



# **Phrasing Technologies, Applications & Challenges**

**Elizabeth D. Liddy, Ph.D.**

**Director, Center for Natural Language Processing  
Professor, School of Information Studies  
Syracuse University**

**June 28, 2001**

---

**Center for NLP**

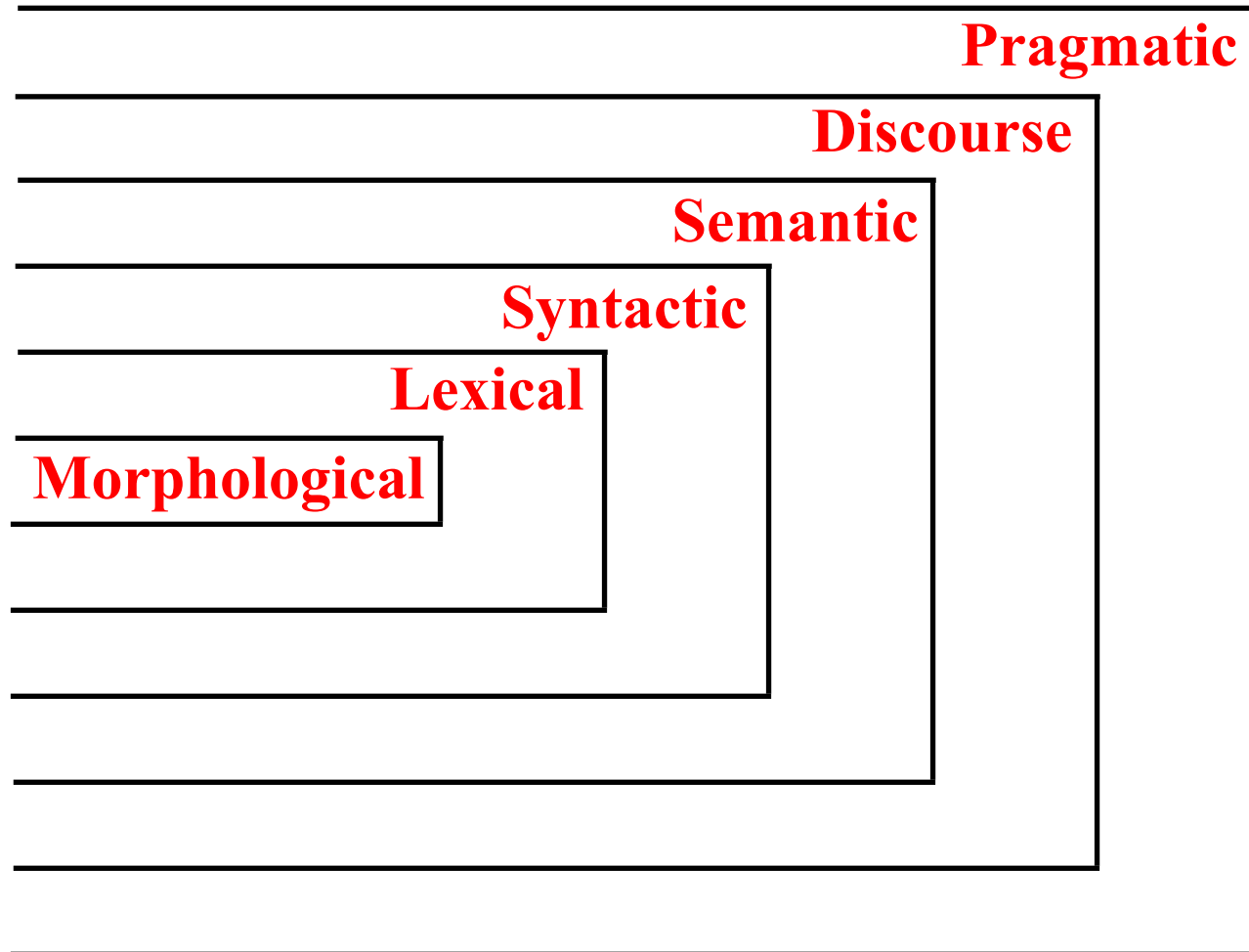


# Overview of CNLP's Approach

---

- Automatically identify and extract **events** involving **entities** and the complex **relations** amongst them
- Using **all** levels of Natural Language Processing
- General technology capability which has been successfully specialized for a wide range of domains and used in a range of applications

# Levels of Language Understanding





# Types of Phrases

---

- **Minimal noun phrases**
- **Maximal noun phrases**
- **Collocations**
- **Co-occurrences**
- **Proper Noun phrases**
- **Verb phrases**



# Applications

---

- Document representation / indexing
- Query representation & expansion
- Automatic summarization
- Results browsing
- Focus group / interview transcript analysis
- Input to visualization tools
- Metadata generation
- Building / adding to Knowledge Bases for use by human & automated reasoners



# Applications

---

- Document representation / indexing
- Query representation & expansion
- Automatic summarization
- Results browsing
- Focus group / interview transcript analysis
- Input to visualization tools
- Metadata generation
- Building / adding to Knowledge Bases for use by human & automated reasoners



# Document Processing

---

*03/14/1999 (AFP)*... the extremist Harkatul Jihad group,  
reportedly backed by Saudi dissident Osama bin Laden ...



# Document Processing

---

*03/14/1999 (AFP)*... the extremist Harkatul Jihad group,  
reportedly backed by Saudi dissident Osama bin Laden ...

... the|**DT** extremist|**JJ** Harkatul|**NP** Jihad|**NP** group|**NN** ,|,  
reportedly|**RB** backed|**VBD** by|**IN** Saudi|**NP** dissident|**NN**  
Osama|**NP** bin|**VB** Laden|**NP** ...

# Document Processing

---

*03/14/1999 (AFP)*... the extremist Harkatul Jihad group,  
reportedly backed by Saudi dissident Osama bin Laden ...

... the|DT extremist|JJ Harkatul|NP Jihad|NP group|NN ,|,  
reportedly|RB backed|VBD by|IN Saudi|NP dissident|NN  
Osama|NP bin|VB Laden|NP ...

... the|DT extremist|JJ **<entity>** Harkatul\_Jihad\_group**</entity>** ,|,  
reportedly|RB backed|VBD by|IN **<entity>** Saudi\_dissident  
**</entity>** **<entity>** Osama\_bin\_Laden **</entity>** ...

# Document Processing

---

03/14/1999 (AFP)... the extremist Harkatul Jihad group, reportedly backed by Saudi dissident Osama bin Laden ...

... the|DT extremist|JJ Harkatul|NP Jihad|NP group|NN ,|, reportedly|RB backed|VBD by|IN Saudi|NP dissident|NN Osama|NP bin|VB Laden|NP ...

... the|DT extremist|JJ *<entity> ref=1; type=terrorist group; Harkatul\_Jihad\_group </entity>* ,|, reportedly|RB backed|VBD by|IN *<entity> ref=2; type=nationality Saudi\_dissident </entity>* *<entity> ref=3; type=person; Osama\_bin\_Laden </entity>* ...

# Document Processing

---

03/14/1999 (AFP)... the extremist Harkatul Jihad group, reportedly backed by Saudi dissident Osama bin Laden ...

... the|DT extremist|JJ Harkatul|NP Jihad|NP group|NN ,|, reportedly|RB backed|VBD by|IN Saudi|NP dissident|NN Osama|NP bin|VB Laden|NP ...

... the|DT extremist|JJ <entity> ref=1; type=*terrorist group*; Harkatul\_Jihad\_group </entity> ,|, reportedly|RB backed|VBD by|IN <entity> ref=2; type=*nationality* Saudi\_dissident </entity> <entity> ref=3; type=*person*; Osama\_bin\_Laden </entity> ...

# Document Processing

---

03/14/1999 (AFP)... the extremist Harkatul Jihad group, reportedly backed by Saudi dissident Osama bin Laden ...

... the|DT extremist|JJ Harkatul|NP Jihad|NP group|NN ,|, reportedly|RB backed|VBD by|IN Saudi|NP dissident|NN Osama|NP bin|VB Laden|NP ...

... the|DT extremist|JJ <entity> ref=1; type=*terrorist group*; Harkatul\_Jihad\_group </entity> ,|, reportedly|RB backed|VBD by|IN <entity> ref=2; type=*nationality* Saudi\_dissident </entity> <entity> ref=3; type=*person*; Osama\_bin\_Laden </entity> ...

# Document Processing

---

03/14/1999 (AFP)... the extremist Harkatul Jihad group, reportedly backed by Saudi dissident Osama bin Laden ...

... the|DT extremist|JJ Harkatul|NP Jihad|NP group|NN ,|, reportedly|RB backed|VBD by|IN Saudi|NP dissident|NN Osama|NP bin|VB Laden|NP ...

... the|DT extremist|JJ <entity> ref=1; type=*terrorist group*; Harkatul\_Jihad\_group </entity> ,|, <EV = SUPPORT> reportedly|RB backed|VBD\_by|IN</EV> <entity> ref=2; type=*nationality* Saudi\_dissident </entity> <entity> ref=3; type=*person*; Osama\_bin\_Laden </entity> ...



# Applications

---

- Document representation / indexing
- Query representation & expansion
- Automatic summarization
- Results browsing
- Focus group / interview transcript analysis
- Input to visualization tools
- Metadata generation
- Building / adding to Knowledge Bases for use by human & automated reasoners



# Query Representation

---

1. “I would like information about indictments against Bosnian war criminals.”

*indictment\* +Bosnian “war\_criminal”*

# Query Representation

---

1. “I would like information about indictments against Bosnian war criminals.”

*indictment\* +Bosnian “war\_criminal”*

2. “I would like information about efforts to bring suspects of Lockerbie bombing to trial.”

*effort\* bring\* suspect\* +”Lockerbie\_bomb\*”  
trial\**



# Query Representation

---

**3. “What Iranian backed terrorist groups exist in the Middle East region?”**

*Iran\* AND back\* AND “terrorist\_group” AND “Middle\_East”*

# Query Representation

---

3. “What Iranian backed terrorist groups exist in the Middle East region?”

*Iran\* AND back\* AND “terrorist\_group” AND “Middle\_East”*

4. “Are there pro Iranian or Islamic fundamentalist terrorist groups within Saudi Arabia?”

*(“pro\_Iran” OR (Islam\* AND fundamental\*)) AND “terrorist\_group” AND “Saudi\_Arabia”*



# Query Phrase expansion

- **Complex Nominals**
  - [Adj\*] [N\*]
  - where adjectives are of semantic class of non-predicating adjectives
    - *Electrical engineer*
    - Vs.
    - *Unhappy engineer*



## Query Phrase expansion (cont'd)

- **Words which co-occur frequently with same heads or same modifiers**
- **Useful for expansion to synonymous phrases**
  - *foreign language -> non-native language*
  - *anticipated demand -> forecast demand*
  - *language software -> communication software*



# Applications

---

- Document representation / indexing
- Query representation & expansion
- **Results browsing**
- Automatic summarization
- Focus group / interview transcript analysis
- Input to visualization tools
- Metadata generation
- Building / adding to Knowledge Bases for use by human & automated reasoners

**Controls:**

[Change View](#)

[Close Visualizer](#)

**Tips:** This folder categorizes document information by topic.

The topic categories displayed are based on the results of the selected query.

The total number of topics for each category appears next to the category title.

Narrow the query results by **selecting** up to four topics in each category.

As you select topics, each category changes to indicate only the topics that appear in the documents that match the current selection.

- Subject Areas (18)**
- Electricity/Electronics (21)**
  - Mechanical Devices (17)**
  - Business Practices (16)**
  - United States (10)**
  - Commerce/Trade (8)**
  - Automobile Industry (6)**

- City (31)**
- San Francisco (8)**
  - San Diego (6)**
  - District of Columbia (5)**
  - Los Angeles (5)**
  - Montreal (5)**
  - New York City (5)**

- Noun Phrases (50)**
- electric vehicle industry (20)**
  - electric vehicles (20)**
  - electric cars (11)**
  - natural gas (6)**
  - air quality (5)**
  - clean air (5)**

- Company (32)**
- General Motors (11)**
  - BC Research (3)**
  - Daimler-Benz AG (3)**
  - Minnesota Mining and Manufacturing (3)**
  - Norvik Traction (3)**
  - U.S. Electricar (3)**

Sum	Full	Rank	Headlines
		1.	Iacocca Selects Unique Mobility CEO for EVG Board
		2.	Advances in electric cars propel clean-air promise
		3.	Statement from Electric Vehicle Association of the Americas on General Motors Dedication Today of First Electric Vehicle Manufacturing Plant
		4.	Statement from Electric Vehicle Association of the Americas on General Motors



# Applications

---

- Document representation / indexing
- Query representation & expansion
- Results browsing
- **Automatic summarization**
- Focus group / interview transcript analysis
- Input to visualization tools
- Metadata generation
- Building / adding to Knowledge Bases for use by human & automated reasoners

## Product - Automatic NLP Indexing of One Document

*Headline:* **Politics & Policy: Lake Quits Fight to Get CIA Post**

*Source:* **The Wall Street Journal (3/18/97)**

*Key Concepts:*

**bipartisan committee; campaign contributors; confirmation hearings; congressional campaigns; continuing dispute; controversial political donor; endless delays; illegal campaign contributions; political circus; presidential campaigns**

*Proper Names:*

***U.S. Fed/Legislative:* Congress; Senate Intelligence Committee**

***U.S. Fed/Executive:* Dept. of Energy; Dept. of Justice; FBI; NSC;  
White House**

***U.S. Fed/Ind Org.:* CIA**

***Political Org.:* Democratic Party; Republican Party**

***Person:* Clinton (President); Donald Fowler (Chairman); Sheila Heslin; Bob Kerrey (Senator); Anthony Lake; Thomas McLarty; David Rogers; Richard Shelby (Chairman); Roger Tamraz**

*Subject Fields:*

**Governmental Institutions; Strategy/Tactics; Elections/Campaigns**



# Applications

---

- Document representation / indexing
- Query representation & expansion
- Automatic summarization
- Results browsing
- **Focus group / interview transcript analysis**
- Input to visualization tools
- Metadata generation
- Building / adding to Knowledge Bases for use by human & automated reasoners



# Focus Group / transcript analysis

---

- **Applied increasing levels of NLP to produce feature sets of consumer dialogues**
- **Goal was to understand differences between groups - OR – different views of same subject before and after use of a product**
- **To target advertising based on improved understanding of the needs of customers and perceived benefits**



# Data – Characteristics

---

- **Typical of transcripts**
  1. Many short incomplete sentences
  2. Many long run-on, train-of-thought sentences
  3. Internal sentence punctuation, e.g. commas, often missing
  4. Words often misspelled (homonyms, like “your” for “you’re”)
  5. Missing words in the middle of sentences due to transcription difficulties
  6. Fair amount of “filler” like “Erm”, “erm”
- **Resulted in our training a new POS tagger**



# 1. Collocations

---

- **Two (or more) consecutive words that occur together more frequently than chance would allow**
- **May have compositional or non-compositional meaning**
  - *High chair*
  - *Skunk works*
- **Many different methods / algorithms available**
- **No generally agreed upon best method**



# Collocation Methods

---

## 1. Frequency

- Of POS tagged words
- Does not take chance into account

## 2. Hypothesis testing (t-test or chi-square)

- Rules out chance co-occurrence of high-frequency words
- Useful in finding collocations that best distinguish two similar terms (i.e. *strong* and *powerful*)

## 3. Mutual Information

- Pointwise Mutual Information
  - $MI = \log (p(x,y)/p(x)*p(y))$
- Proven best for teasing out chance



# Collocation Methods (cont'd)

---

## Mutual information

- Typically used to find associations between **2** words
- We tested extending it to collocations of **3, 4 & 5** words
  - Modified the pointwise MI formula
  - 3-word modification gave reasonable results
  - 4 & 5 word modification did **not**



## 2. Co-occurrences

---

**Co-occurrence: Two (or more) words that frequently occur together**

- not necessarily contiguous
- intervening words may separate them  
(ex: *knock, door*)

**Reviewed and tested potentially appropriate algorithms**

- Hidden Markov Model
- “Jeffrey’s Rule” (Bayesian learning)
- Adjusted Mutual Information

# 3. Minimal noun phrases

---

## Based on POS tagged output

- Potential elements of minimal noun phrase are:
  - Proper nouns – NP
  - Common nouns – NN, NNS
  - Apostrophe/Possessive – POS
  - Adjectives – JJ, JJR, JJS
  - Other – CD, DT, PDT, PRP\$
- Sequence must contain at least one of (NP, NN or NNS)
- Little attention paid to order of tags, except that DT must occur first, if present

### 3. Minimal noun phrases (cont'd)

---

1. a|DT\_Neutralia|NP\_Garnier|NP\_dermo|NN  
\_protection|NN\_healthy|JJ\_hair|NN\_shamp  
oo|NN
  - *a Neutralia Garnier dermo protection  
healthy hair shampoo*
2. long|JJ\_straight|JJ\_glossy|JJ\_hair|NN
  - *long straight glossy hair*
3. the|DT\_long|JJ\_glossy|JJ\_healthy|JJ\_lookin  
g|NN\_hair|NN
  - *the long glossy healthy looking hair*



## 4. Maximal noun phrases

---

- **Maximal noun phrases are built around, but larger than, minimal noun phrases**
- **Maximal noun phrases include one or more of the following tags:**
  - Preposition
  - Conjunction
  - Subordinating conjunction
  - Gerund
- **Each pose difficult parsing problems, but promise to be quite rich and rewarding for fuller phrases**

## 4. Maximal noun phrases (cont.)

---

### Extracted some good phrases

1. < organics|NP\_for|IN\_fine|JJ\_and|CC\_lifeless  
|JJ\_hair|NN />  
– *Organics for fine and lifeless hair*
2. < five|CD\_different|JJ\_set|NNS\_of|IN  
shampoo|NNS\_and|CC\_conditioner|NNS />  
– *Five different sets of shampoo and  
conditioner*

### Extracted problem phrases as well!

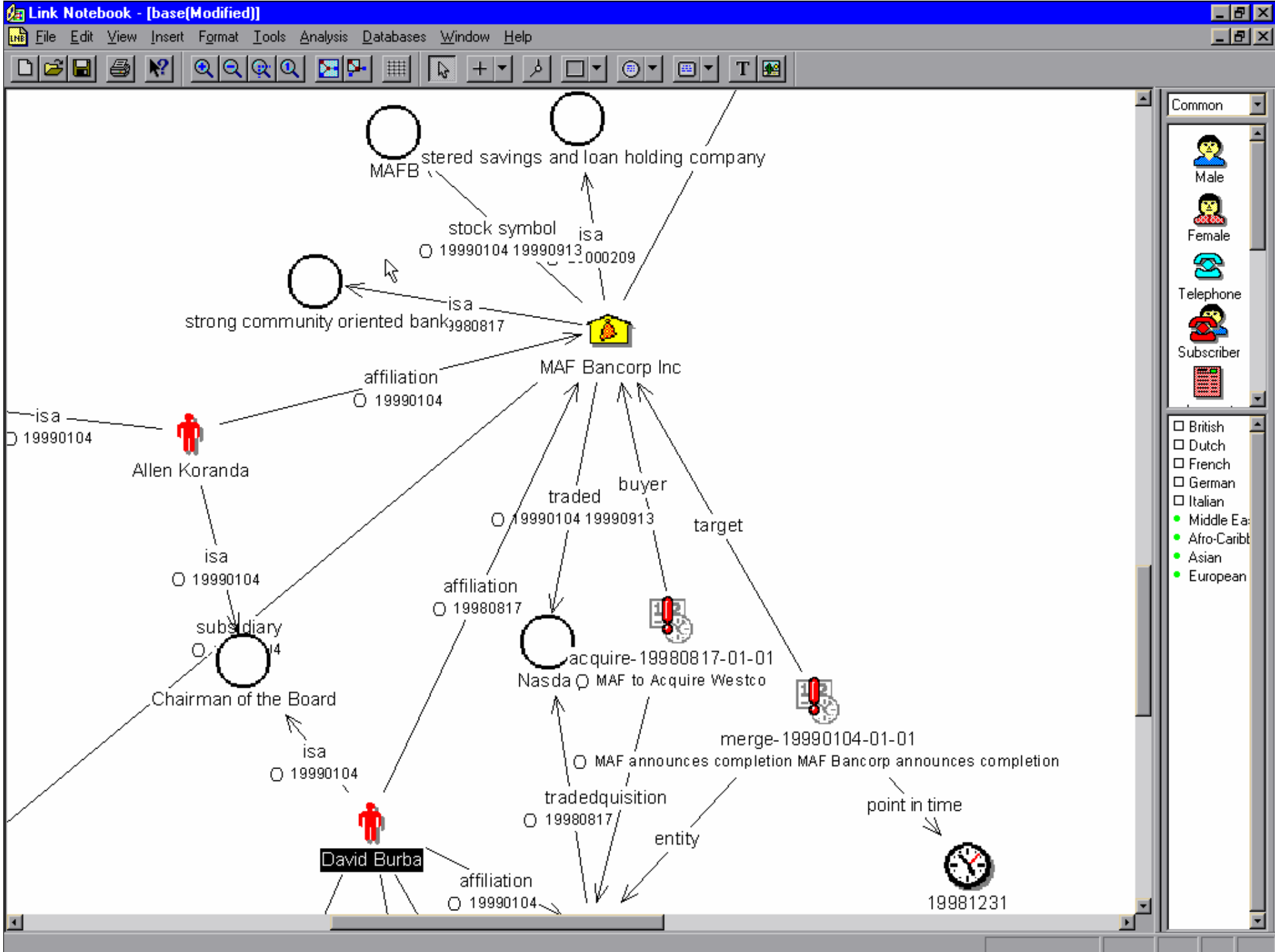


# Applications

---

- Document representation / indexing
- Query representation & expansion
- Automatic summarization
- Results browsing
- Focus group / interview transcript analysis
- **Input to visualization tools**
- Metadata generation
- Building / adding to Knowledge Bases for use by human & automated reasoners







# Applications

---

- Document representation / indexing
- Query representation & expansion
- Automatic summarization
- Results browsing
- Focus group / interview transcript analysis
- Input to visualization tools
- **Metadata generation**
- Building / adding to Knowledge Bases for use by human & automated reasoners

# Educational Resource Example

## ***Stream Channel Erosion Activity***

### **Student/Teacher Background Information:**

Rivers and streams form the channels in which they flow. A river channel is formed by the quantity of water and debris that is carried by the water in it. The water carves and maintains the conduit containing it. Thus, the channel is self-adjusting. If the volume of water, or amount of debris is changed, the channel adjusts to the new set of conditions.

...

### **Student Objectives:**

The student will discuss stream sedimentation that occurred in the Grand Canyon as a result of the controlled release from Glen Canyon Dam.

...

# Metadata generation

**Title (SIE):** Grand Canyon: Flood!  
- Stream Channel Erosion Activity

**Grade Levels (SIE):** 6, 7, 8

**GEM Subjects (TC):** Science--Geology  
Mathematics--Geometry  
Mathematics--Measurement  
Science--Process Skills  
Science--Instructional Issues

**Keywords (TIE):**

Proper Names: Colorado River (river), Grand Canyon (geography / location), Glen Canyon Dam (buildings&structures)

Subject Keywords: channels, conduit, controlled release, dam, reservoir, rivers, sediment, streams, volume of flow

Material Keywords: cookie sheet, roasting pan, cup, sand, clayboard, water, paper towel, pencil, paper

Procedure Keywords: poke a hole, divide, take, hold, pour, make drawing, identify areas, diagram, compare

<b>Pedagogy (TC)</b>	Collaborative learning Hands on learning
<b>Tool For (SIE):</b>	Teachers
<b>Resource Type (TC):</b>	Lesson Plan
<b>Format (SIE):</b>	text/HTML
<b>Placed Online (SIE):</b>	1998-09-02
<b>Name (SIE):</b>	PBS Online
<b>Role (SIE):</b>	onlineProvider
<b>Homepage (SIE):</b>	<a href="http://www.pbs.org">http://www.pbs.org</a>

#### **Metadata Generation Methods:**

**Structured Information Extraction (SIE)**

**Textual Information Extraction (TIE)**

**Text Categorization (TC)**



# Applications

---

- Document representation / indexing
- Query representation & expansion
- Automatic summarization
- Results browsing
- Focus group / interview transcript analysis
- Input to visualization tools
- Metadata generation
- Building / adding to Knowledge Bases for use by human & automated reasoners



# High Performance Knowledge Bases

---

- **KB development technology:**
  - rapidly deployable (within months)
  - large (100k-1M axiom/rule/frame)
  - comprehensive coverage
  - reusable
  - maintainable
- **Development Steps:**
  - building foundation knowledge
  - acquiring domain knowledge
  - developing efficient problem solving



# Extend Knowledge Bases

---

**For example - World Fact Book (WFB):**

- can be enhanced by full-text sources
- NLP extraction will provide:
  - greater depth
  - more currency
  - beyond 'national' information



# Automatically Constructed KB

---

## **SOURCES (26,000+ documents about Iran):**

- LA Times
- New York Times
- Reuters
- AFP
- IET Annotated Pages
- CP Model Fragments

## **KB SIZE:**

- 1,520,903 Unique CRC Triples
- 245,493 Unique Concepts

**“Egyptian President Hosni Mubarak”**

**-> ISA (Hosni Mubarak, President)**

**LOC (Hosni Mubarak, Egypt)**

# Concept-Relation Extraction

---

**HEADLINE:** US Campaign on Sudan has Limited Success at Security Council

**SOURCE:** Washington Post, 04/26/96, John M. Goshko

Egyptian President Hosni Mubarak was attacked by Islamic militants in Addis Ababa. The assassination attempt was made on June 26, 1995.

CG\_1      ASSOC ( militant, Islamic|**religion** )  
            OBJ ( attack, Hosni Mubarak|**person** )  
            LOC ( Islamic militant, Addis Ababa|**city** )  
            LOC ( attack, Addis Ababa|**city** )  
            AGNT ( attack, Islamic militant )

CG\_2      OBJ ( make, assassination attempt )  
            PTIM ( make, June\_26\_,\_1995 )

---

Center for NLP



# Evaluation of Knowledge Base Additions

---

- **High level performance on conceptual phrase extractions**
  - Precision - 91%
  - Recall - 84%



# Challenges

---

- **Dirty data**
- **May need to train genre-specific POS taggers**
- **Phrase-boundary detection**
- **Anaphora**
- **Alias-tracking**
- **Synonymous phrasings**
- **Selection of sub-set of most useful phrases**
- **Evaluation of contribution to a larger task**