



**When You Want an Answer,  
Why Settle for A List?**

**Elizabeth D. Liddy**

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

**May 2, 2001**



# Current Situation

- **Intranet users have a wide range of *types* of information needs**
  - Broad environmental scanning reports
  - Short summaries of events
  - Directory style lists
  - Paragraph or sentence length answers
  - Brief facts
- **Intranets have a single response**
  - a page or a list of pages



# Sources of dissatisfaction

1. **Users have needs which current intranet search engines cannot accept**
  - Information needs are **not** topics
  - ‘*Accounting*’ is **not** an information need
2. **The queries which are accepted by the intranet search engine, return a page or a list of pages – not an answer**



# Difference in Users' Requirements

## Information (*Document*) Retrieval:

- *Topical* match between query & term-based inverted file index of documents
- Less precise matching required

## Question-Answering:

- A very specific response to a richly represented query
- Very precise matching of *entity* and *relation* requirements



# What is a real information need?

- *Not Human Resources*

-- BUT --

- *“Do adoptive fathers qualify for family leave?”*

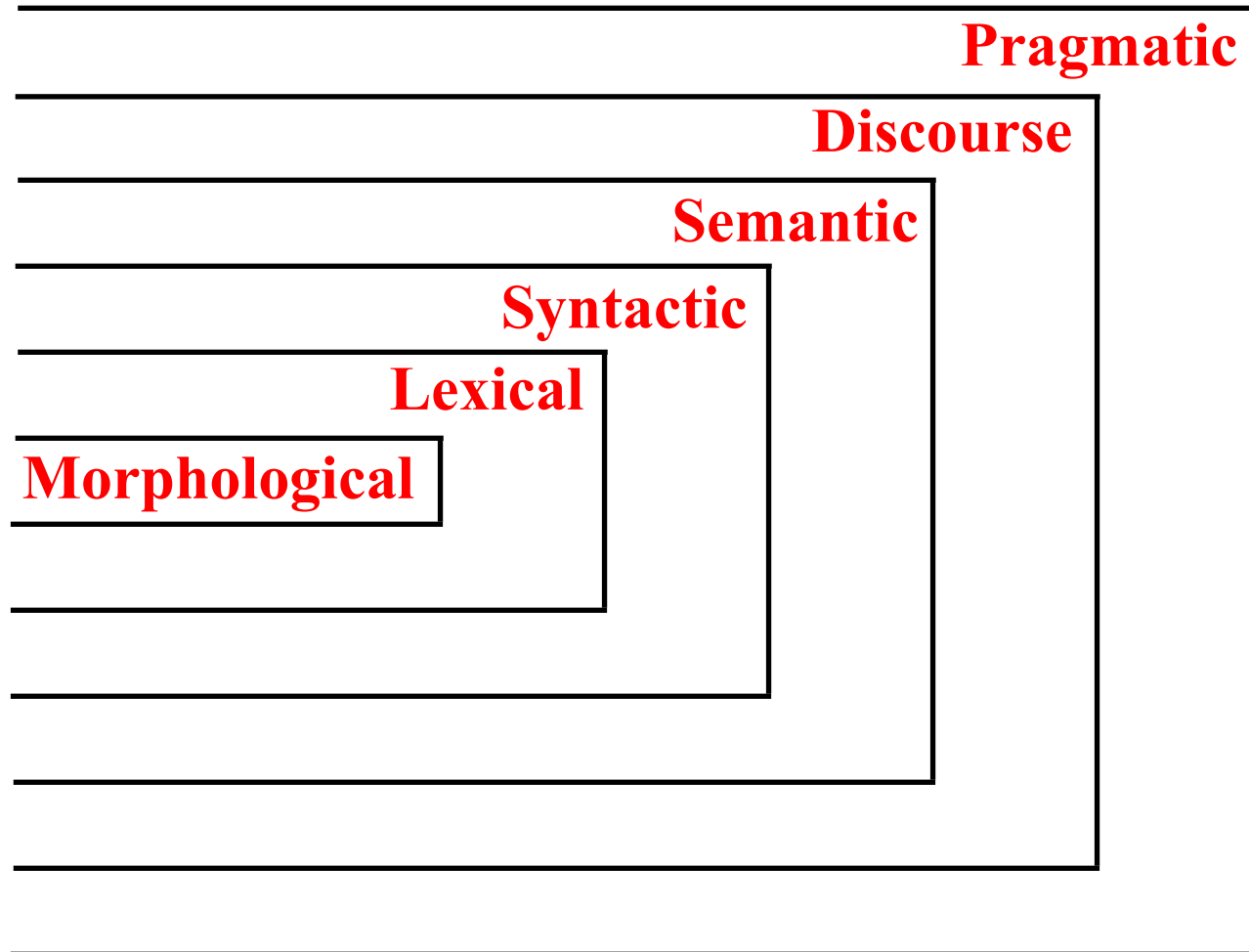


# Core Technology Solution

## Natural Language Processing

- **A technology which enables a system to accomplish human-like understanding of text**
- **Extracts both explicit and implicit meaning**
- **Utilizes all levels of human language understanding when representing the contents of text**

# Levels of Language Understanding





# eQuery

## **An NLP-based, 2-stage Information Access System**

- Interprets important concepts, relations and events in text (*both documents & queries*)
- Can provide document retrieval, question-answering, or visual summarization



## L - 2 - L Query Representation

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

*indictment\* +Bosnian “war criminal”*

**“I want to know about efforts to bring suspects of the  
Lockerbie bombing to trial.”**

*effort\* bring\* suspect\* +Lockerbie bomb\* trial\**



# Comparison of AltaVista to eQuery

HPKB Queries

AltaVista

eQuery

Ave. Precision @ 5

.40

.80



# Query Processing

## 1. Determine question ‘focus’

- *“How long is the rainy season on Maui?”*
  - Number (duration)
- *“When is it fishing season?”*
  - Time (duration)
- Utilize cue phrases, question structure
  - *Who, what, where, when*
- Some questions require more complex processing
  - *“Even if I don’t care to swim, are there other activities I can do in the water?”*
  - *Why, how*



## Query Processing (cont'd)

### 2. Named Entity recognition, boundary detection & entity categorization

- 60% to 80% of queries dealt with named entities

*<Seven\_Oaks\_Golf\_and\_Country\_Club>*

*<Vice\_President\_of\_Business\_Development>*

- Utilize context first, not a database
  - Rules linked to a taxonomy of entity **types**

# Knowledge Organization Structure (KOS)

---

## 0. Life & Living Things (other than human)

### 1. People

#### 1.1 Titles / Positions [Director]

1.1.1 Honorifics [Mr.]

1.1.2 Roles [Dutch Auction Agent]

1.1.3 Military Ranks [Lt.Col.]

#### 1.2 Groups

1.2.1 Organizations [Save the Children]

1.2.1.1 Government Orgs [DARPA]

1.2.1.1 Courts [Bankruptcy Court]

1.2.1.2 Lawmaking groups [Parliament]

1.2.1.3 Military divisions [Boys from Syracuse]

1.2.1.2 Terrorist Groups [Jihad]

.....

1.2.1.20 Organizational subdivisions [Director's Comm.]

#### 1.2.2.Companies [ABC Bancorp]

1.2.2.20 Company subdivisions [Public Relations Dept.]

## 2. Thought, Communication and Communication Channels

2.1 E-mail [liddy@syr.edu]

.....

## 3. Buildings & Structures

## 4. Substances, Materials, Objects, and Equipment

.....



## Query Processing (cont'd)

### 3. Query Expansion

- Expand query terms / phrases into morphological equivalents
- Utilize lexical resources to add synonymous terms / phrases to query representation
- Add ‘answer-indicating’ words to query
  - Reason questions (*why, how*) are expanded with *because, because of, due to, thanks to, since, in order to*

## Query Processing (cont'd)

---

### 4. Query Representation (cont'd)

- *What is the name of the female counterpart to El Nino, which results in cooling temperature and very dry weather?*

<S> what|WP be|VBZ the|DT name|NN of|IN the|DT female|JJ counterpart|NN to|TO El|NP Nino|NP ,|, which|WDT result|VB in|IN cooling |ADJ temperature|NN and|CC very|RB dry|JJ weather|NN ?|. </S>

# Query Processing (cont'd)

## 4. Query Representation (cont'd)

<S> what|WP be|VBZ the|DT name|NN of|IN  
the|DT female|JJ counterpart|NN to|TO  
<El\_Nino|NP> ,|, which|WDT result|VB in|IN  
<cooling\_temperature|NN> and|CC very|RB  
<dry\_weather|NN> ?|. </S>

- female\* counterpart\* +"El Nino\*" result\*  
"cooling temperature\*" "dry weather\*"



## TREC Q & A Track:

---

- A '*competition*' among 50 companies or groups with Q & A technologies
- Task
  - 696 never-before-seen questions
  - 1,000,000 new documents
  - limited time frame for task
  - comparative results amongst participants
- [www.trec.nist.gov](http://www.trec.nist.gov)



# Typical questions:

*Where is Belize located?*

*What type of bridge is the Golden Gate Bridge?*

*What is the population of the Bahamas?*

*How far away is the moon?*

*What is Francis Scott Key best known for?*

*What state has the most Indians?*

*Who invented the paper clip?*

*How many dogs pull a sled in the Iditarod?*

*Where did bocci originate?*

*Who invented the electric guitar?*

*Name a flying mammal?*

*How many hexagons are on a soccer ball?*

*Who is the leader of India?*



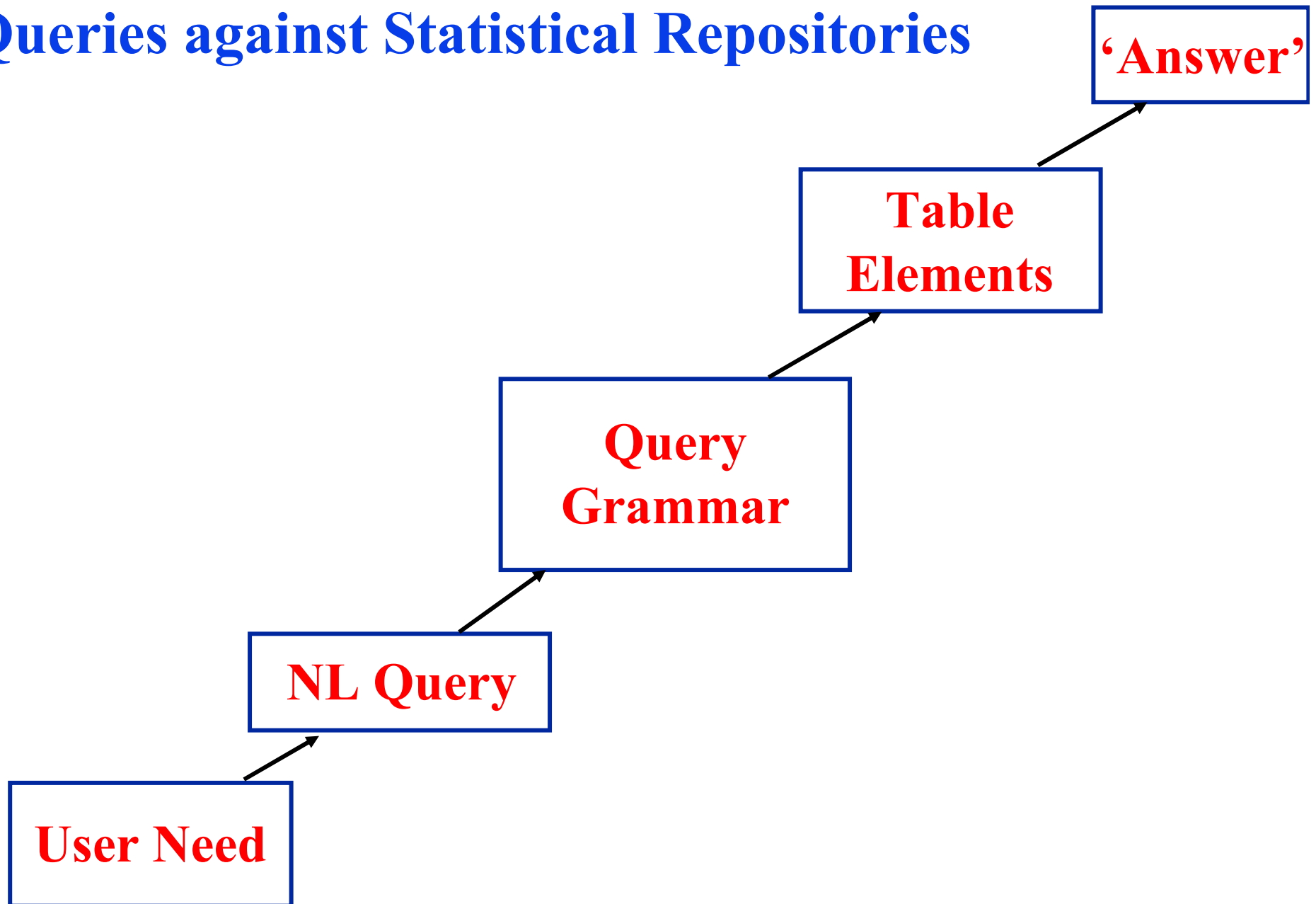
# Specialized Q & A

- **Non-textual**
- **Non SQL**

**- BUT -**

- **Seeking answers from tables and charts**
  - Statistical queries

# Queries against Statistical Repositories





## **Need to Understand:**

---

- **What users ask about:**
  - **Query Dimensions**
- **How they ask:**
  - **Query sublanguage grammar**
- **How NL queries can best be used for retrieving statistical answers:**
  - **Mapping query dimensions into tables' metadata**



# Methodology

---

- **Analyzed 1,000 actual user queries**
  - gathered from logs provided by **Bureau of Labor Statistics**
  - seeking statistical information
- **Developed ontology of query dimensions**
  - data-up from queries
  - extended by values from tables
- **Produced first draft of a query grammar**



# Ontology of Query Dimensions

---

## **WHO**

**Ethnicity**

**African**

**Age**

**Years of age**

**Under the age of 15 years old**

**Gender**

**sex**

**Religion**

**Protestantism**

**Baptist**

**Households**

**Family households**

**Homeless**



# Ontology of Query Dimensions (cont'd)

---

## **WHERE**

**Location**

**Region**

**New England**

**State**

**New York**

**County**

**Onondaga**

**City**

**Syracuse**

**Census Tract**



# Ontology of Query Dimensions (cont'd)

---

## **WHEN**

### **Time**

**1990s**

**present**

**January**



# Ontology of Query Dimensions (cont'd)

---

## **WHAT**

**Income**

**Education**

**Status**

**Level of Education**

**Economic Indicators**

**Consumer Expenditures**

**Unemployment Rate**

**Employment**

**Full-time**

**Occupation**

**White Collar**

**Financial Managers**



# Typically:

---

User wants to know about:

- a quantification *how many*
- a certain population *black women*
- a location *New York City*
- a time period *1999*
- a condition *unemployed*

- “*How many black women living in New York City in 1999 were unemployed?*”



## Sample query structures:

---

- How many **(POPULATION)** were there in **(LOCATION)** in **(TIME)** ?
- What is the **(QUANTITY / RATE)** of **(POPULATION)** with **(CONDITION)** in **(LOCATION)** ?
- What **(LOCATION-CATEGORY)** had the **(CONDITION)** in **(TIME)** ?



## **NLP query analysis:**

---

- **Lexical and ‘focus’ clues pointing to the variable type the user is looking for**
- **Proper Name identification & categorization**
- **Structure of query -> sublanguage analysis**
- **Mapping from user’s vocabulary to organization’s vocabulary using dimension hierarchy**

# Query grammar applied

---

*“In 1996, how many years was a 50 year old woman from the US expected to live?”*

In|**IN** 1996|**CD** ,|, how|**WRB** many|**JJ**  
years|**NNS** was|**VBD** a|**DT** 50|**CD** year|**NN**  
old|**JJ** woman|**NN** from|**IN** the|**DT**  
<**CTRY**> US|**NP** </**CTRY**> expected|**VBD**  
to|**TO** live|**VB** ?|?

## Query grammar applied (cont'd)

---

In **<PTIM> 1996 </PTIM>** ,  
**<HOW MUCH> how many years </HOW MUCH>**  
was a  
**<WHO> 50 year old woman </WHO>**  
from the  
**<LOC> <CTRY> U.S. </CTRY </LOC>**  
**<COND> expected to live </COND> ?**



# Links to Table Elements

---

<b>Where</b>	<b>country</b>	<b>United States</b>
<b>When</b>	<b>pt in time</b>	<b>1996</b>
<b>What</b>	<b>health</b>	<b>life expectancy</b>
<b>Who</b>	<b>sex, age</b>	<b>female, 50 year old</b>
<b>How much</b>		<b>????</b>



## Relationships within Tables:

---

**Where**

**Location**

**Table**

**When**

**Time**

**Column**

**What**

**Condition**

**Row**

**Who**

**Population**

**Sub-row**

**How much**

**Percentage**

**Cell**



## **In Summary:**

- **Q & A is a necessary time-saver for many tasks within an organization.**
- **Q & A is different from document retrieval.**
- **Q & A is what Intranets will need to provide.**
- **Q & A is difficult, but doable!**



## Existing Q & A systems

- **Early systems**
  - **Woods' LUNAR**
  - **Lehnert's QUALM**
- **Systems developed in recent years**
  - **Kupiec's MURAX**
  - **Texttract by Cymfony**
  - **GuruQA by IBM**
  - **eQuery by CNLP**
  - **Falcon by SMU**
  - **<!metaMarker> by solutions-united**