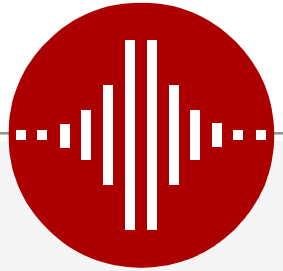
SPOKEN LANGUAGE SYSTEMS



Spoken Computer Conversational Systems

Stephanie Seneff

CSAIL, MIT

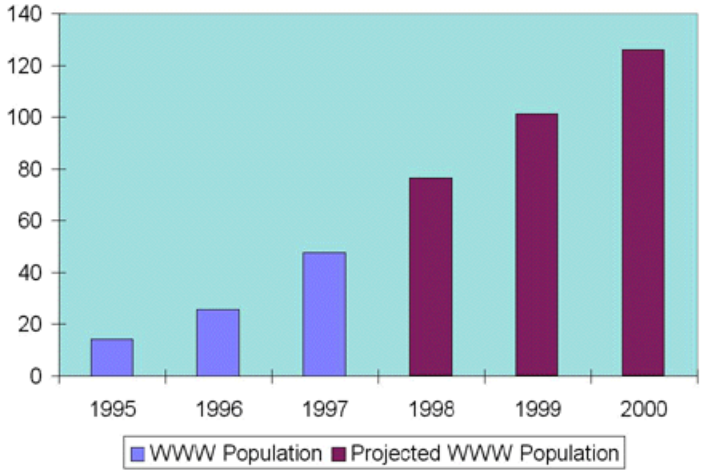
April 12, 2004

Outline

- Introduction and historical context
- Speech understanding
- Discourse and dialogue modeling
- Data collection and evaluation
- Rapid development of new domains
- Flexibility and personalization
- Future research challenges

The Premise:

Everybody wants Information



For North America
CommerceNet
Research Center (1999)

Even when they are on the move

Devices must be small

Need new interfaces

The interface must be easy to use

Speech is It!

What Are Conversational Systems?

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

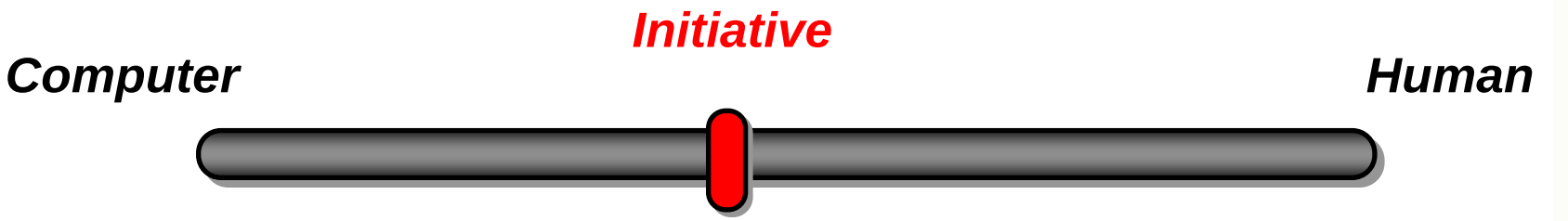
- *Understand* verbal input, using
  - Speech recognition
  - Language understanding (in context)
- *Verbalize* response, using
  - Language generation
  - Speech synthesis
- Engage in *dialogue* with a user during the interaction

An Attractive Strategy

- **Conduct R&D of human language technologies within the context of real (and useful) application domains**
  - Flight schedules and status, weather, restaurant or hotel guide, calendar management, city navigation, traffic reports, sports updates, etc.
- **Forces us to confront critical technical issues (e.g., error recovery, new word problem)**
- **Provides a rich and continuing source of useful data**
- **Demonstrates the usefulness of the technology**
- **Facilitates technology transfer**

Dialogue Interaction Modes

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



- **Computer maintains tight control**
- **Human is highly restricted**

- **Human takes complete control**
- **Computer is totally passive**

C: ***Please say the departure city.***

H: ***I want to visit my grandmother.***

Mixed Initiative Dialogue

The Nature of Mixed Initiative Interactions
(A Human-Human Example)

.....		disfluency
Customer:	**Yeah, [um] I'm looking for the Buford Cinema.**	
Agent:	**OK, and you're wanting to know what's showing there or .**	interruption, overlap
Customer:	**Yes, please.**	confirmation
Agent:	**Are you looking for a particular movie?**	clarification
Customer:	**[um] What's showing.**	
Agent:	**OK, one moment.**	back channel
.....		clarification subdialogue
Agent:	**They're showing *A Troll In Central Park*.**	
Customer:	**No.**	inference
Agent:	***Frankenstein*.**	ellipsis
Customer:	**What time is that on?**	co-reference
Agent:	**Seven twenty and nine fifty.**	
Customer:	**OK, and the others?**	complex quantifier

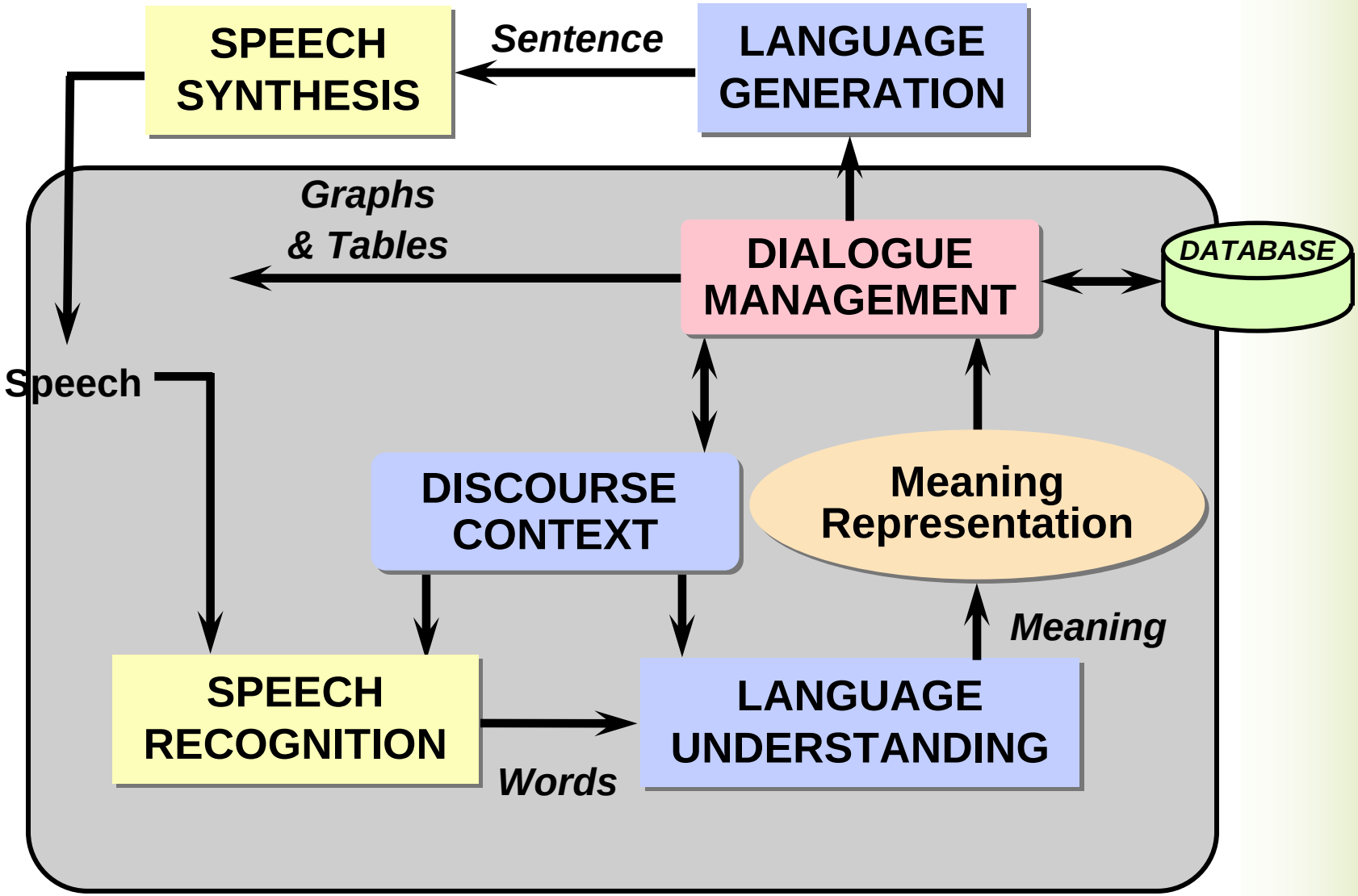
Current Landscape

- **Simple, *directed dialogue* systems are being deployed commercially**
- **Application-specific, *mixed-initiative* spoken dialogue systems are emerging from universities and other research institutions**
- **Research on *multi-modal* mixed-initiative dialogue systems is beginning**

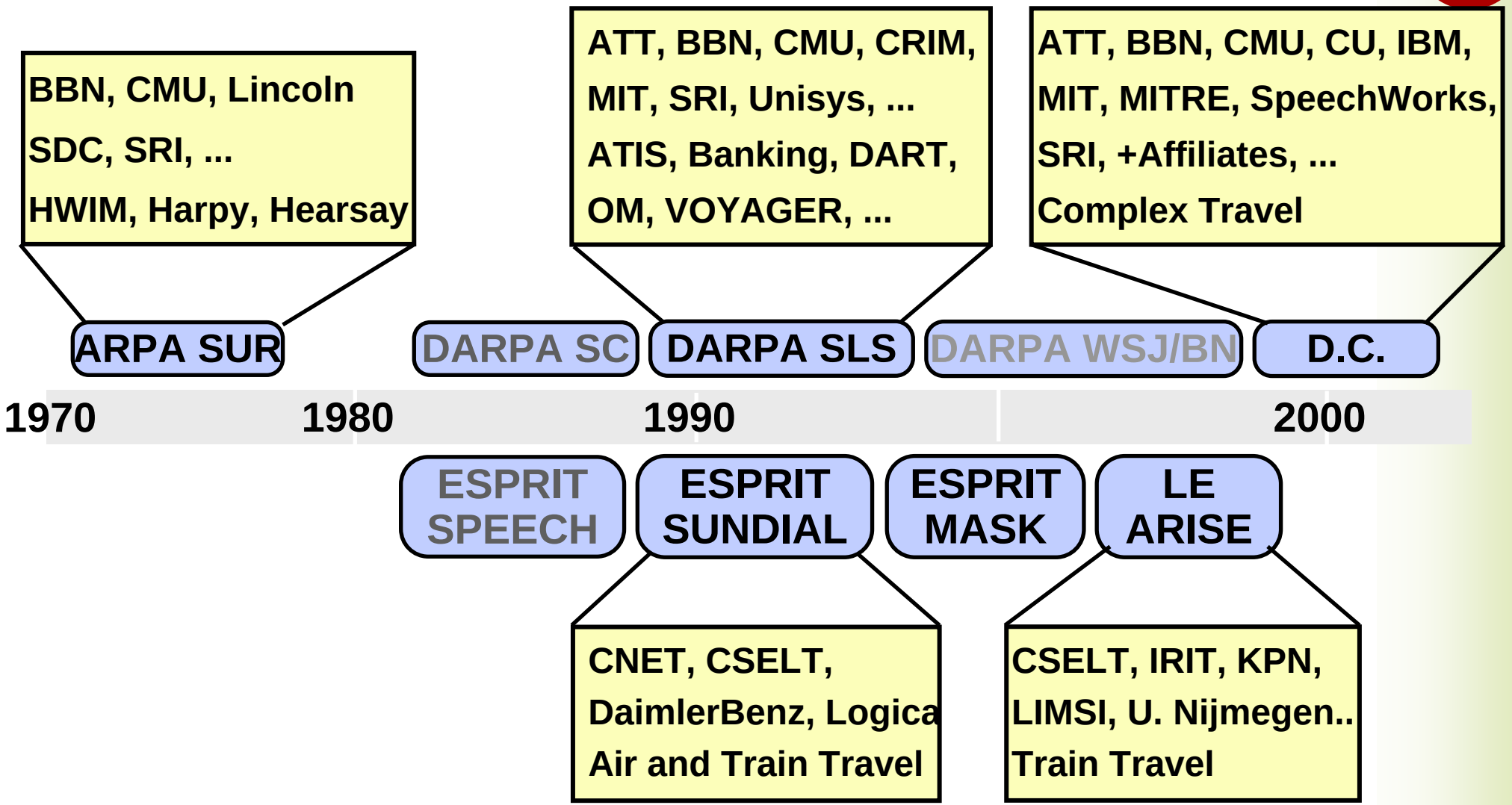
Dialogue Management Strategies

- ***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**
  - User has flexibility to specify constraints in any order
  - System can “back off” to a directed dialogue under certain circumstances
  - Mixed-initiative dialogue systems are mainly research prototypes

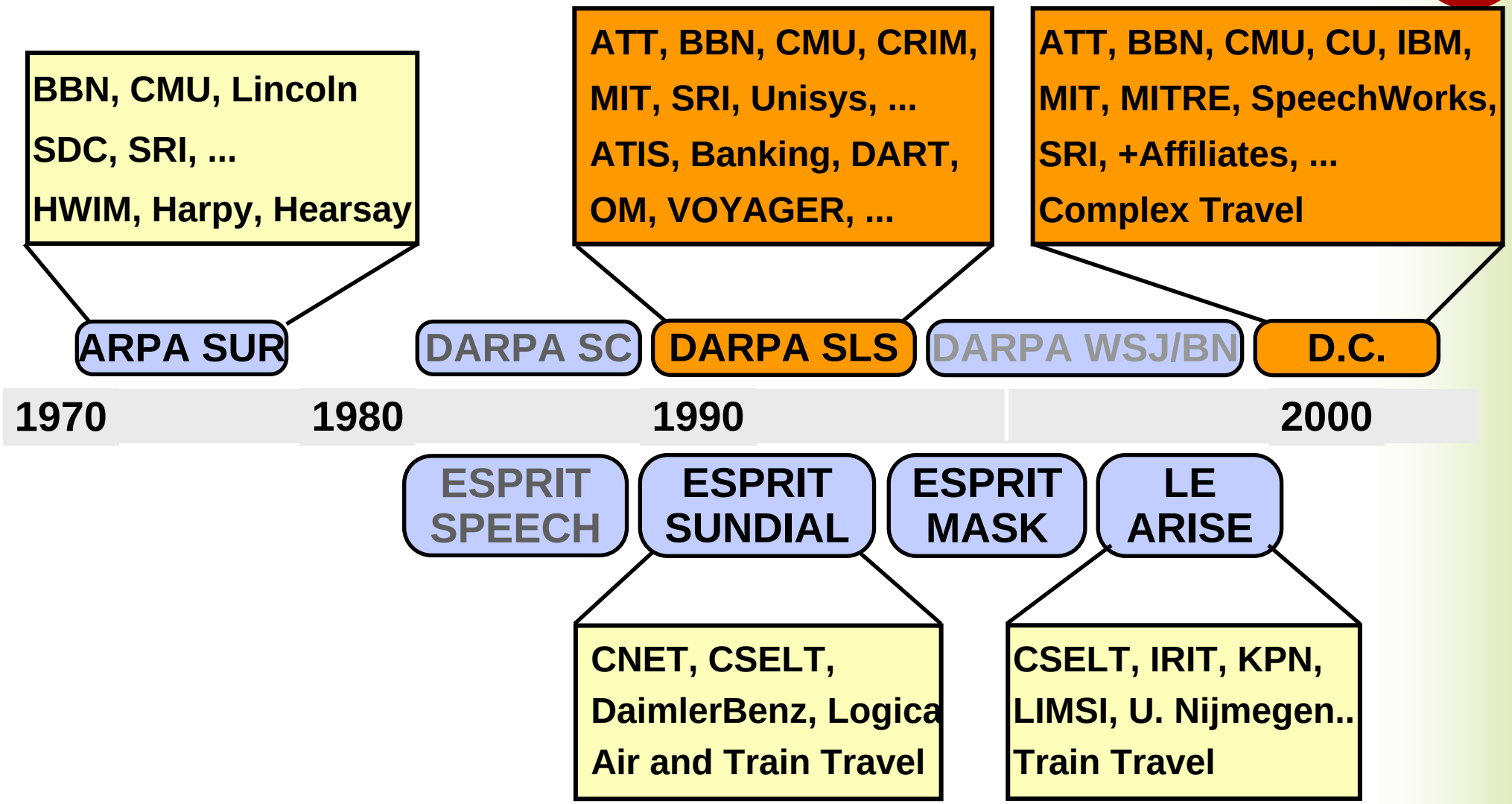
Components of a Conversational System



Some Speech-Related Government Programs



Some Speech-Related Government Programs



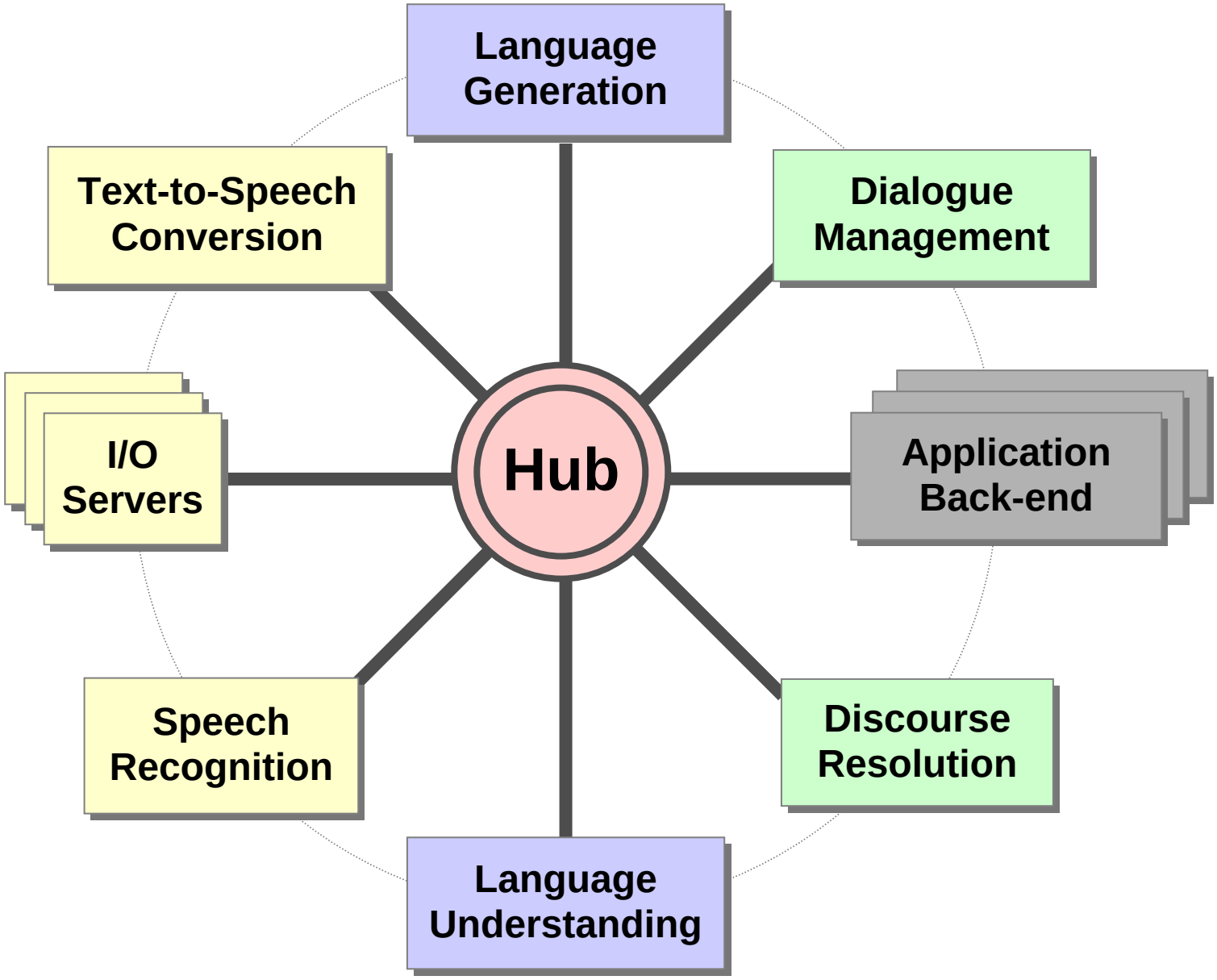
The U.S. DARPA SLS Program (1990-1995)

- **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)
- **All systems could handle continuous speech from unknown speakers (~2,000 word vocabulary)**
- **Research driven by five annual common evaluations**
  - CAS evaluation methodology – heavy dependence on strict manually provided "correct" answers
  - restricted system design to passive mode of interaction
  - shifted focus away from user interface

The U.S. DARPA Communicator Program (1998-2003)

- **Still focusing on flight scheduling domain**
  - Sites were tasked with finding their own *real* flight database
- **Emphasis on common architecture with plug-and-play capabilities: Galaxy Communicator**
- **Generally much larger set of cities than in ATIS (>500), and covering at least major airports world-wide**
- **Sites were free to organize dialogue interaction in any way they chose**
  - Encouraged mixed-initiative dialogue development
- **Evaluation was conducted on per-site basis and depended critically on user exit polls**
  - Users were frequent travelers booking their real travel arrangements

Galaxy Communicator Architecture



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)**

Some other Spoken Dialogue Systems

Asia

- Canon TARSAN (Japan
  - Info retrieval from C
- InfoTalk (Cantonese)
  - Transit fare
- KDD ACTIS (Japanese)
  - Area-codes, country and time-difference
- NEC (Japanese)
  - Ticket reservation
- NTT (Japanese)
  - Directory assistance
- SpeechWorks (Chinese
  - Stock quotes
- Toshiba TOSBURG (Ja
  - 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 Routin
- MIT Jupiter, Voyager, Pega
  - Weather, navigation, flig
- Nuance Finance, Travel,...
- OGI CSLU Toolkit
- SpeechWorks Finance, Tra
- 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, ScanSoft (SpeechWorks)

Outline

- Introduction and historical context
- Speech understanding
- Context resolution and dialogue modeling
- Data collection and evaluation
- Rapid development of new domains
- Flexibility and personalization
- Future research challenges

Natural Language Processing Components

SLS

- **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 Context 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.

Users can be very creative with language, especially when frustrated

Examples from ATIS domain:

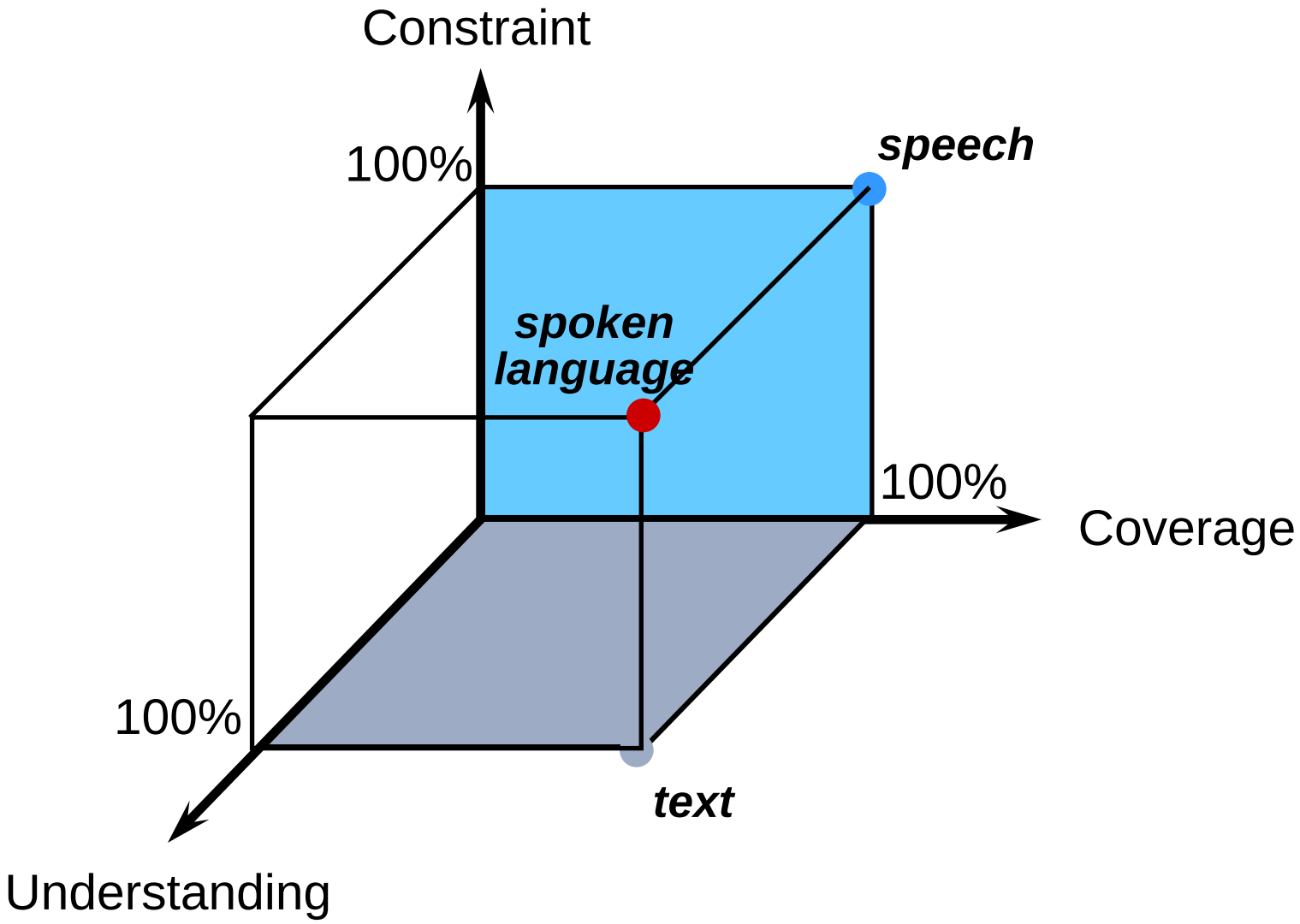
- I would like to find a flight from Pittsburgh to Boston on Wednesday and I have to be in Boston by one so I would like a flight out of here no later than 11 a.m.
- I'll repeat what I said before I'm on scenario 3 I would like a 727 flight from Washington DC to Atlanta Georgia I would like it during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

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

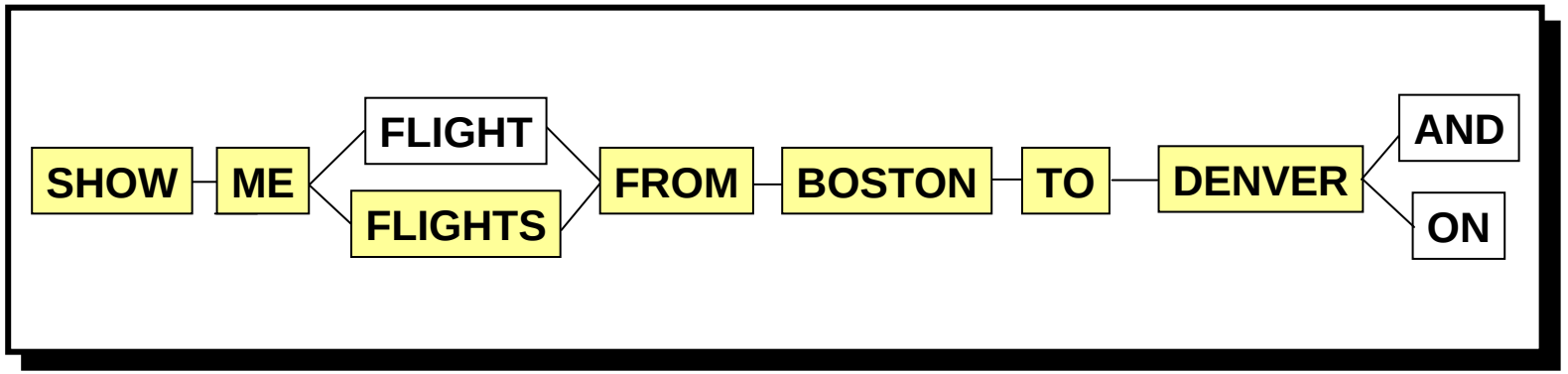
Spoken Language Understanding

- **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**

Multiple Roles for Natural Language Parsing in Spoken Language Context



Input Processing: Understanding



Speech Waveform

SPEECH RECOGNITION

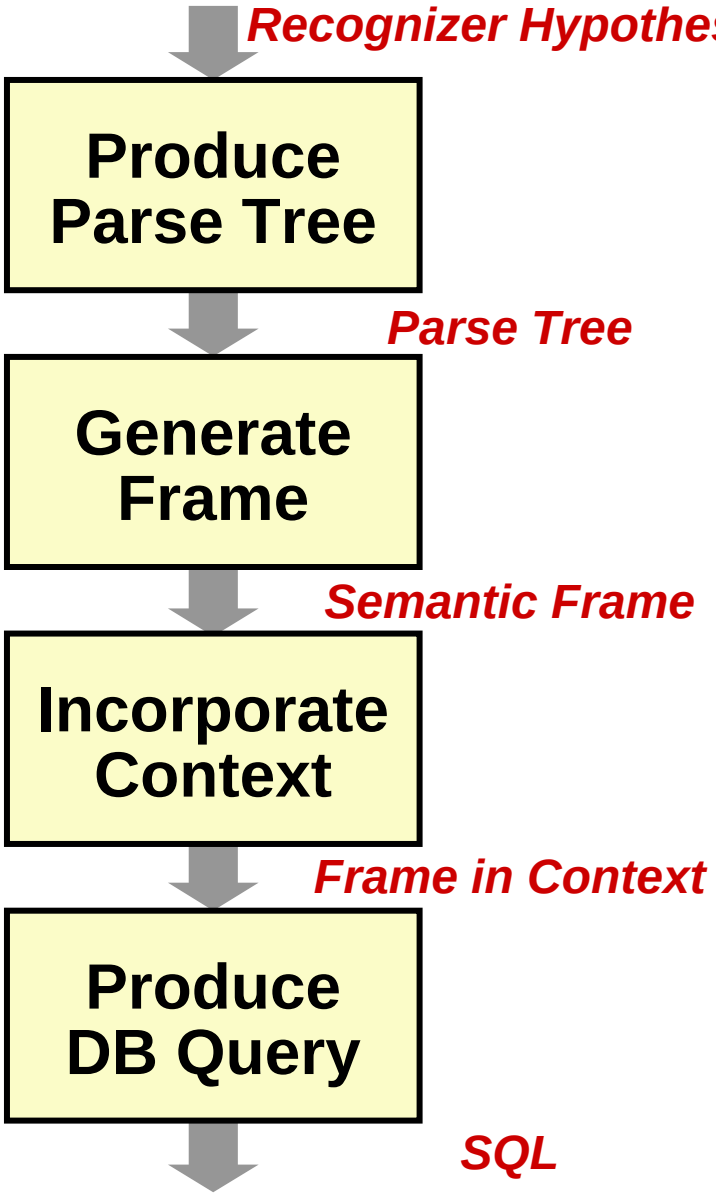
Sentence Hypotheses

LANGUAGE UNDERSTANDING

Semantic Representation

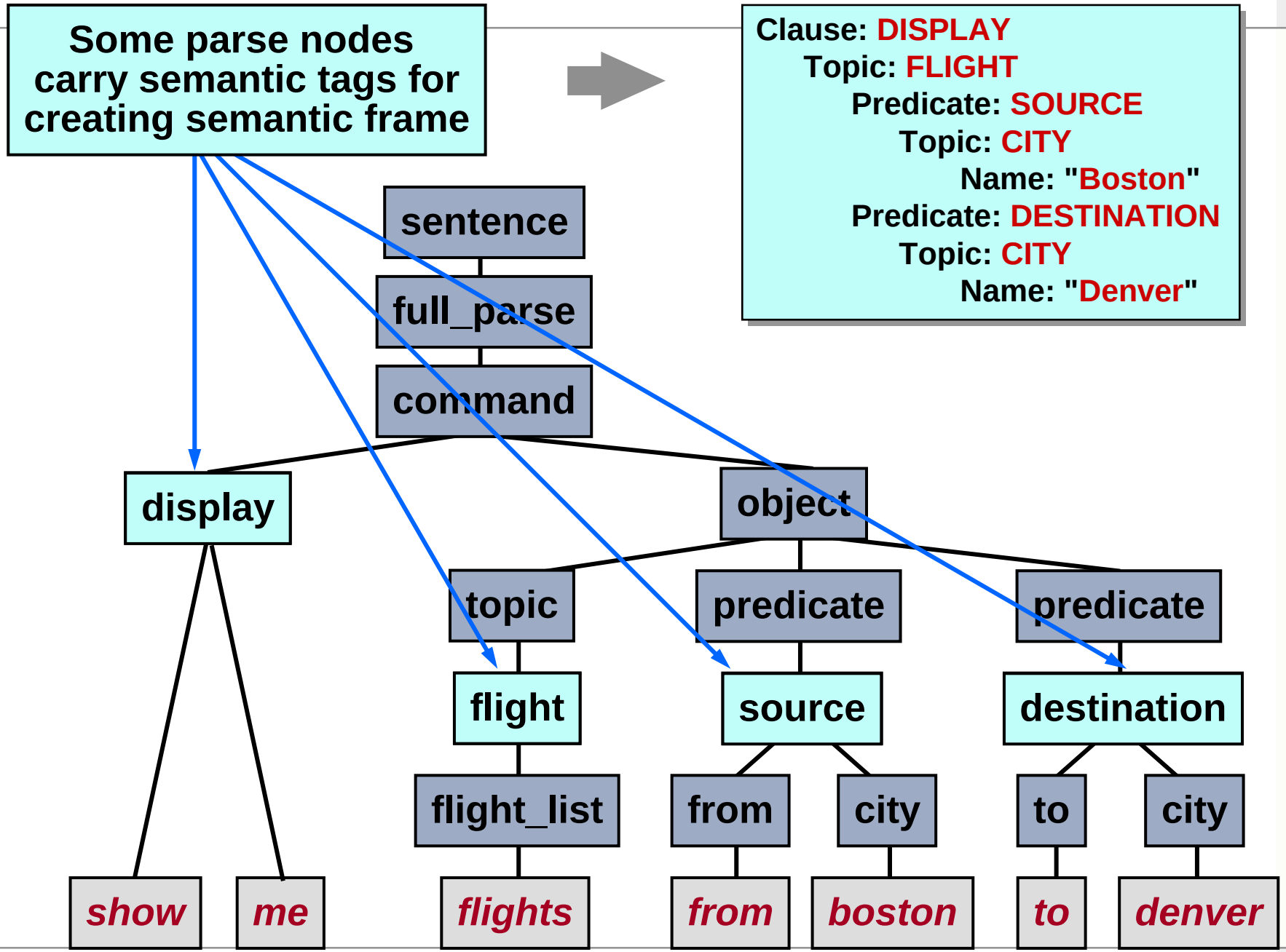
```
Clause: DISPLAY
    Topic: FLIGHT
        Predicate: FROM
            Topic: CITY
                Name: "Boston"
        Predicate: TO
            Topic: CITY
                Name: "Denver"
```

Typical Steps in Transforming User Query



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

Example of Natural Language Understanding



Context Free Rules for Example

SLS

Show me flights from Boston to Denver

sentence	→	full_parse [robust_parse]
full_parse	→	(command question statement …)
command	→	display object
object	→	[determiner] topic [predicate] [predicate]
predicate	→	(source destination depart_time …)
source	→	from (city airport)
destination	→	to (city airport)
display	→	*show me*
city	→	(*boston dallas denver* …)
determiner	→	(*a the*)
...		

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

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**

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

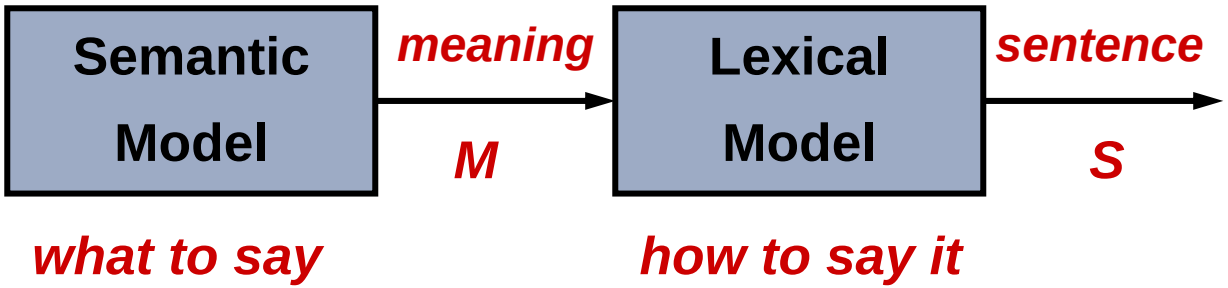
"six" could mean:
- A time
- A count
- A flight number

- **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**

MIT's TINA NL System

- **TINA was designed for speech understanding**
  - Grammar rules intermix syntax and semantics
  - Probabilities are trained from user utterances
  - Parse tree converted to a semantic frame that encapsulates meaning
- **TINA enhances its coverage through robust parsing**
  - "Full parse" hypotheses are preferred
  - Backs off to parsing fragments and skipping unimportant words
  - Fragments are combined into a full semantic frame
  - When all else fails, system backs off to phrase spotting

Stochastic Approaches to Language Understanding

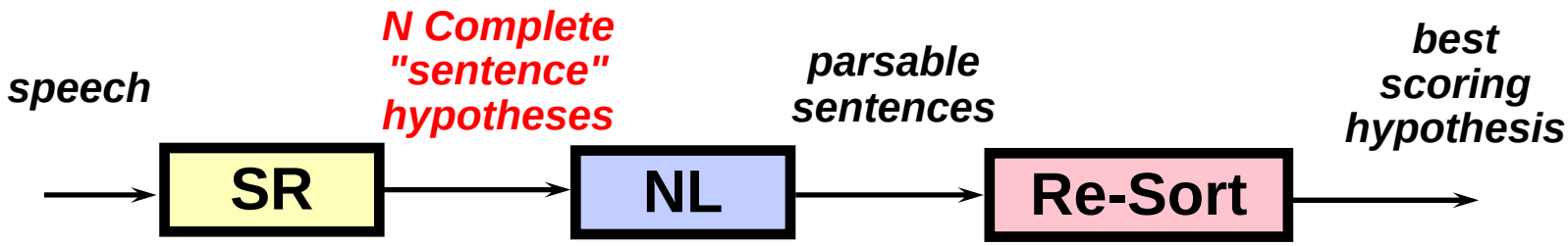


- **Choose among all possible meanings the one that maximizes:**

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

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

SR/NL Integration via *N*-Best Interface



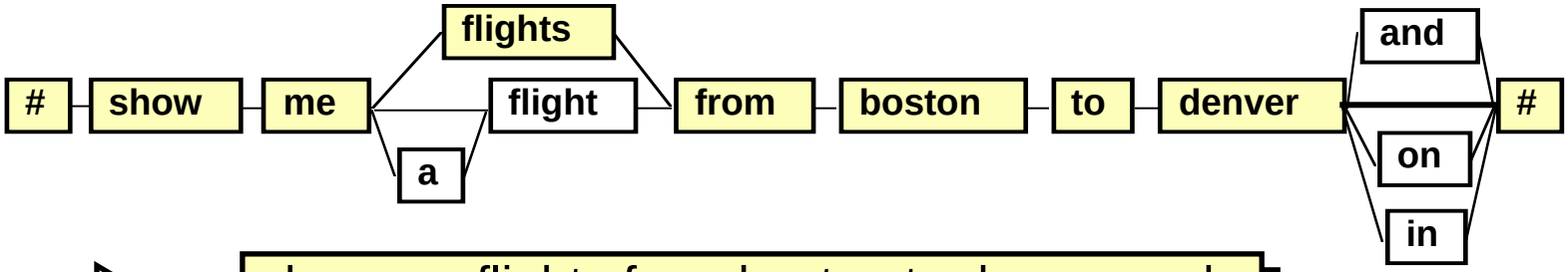
show me flights from boston to denver and
show me flights from boston to denver
show me flights from boston to denver on
show me flight from boston to denver and
show me flight from boston to denver
show me flight from boston to denver on
show me flights from boston to denver in
show me a flight from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

Word Networks: Efficient Representation of N-best Lists

SLS



If the parser can propose probabilities for next-word theories in the network, then it can be used to adjust theory scores in second pass through the network

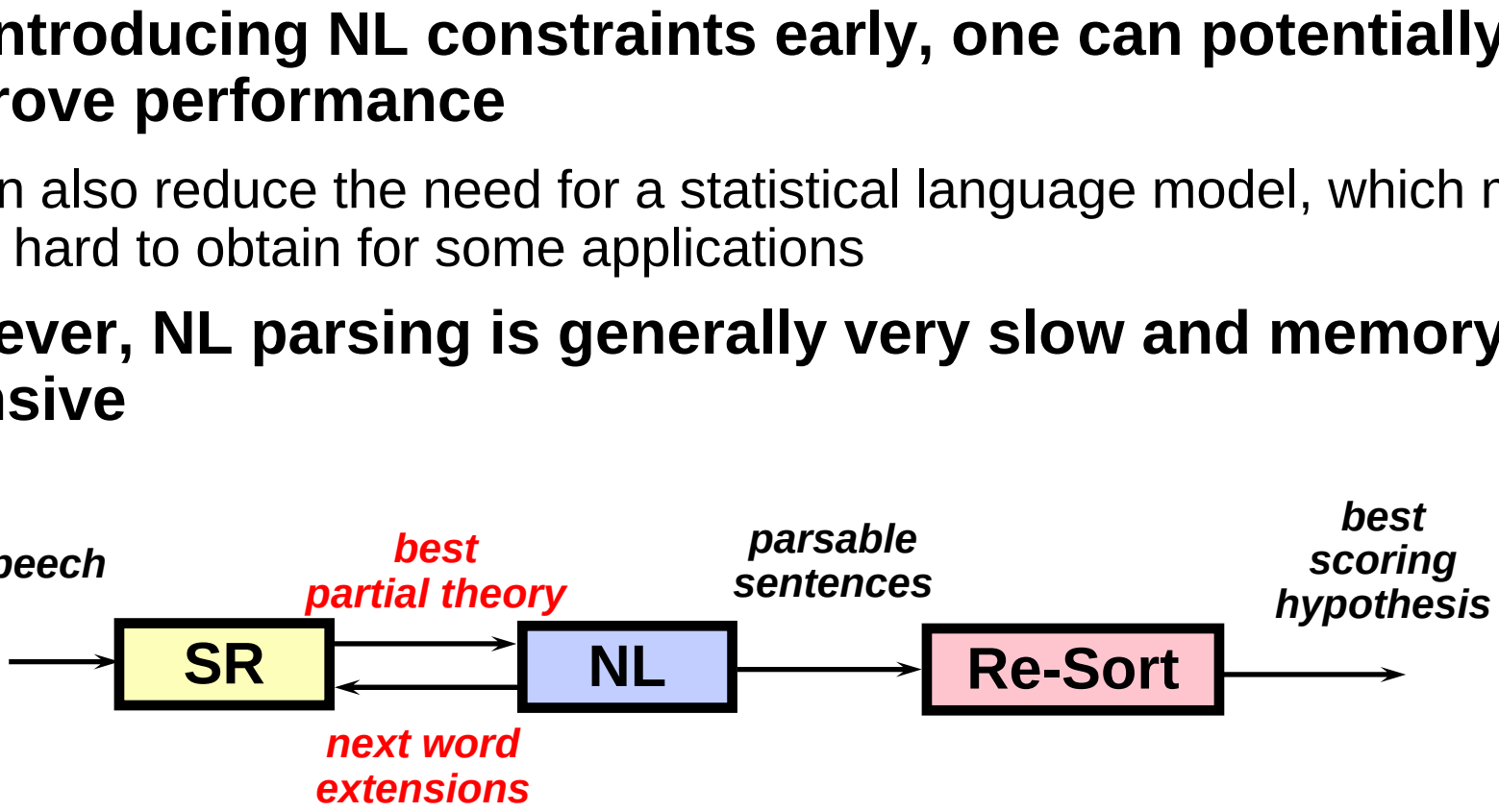
answer

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

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

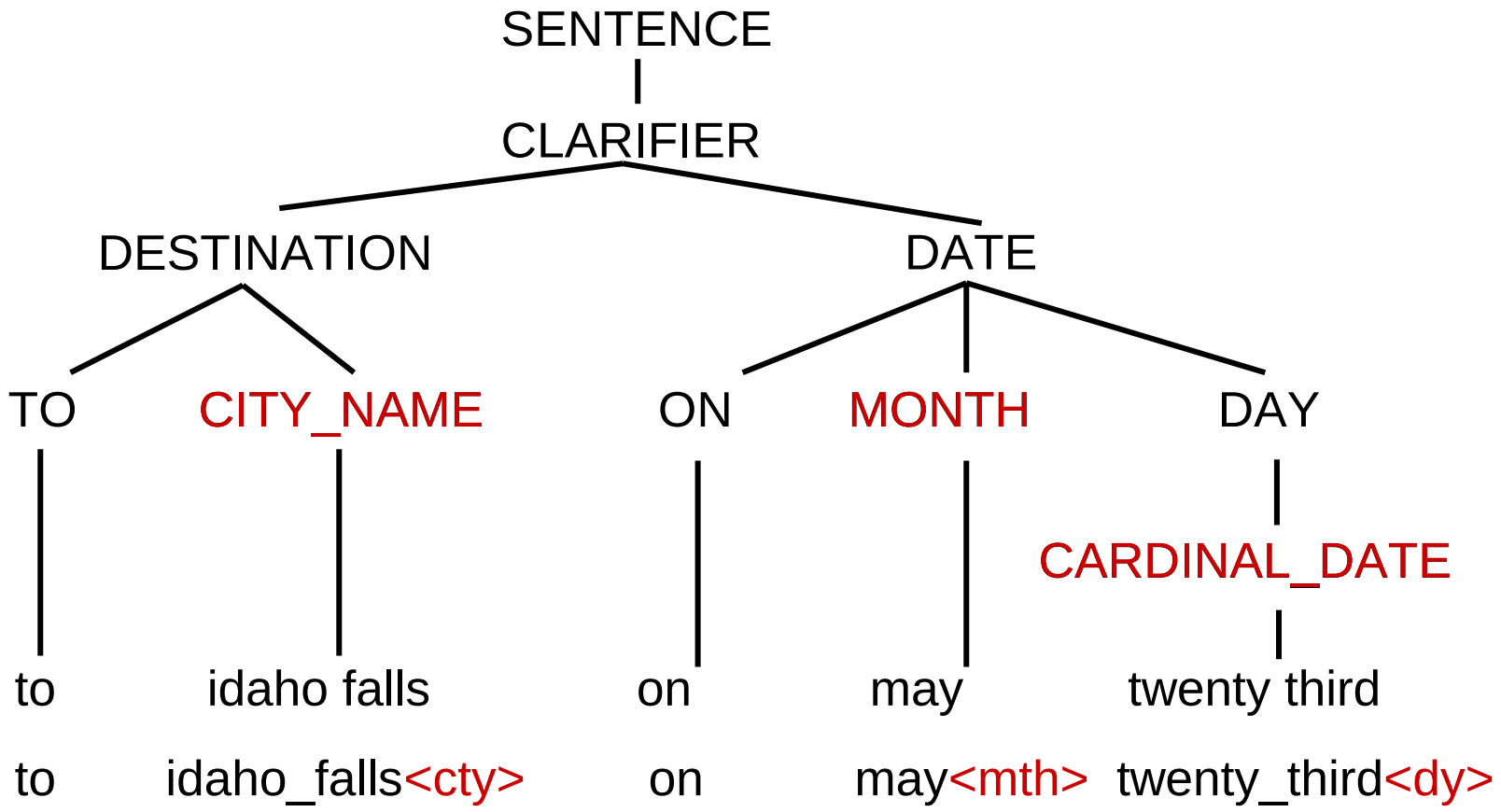


- **By introducing NL constraints early, one can potentially improve performance**
  - can also reduce the need for a statistical language model, which may be hard to obtain for some applications
- **However, NL parsing is generally very slow and memory intensive**

Incorporating Soft NL Constraints into Recognizer

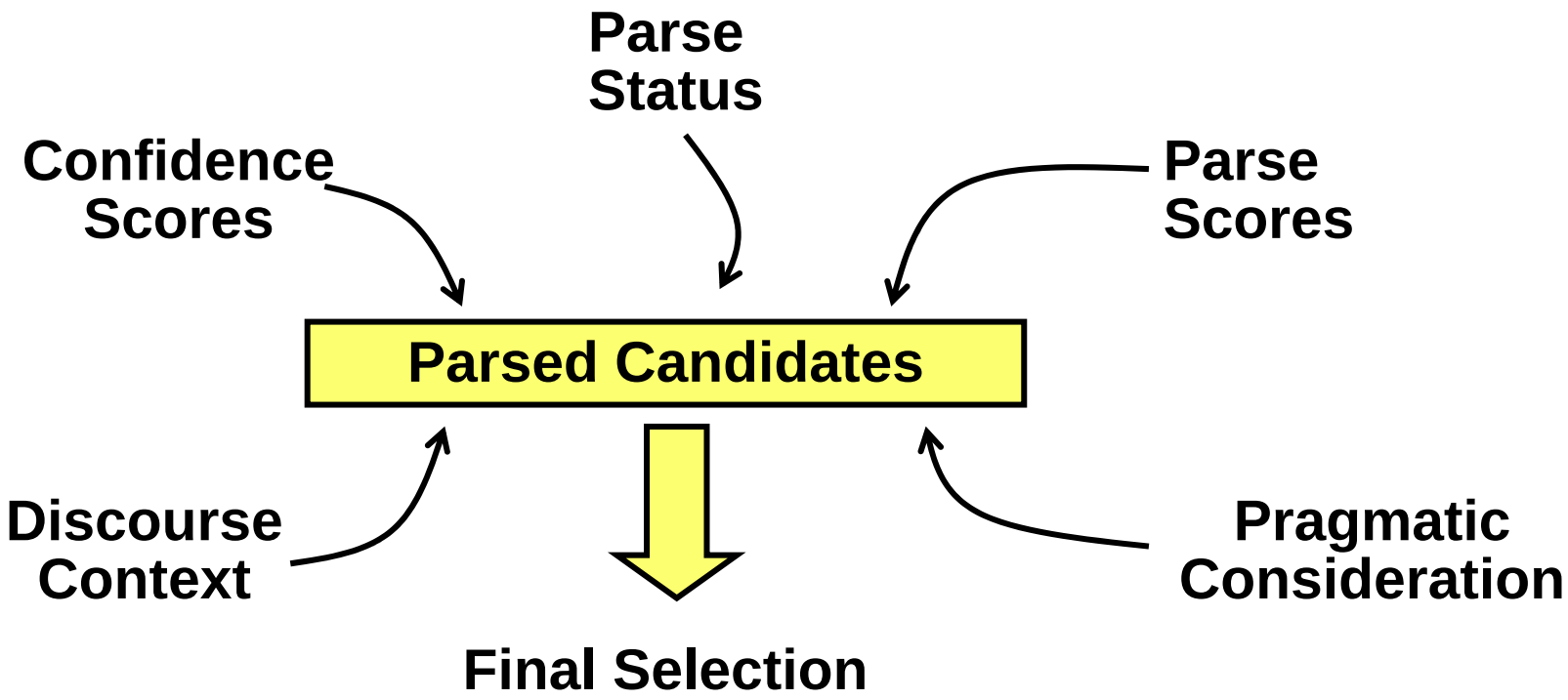
SLS

- **Class *n*-gram can be automatically derived from NL Grammar**



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

Final Hypothesis Selection Process



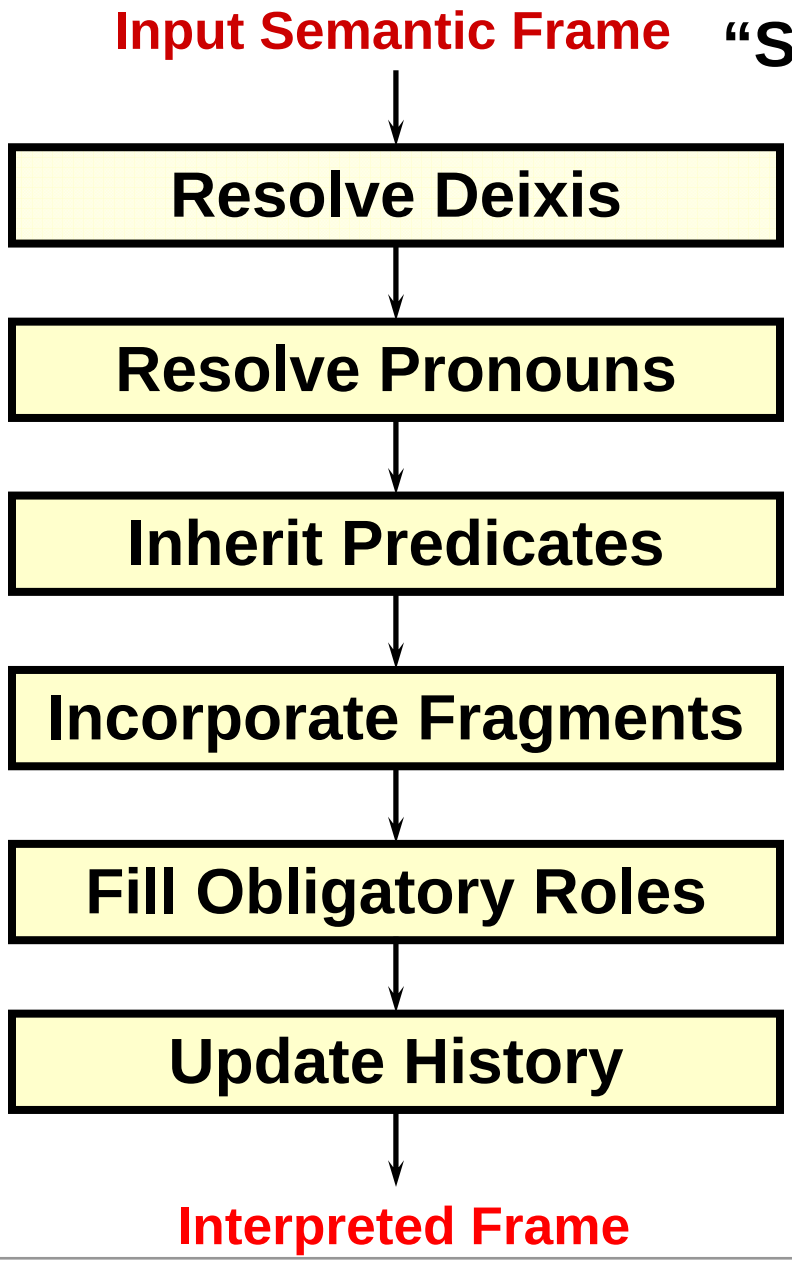
Outline

- Introduction and historical context
- Speech understanding
- Context resolution and dialogue modeling
- Data collection and evaluation
- Rapid development of new domains
- Flexibility and personalization
- Future research challenges

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?

Typical Context Resolution Tasks



"Show me restaurants in Cambridge."

"What does this one serve?"

"What is their phone number?"

"Are there any on Main Street?"

"What about Mass Ave?"

"Give me directions from MIT."

Stages of Dialogue Interaction

- **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: What date will you be traveling?
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

Multiple Roles of Dialogue Modeling

- **For each turn, prepare 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
  - Update discourse context

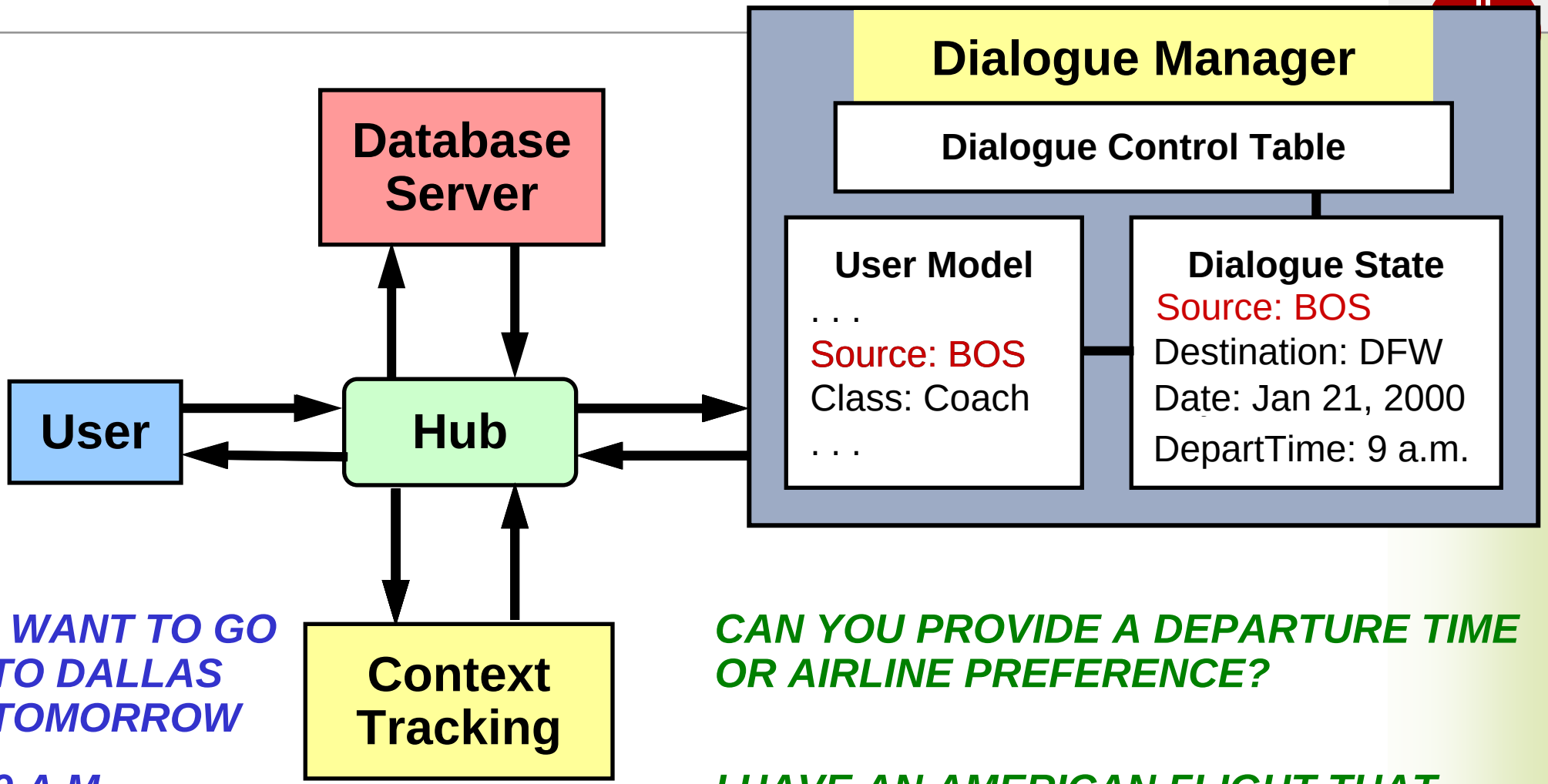
Table-Driven Dialogue Control

- **Set of operations perform specialized tasks**
- **Ordered set of rules specify active operations**
- **Dynamic set of state variables drive rule execution**
- **A rule fires when conditions are met:**
  - Simple arithmetic, string, and Boolean tests on state variables
- **Operations typically alter state variables**
- **Operations specify one of three possible moves: Continue, Stop, Start Over**
- **Several rules apply in a single turn**

Representative Entries from Flight Domain

:clause requestkeypad	**→ KeypadDate**
:week | :day | :reldate	**→ ResolveRelativeDate**
!:destination	**→ NeedDestination**
:clause book & :numfound =1	**→ AddFlightToItinerary**
:nonstops & :arrivaltime	**→ SpeakArrivalTimes**

An Illustrative Example



I WANT TO GO TO DALLAS TOMORROW

9 A.M.

BOOK IT

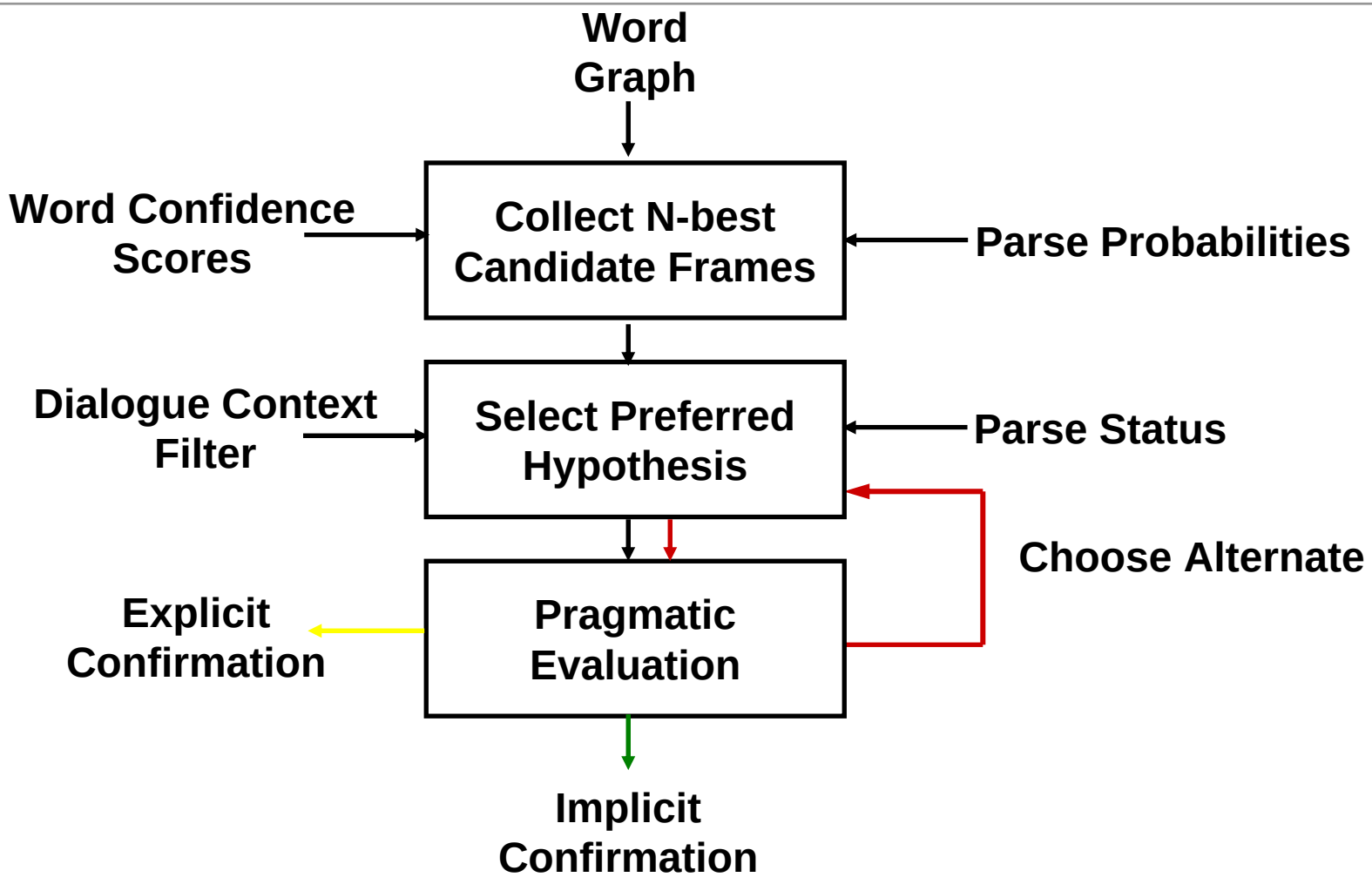
DEPARTING 9 A.M.

SELECTED FLIGHT

CAN YOU PROVIDE A DEPARTURE TIME OR AIRLINE PREFERENCE?

I HAVE AN AMERICAN FLIGHT THAT LEAVES AT 9:30. WOULD THAT WORK?

Hypothesis Selection Process

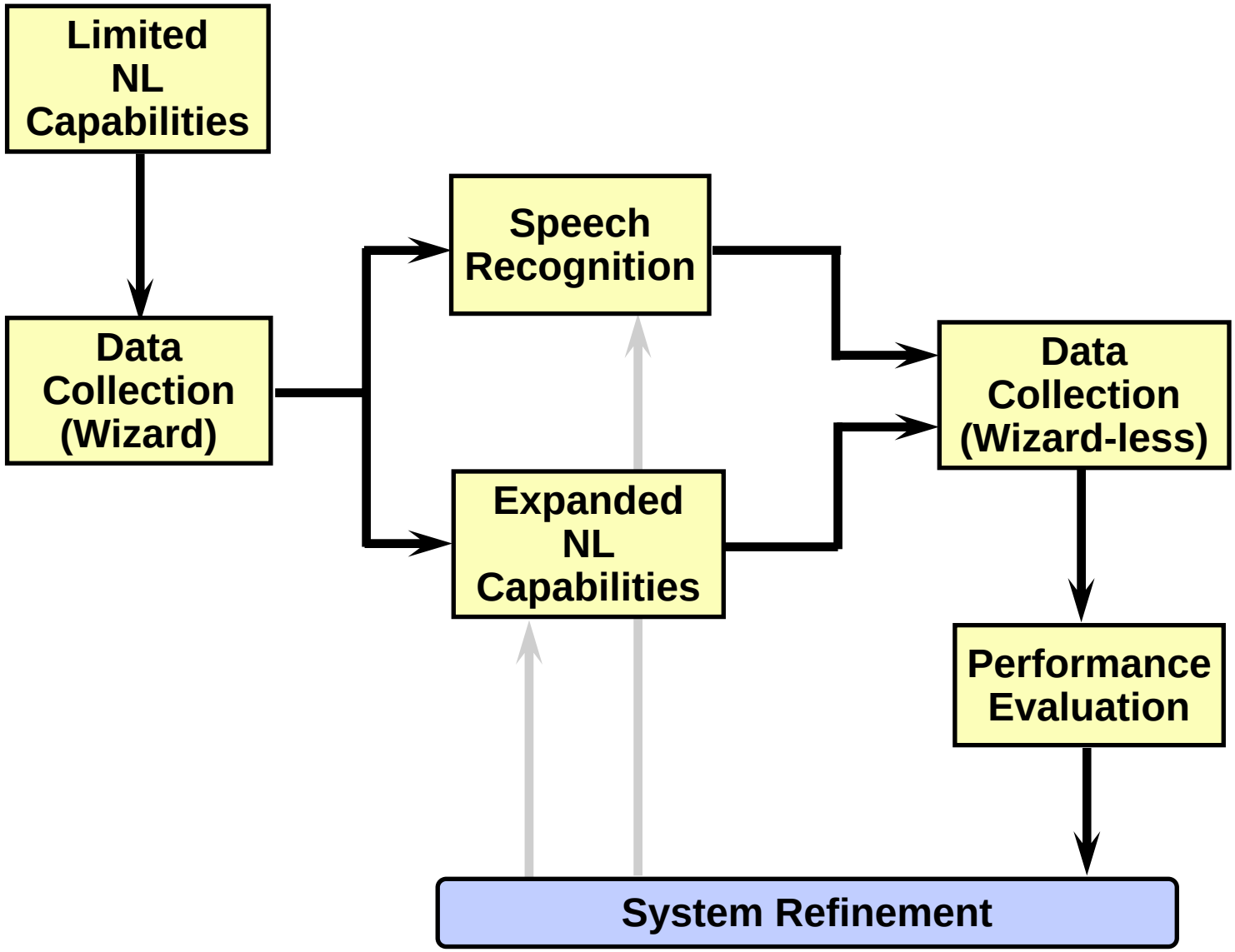


- **Involves recognizer, parser, and dialogue manager**
- **Control specified in hub program**

Outline

- Introduction and historical context
- Speech understanding
- Context resolution and dialogue modeling
- Data collection and evaluation
- Rapid development of new domains
- Flexibility and personalization
- Future research challenges

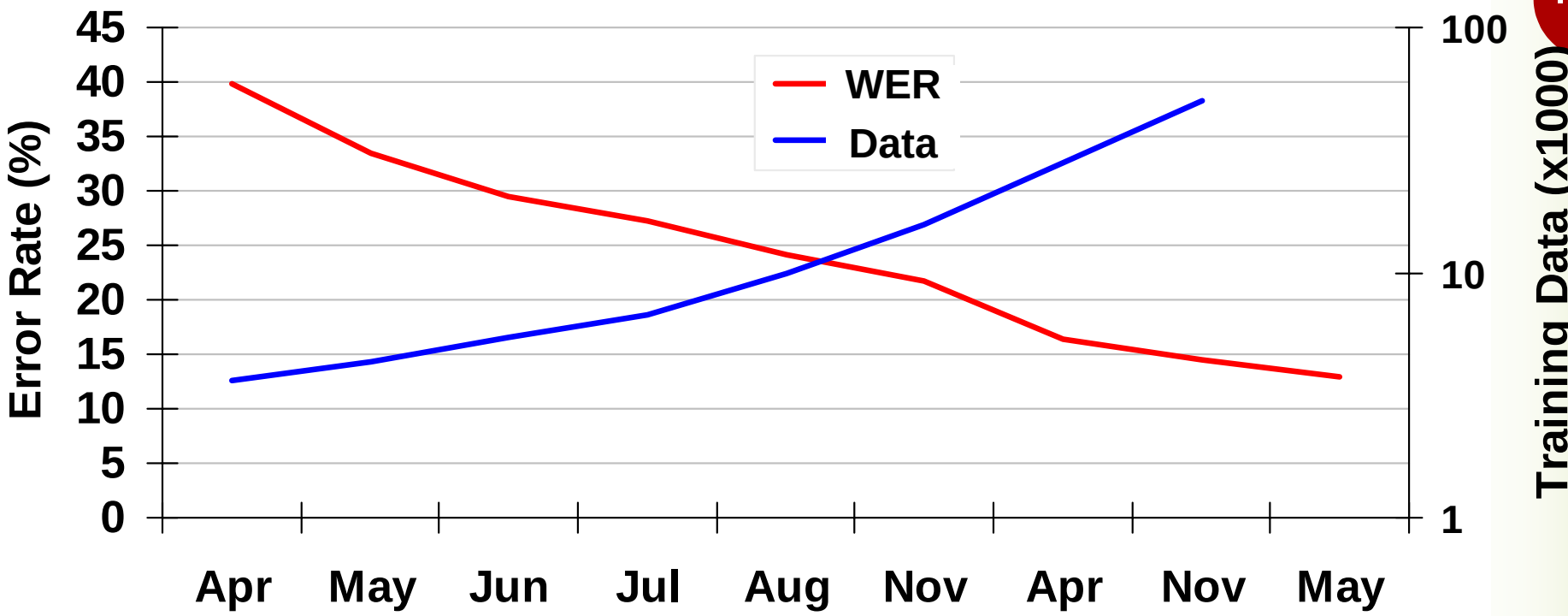
System Development Cycle



Data Collection

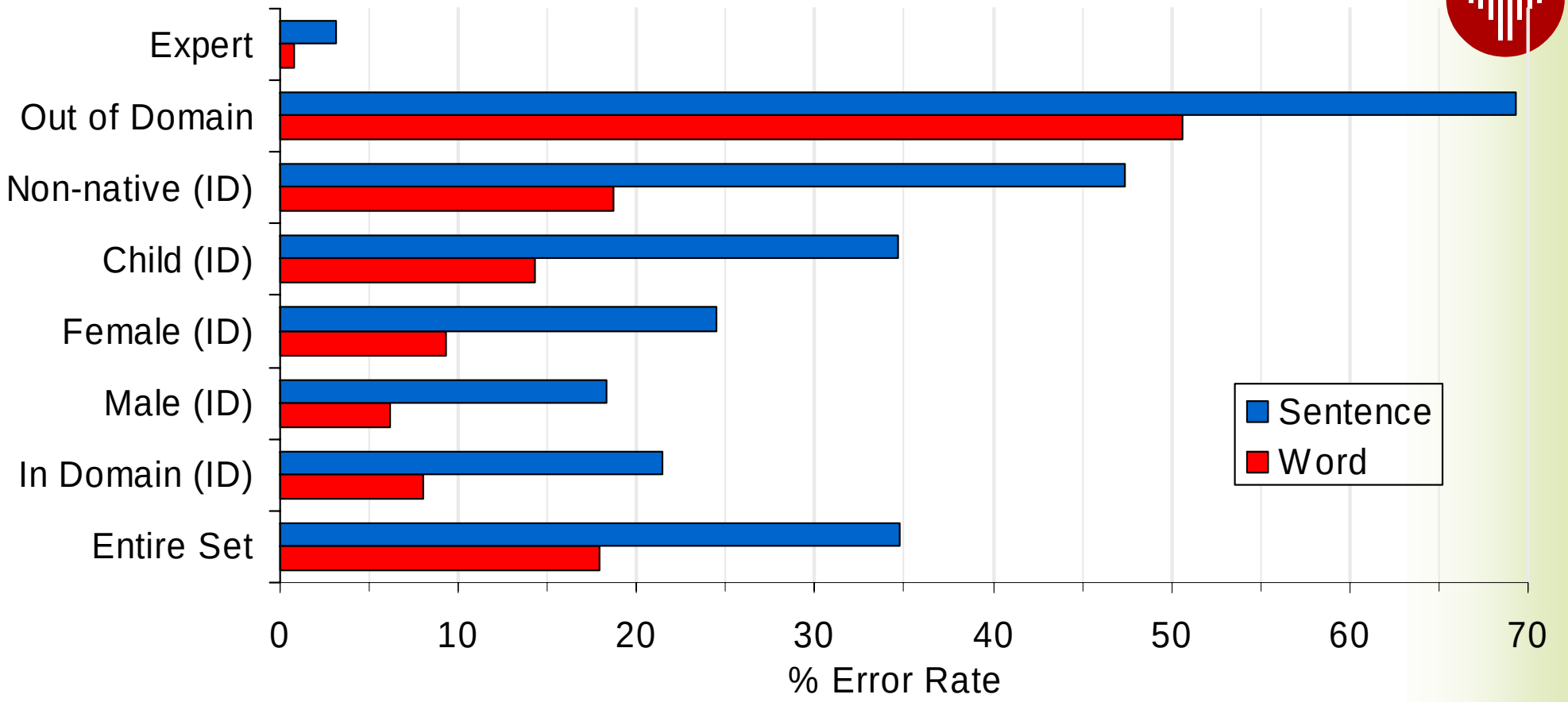
- **System development is chicken & egg problem**
- **Data collection has evolved considerably**
  - Wizard-based → system-based data collection
  - Laboratory deployment → public deployment
  - 100s of users → thousands → millions
- **Data from real users solving real problems accelerates technology 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**

Some Numeric Evaluation Metrics in Flight Domain

SLS

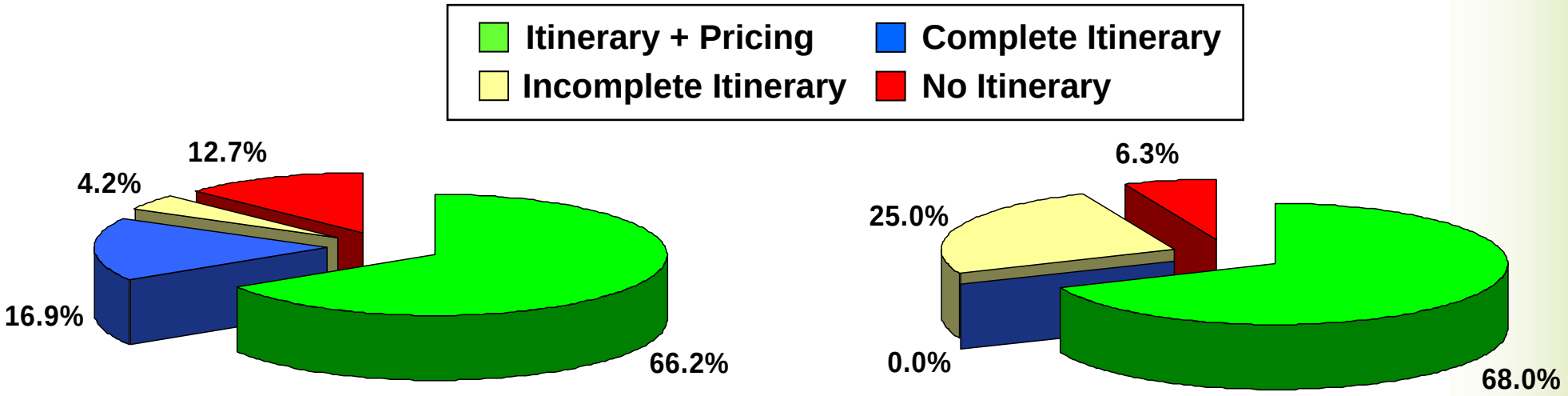
"MIT" Data

(66 calls and 1,150 utterances)

"NIST" Data

(55 calls and 826 utterances)

Task Completion Rate



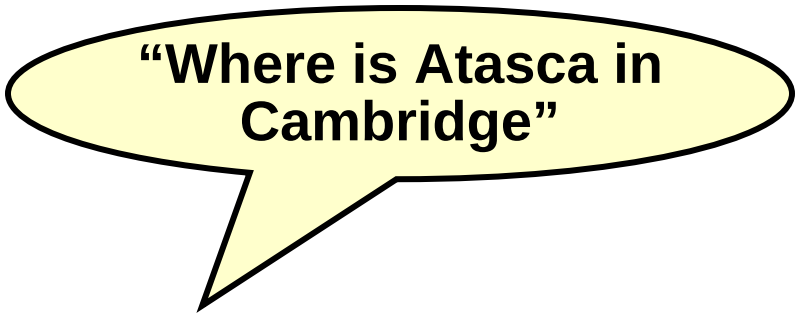
	MIT	NIST
Task Completion Time (s.)	299.3	353.3
Word Error Rate (%)	15.3	14.4
Concept Error Rate (%)	11.0	10.6
Transcription Parse Coverage (%)	89.7	94.1

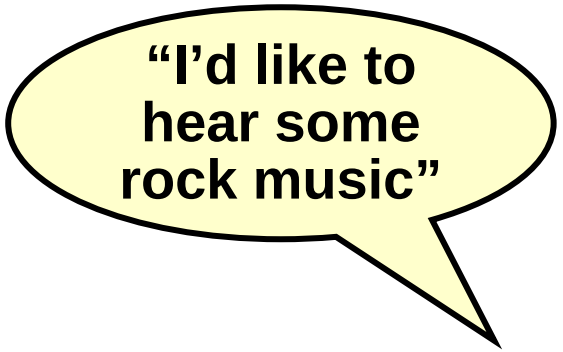
Outline

- Introduction and historical context
- Speech understanding
- Context resolution and dialogue modeling
- Data collection and evaluation
- Rapid development of new domains
- Flexibility and personalization
- Future research challenges

Flexible Conversational Interaction

- **Conversational interfaces will be much more effective if they can adapt to user requests and changing database content**

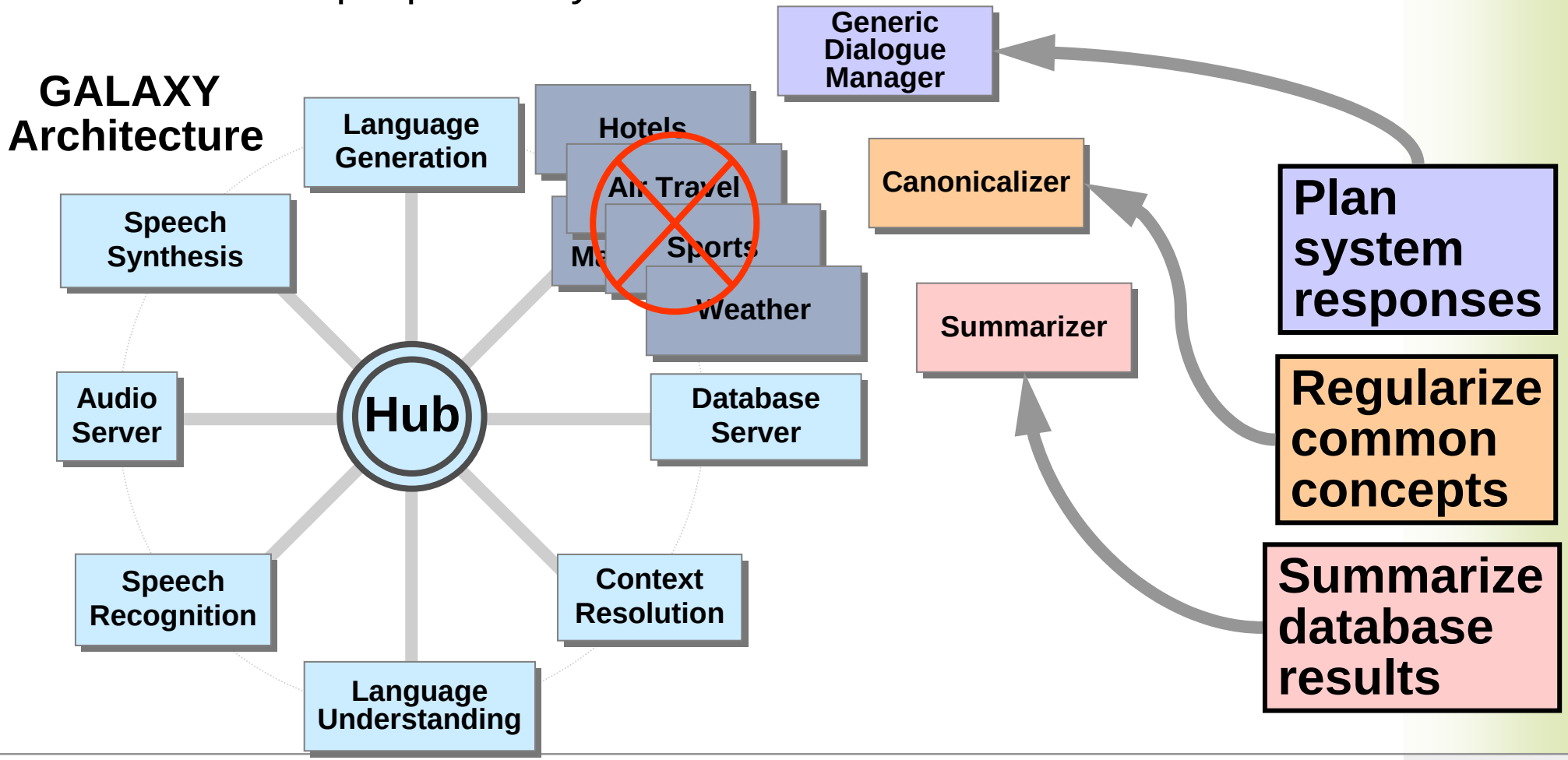




- **To achieve this flexibility requires progress in many areas:**
  - **Dialogue manager needs to be domain-independent**
  - **Responses should reflect distributed properties of database retrievals**
  - **Systems must be able to acquire and use new words and concepts**

Generic Dialogue Modeling

- **Dialogue management is abstracted into set of utility functions that serve as component building blocks**
  - Allows extensive sharing of software among broad class of domains
  - Will enable rapid portability to new domains



Summarization

- **Summarize information returned from database query**
  - Use clustering methods to dynamically determine response
  - Guides user towards appropriate disambiguation constraints

"Are there any restaurants in Quincy?"

Cuisine: 4 American, 2 Brazilian, 2 Indian, …
Price: 7 inexpensive, 4 medium, …
Street: 4 Hancock St., 2 Adams St., …

"There are 14 restaurants. Some of the options are American, Brazilian, Indian, and Italian. None of them are expensive. Many of them are on Hancock Street, Adams Street, Billings Road, and Franklin Street."

Example Dialogue in Restaurant Domain

SLS

"Tell me about restaurants in Boston."

"I have found 404 restaurants in Boston. They are located in Boston and South Boston. There are 29 choices for cuisine. I also have information about price range."

"How about a high end restaurant in Back Bay?"

"I have found 10 high end restaurants in Back Bay. Some of the options are Nouvelle Cuisine, American and Italian."

"Italian"

"There are 2 high end Italian restaurants in Back Bay. Cafe Louis on Berkeley Street and Via Matta in the Park Plaza."

"What restaurants do you know in Somerville?"

"There are 35 restaurants in Somerville. There are 13 choices for cuisine. Many of them are on Broadway, Washington St., Elm St. and Holland St."

"Holland Street"

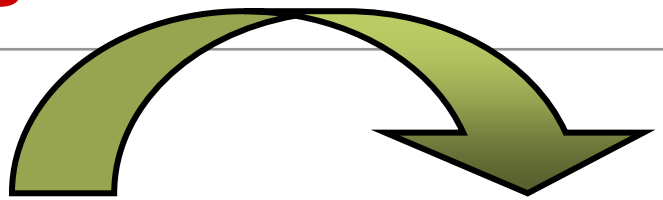
"There are four restaurants in Somerville on Holland Street: Orleans, Mr. Crepe, Johnny D's Uptown and Sabur."

"What kind of food do they serve?"

"Orleans serves American food, Mr. Crepe serves bakery items, Johnny D's Uptown serves barbecue and Sabur serves Mediterranean food."

Backing off to Suggestions

SLS



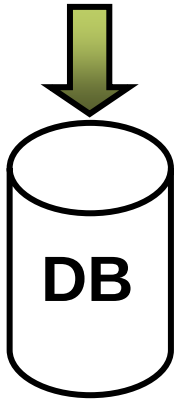
I'm looking for a Chinese restaurant on Main Street in Cambridge

Meaning Representation

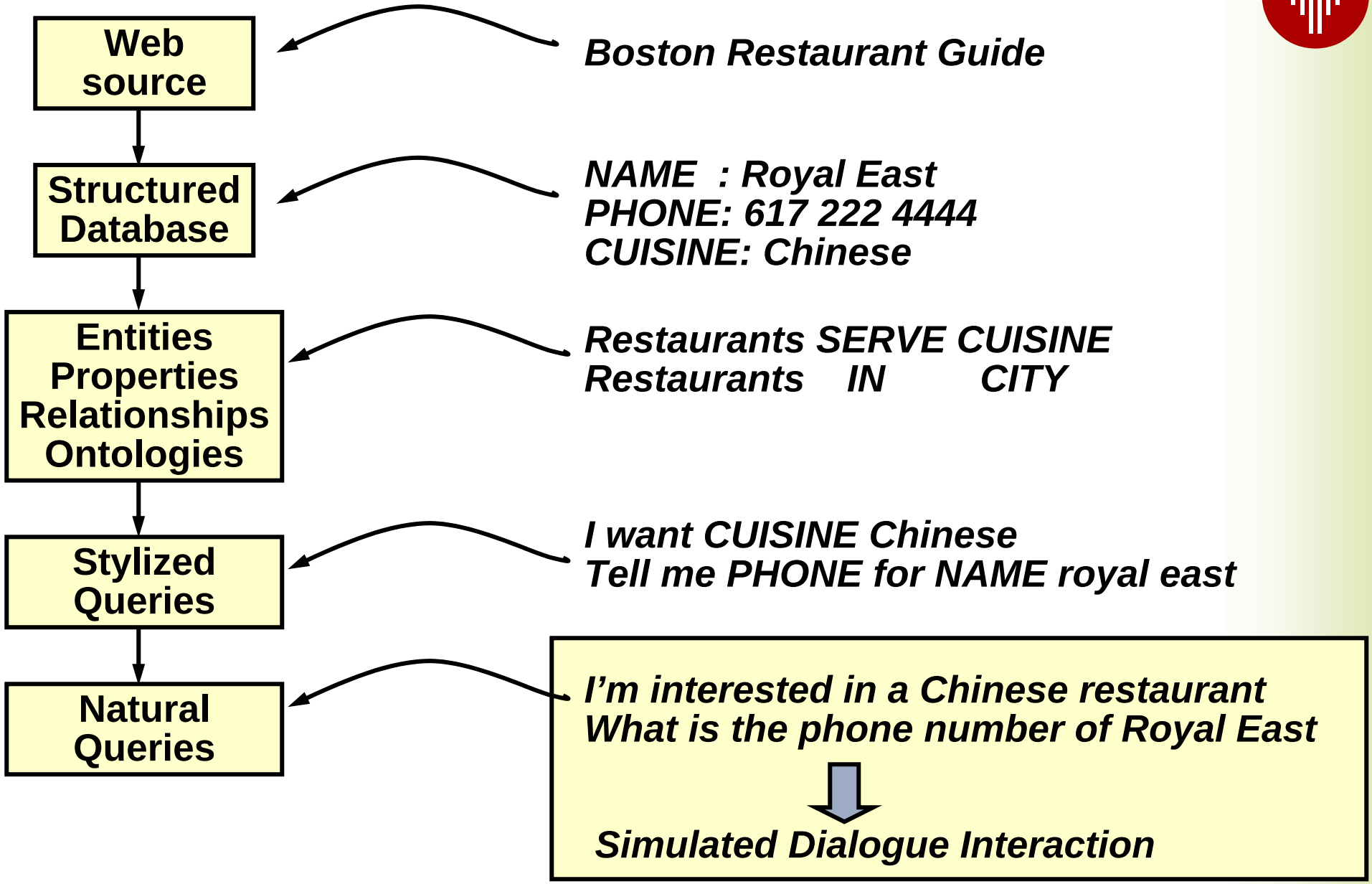
clause:	locate
category:	restaurant
cuisine:	chinese
city:	cambridge
street:	main

I don't know of any Chinese restaurants on Main Street in Cambridge. However, I know of 12 Chinese restaurants in Harvard Square and Central Square.

12 entries



Simulating the User



Outline

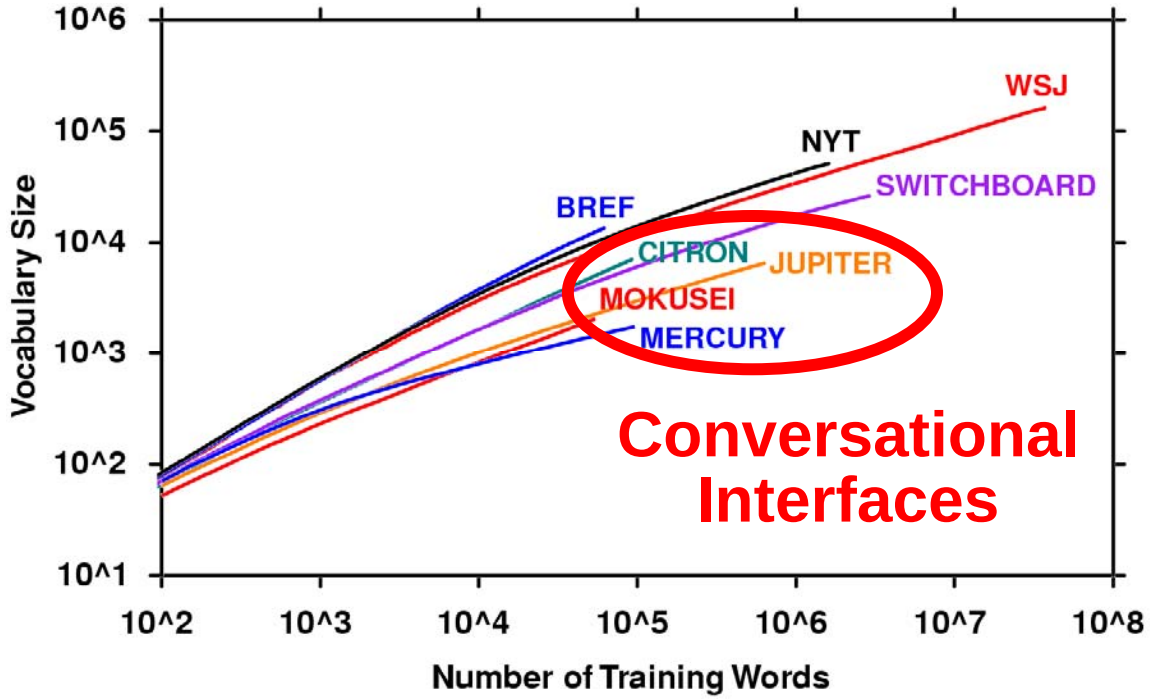
- Introduction and historical context
- Speech understanding
- Context resolution and dialogue modeling
- Data collection and evaluation
- Rapid development of new domains
- Flexibility and personalization
- Future research challenges

Flexible Vocabulary

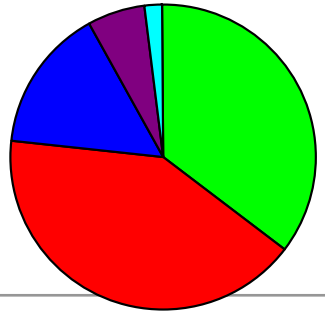
- **Impractical to support all proper names all the time**
  - Several hundred thousand hotel names in the U.S.
  - Issues of recognition accuracy and computational load
- **Solution is two-fold**
  - Support the ability for the user to explicitly enter names when appropriate
  - Adapt the system's working vocabulary to dynamically reflect information it presents to the user

Learning New Words

- **Conversational interfaces must be able to learn new words**
  - **Vocabulary growth is unbounded across a wide variety of tasks**



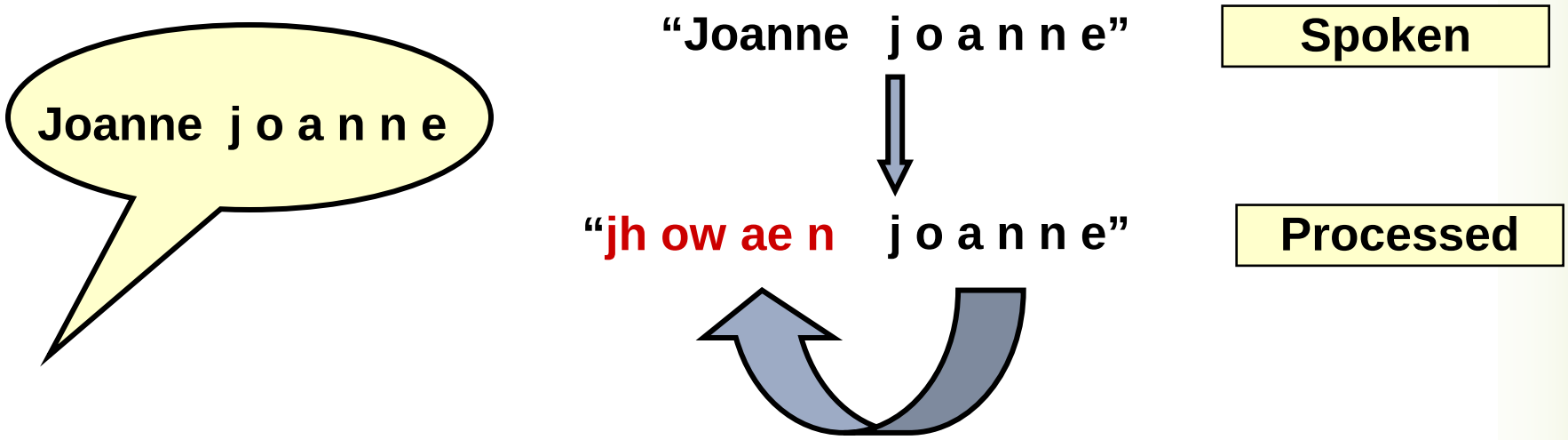
  - **Many new words are important content words (i.e., 75% nouns)**



NAME
NOUN
VERB
ADJECTIVE
ADVERB

Acquiring New Words: Proper Names

- **Initial research based on acquiring unknown user names**
  - User is asked to speak and spell their first and last names



- **Obtains both pronunciation and spelling of unknown word**
  - **Integrated sound-to-letter constraints reduce letter error rate by 44%**
- **Used in enrollment phase of a task delegation domain (Orion)**
  - **New users can register over the telephone**
  - **System automatically incorporates information for subsequent use**

Illustration of User Name Enrollment

- **Enrollment dialogue has been simplified for illustrative purposes**
  - User prompted for name, cell phone number, and time zone
- **If user confirms spellings of first and last names, vocabulary is automatically augmented to support them**
- **(Not illustrated) System backs off to keypad entry when spoken information incorrectly interpreted**

Envisioned Future Extension

USER: Can you tell me the phone number of the Thaiku Restaurant in Seattle?

SYSTEM: I may not know the name of the restaurant. Can you spell it for me?

USER: T H A I K U

SYSTEM: The phone number of Thaiku is 206 706 7807

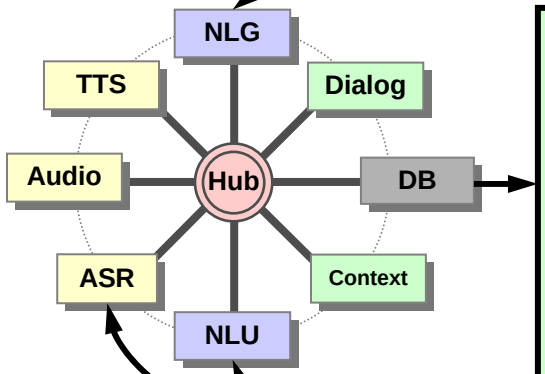
Dynamic Vocabulary Understanding

SLS

Dynamically alter vocabulary based on dialogue context

"Tell me about restaurants in Arlington."

"There are 11 restaurants in Arlington. Some of the options are…"



Arlington Diner
Blue Plate Express
Tea Tray in the Sky
Asiana Grille
Bagels etc
Flora
….

"What's the phone number for Flora?"

"The telephone number for Flora is …"

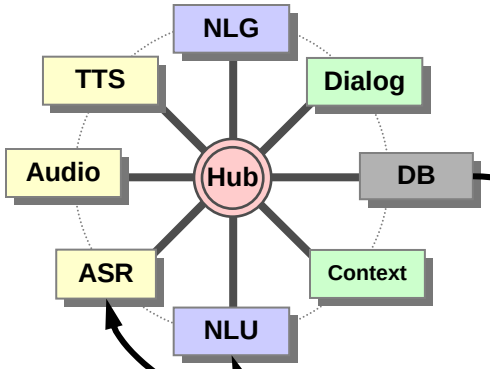
Dynamic Vocabulary Understanding: II

SLS

- **Dynamically alter vocabulary within a single utterance**

"What's the phone number for Flora in Arlington."

What's the phone number of Flora in Arlington



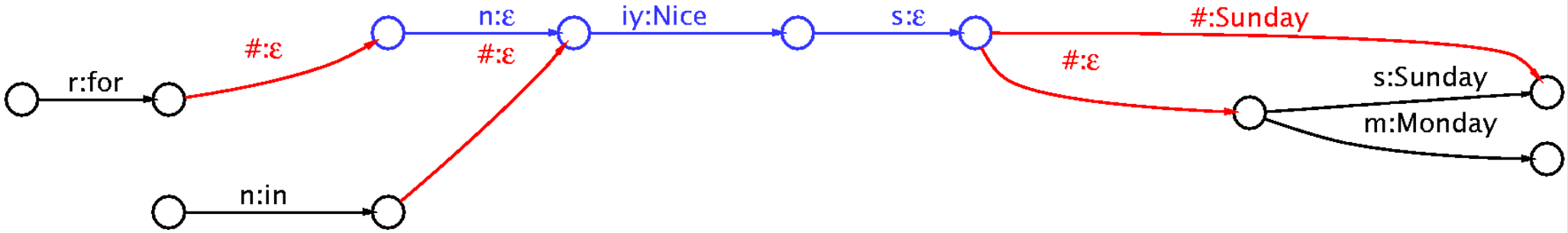
Clause:	wh_question
Property:	phone
Topic:	restaurant
Name:	Flora
City:	Arlington

Arlington Diner
Blue Plate Express
Tea Tray in the Sky
Asiana Grille
Bagels etc
Flora
….

"The telephone number for Flora is …"

Dynamic Vocabulary Recognition

- Recognizer search space represented as a finite-state transducer containing static and dynamic components
- Dynamic word classes are pre-specified (e.g., CITY)
- New vocabulary words determined by dialogue (e.g., Nice)
- Graph splices determined by phonological constraints
- Phantom word-class marker not used for recognition



Audio Clip: Rapid Development and Flexible Vocabulary

SLS

- **Domain-independent dialogue manager**
  - Domain-specific information encoded in external tables
- **No real user data available for training**
  - Generate thousands of utterances through dialogue simulation
  - Train recognizer and NL components on simulated utterances
- **Responses tailored to properties of retrieved database entries**
  - Enumerate short lists
  - Provide succinct summary of long lists
- **Recognizer vocabulary of restaurant names dynamically adjusted as dialogue unfolds**

Example Interaction in Restaurant Domain

SLS

- System knows NO restaurants by name upon start-up
- Two-pass processing recognizes “Royal East” in first query
- Level of detail in summaries dependent on data sets
- Database retrievals license more restaurant names
- “Bollywood Café” recognized in first pass
- This approach can in principal handle an unlimited number of restaurant names, worldwide

Outline

- **Introduction and historical context**
- **Speech understanding**
- **Context resolution and dialogue modeling**
- **Data collection and evaluation**
- **Rapid development of new domains**
- **Flexibility and personalization**
- **Future research challenges**

Research Issues: Speech Understanding

- **We need to do more than just understand the words**
  - Confidence scoring (utterance & word levels)
  - Utilizing timing information
  - Modeling non-speech events and disfluencies
  - Out-of-vocabulary word detection & addition
- **Acquisition of linguistic knowledge**
  - Still don't know how to rapidly develop effective language models that yield high coverage of within-domain linguistic space
- **Robustness to environments and speakers**
  - Adaptation and personalization
- **Other challenges:**
  - Detecting and utilizing non-linguistic information such as speaker identity and emotional state

Research Issues: Dialogue Modeling

- **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?
- **Dialogue strategies**
  - When to use explicit *vs.* implicit *vs.* no confirmation?
  - When to back off to alternatives such as typing or keypad entry?
  - How to model help mechanisms to inform users of system capabilities?
  - How to recognize and recover from errors?
  - How to enable the above capabilities across diverse domains
- **Producing and responding to back-channel**
  - Would likely have striking effect if properly implemented
- **How can system learn user preferences through repeated interactions?**

Rapid Development of Flexible Systems

A Challenge:

- Given an unstructured knowledge source, how long would it take to create a dialogue system capable of providing access to that information space through natural spoken interaction?
- Which aspects of system development consume the most resources?

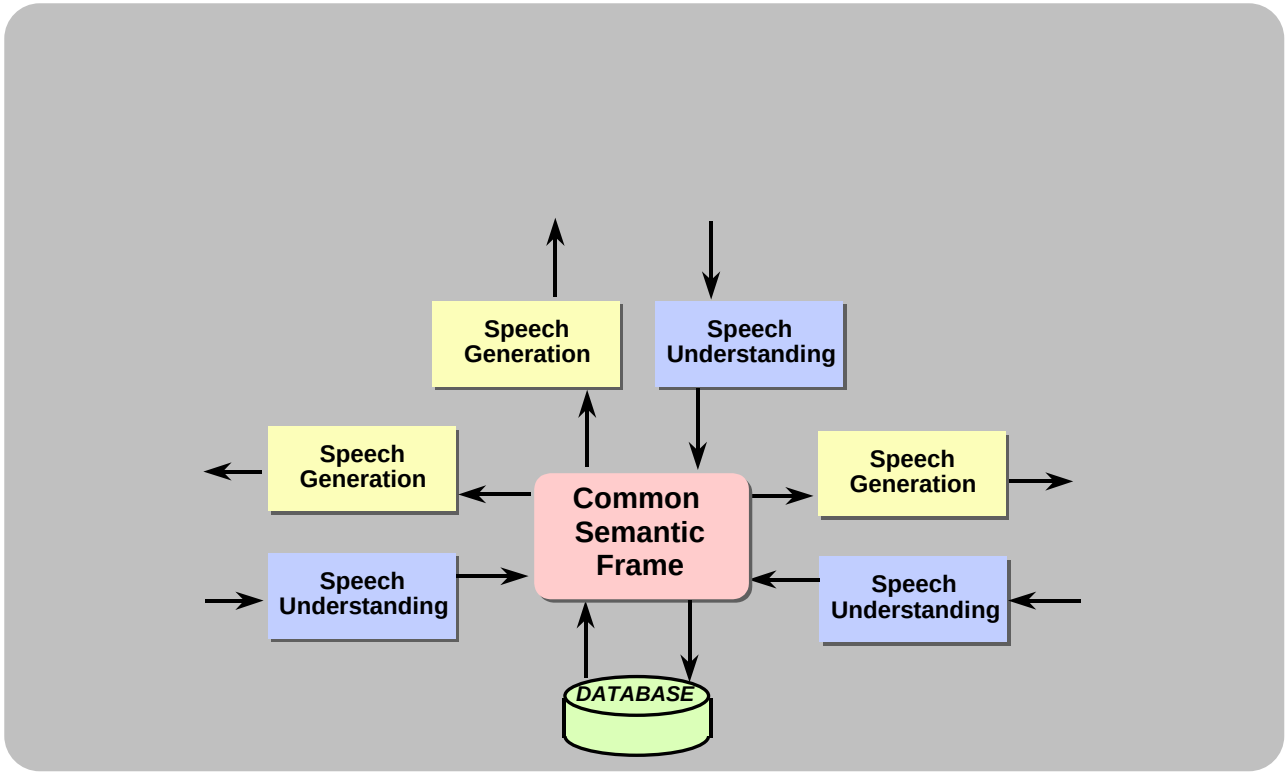
- **Speech Understanding:**
  - How to obtain language models adequately capturing the linguistic space of the domain?
  - How to exploit dialogue context to adjust the vocabulary and language models
- **Dialogue Management/Response Planning:**
  - How to efficiently encode all the appropriate system responses, including help mechanisms and error recovery?
  - How to separate out domain-specific aspects so that new domains can leverage code developed for other applications?

The Role of Simulated Dialogues?

- **We cannot rely solely on live human-computer dialogue to stress test our systems**
- **Somewhat effective strategy:**
  - Batch mode reprocessing of previously recorded dialogues
- **However, prior mixed initiative dialogues quickly become incoherent as systems evolve**
- **Proposal: *simulate* user half of the conversation**
  - Randomly generate an appropriate response in reaction to each system turn
- **Simulated Input as text strings or *spoken utterances***
  - selection from library of user utterances or
  - through speech synthesis

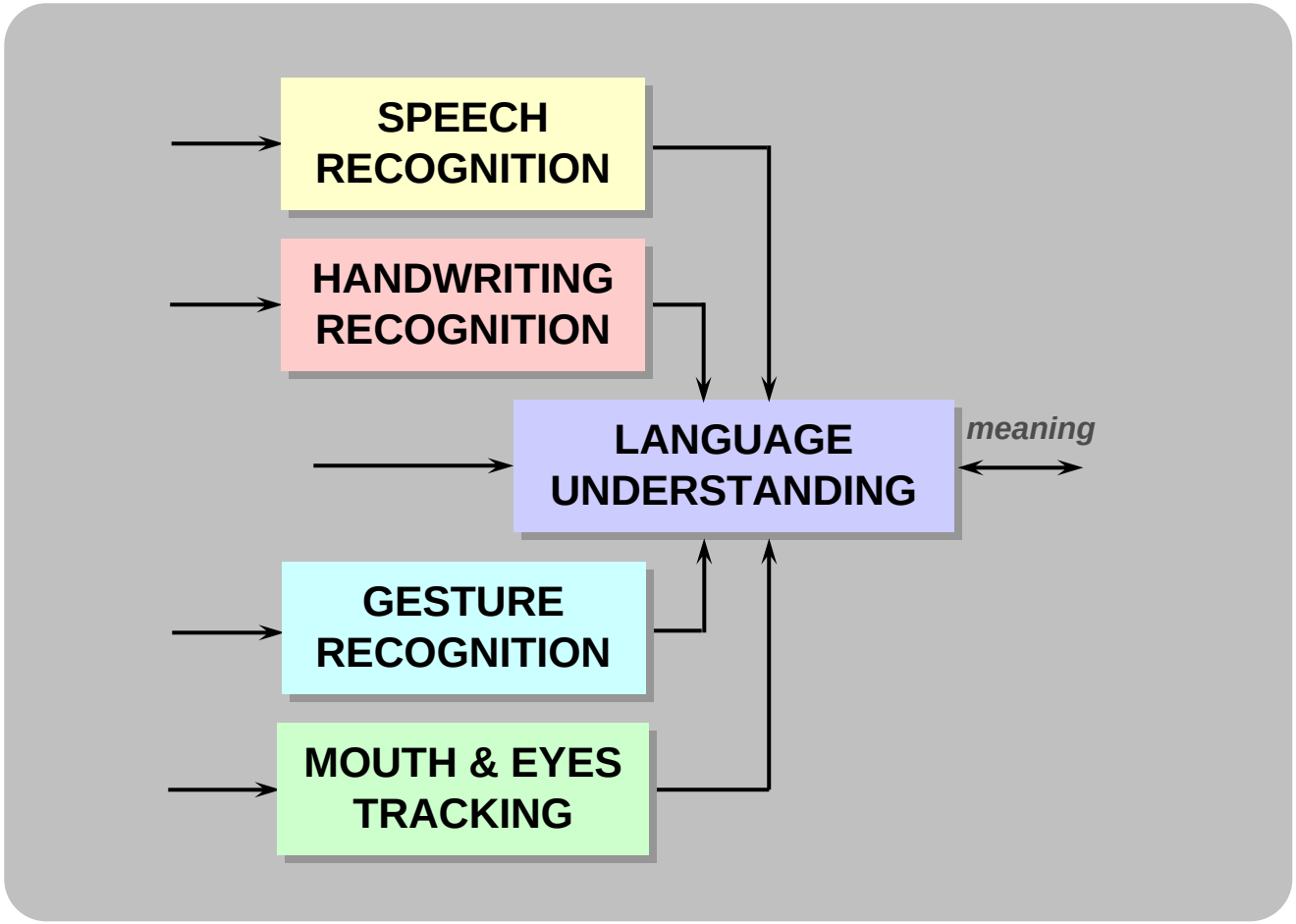
Monolingual → Multilingual

- **Language transparent design:**
  - it is crucial that we seek solutions that will easily port to other languages besides English
- **Can we develop tools to automatically derive linguistic structure (e.g., parsing rules) from aligned corpora?**



Speech only → Multimodal Interactions

- **Typing, pointing, clicking can augment/complement speech**
- **A picture (or a map) is worth a thousand words?**



Research Issues: Multimodal Interactions

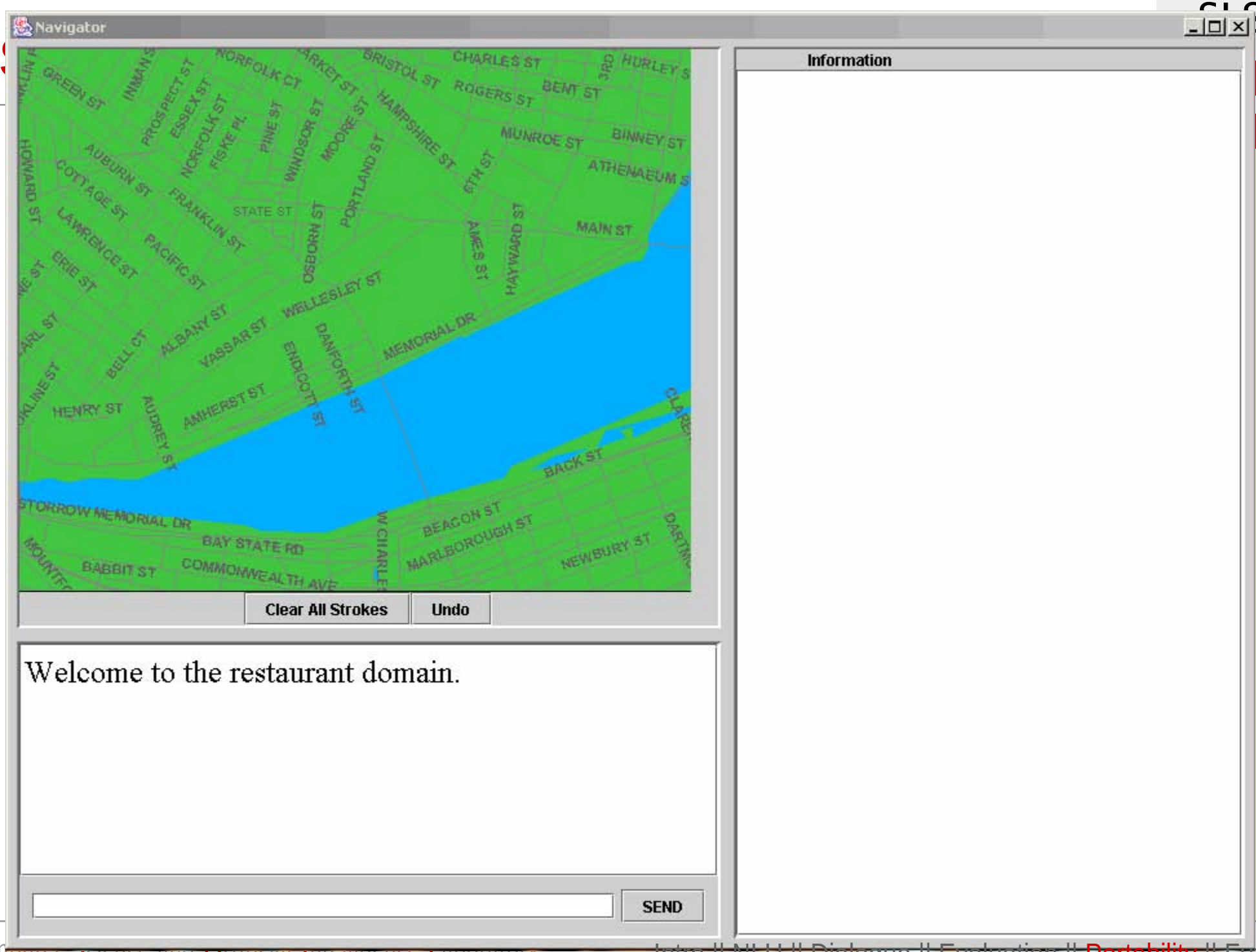
- **What is the unifying linguistic framework that can adequately describe multi-modal interactions?**
- **What is the optimal design for system configuration?**
  - E.g., timing constraints less stringent when signals are more robust
- **What are the appropriate integration and delivery strategies?**
  - How are modalities affected by presence of alternative modes?
    - * Graphical interface alters parameters of response decisions
    - * Terse *vs.* verbose spoken responses depend on existence of ancillary graphics
  - When to utilize which modality
  - trans-modal interfaces (e.g., read email over the phone)

Conclusions

- **Spoken dialogue systems are needed, due to**
  - Miniaturization of computers
  - Increased connectivity
  - Human desire to communicate
- **To be truly useful, these interfaces must behave naturally**
  - Embody linguistic competence, both input and output
  - Help people solve real problems efficiently
- **Conversational interfaces must be able to learn from user interaction and database content**
- **To achieve this flexibility requires progress in key areas:**
  - Response planning needs to be flexible and content-driven
  - New concepts must be acquired naturally during interaction

Short Video Clip

- First two turns illustrate different summarizations of database results for two different cities (automatically determined)
- Third turn shows multi-modal interaction (speech plus pen)
- In last turn, user refers to restaurant by name, but the name was unknown to the recognizer at the beginning of the dialogue (flexible vocabulary)


Navigator
Information
Clear All Strokes
Undo
Welcome to the restaurant domain.
SEND

MIT's Discourse Module Internals