

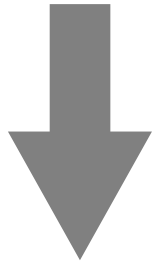
# MIT Conversational Systems\*: Advances and Challenges

- Introduction
- Speech Understanding
  - Natural Language Understanding
  - Discourse Resolution
  - Dialogue Modeling
- Development Issues
- Recent Progress
- Future Challenges
- Summary

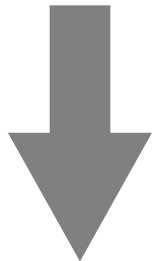
\* AKA spoken language systems or spoken dialogue systems  
See article by Zue and Glass (2000)

# MIT The Premise:

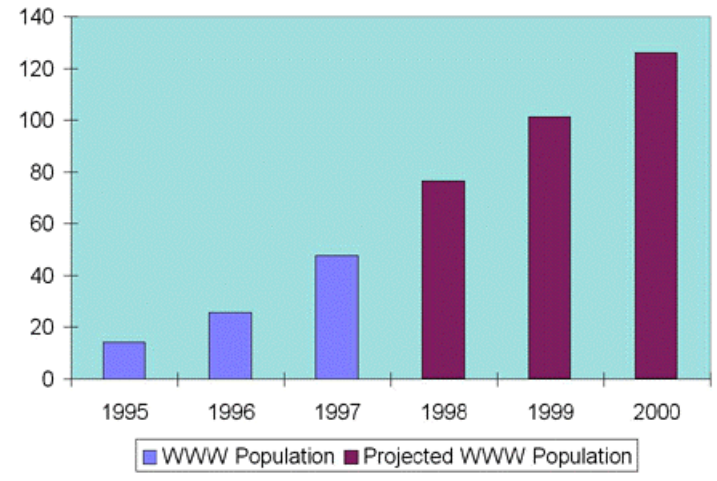
**Everybody  
wants  
Information**



**Even when  
they are  
on the move**



**The interface  
must be  
easy to use**



For North America  
CommerceNet  
Research Center (1999)

**Devices  
must be  
small**



**Need new  
interfaces**

**Speech is It!**

# MIT What Are Conversational Systems?

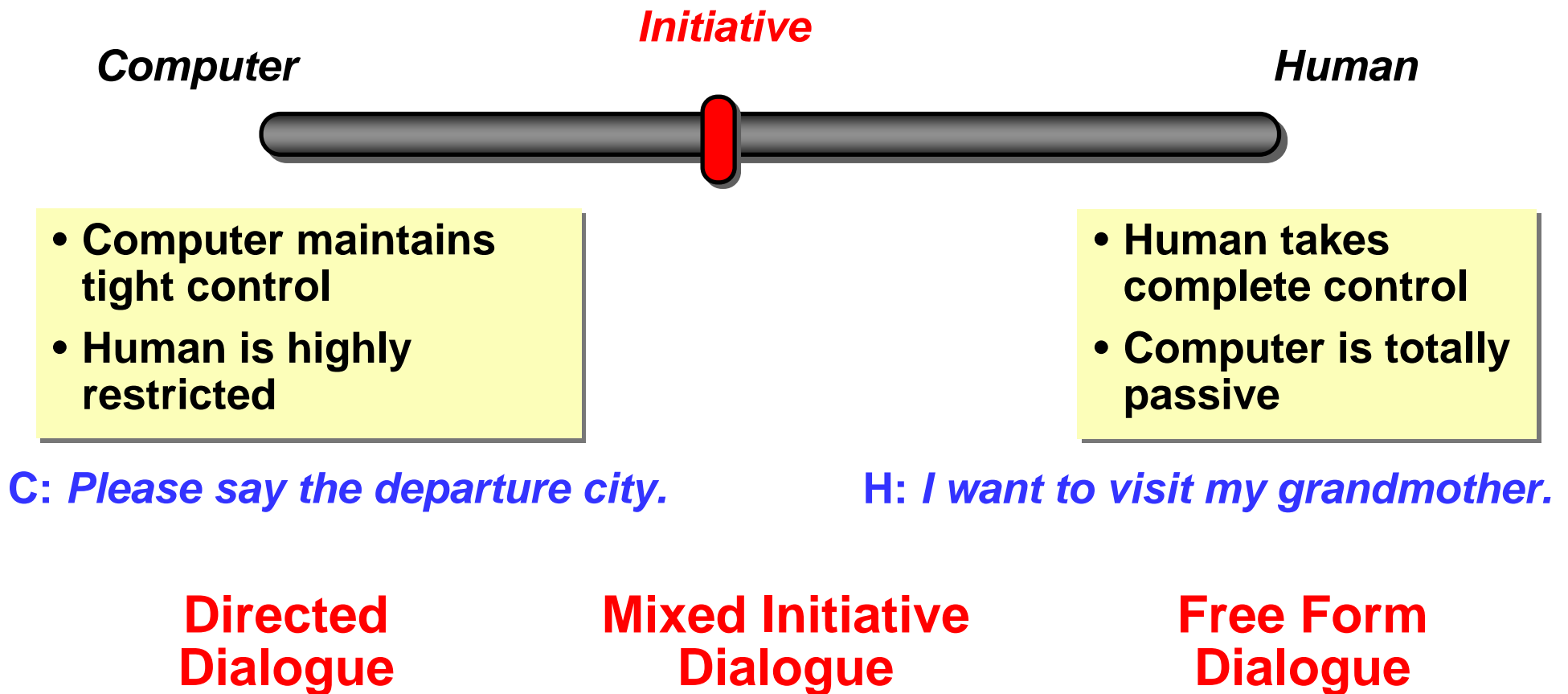
Systems that can communicate with users through a ***conversational*** paradigm, i.e., they can:

- ***Understand*** verbal input, using
  - \* Speech recognition
  - \* Language understanding (in context)
- ***Verbalize*** response, using
  - \* Language generation
  - \* Speech synthesis
- Engage in ***dialogue*** with a user during the interaction

# MIT

## Defining the Context

- Conversational systems differ in the degree with which human or computer takes the initiative



# The Nature of Mixed Initiative Interactions

## (A Human-Human Example)

..... **disfluency**

**C:** Yeah, [um] I'm looking for the Buford Cinema.

**A:** OK, and you're wanting to know what's showing there or

... **interruption, overlap**

**C:** Yes, please. **confirmation**

**A:** Are you looking for a particular movie? **clarification**

**C:** [um] What's showing.

**A:** OK, one moment. **back channel**

.....

**A:** They're showing *A Troll In Central Park*.

**C:** No. **inference**

**A:** *Frankenstein*. **ellipsis**

**C:** What time is that on? **co-reference**

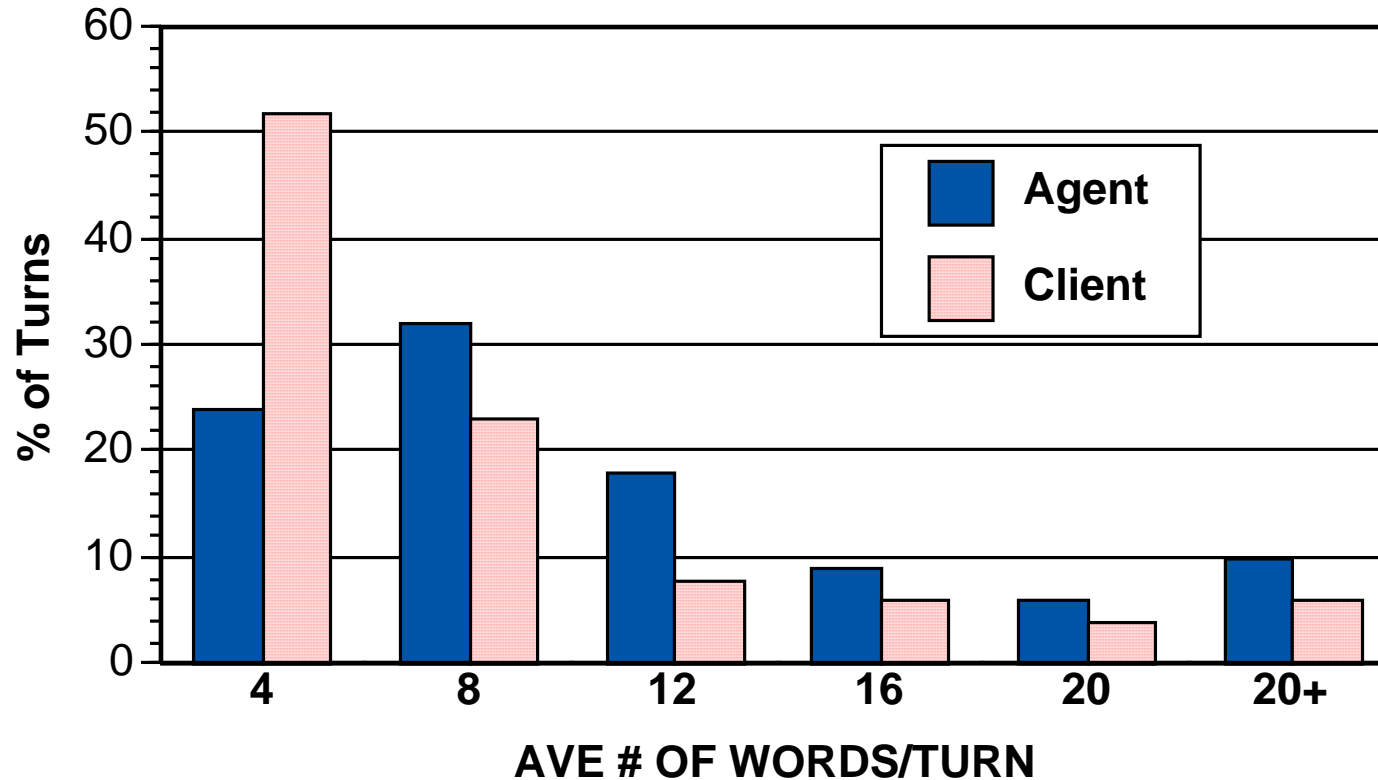
**A:** Seven twenty and nine fifty.

**C:** OK, any others?



Media Clip

# Study of human-human interactions can lead to good insights in building human-machine systems



- **Over 1,000 dialogues in many domains (Flammia '98)**
- **Some lessons learned (about clients):**
  - **More than 80% of utterances are 12 words or less**
  - **Most short utterances are confirmation and back channel communications**

# MIT Dialogue Management Strategies

- ***Directed dialogues*** can be implemented as a directed graph between dialogue states
  - Connections between states are predefined
  - User is guided through the graph by the machine
  - Directed dialogues have been successfully deployed commercially
- ***Mixed-initiative dialogues*** are possible when state transitions determined dynamically
  - Transitions can be determined, e.g., by E-form variable values
  - User has flexibility to specify constraints in any order
  - System can “back off” to a directed dialogue if desired
  - Mixed-initiative dialogues mainly research prototypes

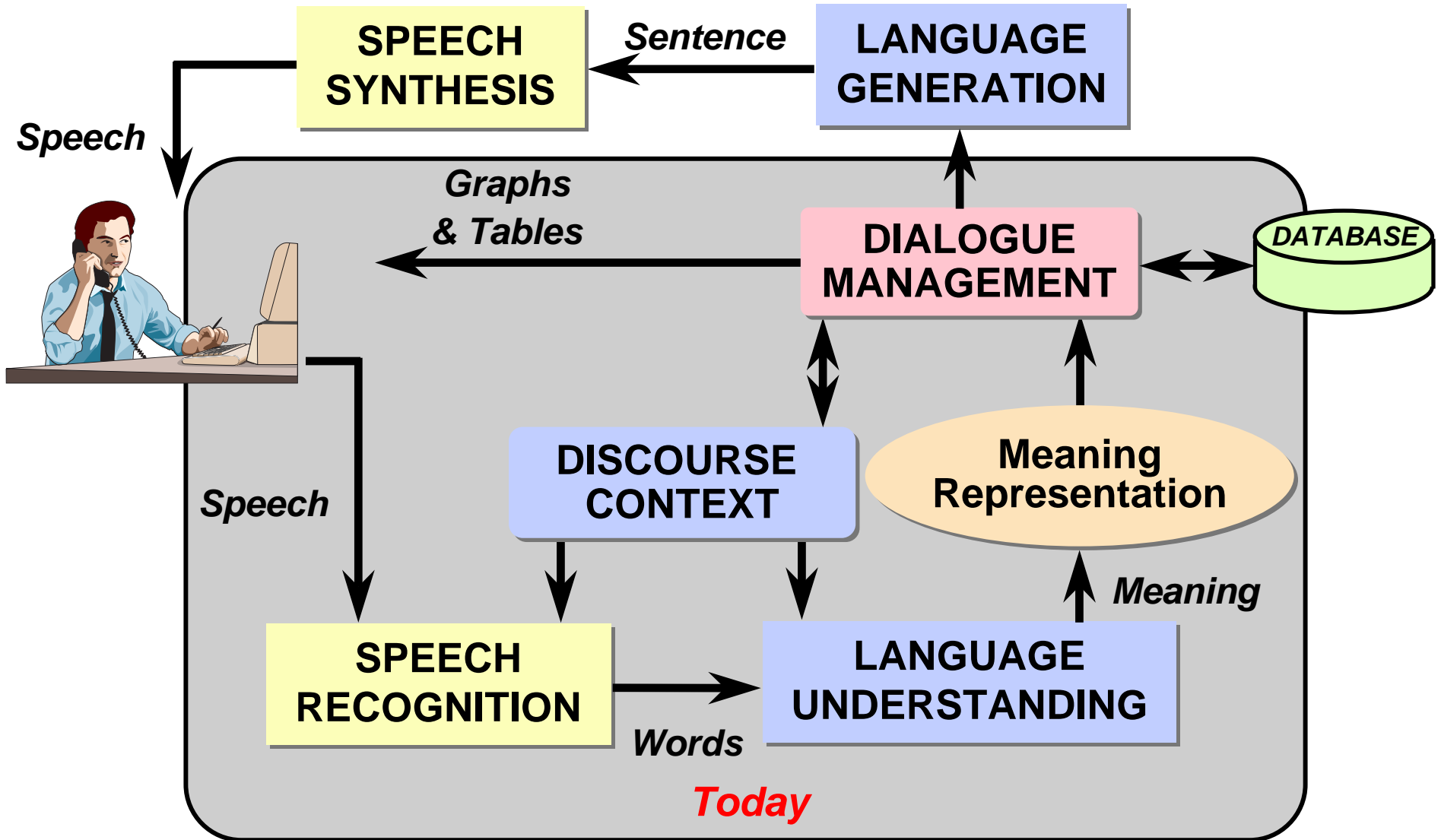
# Example of MIT's Mercury Travel Planning System

- **New user calling into Mercury flight planning system**
- **Illustrated technical issues:**
  - **Back-off to directed dialogue when necessary (e.g., password)**
  - **Understanding mid-stream corrections (e.g., “no Wednesday”)**
  - **Soliciting necessary information from user**
  - **Confirming understood concepts to user**
  - **Summarizing multiple database results**
  - **Allowing negotiation with user**
  - **Articulating pertinent information**
  - **Understanding fragments in context (e.g., “4:45”)**
  - **Understanding relative dates (e.g., “the following Tuesday”)**
  - **Quantifying user satisfaction (e.g., questionnaire)**





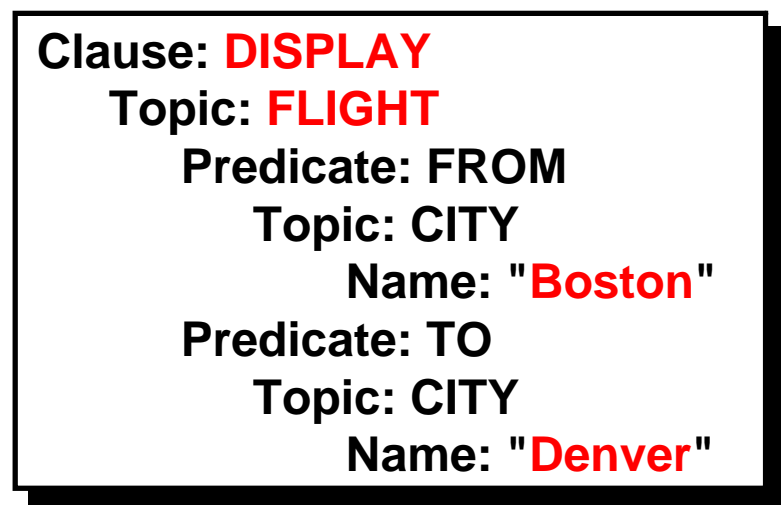
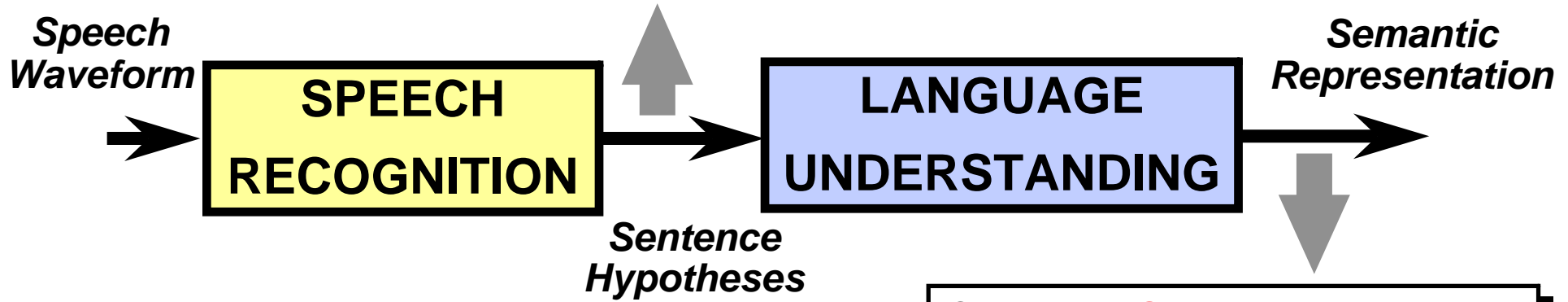
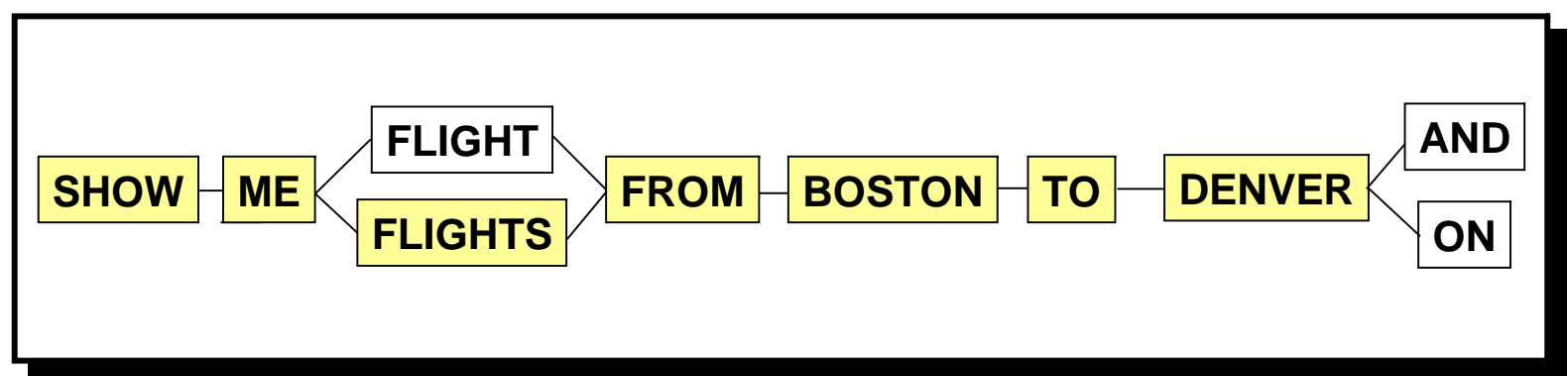
# Components of a Conversational System



# Natural Language Processing Components

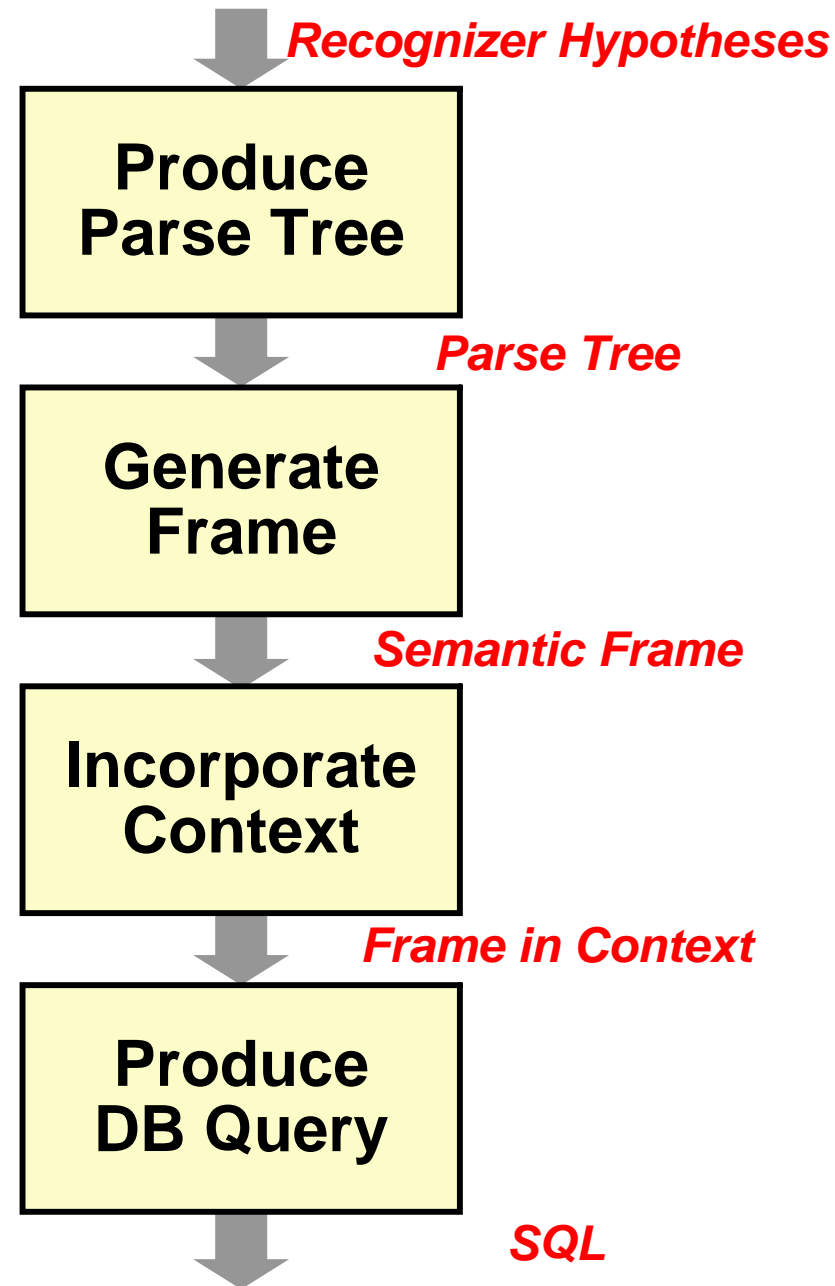
- **Understanding:**
  - Parse input query into a meaning representation, to be interpreted for appropriate action by application domain
  - Select best candidate from proposed recognizer hypotheses
- **Discourse Resolution**
  - Interpret each query in context of preceding dialogue
- **Dialogue Management**
  - Plan course of action under both expected and unexpected conditions; compose response frames.
- **Generation**
  - Paraphrase user queries into same or different language.
  - Compose well-formed sentences to speak the (sequence of) response frames prepared by the dialogue manager.

## Input Processing: Understanding



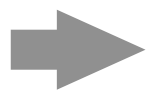
# MIT Typical Steps in Transforming User Query

- **Parsing**
  - Establishes syntactic organization and semantic content
- **Translation to a Semantic Frame**
  - Produces meaning representation identifying relevant constituents and their relationships
- **Incorporation of discourse context**
  - Deals with fragments, pronominal references, etc.
- **Translation to a database query**
  - Produces SQL formatted string for database retrieval

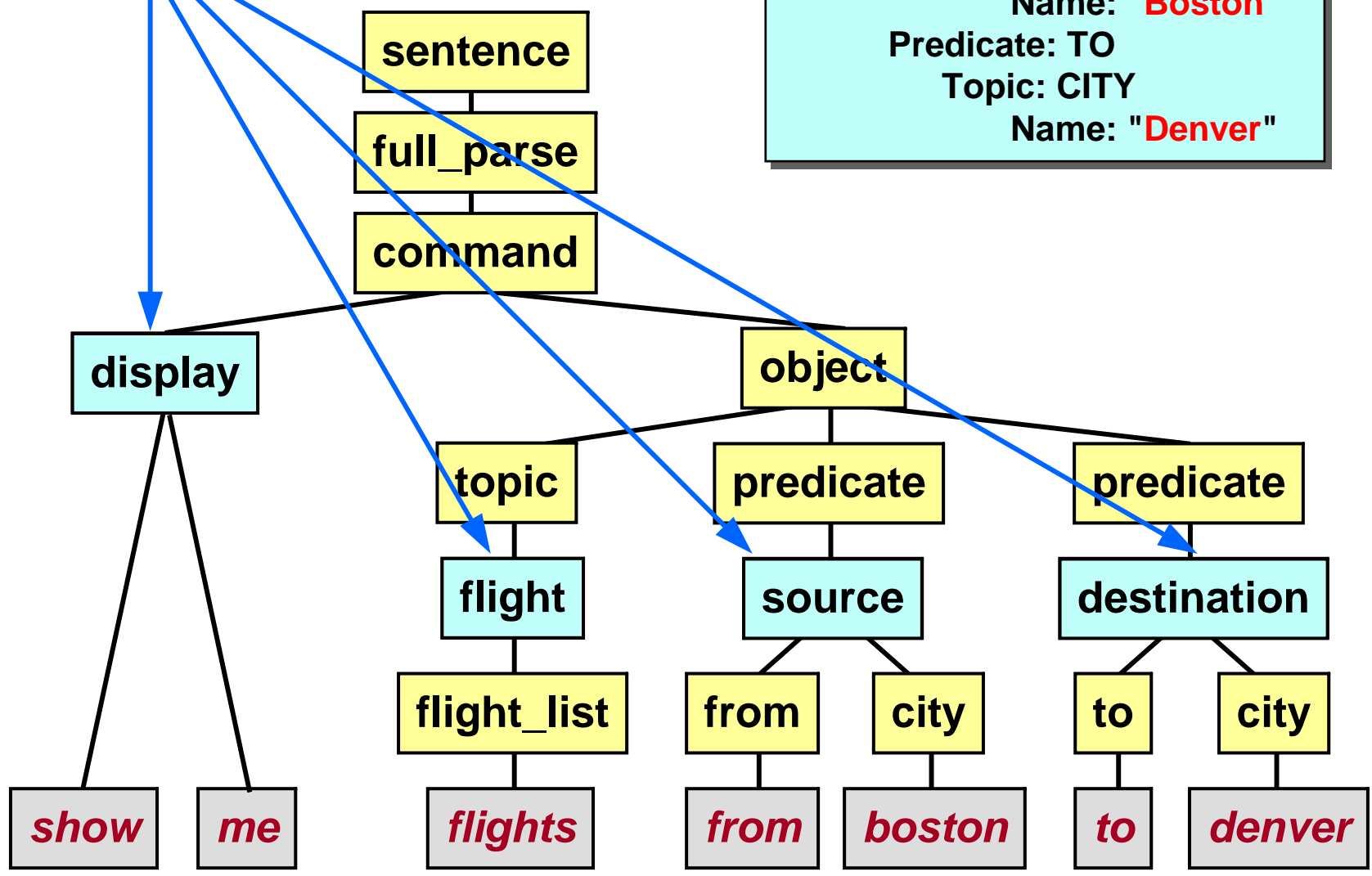


# Natural Language Understanding

Some syntactic nodes carry semantic tags for creating semantic frame



Clause: **DISPLAY**  
Topic: **FLIGHT**  
Predicate: FROM  
Topic: CITY  
Name: "**Boston**"  
Predicate: TO  
Topic: CITY  
Name: "**Denver**"



# Context Free Rules for Example

*Show me flights from Boston to Denver*

sentence	→	(display-clause truth-clause ...)
display-clause	→	display direct-object
direct-object	→	[determiner] (flight-event fare-event ...)
flight-event	→	flight [from-place] [to-place]
from-place	→	from a-city
to-place	→	to a-city
display	→	show-me
show-me	→	[ <i>please</i> ] <i>show</i> [ <i>me</i> ]
a-city	→	( <i>boston dallas denver</i> ...)
determiner	→	( <i>a the</i> )
...		

- **Context free:** left hand side of rule is single symbol
- **brackets [ ]:** *optional*
- **Parentheses ( ):** *alternates.*
- **Terminal words** in italics

# MIT

## What Makes Parsing Hard?

- **Must realize high coverage of well-formed sentences within domain**
- **Should disallow ill-formed sentences, e.g.,**
  - the flight that arriving in the morning
  - what restaurants do you know about any banks?
- **Avoid parse ambiguity (redundant parses)**
- **Maintain efficiency**

# MIT

## Understanding Words in Context

- **Subtle differences in phrasing can lead to completely different interpretations**

- Is there a six A.M. flight?
- Are there six A.A. flights?
- Is there a flight six?
- Is there a flight at six

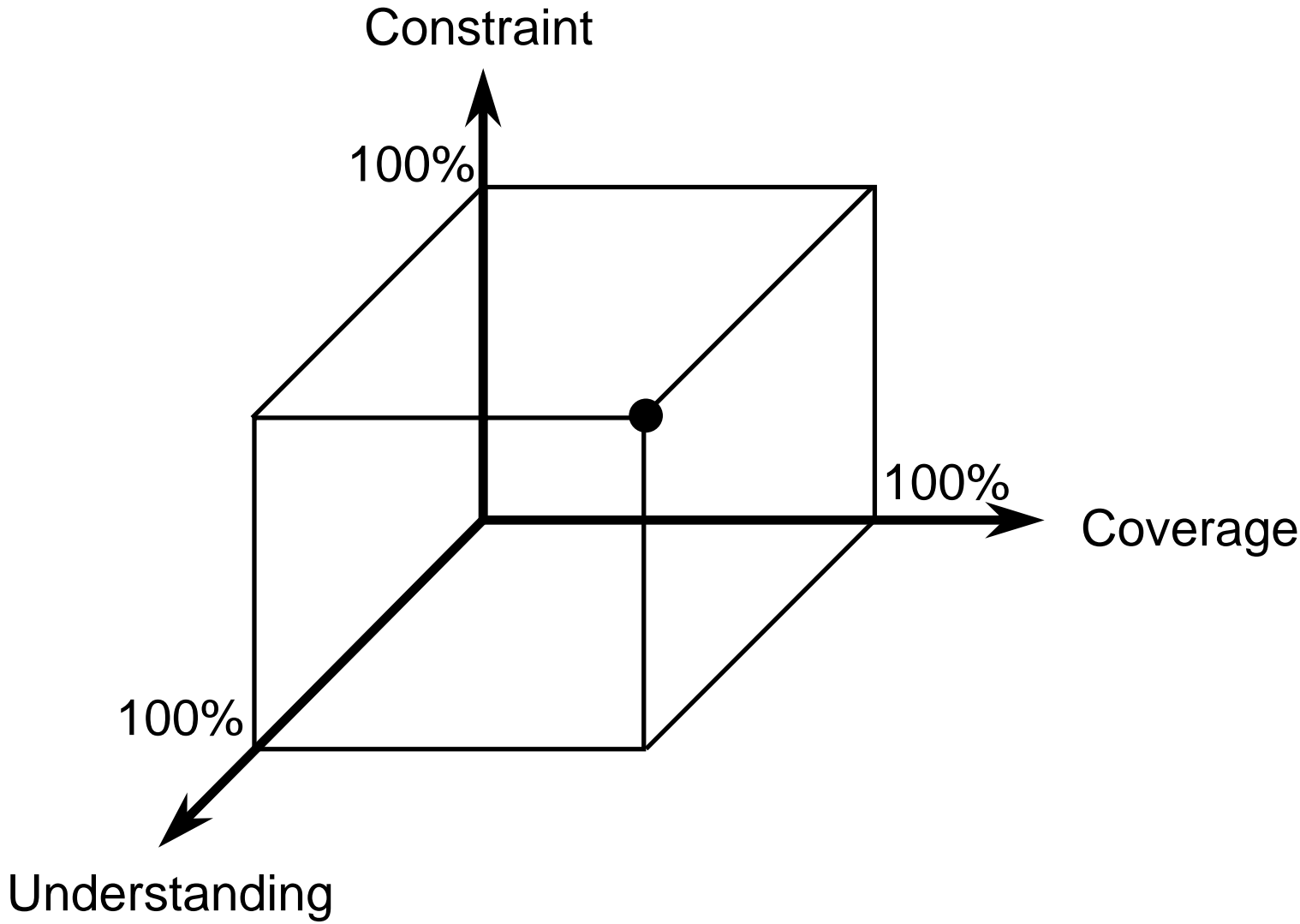
**“six” could mean:**

- A time
- A count
- A flight number

- **The possibility of recognition errors makes it hard to rely on features like the article “a” or the plurality of “flights.”**
- **Yet insufficient syntactic/semantic analysis can lead to gross misinterpretations**



# Multiple Roles for Natural Language Parsing in Spoken Language Context



# Contrasting Language Models for Speech Recognition and Natural Language Understanding

**Statistical language models (i.e.,  $n$ -grams) used for speech recognition are inappropriate for speech understanding applications, because they don't provide a meaning representation**

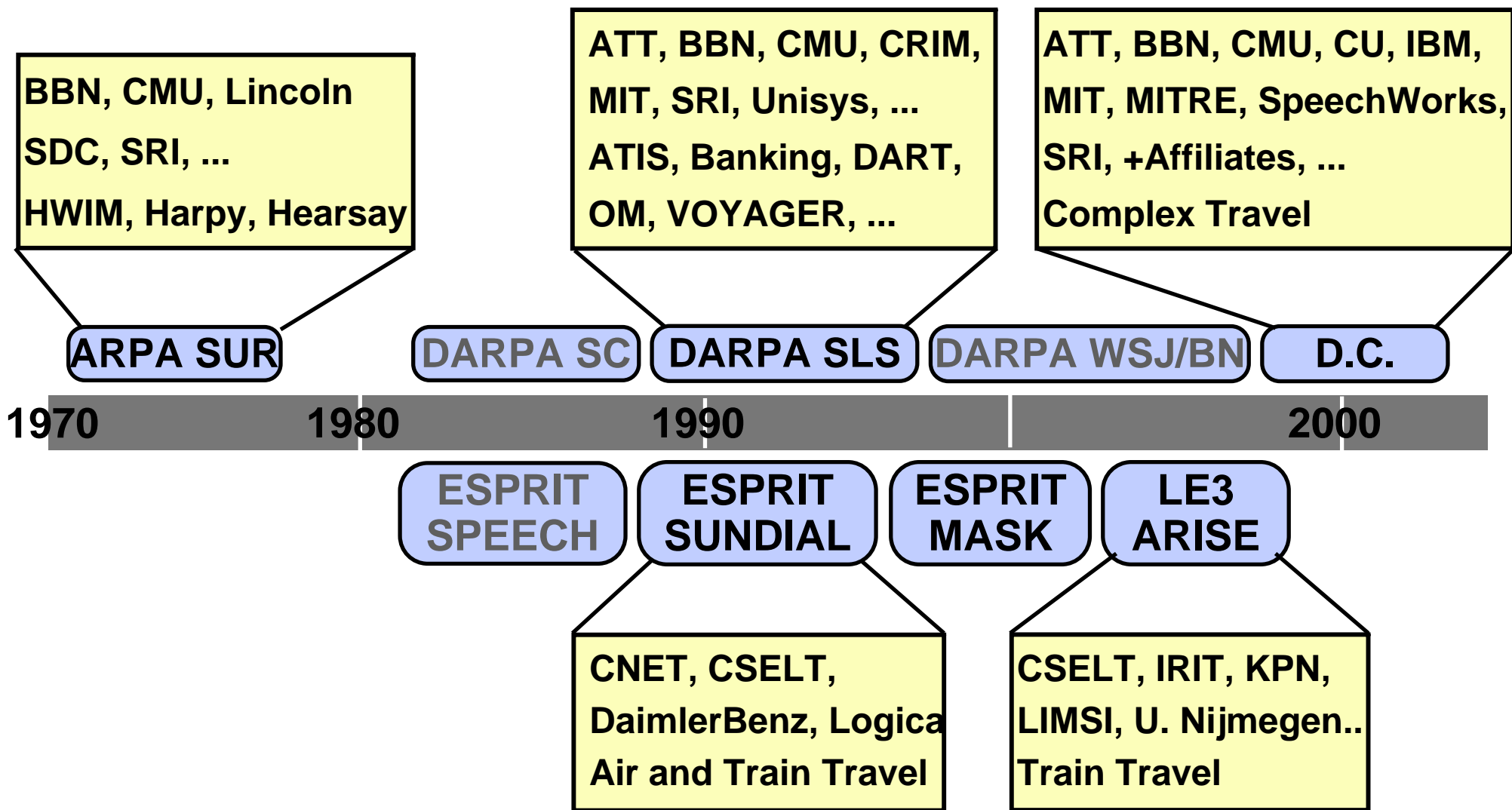
**Text based natural language processing systems may not be well suited for speech understanding applications, because they typically assume that:**

- Word boundaries are known with certainty
- All words are known with certainty
- Sentences are well formed
- Constraints are unnecessary

# Spoken Language Understanding

- **Spoken input differs significantly from text**
  - False starts
  - Filled pauses
  - Agrammatical constructs
  - Recognition errors
- **We need to design natural language components that can both constrain the recognizer's search space and respond appropriately even when the input speech is not fully understood**

## Some Speech-Related Government Programs



# The U.S. DARPA-SLS Program (1990-1995)

- **The Community adopted a common task (Air Travel Information Service, or ATIS) to spur technology development**
- **Users could verbally query a *static* database for air travel information**
  - 11 cities in North America (**ATIS-2**)
  - Expanded to 46 cities in 1993 (**ATIS-3**)
  - Mostly flights and fares
- **All systems could handle continuous speech from unknown speakers (~2,000 word vocabulary)**
- **Infrastructure for technology development and evaluation was developed**
- **Five annual common evaluations took place**

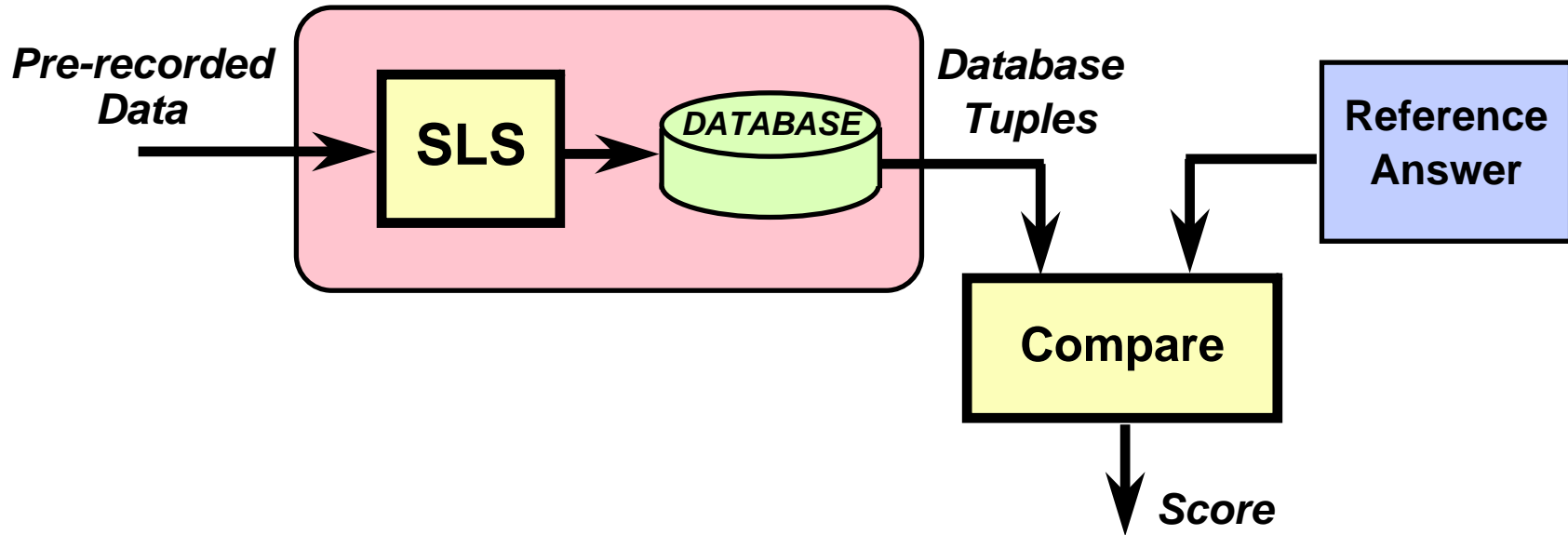
# MIT ATIS Data Collection Status

- Over 25,000 utterances were collected (from AT&T, BBN, CMU, MIT, NIST, and SRI)
- About 80% of the collected data (speech and transcriptions) were distributed for system development and training
- Over 11,000 of training utterances were annotated with database “reference” answer
- About 40% of the data from ATIS-3 (more cities)

Data Set	Class A	Class D	Class X
ATIS-2	43%	33%	24%
ATIS-3	49%	33%	18%

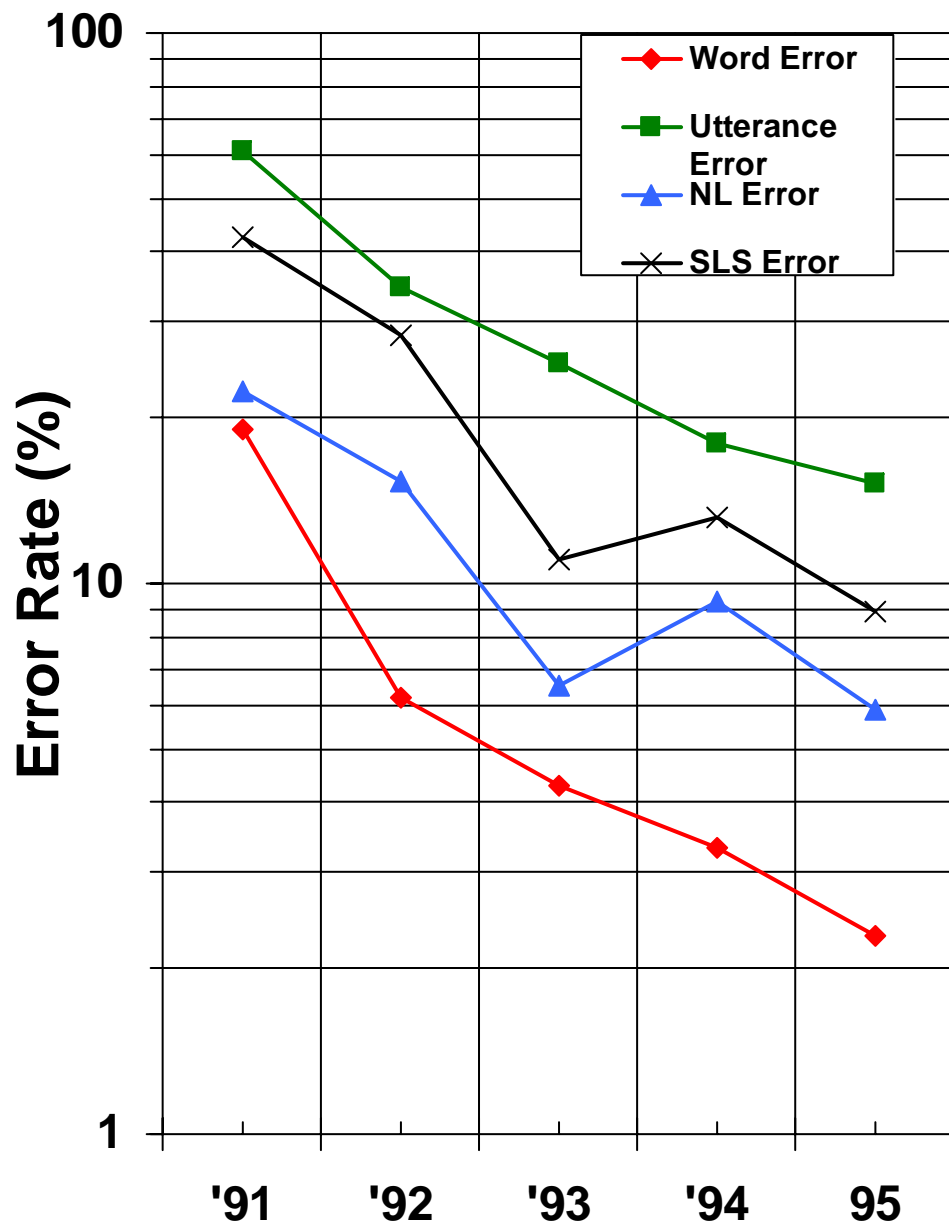
**A: Context-independent queries**  
**D: Context-dependent queries**  
**X: Un-answerable queries**

# Evaluation of SLS Using Common Answer Specification (CAS)



- **Evaluation is automatic (i.e., easy), once we have:**
  - Principles of interpretation (e.g., “red-eye”)
  - Properly annotated data, and
  - Comparator
- **But it is costly, and does not address important research issues such as dialogue modeling and system usefulness**

# State of the Art (The ATIS Domain)



- Word (also utterance) error rate (ER) for spontaneous speech approaching that for read speech
- Understanding ER <10% for text input; complete NL analyses not required
- ER for speech input only ~2-3% higher than for text input
- Many more sentences understood than correctly recognized
- In most cases, ER cut by about half every two years
- Real-time performance achieved using high-end workstations
- Results for “answerable” queries only






# Example Sentences Some Systems Can Handle

- I WOULD LIKE TO FLY FROM SAINT PAUL TO SAN JOSE MONDAY MORNING FROM SAN JOSE TO HOUSTON TUESDAY MORNING AND FROM HOUSTON TO SAINT PAUL ON WEDNESDAY MORNING
- [UM] I WOULD LIKE TO FIND OUT WHAT FLIGHTS THERE ARE ON FRIDAY JUNE ELEVENTH FROM SAINT PETERSBURG <TO> M- TO M- MILWAUKEE AND THEN FROM MILWAUKEE TO TACOMA THANK YOU

# MIT

## Difficult, But Real, Sentences

-  I would like to find a flight from Pittsburgh to Boston on Wednesday and I have to be in Boston by one so I would like a flight out of here no later than 11 a.m.
-  I'll repeat what I said before on scenario 3 I would like a 727 flight from Washington DC to Atlanta Georgia I would like it during the hours of from 9 a.m. till 2 p.m. if I can get a flight within that time frame and I would like it for Friday
-  Some database I'm inquiring about a first class flight originating city Atlanta destination city Boston any class fare will be all right

**We cannot expect any natural language system to be able to fully parse and understand all such sentences**

# Historical Perspective on Key Players in ATIS Effort

- **CMU:** Strictly semantic grammar, syntactic information mostly ignored
- **MIT:** Grammar rules interleave syntactic and semantic categories
- **BBN, SRI:**
  - Initial systems used syntactic grammars based on unification framework, with parallel semantic rules
  - Both sites now have a strictly semantic grammar as well
  - SRI combines two outputs into one system; BBN has separate competing systems
- **ATT, BBN, IBM:** Stochastic approaches using HMM

# CMU's Approach

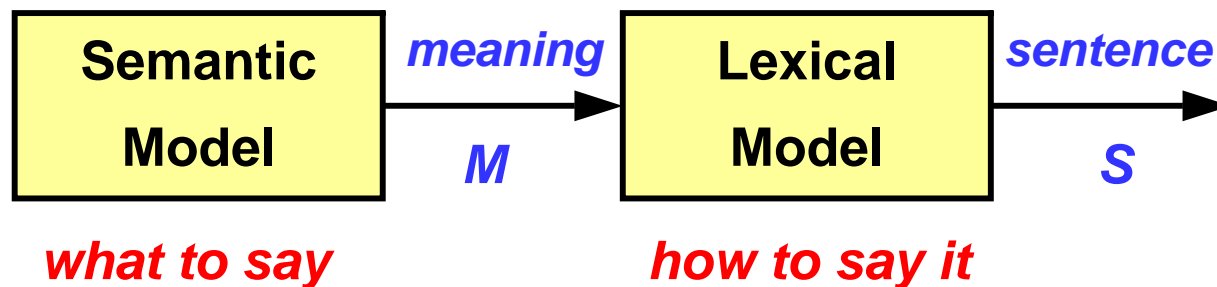
- Grammar consists of ~70 autonomous semantic concepts (e.g., **Depart Location**)
- Each concept is realized as a set of possible word class sequences, e.g.,  
**Depart Location => [FROM] [LOC]**  
which are specified through recursive transition networks (RTNs)
- Semantic frame is a flat structure of key-value pairs as defined by the concepts
- Syntactic structure is ignored
- Recognizer only produces a single theory

## Example

okay the next uh uh (i'm going to need) a (from denver) (about two o'clock) and (go to atlanta)

- **TINA was designed for speech understanding**
  - Grammar rules intermix syntax and semantics
  - Probabilities are trained from user utterances
  - Parse tree is converted to a semantic frame that encapsulates the meaning
- **TINA enhances its coverage through a robust parsing strategy**
  - Sentences that fail to parse are subjected to a fragment parse strategy
  - Fragments are combined into a full semantic frame
  - When all things fail, resort to word spotting

# MIT Stochastic Approaches

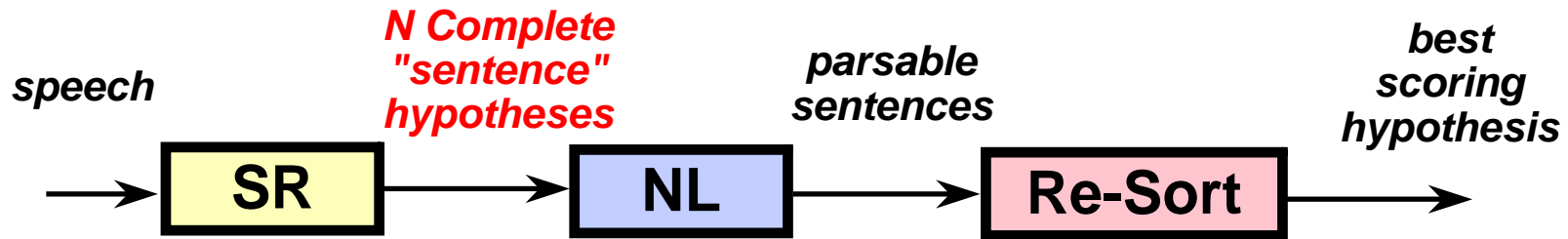


- Choose among all possible meanings the one that maximizes:

$$P(M | S) = \frac{P(S | M)P(M)}{P(S)}$$

- HMM techniques have been used to determine the meaning of utterances (ATT, BBN, IBM)
- Encouraging results have been achieved, but a large body of annotated data is needed for training

# SR/NL Integration via *N*-Best Interface



show me flights from boston to denver and  
**show me flights from boston to denver**  
show me flights from boston to denver on  
show me flight from boston to denver and  
show me flight from boston to denver  
show me flight from boston to denver on  
show me flights from boston to denver in  
show me a flight from boston to denver and  
show me a flight from boston to denver  
show me a flight from boston to denver on

**Answer**

- ***N*-Best resorting has also been used as a mechanism for applying computationally expensive constraints**

# Some Issues Related to Search

- An  $A^*$  algorithm is often used to construct the top- $N$  sentence hypotheses

$$f^*(p) = g(p) + h^*(p)$$

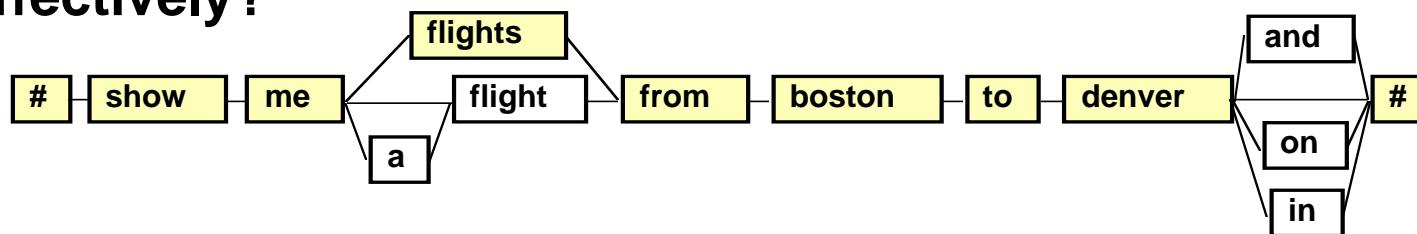
where:  $f^*(p)$  is the estimated score of the best path containing partial path  $p$

$g(p)$  is the score from the beginning to the end of the partial path  $p$ , and

$h^*(p)$  is the estimated score of the best-scoring extension of  $p$

- Questions:

- How can information in the  $N$ -best list be captured more effectively?



- What are some computationally efficient choices of  $h^*(p)$ , even if inadmissible?



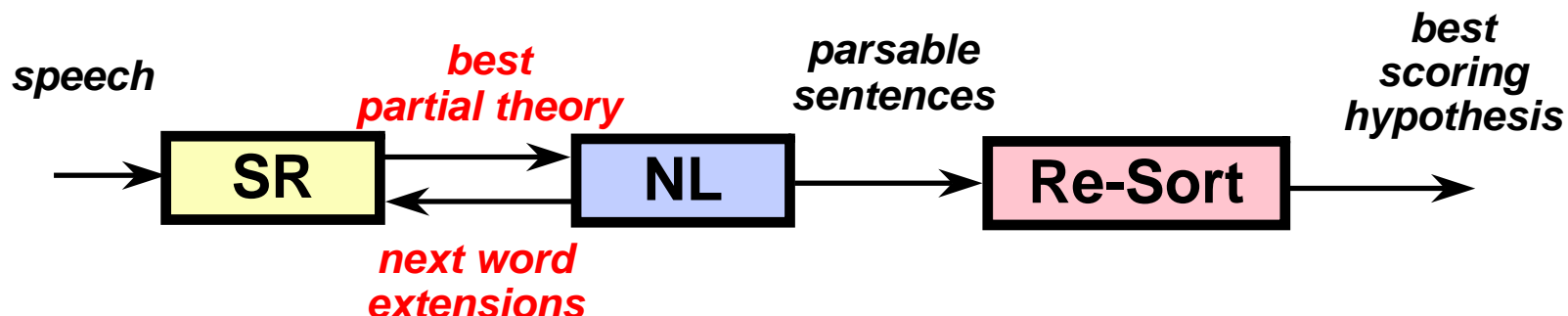
# MIT

## Tighter SR/NL Integration

- Natural language analysis can provide long distance constraints that  $n$ -grams cannot
- Examples:
  - What is the flight serves dinner?
  - What meals does flight two serve dinner?
- **Question:** How can we design systems that will take advantage of such constraints?

# Alternatives to N-Best Interface

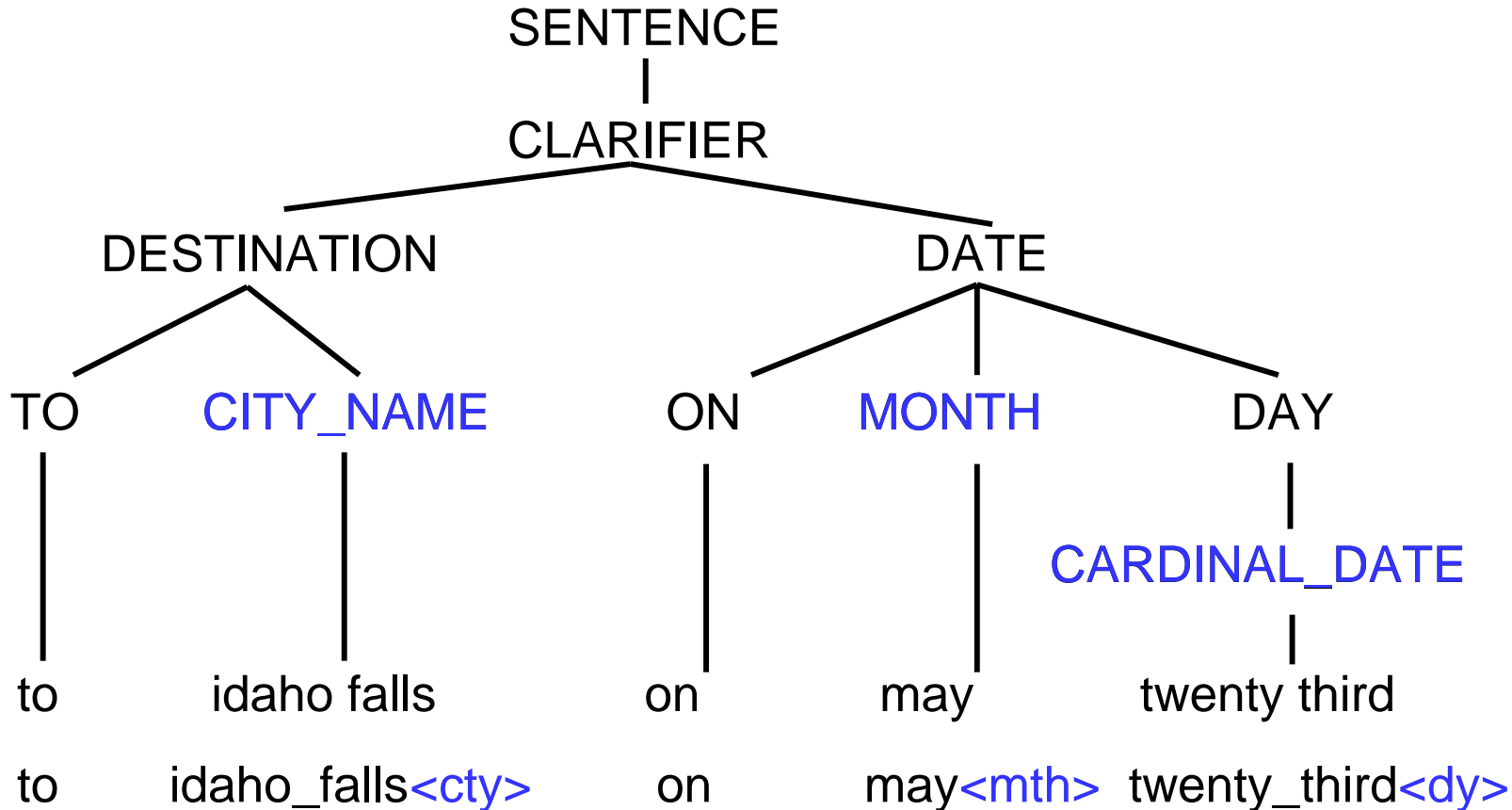
- By introducing NL constraints earlier, one can potentially reduce computation while improving performance



- Early integration can also remove the need for a statistical language model, which may be hard to obtain for some applications
- As the vocabulary size increases, we must begin to explore alternative search strategies
  - Parallel search
  - Fast search to reduce word candidate list

# Generating $n$ -grams from Parse Trees

- NLU can help generate a consistent class  $n$ -gram



- Developer identifies parse categories for class  $n$ -gram
- System tags words with associated class labels

# MIT Some SR/NL Coupling Experiments (ATIS Domain)

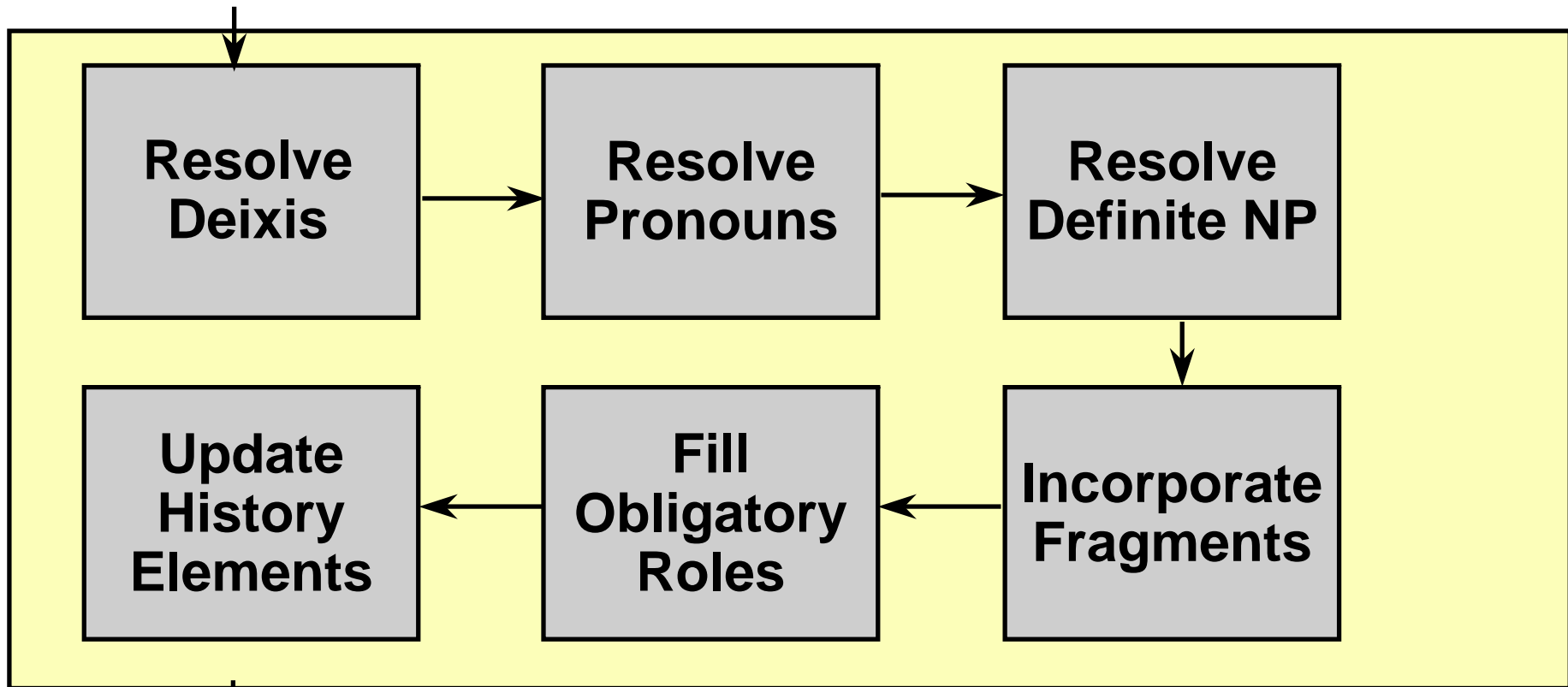
- **MIT (Goddeau, 1992)**
  - Probabilistic LR parser
  - Integrated into recognizer  $A^*$  search
  - Achieved comparable recognition accuracy to  $N$ -best resorting, but with considerably more efficiency
- **CMU (Ward, 1993)**
  - Modeled semantic concept sequences through trigram; and terminal word sequences through bigram
  - Integrated into recognizer  $A^*$  search
  - Reduced understanding (CAS) error by 10%
- **SRI (Moore, 1995)**
  - Modelled semantically meaningful fragments through trigram; and word classes through 4-gram
  - The NL score is added to the basic recognition score
  - Achieved ~15% word error reduction

# Typical Discourse Phenomena in Conversational Systems

- **Deictic (verbal pointing) and anaphoric (e.g., pronominal) reference:**
  1. Show me the restaurants in Cambridge.
  2. What is the phone number of **the third one**?
  3. How do I get **there** from the nearest subway stop?
- **Ellipsis:**
  1. When does flight twenty two arrive in Dallas?
  2. What is the departure time **()**?
- **Fragments:**
  1. What is the weather today in Denver?
  2. **How about** Salt Lake City?

# MIT's Discourse Module Internals

**Input Frame  
Displayed List**



**Interpreted Frame**

# MIT

## Different Roles of Dialogue Management

- **Pre-Retrieval: Ambiguous Input => Unique Query to DB**

U: I need a flight from Boston to San Francisco

C: Did you say Boston or Austin?

U: Boston, Massachusetts

C: I need a date before I can access Travelocity

U: Tomorrow

C: Hold on while I retrieve the flights for you

**Clarification  
(recognition errors)**

**Clarification  
(insufficient info)**

- **Post-Retrieval: Multiple DB Retrievals => Unique Response**

C: I have found 10 flights meeting your specification.  
When would you like to leave?

U: In the morning.

C: Do you have a preferred airline?

U: United

C: I found two non-stop United flights leaving in the morning ...

**Help the user narrow  
down the choices**

# MIT

## Multiple Roles of Dialogue Modeling

- **Our definition:** For each turn, preparing the system's side of the conversation, including responses and clarifications
- **Resolve ambiguities**
  - Ambiguous database retrieval (e.g. London, England or London, Kentucky)
  - Pragmatic considerations (e.g., too many flights to speak)
- **Inform and guide user**
  - Suggest subsequent sub-goals (e.g., what time?)
  - Offer dialogue-context dependent assistance upon request
  - Provide plausible alternatives when requested information unavailable
  - Initiate clarification sub-dialogues for confirmation
- **Influence other system components**
  - Adjust language model due to dialogue context
  - Adjust discourse history due to pragmatics (e.g., New York)

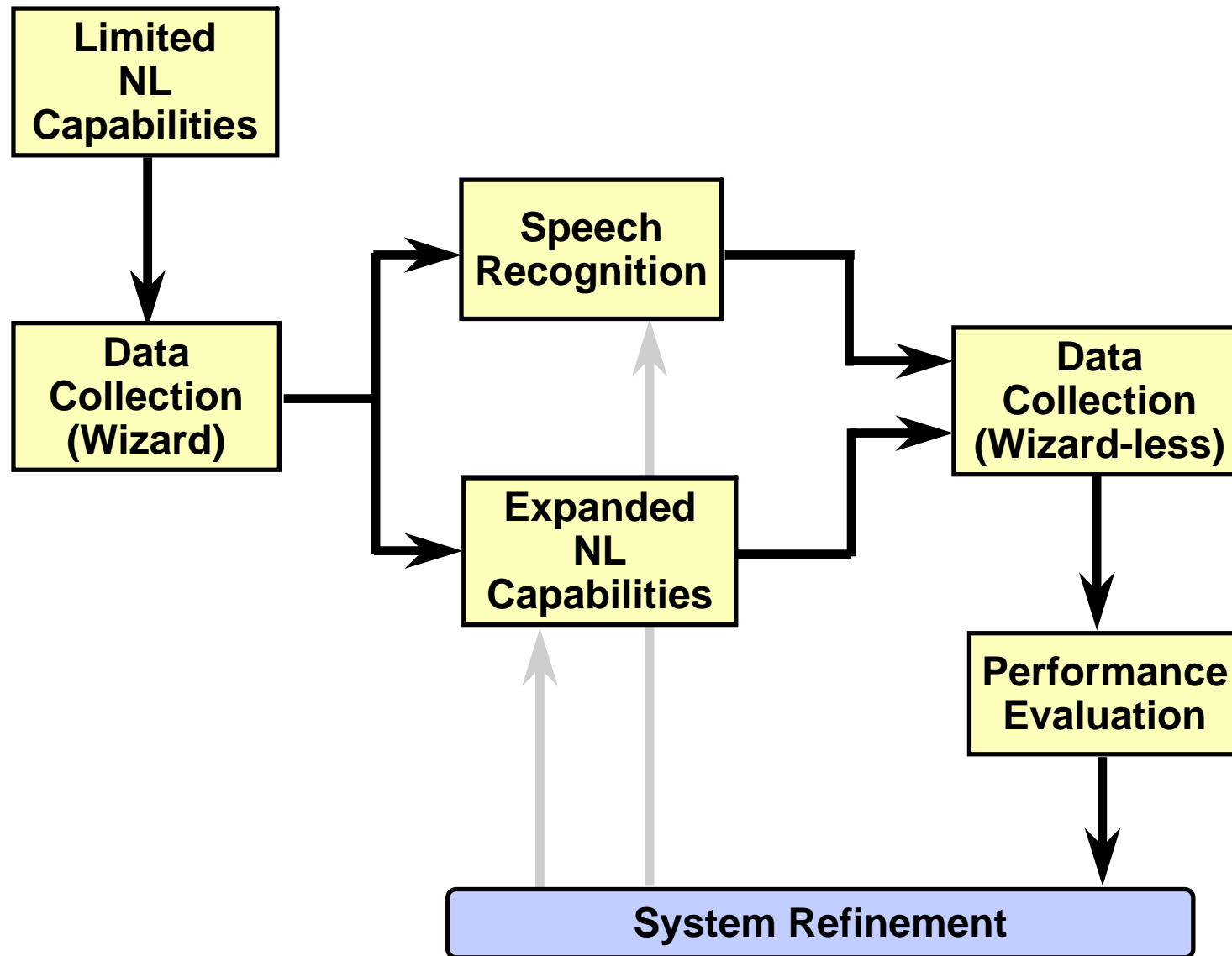


# MIT

## An Attractive Strategy

- **Conduct R&D of human language technologies within the context of real application domains**
  - **Forces us to:**
    - \* Confront critical technical issues (e.g., rejection, new word problem) and
    - \* Set priorities (e.g., better match technical capabilities with useful applications)
  - **Provides a rich and continuing source of useful data**
    - \* Real data from real users are invaluable
  - **Demonstrates the usefulness of the technology**
  - **Facilitates technology transfer**

# System Development Cycle

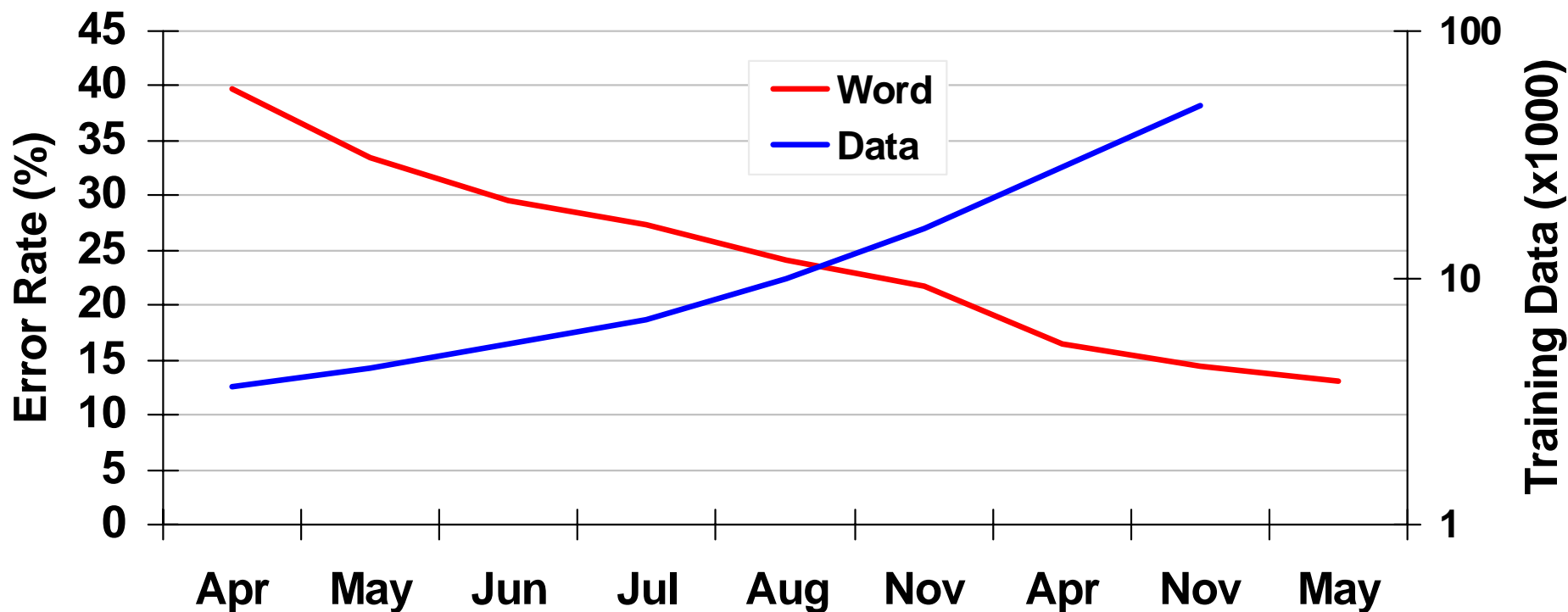


# MIT

## Data Collection

- **System development is chicken & egg problem**
- **Data collection has evolved considerably**
  - Wizard-based → system-based data collection
  - Laboratory deployment → public deployment
  - 100s of users → thousands → millions
- **Data from **real** users solving **real** problems accelerates technology development**
  - Significantly different from laboratory environment
  - Highlights weaknesses, allows continuous evaluation
  - But, requires **systems** providing **real** information!
- **Expanding corpora will require unsupervised training or adaptation to unlabelled data**

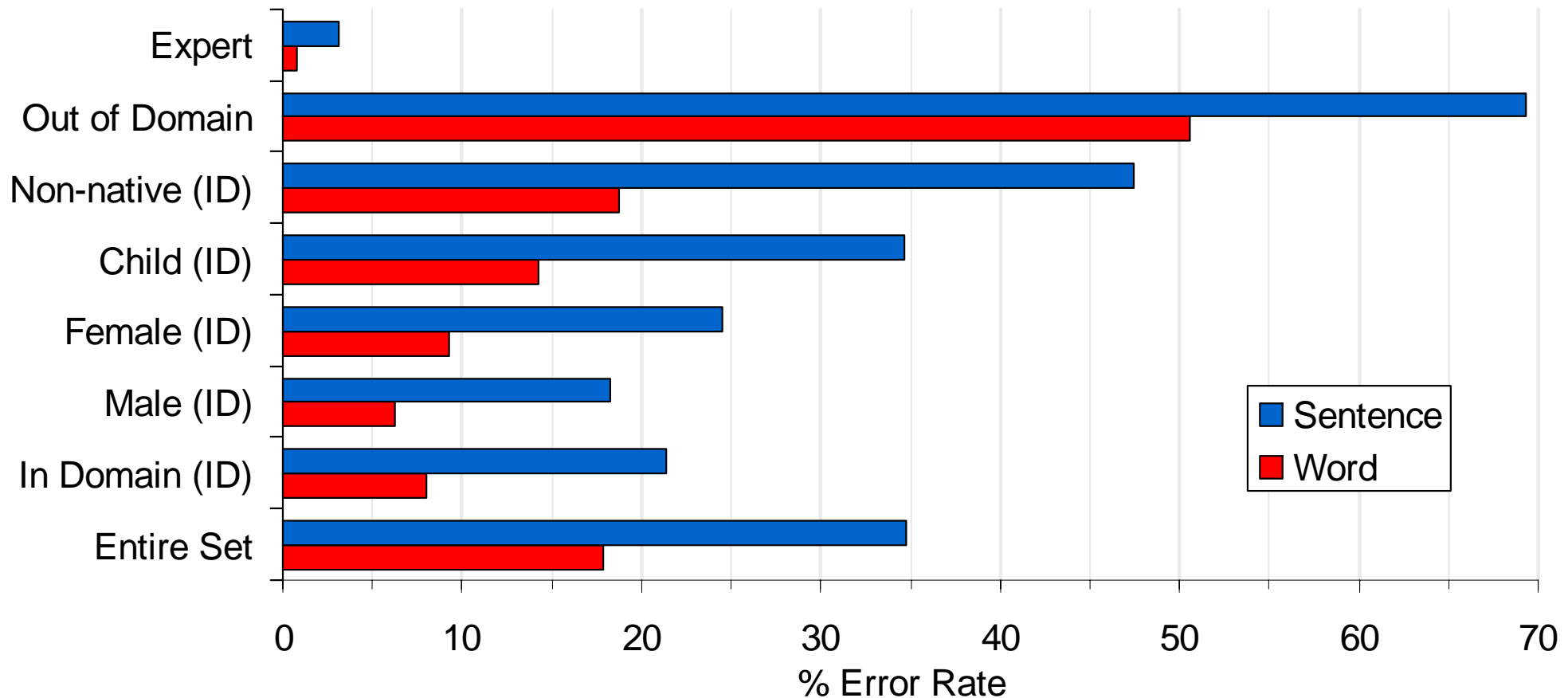
## Data vs. Performance (Weather Domain)



- Longitudinal evaluations show improvements
- Collecting real data improves performance:
  - Enables increased complexity and improved robustness for acoustic and language models
  - Better match than laboratory recording conditions
- Users come in all kinds



# ASR Error Analysis (Weather Domain)



- **Male ERs are better than females (1.5x) and children (2x)**
- **Strong foreign accents and out-of-domain queries are hard**
- **Experienced users are 5x better than novices**
- **Understanding error rate is consistently lower than SER**

# Examples of Spoken Dialogue Systems

## Asia

- Canon **TARSAN** (Japanese)
  - Info retrieval from CD-ROMs
- InfoTalk (Cantonese)
  - Transit fare
- KDD **ACTIS** (Japanese)
  - Area-codes, country codes, and time-difference
- NEC (Japanese)
  - Ticket reservation
- NTT (Japanese)
  - Directory assistance
- SpeechWorks (Chinese)
  - Stock quotes
- Toshiba **TOSBURG** (Japanese)
  - Fast food ordering

## U.S.

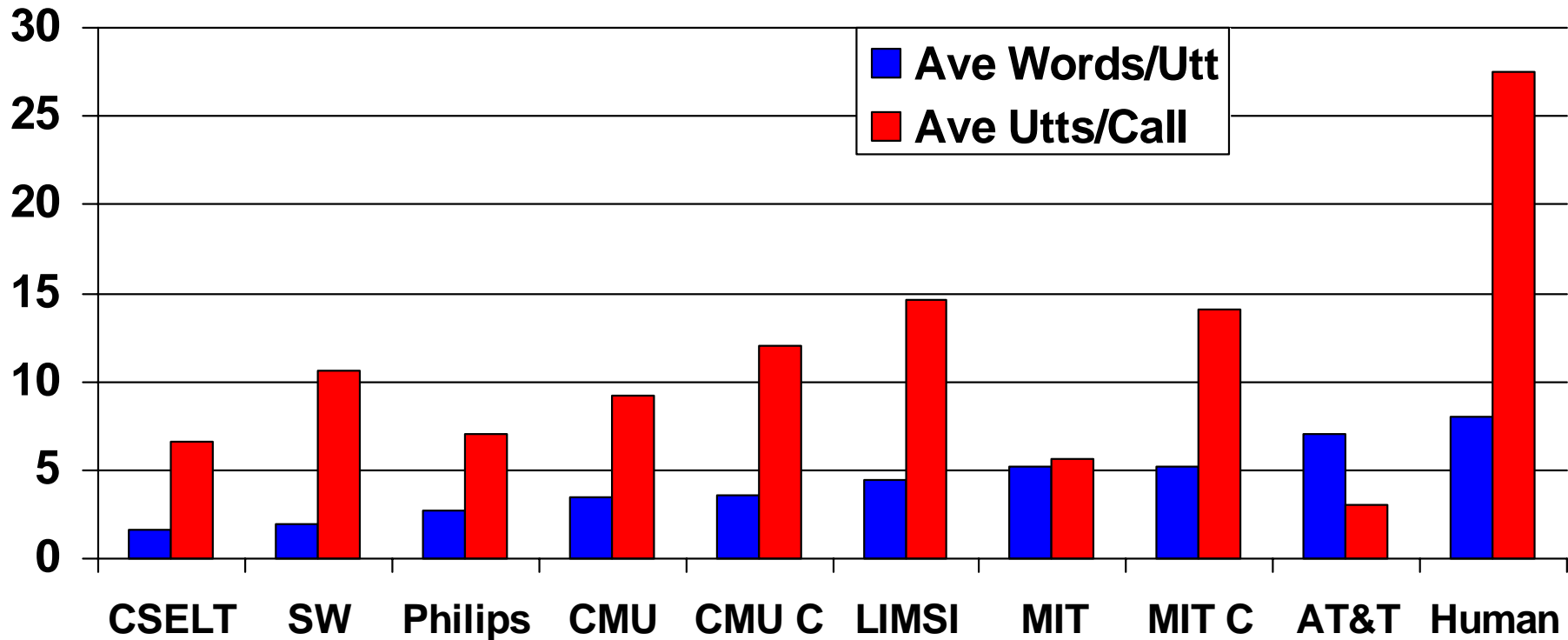
- AT&T **How May I Help You?**
- BBN **Call Routing**
- CMU **Movieline, Travel,...**
- Colorado U **Travel**
- IBM **Mutual funds, Travel**
- Lucent **Movies, Call Routing**
- MIT **Jupiter, Voyager, Pegasus**
  - Weather, navigation, flight
- Nuance **Finance, Travel,...**
- OGI **CSLU Toolkit**
- SpeechWorks **Finance, Travel**
- UC-Berkeley **BERP**
  - Restaurant information
- U Rochester **TRAINS**
  - Scheduling trains

## Europe

- CSELT (Italian)
  - Train schedules
- KTH **WAXHOLM** (Swedish)
  - Ferry schedule
- LIMSI (French)
  - Flight/train schedules
- Nijmegen (Dutch)
  - Train schedule
- Philips (Dutch,Fr.,German)
  - Flight/Train schedules
- Vocalis **VOCALIST** (English)
  - Flight schedules

- Large-scale deployment of some dialogue systems
  - e.g., CSELT, Nuance, Philips, SpeechWorks

# Example Dialogue Systems



- **Vocabularies typically have 1000s of words**
- **Widely deployed systems tend to be more conservative**
- **Directed dialogues have fewer words per utterance**
- **Word averages lowered by more confirmations**
- **Human-human conversations use more words**

# Some Speech Recognition Research Issues

- **Widespread robustness to environments & speakers**
  - **Channel conditions:**
    - \* Wide-band → telephone → cellular
    - \* Wide-band → microphone arrays (echo cancellation)
  - **Conversational speech phenomena**
  - **Speaker variation (native → non-native)**
- **Knowing what you don't know**
  - **Confidence scoring (utterance & word)**
  - **Out-of-vocabulary word detection & addition**
- **Beyond word  $n$ -grams?**
  - **Providing coverage, constraint, and a platform for understanding**
- **Other challenges:**
  - **Adaptation (long-term → short-term)**
  - **Domain-independent acoustic and language modelling**



- **Variety of methods explored to achieve robust understanding**
  - Full grammars with back-off to robust parse (e.g, Seneff)
  - Semantic grammars, template-based approaches (e.g., Ward)
  - Stochastic speech-to-meaning models (e.g., Miller, Levin et al.)
  - Ongoing work in automatic grammar acquisition (e.g., Roukos et al., Kuhn et al.)
- **Interface mechanisms**
  - Two-stage *N*-best/word-graph vs. coupled search
  - How do we achieve understanding during decoding?
- **Ongoing challenges:**
  - Domain-independent language understanding
  - Will current approaches scale to more complex or general understanding tasks?
  - Integration of multimodal inputs into a common understanding framework (e.g., Cohen, Flanagan, Waibel)

# MIT

## Some Dialogue Research Issues

- **Modeling human-human conversations?**
  - Are human-human dialogues a good model for systems?
  - If so, how do we structure our systems to enable the same kinds of interaction found in human-human conversations?
- **Implementation strategies:**
  - Directed vs. mixed-initiative with back-off (e.g., Lamel et al.)
  - Machine-learning of dialogue strategies (e.g., Levin et al.)
- **Handling user dialogue phenomena**
  - Interruptions (via barge-in), anaphora, ellipsis
  - Barge-in can increase complexity of discourse
- **Modeling agent dialogue phenomena**
  - Back-channel (e.g., N. Ward)
- **Other issues:**
  - Detecting and recovering from errors (e.g., Walker et al.)
  - Matching capabilities with expectations

- **Spoken dialogue systems are needed, due to**
  - Miniaturization of computers
  - Increased connectivity
  - Human desire to communicate
- **To be truly useful, these interfaces must be conversational in nature**
  - Embody linguistic competence, both input and output
  - Help people solve real problems efficiently
- **Systems with limited capabilities are emerging**
- **Much research remains to be done**