topics and events that are of particular interest, query in detail, store information, and begin a new query path. The system presented here incorporates some standard tools, such as document clustering and named-entity recognition, as well as new tools. The innovations we present here are in both the technical arena and the user interface. The innovative technical aspects include the creation of an event ontology based on classes defined from WordNet. Additionally, we have created an ontology of rhetorical relations that can help define relationships between events or classes of events.

Informed by the cognitive process that analysts use to gather and disseminate intelligence information, our system design is innovative in that it mirrors the process that analysts use to digest and construct intelligence information. In order to validate our design concepts, we will soon start performing usability evaluations with Battelle intelligence analysts. A two-stage process shall be used. Before exposing the analysts to our environment, we shall perform a heuristic evaluation of the interface to ensure consistency, flexibility, and efficiency of use. Next, analysts shall interact with the environment in order complete predefined tasks. Observation, protocol analysis, and questionnaires shall be used to investigate their perceptions of the system. Any inadequacies found in the system shall be redressed through a process of re-design and reimplementation, as part of our iterative design cycle.

Acknowledgments

The work described in this paper was developed within the context of the *National Visualization and Analytics Center*[™] (NVAC[™])—one of several research thrusts in the *Threat Vulnerability, Testing and Assessment* program at the Department of Homeland Security (DHS) Science and Technology Directorate. We would like to thank Alan Turner and Jim Thomas for continuous feedback and support, Sandy Landsberg for championing this work, and Dave Shepherd, Frank Greitzer, and Dave Thurman for helpful comments on previous drafts of this paper.

References

Chappell, A. R., Cowell, A. J., Thurman, D. A., and Thomson, J. R. 2004. Supporting Mutual Understanding in a Visual Dialogue between Analyst and Computer. *Proceedings of the 48th Annual Meeting of the Human Factors and Ergonomics Society*, New Orleans, LA.

Dearth, D., Goodden, R editors. 1995. *Strategic Intelligence: Theory and Application*, 2nd ed., Washington, DC: Joint Military Intelligence College.

Fikes R., Ferrucci, D., and Thurman D. 2005. Knowledge Associates for Novel Intelligence (KANI). To appear in *Proceedings of the 2005 International Conference on Intelligence Analysis*, McLean, VA.

Heuer, Richard J., Jr. 1999. *Psychology of Intelligence Analysis*. Center for the Study of Intelligence, Central Intelligence Agency, Washington, DC.

Kipper K., Dang H., Palmer, M. 2000. Class-Based Construction of a Verb Lexicon. *Proceedings of the Seventeenth National Conference on Artificial Intelligence*, Austin, TX.

Krizan, L. 1999. *Intelligence Essentials for Everyone* Joint Military Intelligence College: Washington, DC.

Marcu, D. and Echihiabi, A. 2002. An Unsupervised Approach to Recognizing Discourse Relations. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002)*, Philadelphia, PA, July 7-12.

Merlino, A. and Maybury, M. 1999. An Empirical Study of the Optimal Presentation of Multimedia Summaries of Broadcast News. Mani, I. and Maybury, M. (eds.) *Automated Text Summarization*. MIT Press. pp. 391-401.

Miller, G. 1956. The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *The Psychological Review*, 1956, vol. 63, pp. 81-9.

Miltsakaki, Eleni, Rashmi Prasad, Aravind Joshi and Bonnie Webber. 2004. The Penn Discourse TreeBank. *Proceedings of the Language Resources and Evaluation. Lisbon, Portugal*. 2004.

Resnik, P. 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Vol. 1,448-453, Montreal, August 1995.

Ruppenhofer, Josef, Collin F. Baker and Charles J. Fillmore. 2002. The FrameNet Database and Software Tools. Braasch, Anna and Claus Povlsen (eds.), *Proceedings of the Tenth Euralex International Congress*. Copenhagen, Denmark. Vol. I: 371-375.

Thurman, D. A., Cowell, A. J., Andrew, A. H. & Chappell, A. R. 2003. Human-information Interaction with Knowledge Associates. *Proceedings of the 47th Annual Meeting of the Human Factors and Ergonomics Society*, Denver, CO.