Multimodal Interaction Involving Speech and Language Technologies

June 23, 2008

Alex Waibel
International Center for Advanced Communication Technologies
Carnegie Mellon University
University of Karlsruhe
http://www.interact.cs.cmu.edu

InterACT Center

- InterACT
  - International Center for Advanced Communication Technologies
  - Joint Center between Carnegie Mellon and University of Karlsruhe
  - Emerged from 15 year Collaboration
  - Launched January, 2004
- Mission of Center
  - To Develop Advanced Communication Technologies
  - To Facilitate Student Exchange and Training
- Major Ongoing Projects
  - CHIL – Computers in the Human Interaction Loop
  - TC-STAR & STR-DUST & TRANSTAC & GALE – Speech Translation
  - TIDES & ASSIST &… - Text, Image Translation

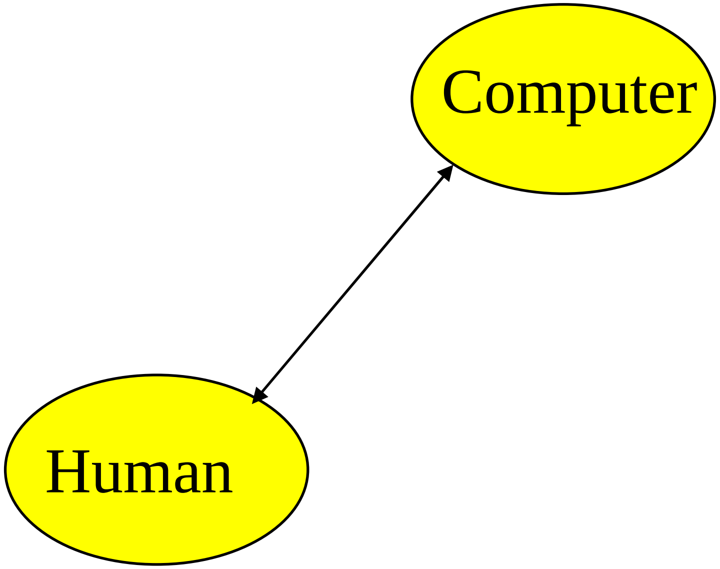
Acknowledgment of Support

- Infrastructure:
  - – InterACT Center Support:
    CMU and State of Baden-Wuerttemberg
  - – C-STAR: Consortium for Speech Translation Research
- Research Projects:
  - – In the US:
    - STR-DUST (NSF-ITR)
    - TIDES (DARPA)
    - GALE (DARPA
    - Babylon/Caste/Transtac (DARPA), Laser-ACTD
  - – In Europe:
    - CHIL (European Commission)
    - TC-STAR (European Commission)
    - PF-STAR, FAME (European Commission)

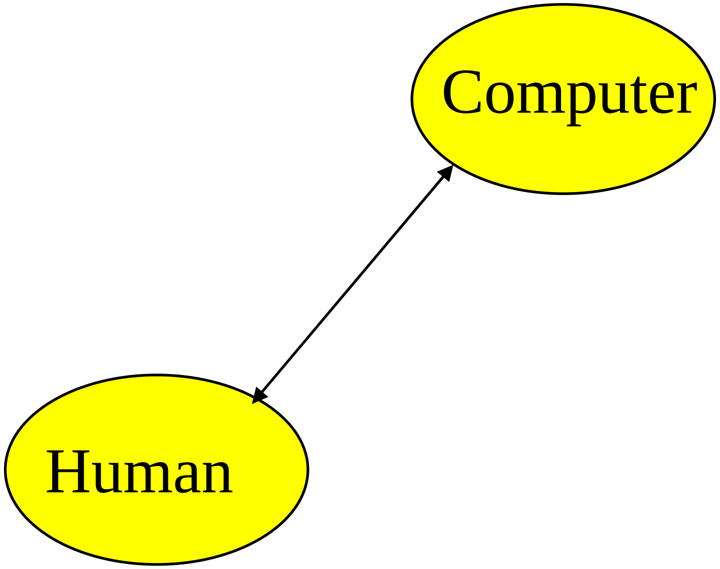
InterACT – Tech Transition

- Universities
  - Carnegie Mellon, #1 in CS in USA
  - U. of Karlsruhe, #1 in CS in Germany
- Corporations
  - Multicom Inc. – Speech Datacollection (closed)
  - ISI – Speech Recognition (sold)
  - SMI – Handwriting Recognition (active)
  - AMI – Japanese Speech Recognition (active, IPO)
  - Multimodal – Speech Transcriptions in Health Care (active)
  - **Ichibel / Mobile Technologies – Speech Translation (active, growing)**

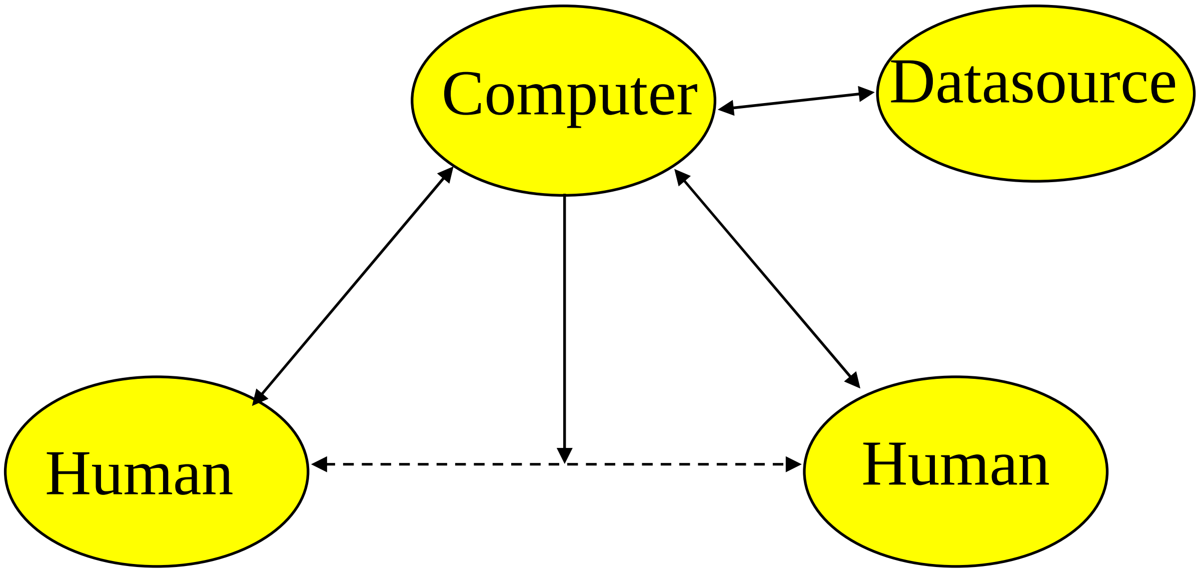
Classical Human-Computer Interaction



Classical Human-Computer Interaction



New Roles for Humans and Computers



Multimodal Interfaces

- Exploit All Human Communication Modalities
- Advantage:
  - Complementarity
  - Redundancy
  - Robustness
  - Naturalness
  - Flexibility.. “Fleximodal”

Interface Types

Multimodal Interfaces

1. Human -> Machine: Dictation
2. Human <-> Machine: Interactive Dialog
3. Human <-> Multimedia Data: Interactive Retrieval
4. Human <-> Machine <-> Human: Mediation, Interpretation
5. Human <-> Human, Machine Assistance: CHIL

Case 1

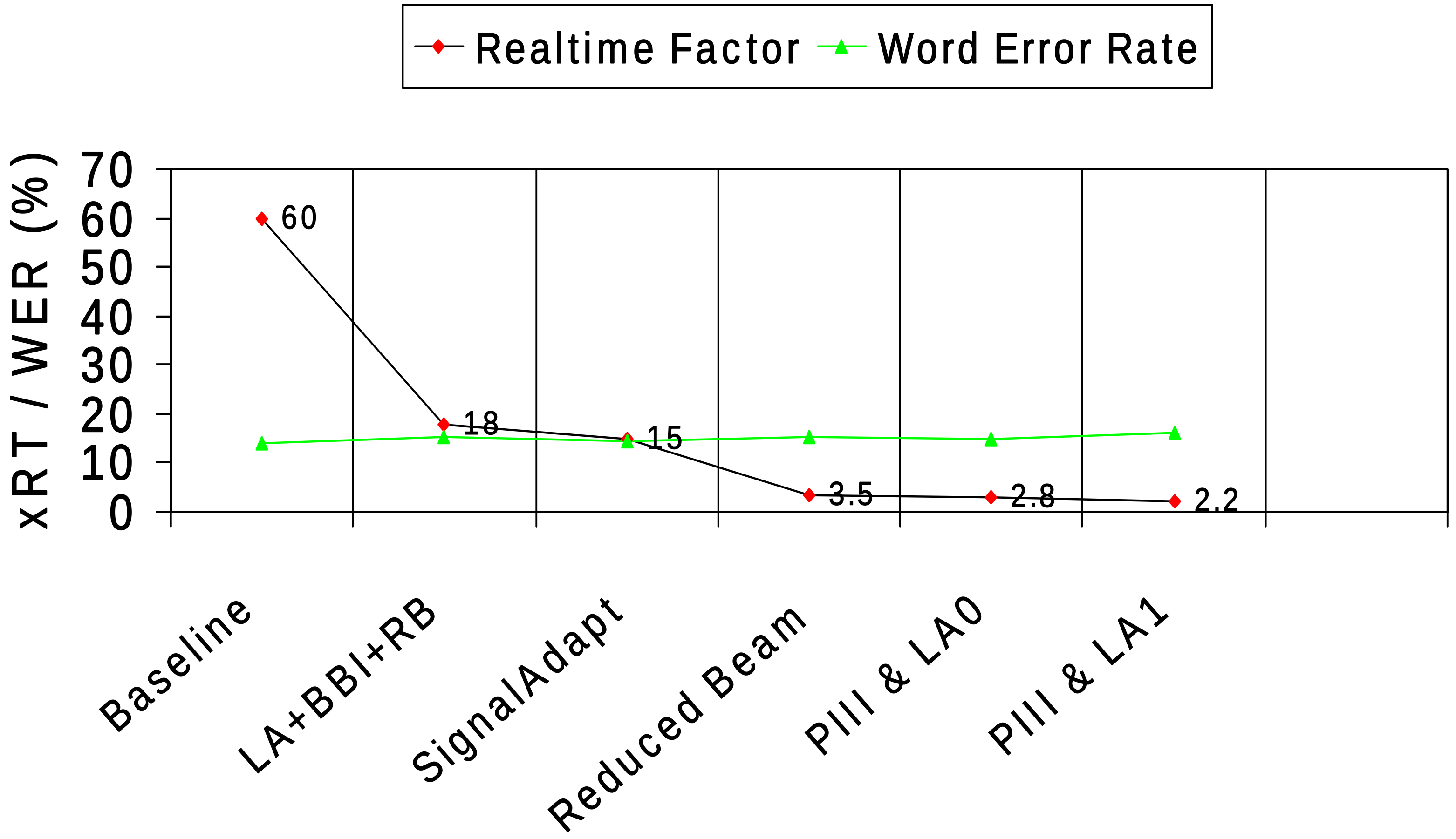
Human →Machine

Dictation

System Characteristics

- Close Speaking Mic
  - Low Noise
- Speaker:
  - Single or Few Talkers
  - Cooperative
  - Read Speech
- Issues:
  - Vocabulary Maintenance
  - Perplexity Control
  - Speed
  - Human Factors

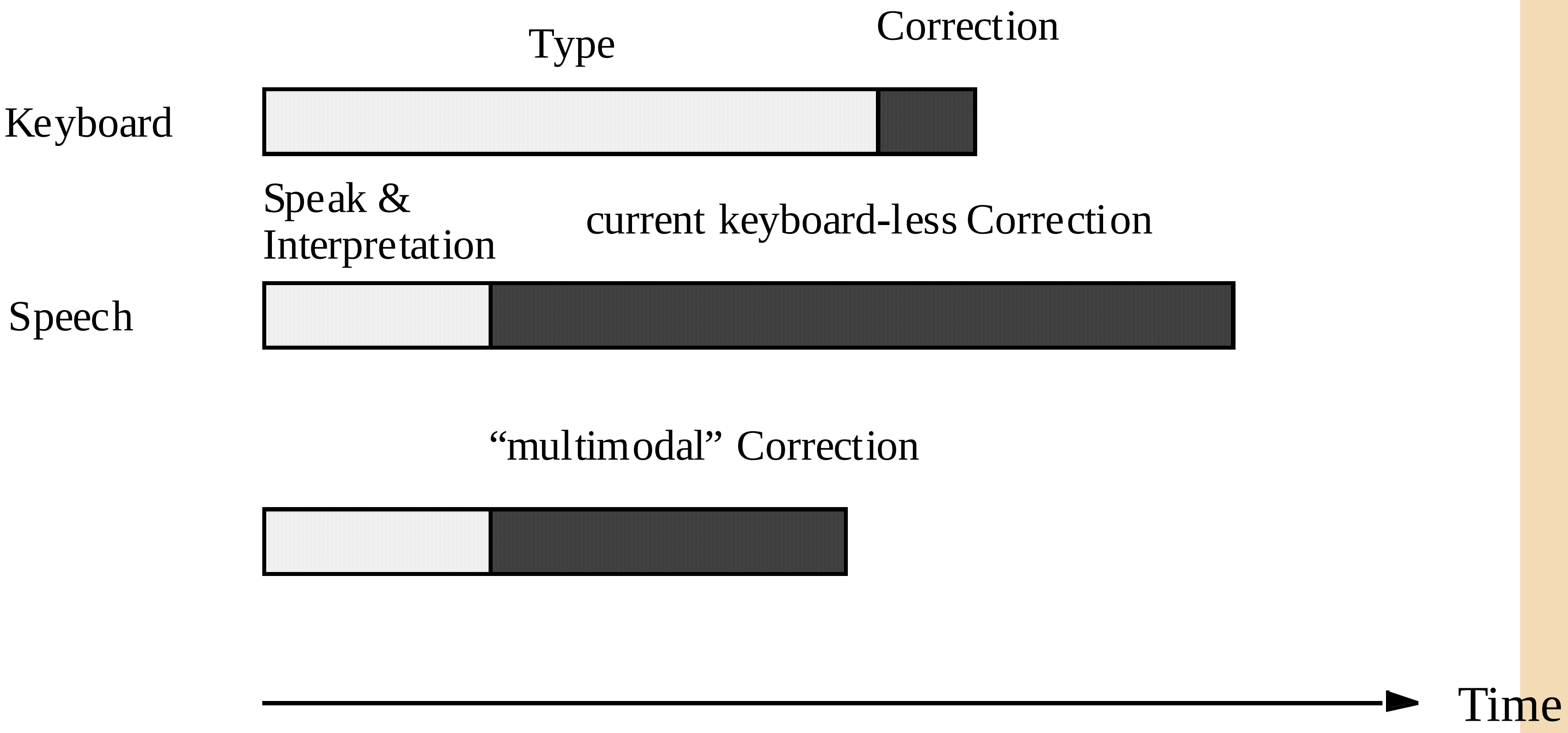
Improving Speed on Cooperative Speech



Dictionaries & Language Models

- Grammars for Search Control Impractical
- Language Models:
  - Predict Next Word based on History (N-Grams)
- Dictionary: Use Large 60,000+ Dictionary
- Problem:
  - Suitable Vocabularies and Language Models Vary for each User
  - How to Deal with Machine and Human Errors
- Solution:
  - Provide Tools to Adapt Dictionaries and Language Models
  - Provide Better Error Correction Tools

Voice, Multimodal, Keyboard Input



Keyboard

Type

Correction

Speech

Speak & Interpretation

current keyboard-less Correction

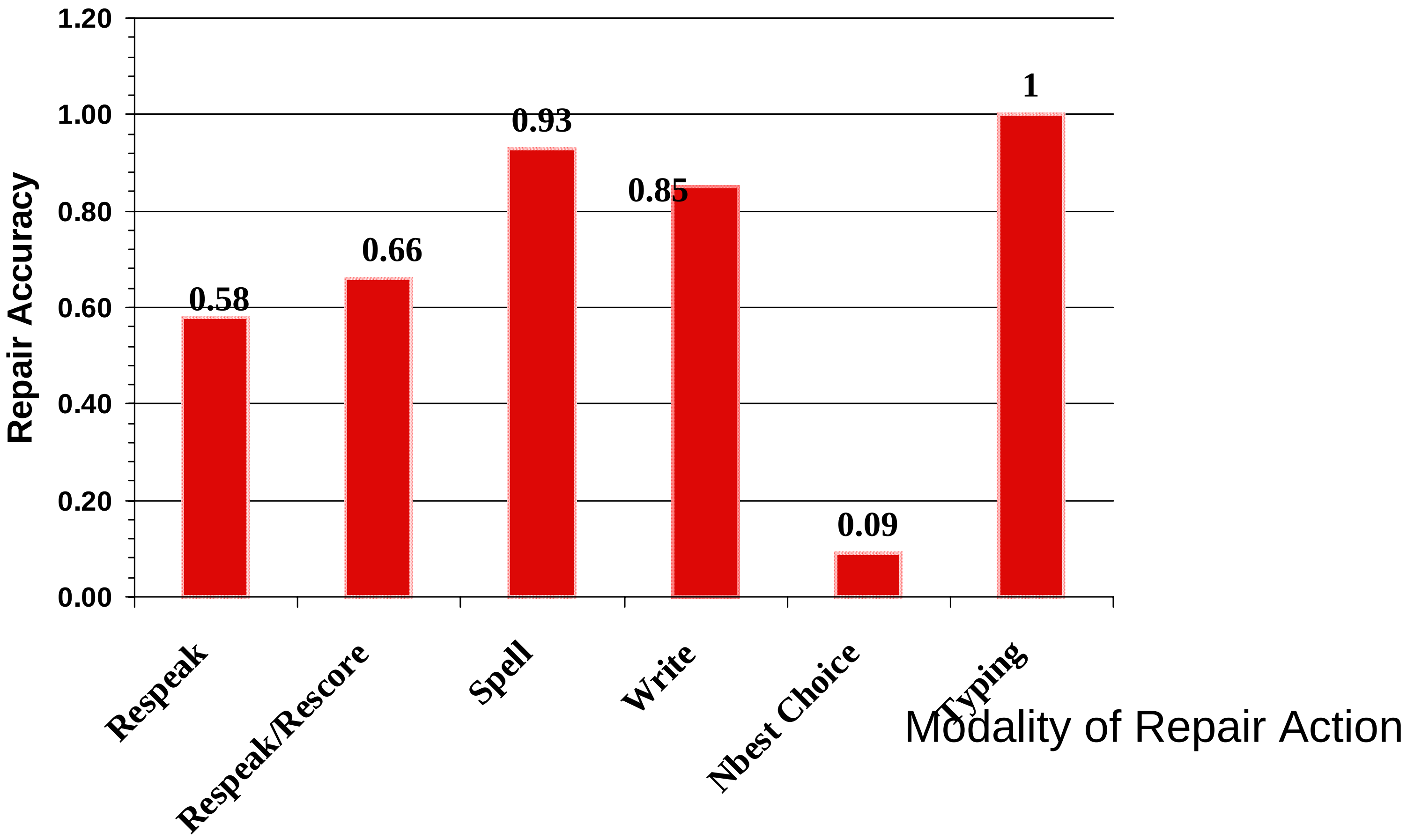
“multimodal” Correction

Time

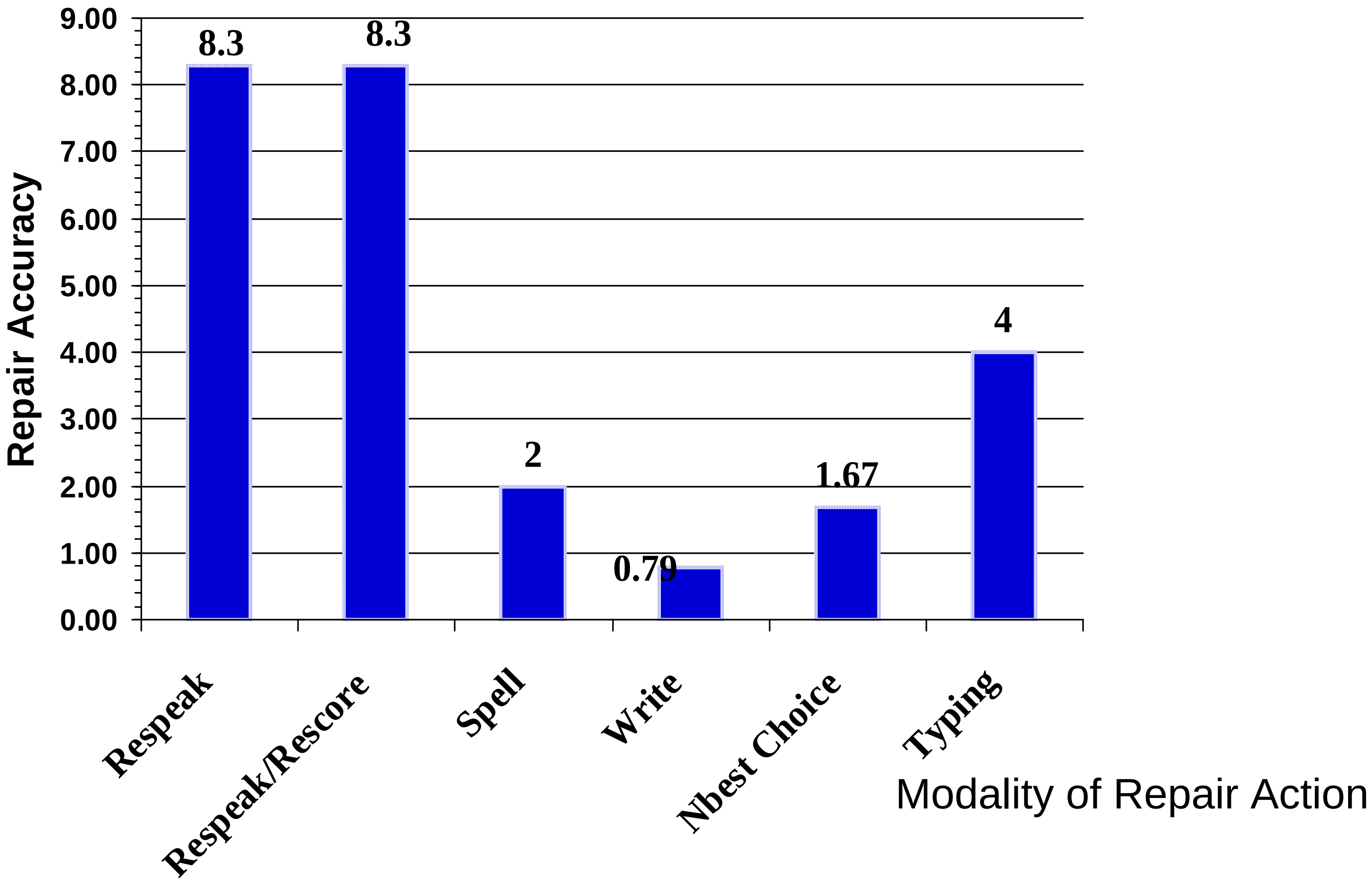
Interactive Cross-Modal Repair

- Exploit ***alternate, complementary*** Modalities to Correct Errors
- Possible Modalities:
  - Speaking,
  - Respeaking,
  - Spelling,
  - Pointing,
  - Gesturing,
  - Handwriting,
  - N-best Lists,
  - Paraphrase
  - Semantic Repair Dialog

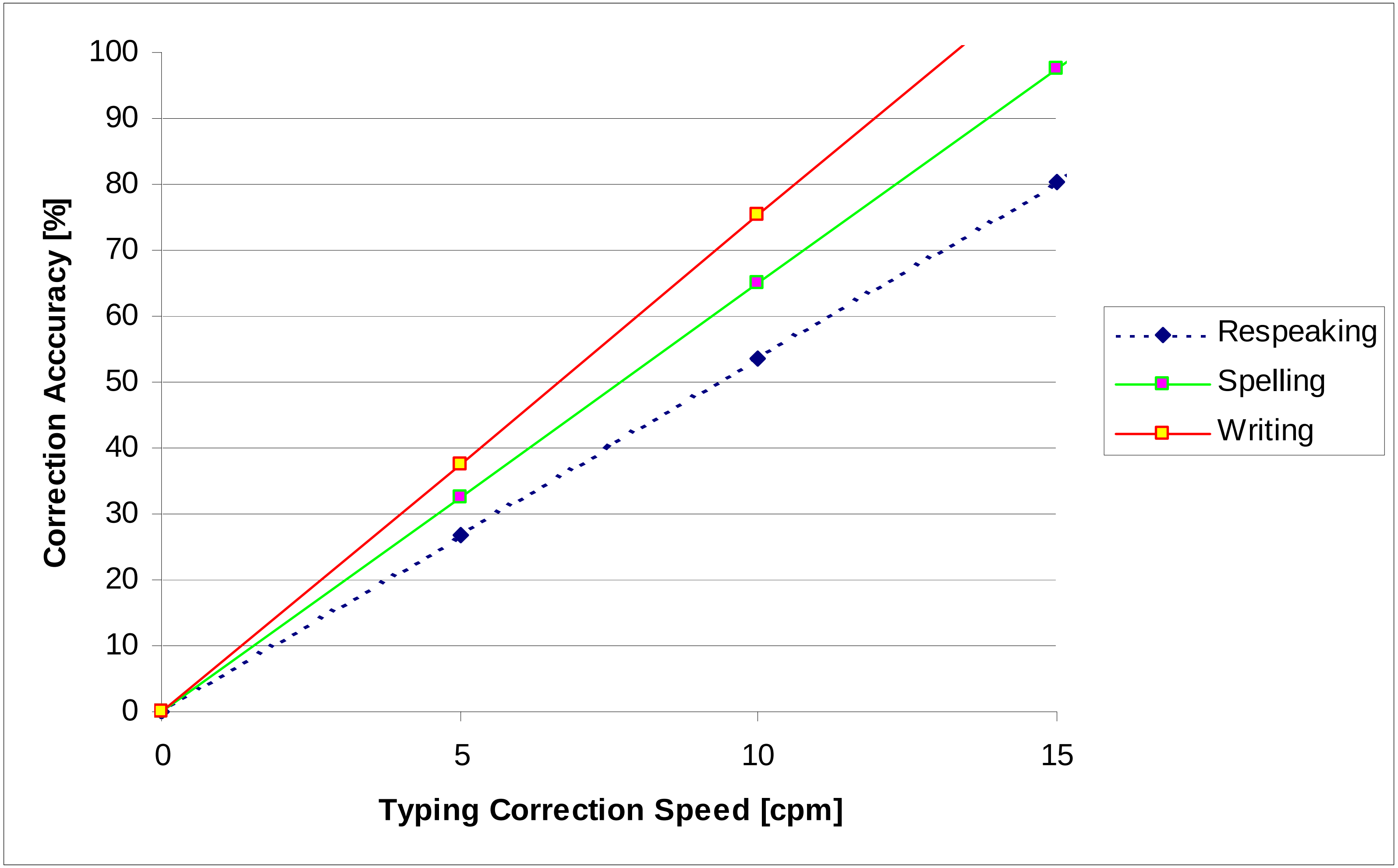
Accuracy of Repair



Speed of Repair



Correction Accuracy to beat Typing in Correction Speed



Slow Typing

Fast Secretarial

Average Computer Worker

Multimodal Error Repair

Interface Type 2

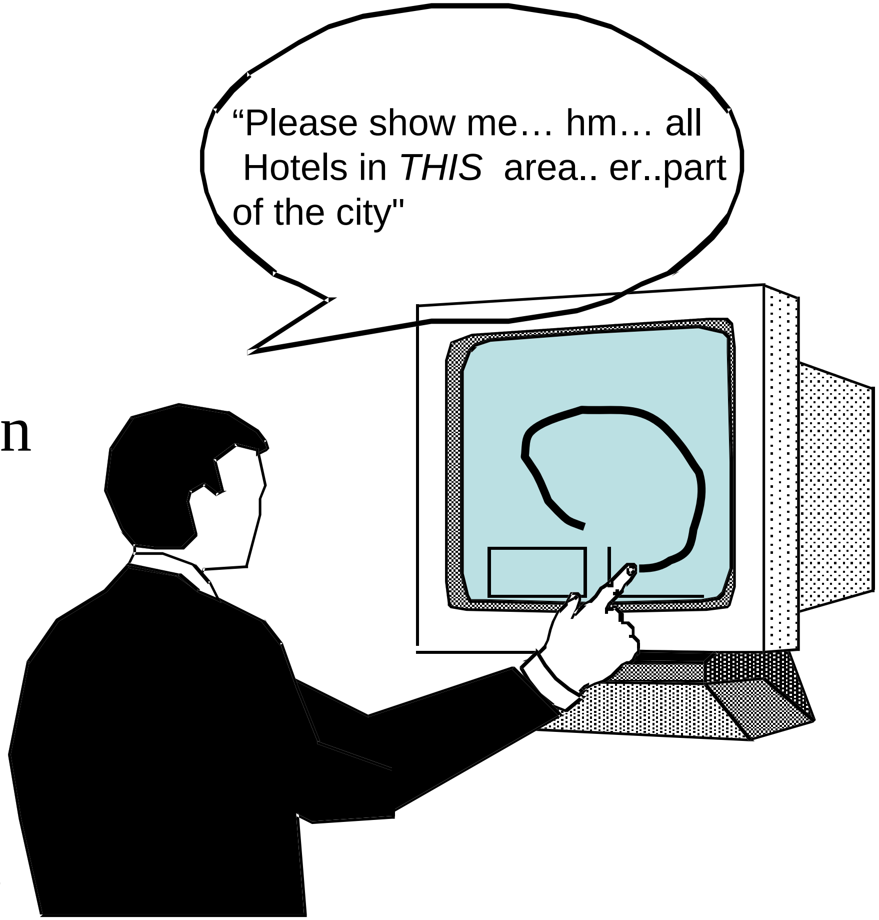
Human ←→ Machine

Interactive Dialog

Navigation

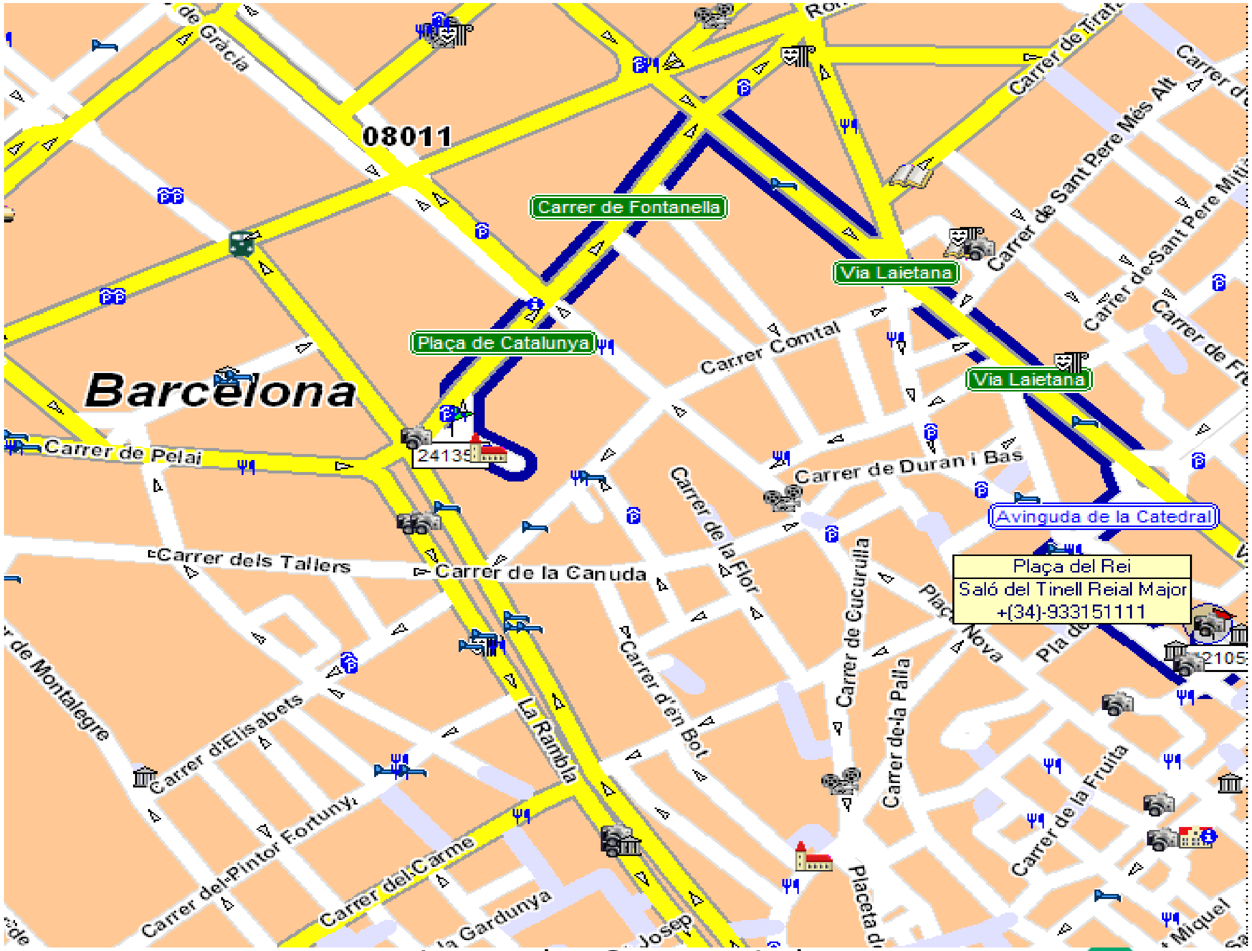
Multimodal Dialog: Human-Machine Interaction

- Speaking
- Pointing,
- Gesturing
- Hand-Writing
- Drawing
- Presence/Focus of Attention
- Combination
  - Sp+HndWrtg+Gestr.
  - Repair
- Response Generation:
- Multimodal NLP & Dialog

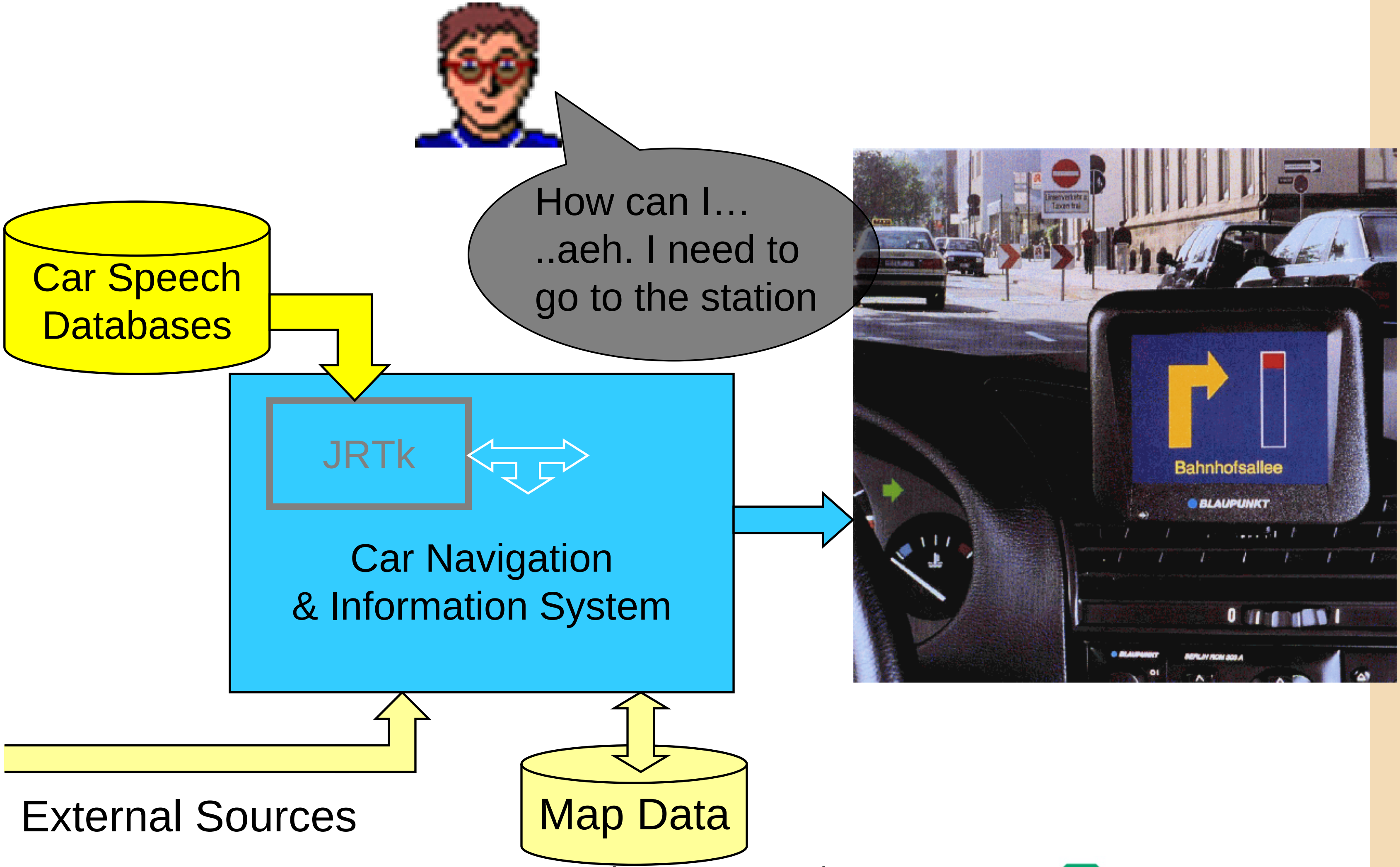

“Please show me… hm… all Hotels in *THIS* area.. er..part of the city"

Application: Navigation

“How do I get to the Plaza Catalunya ?”



Car Navigation



How can I…
..aeh. I need to
go to the station

Car Speech Databases

JRTk

Car Navigation
& Information System

External Sources

Map Data

System Characteristics

- Possibly Good Recording Conditions
  - Sometimes Close Speaking Mic, Low Noise
- Speaker:
  - Few Dominant Talkers, No Cross-Talk
  - Clear Cooperative Speaking Style
- Task:
  - Usually Restricted
  - Perplexity and Vocabulary Limited
- Issues:
  - With Remote Mics, Severe Noise Degradation (Driving Noise)
  - Spontaneous Speech
  - Dialog Management and Control
  - Modalities other than Speech

Dialog Modeling

So Far:

• Speech recognition:	sentence by sentence
• Language modeling:	within sentence constraints only
• Parsing:	one sentence at a time

Dialog Modeling:

- What Information Connects Individual Utterances
- Manage Human-Machine Interaction
- How should the Machine Respond ?
  - How to Optimize for Task Completion
- Who Takes the Initiative ?
  - Prompted, Free, Mixed

Research on Dialogue Systems

Goals

- Cooperative task-oriented dialogue
- Develop algorithms to support a computer's participation in a cooperative dialogue

Approaches

- Plan-based models
- Joint action theories of dialogue
- Dialogue grammars
- Frame-Based Systems
- Statistical Learning Systems

Problems

- Grammar Writing Effort
- Data-Collection Effort
- Domain Coverage

Interface Type 3

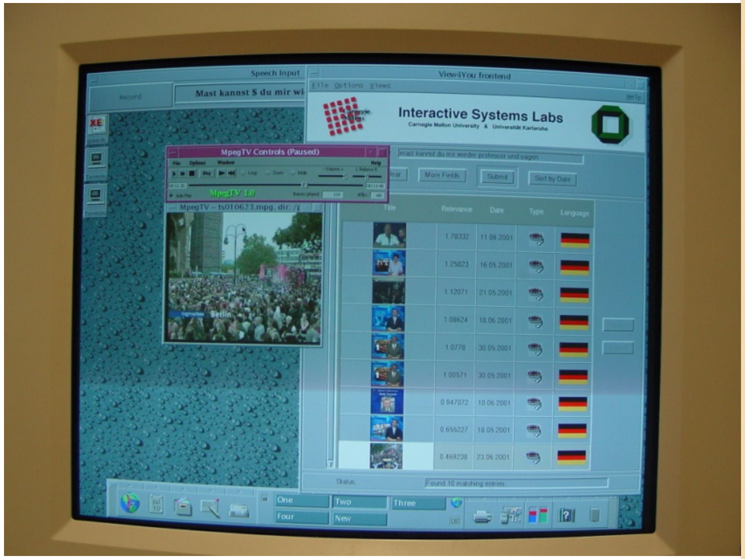
Human ←→ Machine

Human-Data ←→ Machine

Video-on-Demand

The View4You System

Was gibt’s denn heute Neues aus Karlsruhe ?



“View4You”: Video-on-Demand

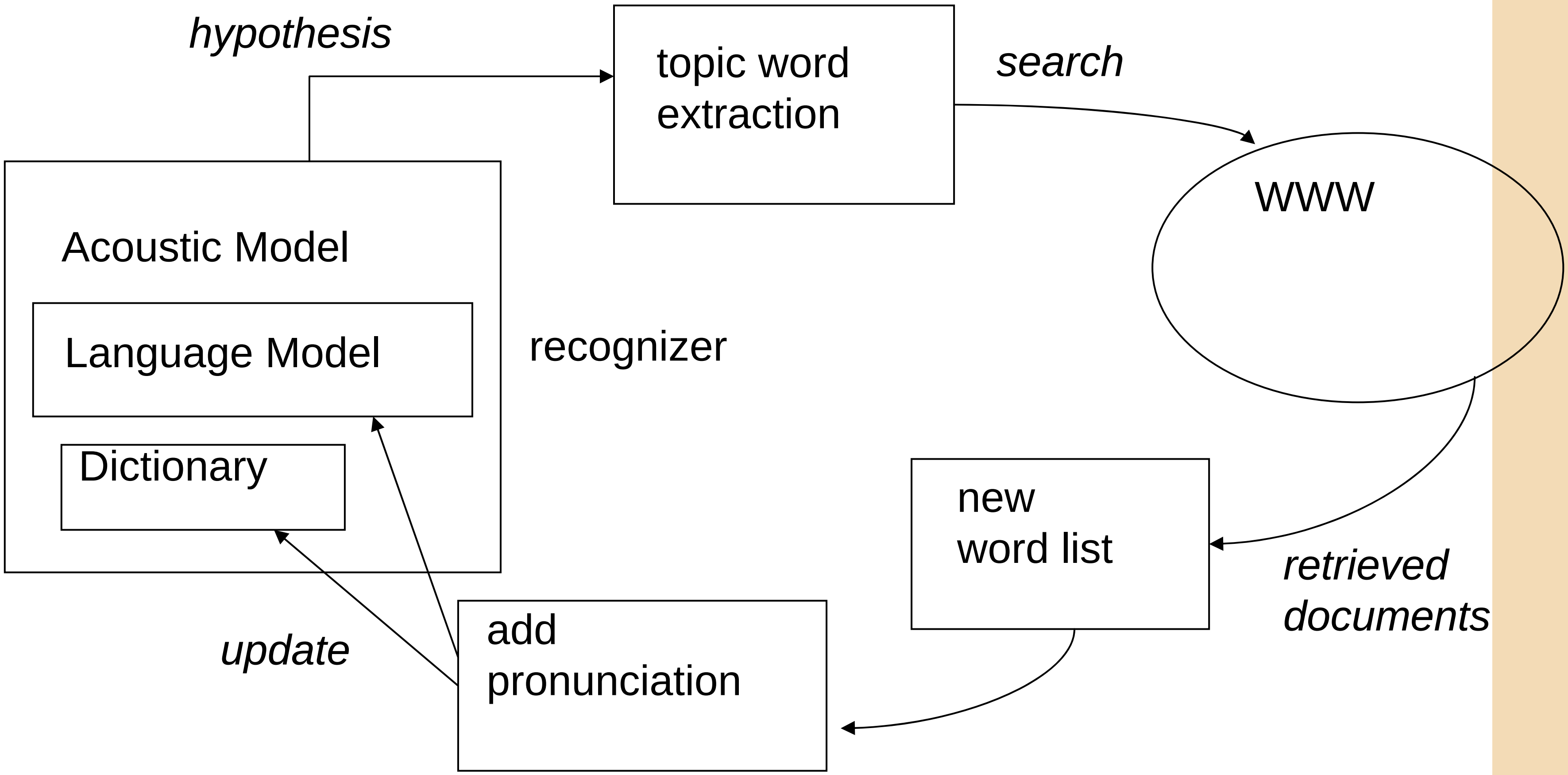
System Characteristics

- Recording Conditions
  - In case of TV, Mostly Low Noise but Varied (Correspondents, etc.)
- Speaker:
  - Few Dominant Talkers
  - Mostly Read Speech
- Issues:
  - No Interaction with Speaker, Cannot Influence Behavior
  - Vocabulary Maintenance
  - Perplexity Control
  - System Integration
- Problems:
  - Read-Speech → Conversational Speech (TV → Lcetures/Meetings)

Language Models

- Language Models: Use N-Grams
- Dictionary: Use Large 60,000+ Dictionary
- Problem:
  - News is Dynamic and Vocabularies Change
  - System Integration: Is Speech Recognition Good Enough ?
- Solution:
  - Adapt Dictionaries and Language Models Dynamically
  - Information Retrieval Can Accept Limited Reco Performance (even up to 30% WER !)

Dynamic Adaptation



Experiments

- Using mutual information extract keywords around key topics.
- Using keywords search for relevant documents on WWW
- Identify ‘new’ words in the new found documents
- Augment dictionary by new words
- Use Text-to-Speech Synthesis to get pronunciation.
- Result:
  - With 46k base dictionary + 7k token (0.5k word type) text, 11 / 23 OOV words are retrieved.

Acoustic Adaptation

- Use Recognition Runs over Past TV Shows
  - Recognizer ‘Listens’ to and ‘learns’ from TV all the time
- Assuming Recognition is Correct:
  - Adapt Acoustic Models
- Use Confidence Measures
  - To Weight Transcripts According to Assumed Reliability

Interface Type 5

Human ←→ Human

Machine Assisted Interaction

CHIL

Present Human-Computer Interaction

Human-Human Interaction

Humans Interacting With Humans

Multilingual Communication

Human-Human Interaction Support

- CHIL – Computer in the Human Interaction Loop
  - Rather than Humans in the Computer Loop
  - Explicit Computing Complemented by Implicit Support
- Implicit Computing Services
  - Support Human-Human Interaction Implicitly
  - Increasingly Powerful Computing Services
  - Implicit Services Observe Context and Understanding
  - Reduction in Attention to Technological Artifact,
    → Increased Productivity
  - Computer Learns from Human Activity Implicitly

Project CHIL

- **Integrated Project** (IP) in 6th Framework Program of the EC
  - One of three IP’s in the first call Multimodal/Multilingual:
- **International Consortium**:
  - 15 Partners from 9 countries
    in Europe (12) and the US (3)
- **Budget**
  - CHIL: 25 Million Euro Cost Volume for three Years
- **Other Projects:**
  - Integrated Projects: AMI, TC-STAR
  - DARPA: CALO

The CHIL Project

Coordination:

- Scientific Coordinator: Univ. Karlsruhe, Prof. A. Waibel, R. Stiefelhagen
- Financial Coordinator: Fraunhofer IITB, Prof. Steusloff, K. Watson

The CHIL Team:

IITB

Universität Karlsruhe (TH)

Fraunhofer Institut Informations- und Datenverarbeitung

ITC irst

DaimlerChrysler

TU/e technische universiteit eindhoven

ATHENS INFORMATION TECHNOLOGY
CENTER OF EXCELLENCE FOR RESEARCH AND GRADUATE EDUCATION

IBM

TALP UPC
Centre de Tecnologies i Aplicacions del Llenguatge i la Parla
UNIVERSITAT POLITÈCNICA DE CATALUNYA

CNRS LIMSI

KTH VETENSKAP OCH KONST

STANFORD UNIVERSITY

Carnegie Mellon

INRIA RHÔNE-ALPES

ELDA
Evaluations and Language resources Distribution Agency

Carnegie Mellon

Universität Karlsruhe (TH)

Examples of
Human-Human Communication
Problems Requiring Computer Support

Phone Calls During Meetings

Phone Calls During Meetings

Memory Jog

….What was his name? …Where did I meet him? …What did we discuss last time?

interACT

Language Support

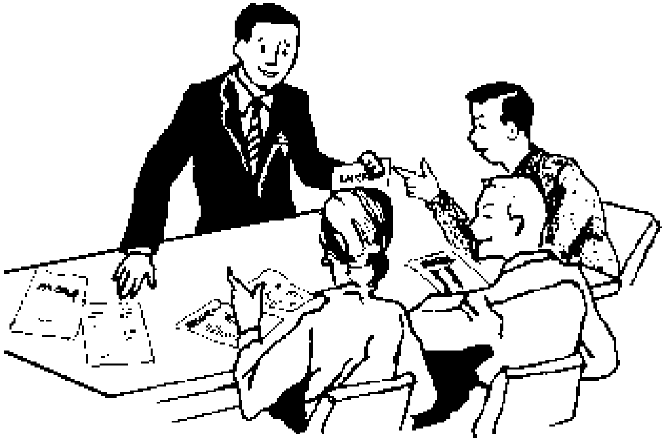
Carnegie Mellon

Universität Karlsruhe (TH)

Interpreting Human Communication

"Why did Joe get angry at Bob about the budget ?"

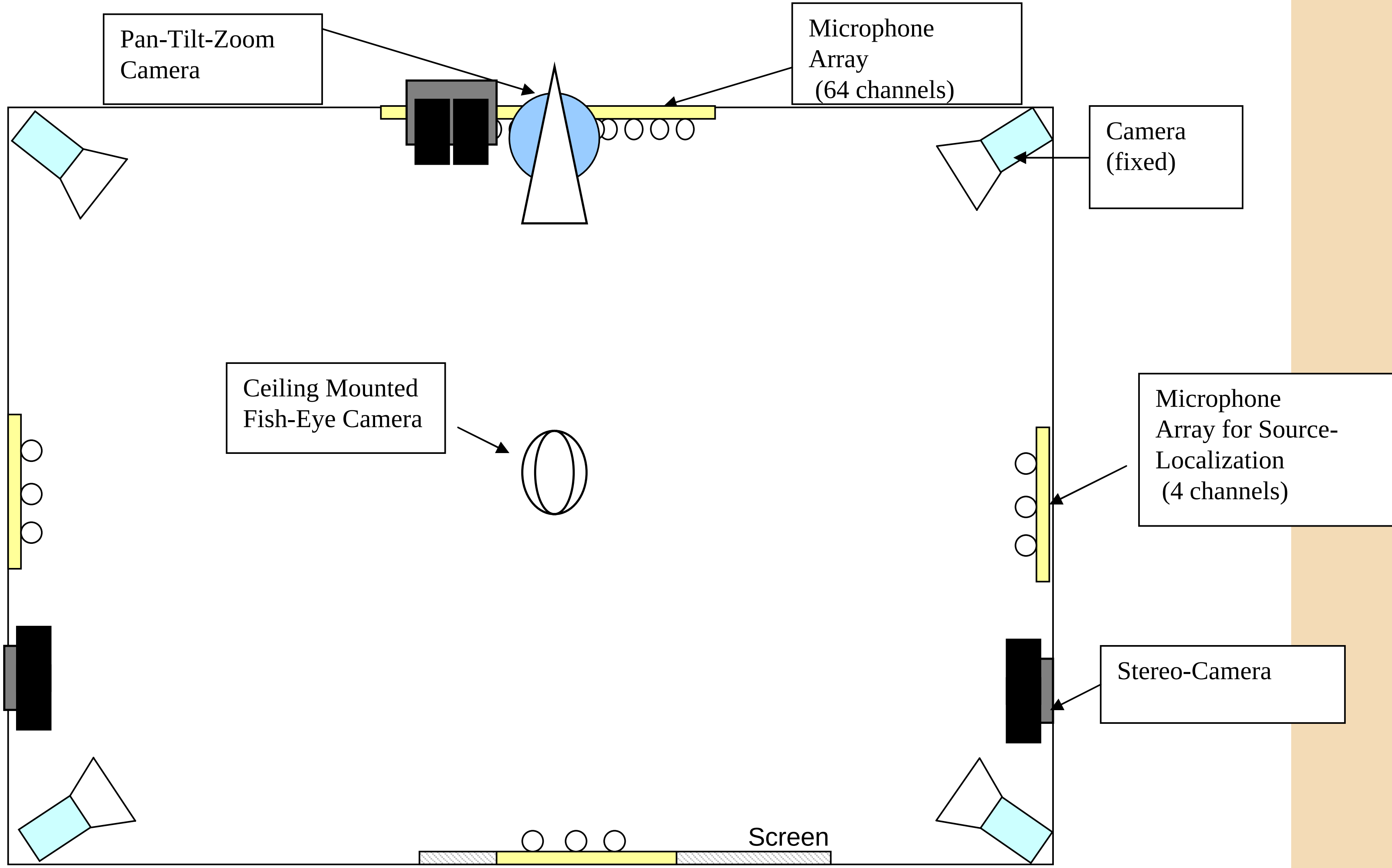
Need Recognition and Understanding of Multimodal Cues



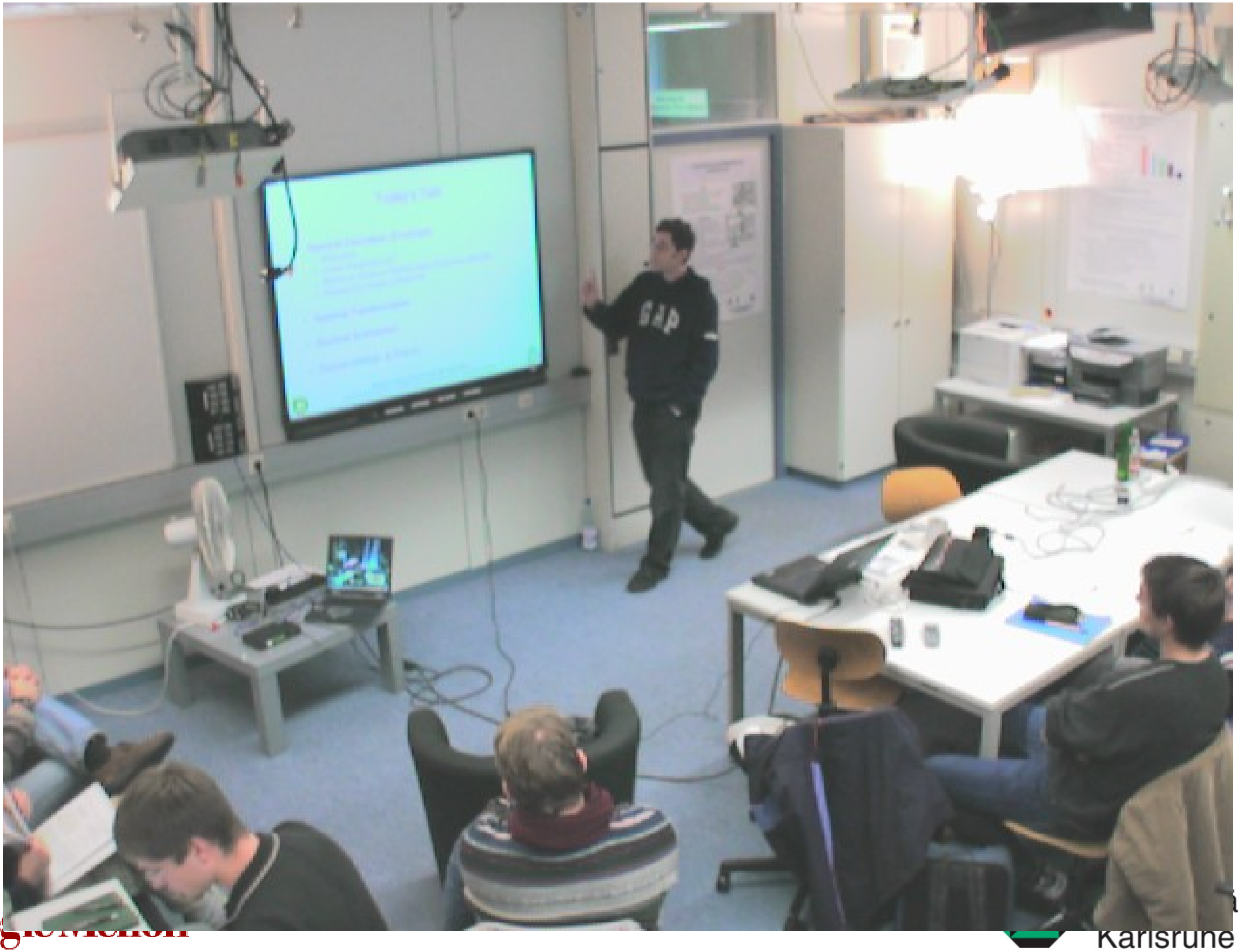
- Verbal:
  - Speech
    - Words
    - Speakers
    - Emotion
    - Genre
  - Language
  - Summaries
  - Topic
  - Handwriting
- Visual
  - Identity
  - Gestures
  - Body-language
  - Track Face, Gaze, Pose
  - Facial Expressions
  - Focus of Attention

We need to understand the: **Who, What, Where, Why** and **How !**

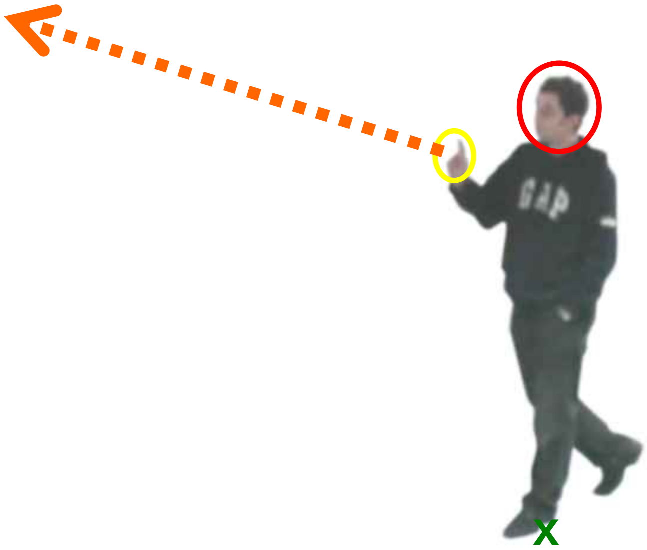
Sensors in the CHIL Room


Pan-Tilt-Zoom Camera
Microphone Array (64 channels)
Camera (fixed)
Ceiling Mounted Fish-Eye Camera
Microphone Array for Source-Localization (4 channels)
Stereo-Camera
Screen

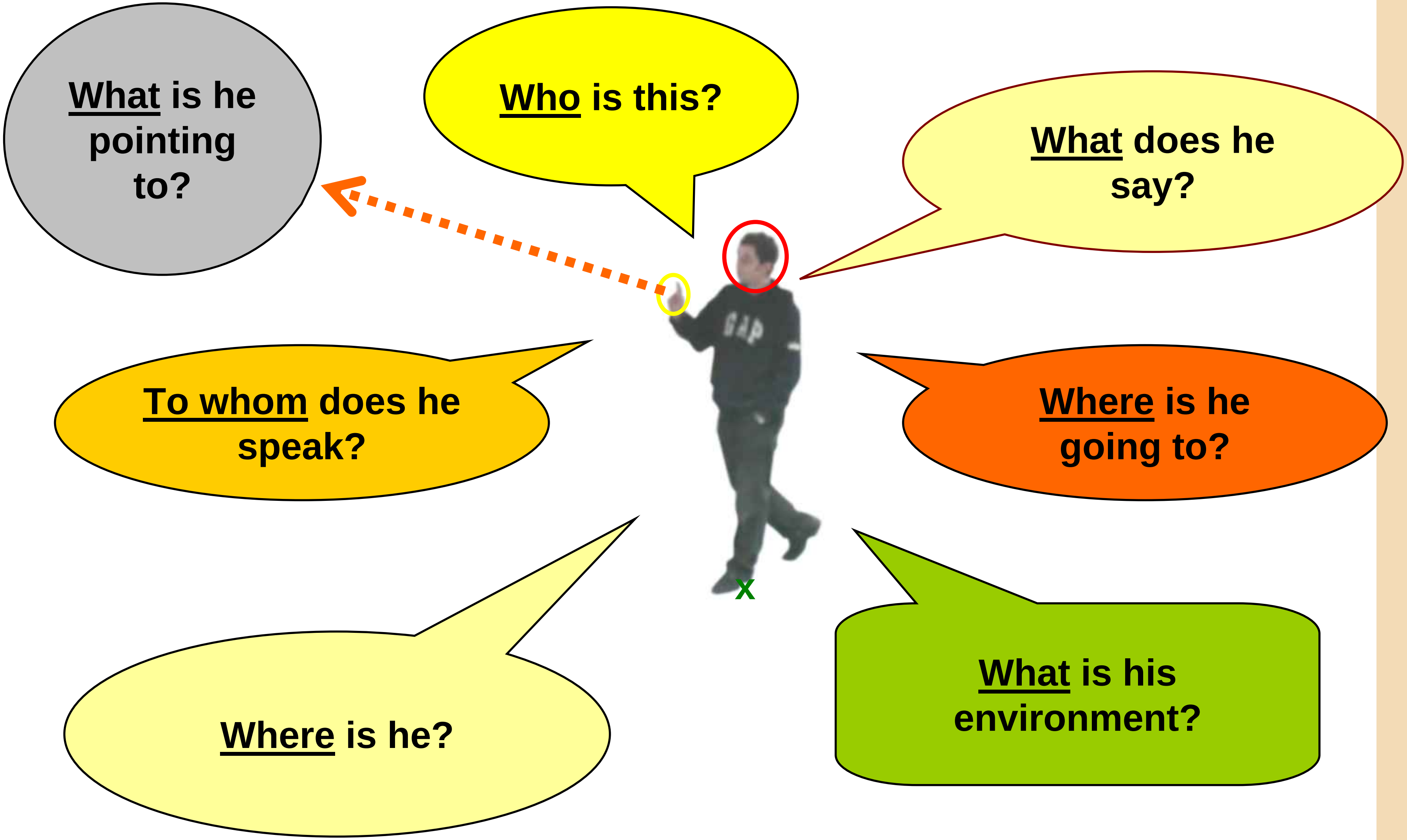
Describing Human Activities



Describing Human Activities



Technologies/Functionalities



What is he pointing to?

Who is this?

What does he say?

To whom does he speak?

Where is he going to?

Where is he?

What is his environment?

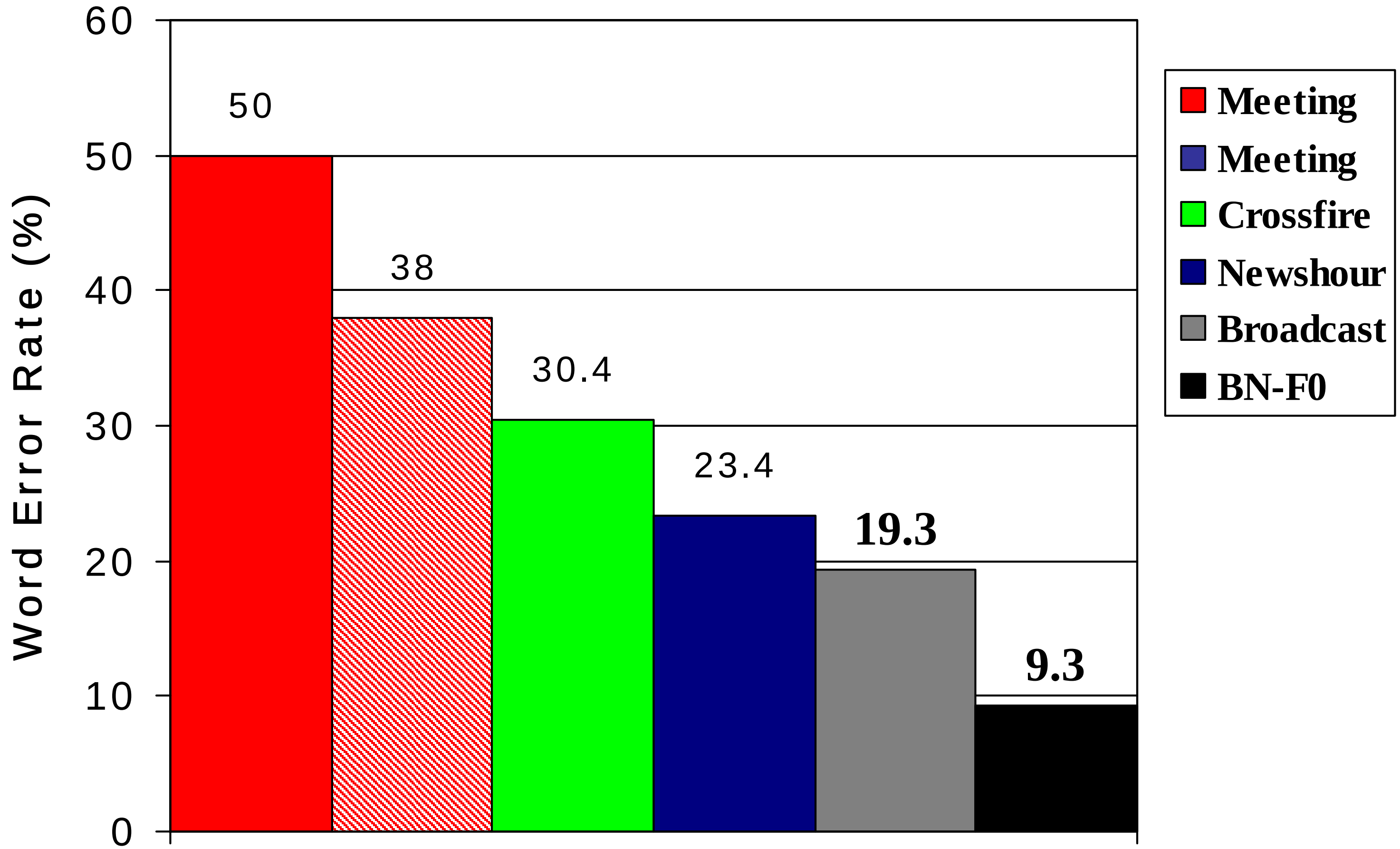
Technologies & Fusion

- **Who & Where ?**
  - Audio-Visual Person Tracking
  - Tracking Hands and Faces
  - AV Person Identification
  - Head Pose / Focus of Attention
  - Pointing Gestures
  - Audio Activity Detection
- **What ?** (Input)
  - Far-field Speech Recognition
  - Far-field Audio-Visual Speech Recognition
  - Acoustic Event Classification
- **What ?** (Output)
  - Animated Social Agents
  - Steerable targeted Sound
  - Q&A Systems
  - Summarization
- **Why & How ?**
  - Classification of Activities
  - Emotion Recognition
  - Interaction & Context Modelling
  - Vision-based posture recognition
  - Topical Segmentation

Special New Challenges & Opportunities

- Require: Performance, Robustness, Realism
  - Distant, Remote Microphones
  - Hands-Free, Always On → Segmentation
  - Sloppy Speech
  - Cross-Talk
  - Noise
  - Disfluencies, Prosody, Structuring Discourse
  - Communication by Other Modalities
  - Other Elements of Speech (Emotion, Direction, Scene Analysis
  - Multimodal People ID
  - Free People Movement
  - Focus of Attention and Direction
  - Named Entities, OOV's
  - Adaptation and Evolution
  - Summarization
- Now rapid Progress by Way of Competitive Evaluations

Recognition of Conversational Speech



TED: Lecture Speech Reco (WER)

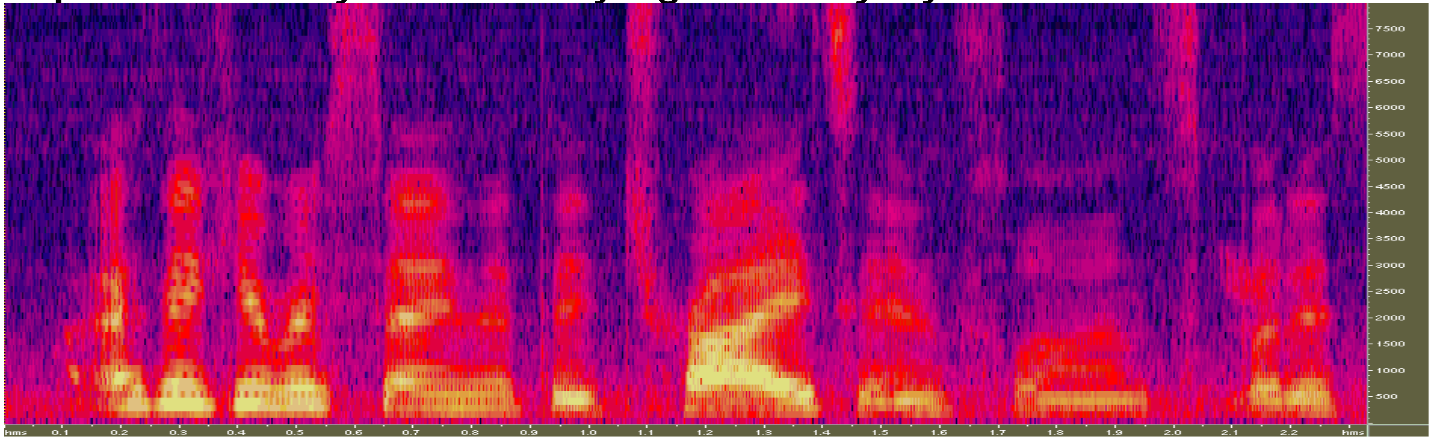
30%	Danish
35%	English
20%	French
40%	Italian
65%	French
45%	Japanese
35%	French
20%	German

interACT

Sloppy Speech in Meetings/Lectures

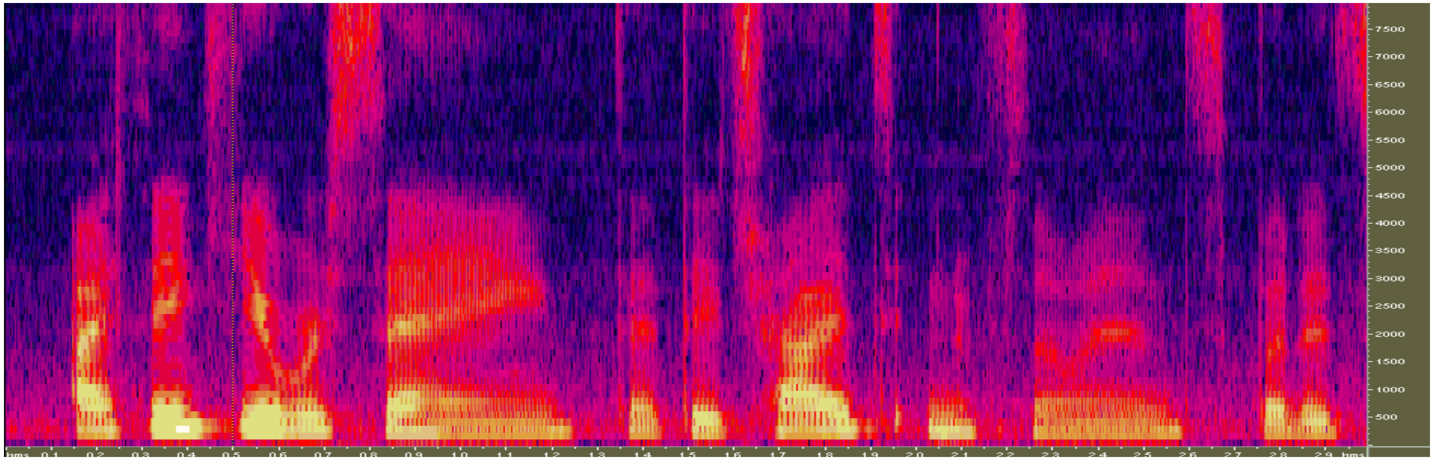
Actual Input: “*I think you were saying that they try to influence …*”

Conver-
Sational
Speech



Recognition: “*I think you insanity tries influence …*”

Read
Speech

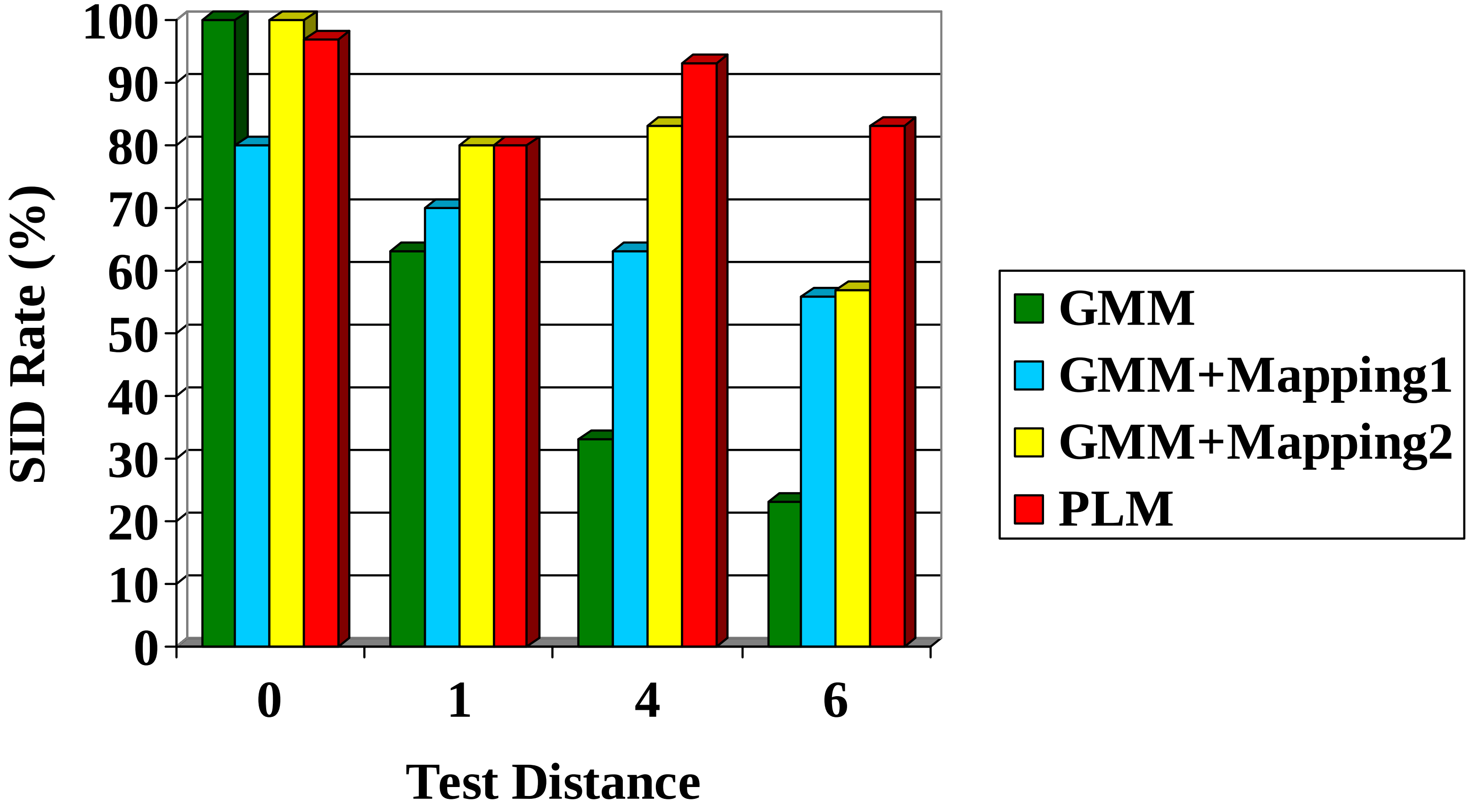


Recognition: “*I think you were saying that they tried to influence …*”

Carnegie Mellon

Universität
Karlsruhe (TH)

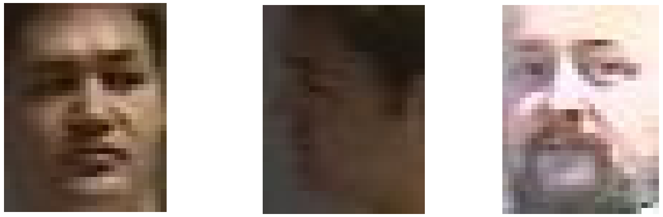
Speaker ID at Varying Mic Distances



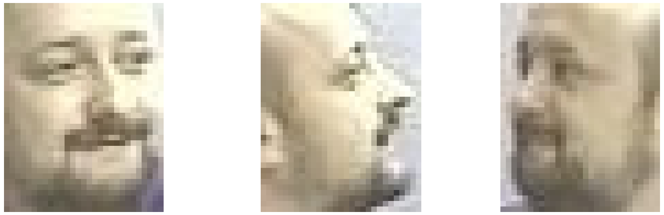
Face Identification

Low quality

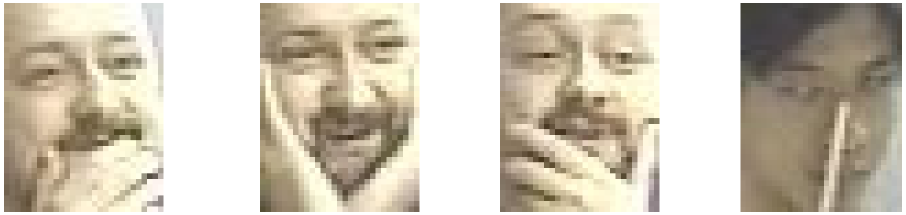
Illumination



Head pose



Occlusion



Evaluation: International Effort

- NIST and EC Programs Join Forces
  - RT-Meeting’06 – Rich Transcription
    - Emerges from established DARPA activity
    - MLMI Workshops, AMI/CHIL
    - Evaluated Verbal Content Extraction
    - Chair: Garofolo (NIST)
  - CLEAR’06 –
    Classification of Locations, Events, Activities, Relationships
    - Emerging from European program efforts (CHIL, etc.) and US-Programs (VACE,..)
    - First Joint Workshop to be Held in Europe
      after Face & Gesture Reco WS, April 13 & 14, Southampton
    - Chair: Stiefelhagen (UKA)

Participants in CLEAR'07

- CHIL (6)
  - AIT, UKA, FBKIRST, UPC, LIMSI, CMU
- VACE (6)
  - Pittsburgh Pattern Recognition
  - Univ. Illionous Urbana Champaign (T. Huang)
  - Univ. Southern California (R. Nevatia)
  - Univ. Maryland (L. Davis)
  - Univ. Central Florida (
  - Sarnoff
- AMI (1)
  - IDIAP
- Others (4)
  - MIT Lincoln Labs
  - Technical Univ. of Tampere, Finnland
  - Tsinghua University, China
  - Queen Mary University, UK

2007 CLEAR Tasks & Data Sets

		Source Data					
Task	Sub-Condition	Interactive Seminars (Meetings)			UKA Head Pose	VACE Surveillance	U A V
		VACE	CHIL	AMI			
3D Person Tracking	Video		X				
	Audio		X				
	Audio+Video		X				
2D Person Tracking						X	X
2D Face Tracking		X	X				
2D Vehicle Tracking						X	
Person ID	Video		X				
	Audio		X				
	Audio+Video		X				
Head Pose Estimation				X	X		
Acoustic Event Detection			X				

Total Tasks & Sub-Tasks:	15
CHIL Sponsored:	9
VACE Sponsored:	5
AMI Sponsored:	1

Carnegie Mellon

Universität Karlsruhe (TH)

2007 CLEAR: #Participants p. Task

		Source Data					
		Interactive Seminars (Meetings)			UKA Head Pose	VACE Surveillance	UAV
Task	Sub-Condition	VACE	CHIL	AMI			
3D Person Tracking	Video		4				
	Audio		5				
	Audio+Video		4				
2D Person Tracking						6	2
2D Face Tracking		3	3				
2D Vehicle Tracking						6	
Person ID	Video		5				
	Audio		6				
	Audio+Video		4				
Head Pose Estimation				2	5		
Acoustic Event Detection			7				

Total Tasks & Sub-Tasks: 15

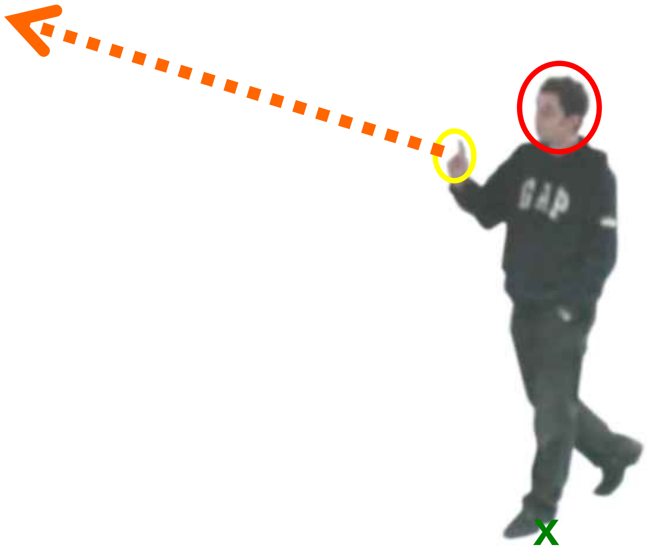
CHIL Sponsored: 9
VACE Sponsored: 5
AMI Sponsored: 1

CLEAR 2007 Results (best systems)

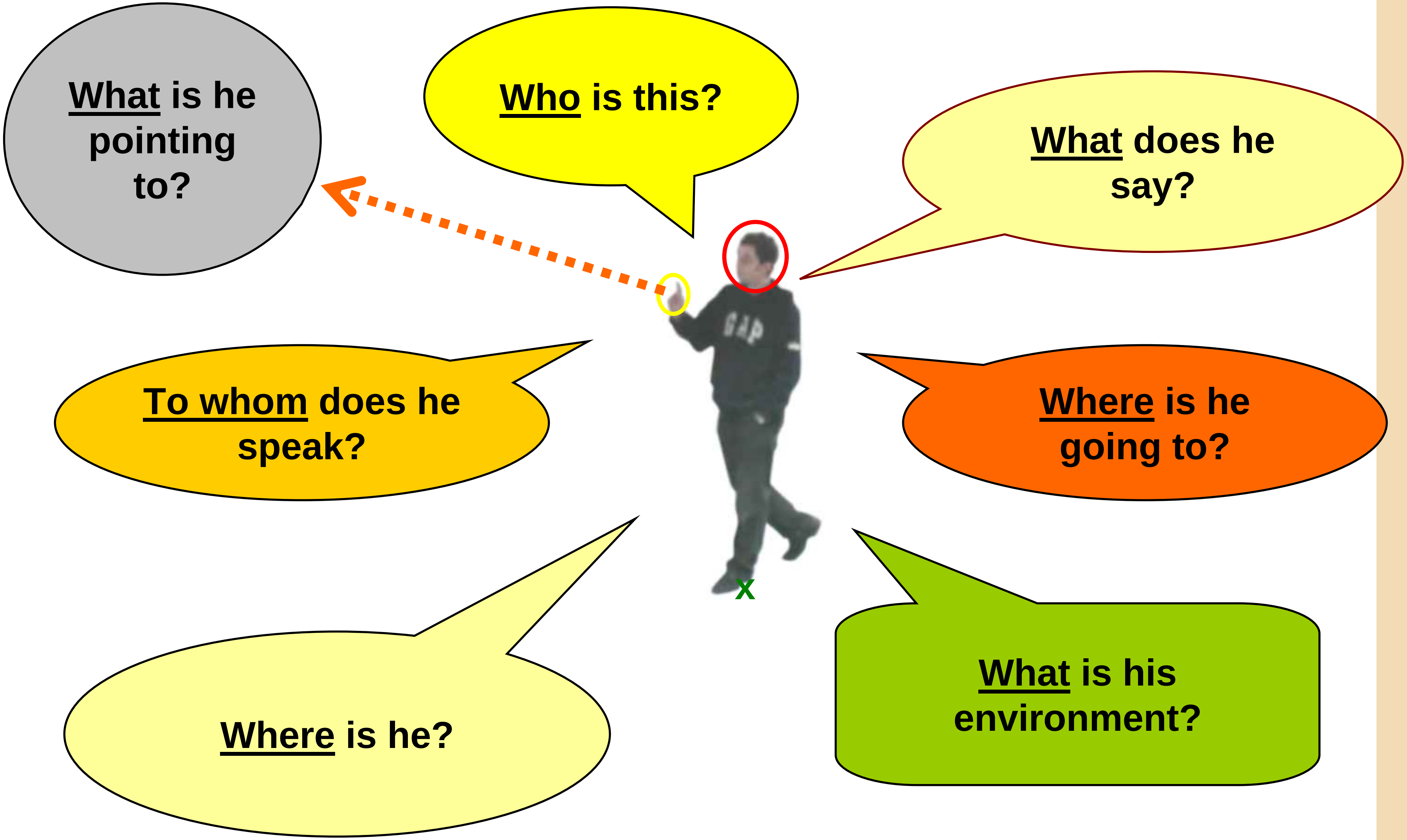
(not yet complete)

		Source Data					
Task	Sub-Condition	Meetings			CHIL Lectures	VACE Surveillance	UAV
		VACE	CHIL	AMI			
3D Person Tracking	Video		78% MOTA 9cm MOTP				
	Audio		54% MOTA 14cm MOTP				
	Audio+Video		58% MOTA 11cm MOTP				
2D Person Tracking						~62% MOTA ~57% MOTP	x
2D Face Tracking		~89% MOTA ~61% MOTP	x				
2D Vehicle Tracking						~71% MOTA ~61% MOTP	
Person ID	Video		85-96%				
	Audio		80-100%				
	Audio+Video		89-100%				
Head Pose Estimation				7°/9°/4° mean error	9°/9°/10° m. error		
Acoustic Event Detection			36%				

Describing Human Activities



Technologies/Functionalities



What is he pointing to?

Who is this?

What does he say?

To whom does he speak?

Where is he going to?

Where is he?

What is his environment?

interACT

Results, June 2004

Face Recognition (7 subjects)
- 76% with manual alignment
- 15% fully automatic

Head Detection:
- 78% correct (error < 15 pixel)

Head Orientation:
- Mean error ca. 10°

Hand Tracking:
- 73% correct

3D Pointing Gestures:
- 75% Recall
- 77% Precision

Body Tracking:
- 80,7% correct (error < 30 cm)
- mean error: 22 cm

Speech Recognition
- Close talking: 37% WER
- Far-field: 65% WER

Speech Detection
- 9% Mismatch rate (CTM)
- 12.5% far field

Source Localization:
- 11° root mean square error

Speaker ID:
- 100% correct, after 30s

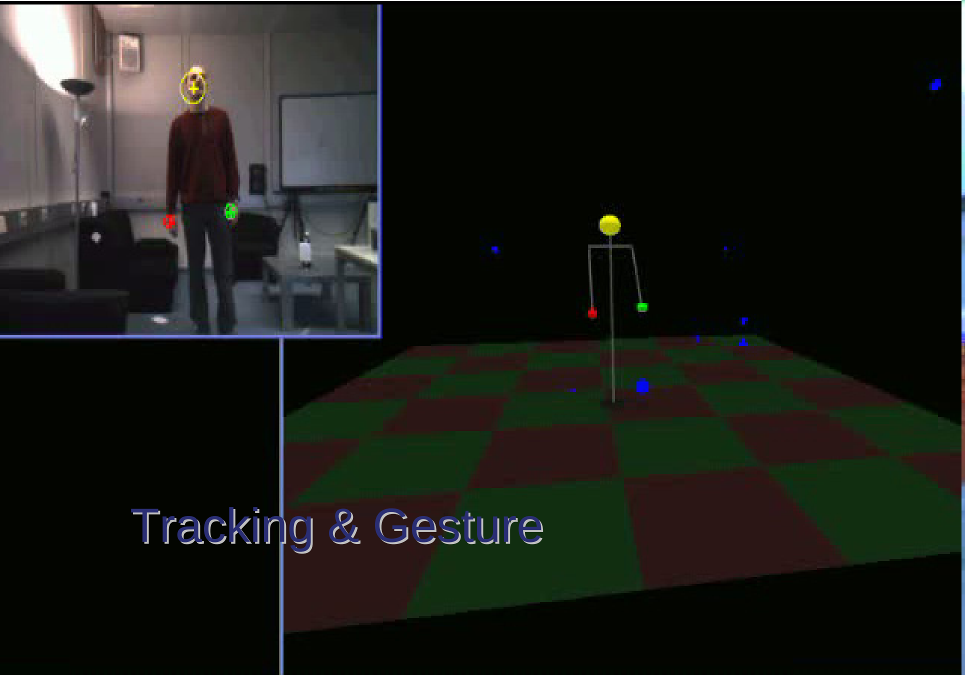
Accoustic event classification (25 classes)
- 38,4% error

Carnegie Mellon

Universität
Karlsruhe (TH)

Localization

Identification



Tracking & Gesture

Focus of Attention

Visual Components

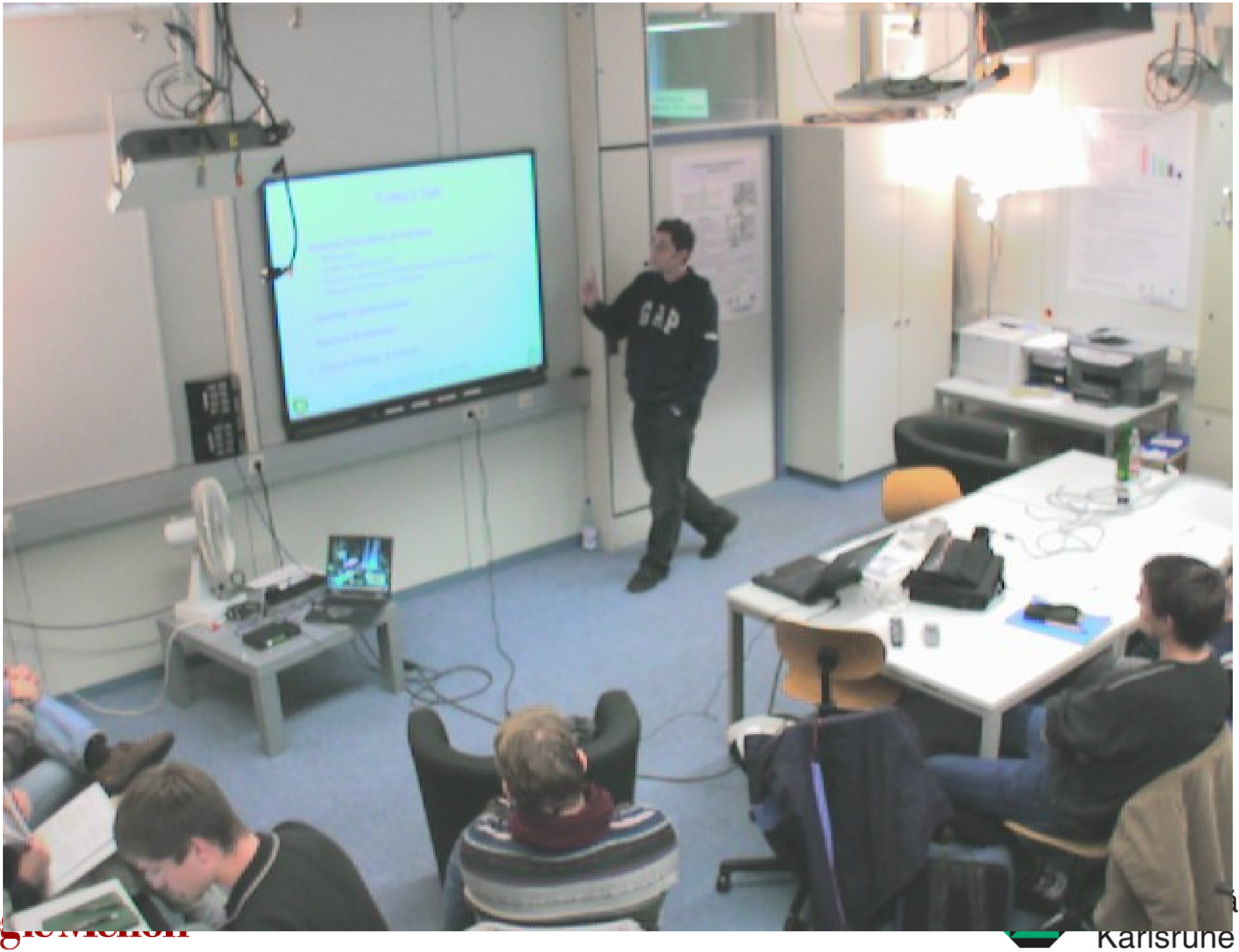
- Tracking
- Focus of Attention
- Face ID
- Gesture Recognition
- Multimodal Fusion
  - Multimodal People ID
  - Activity Analysis

Fusion/Integration:
People ID

Multimodal Fusion: Activity Analysis

Human Robot Interaction

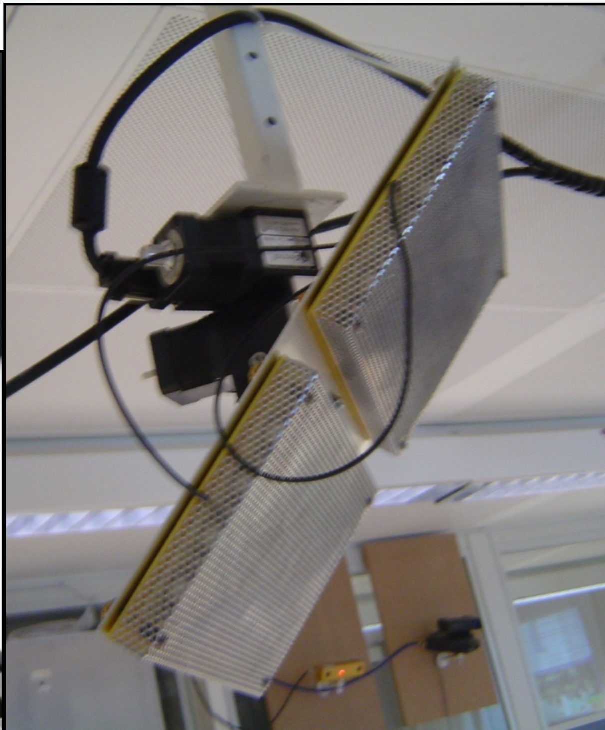
Describing Human Activities



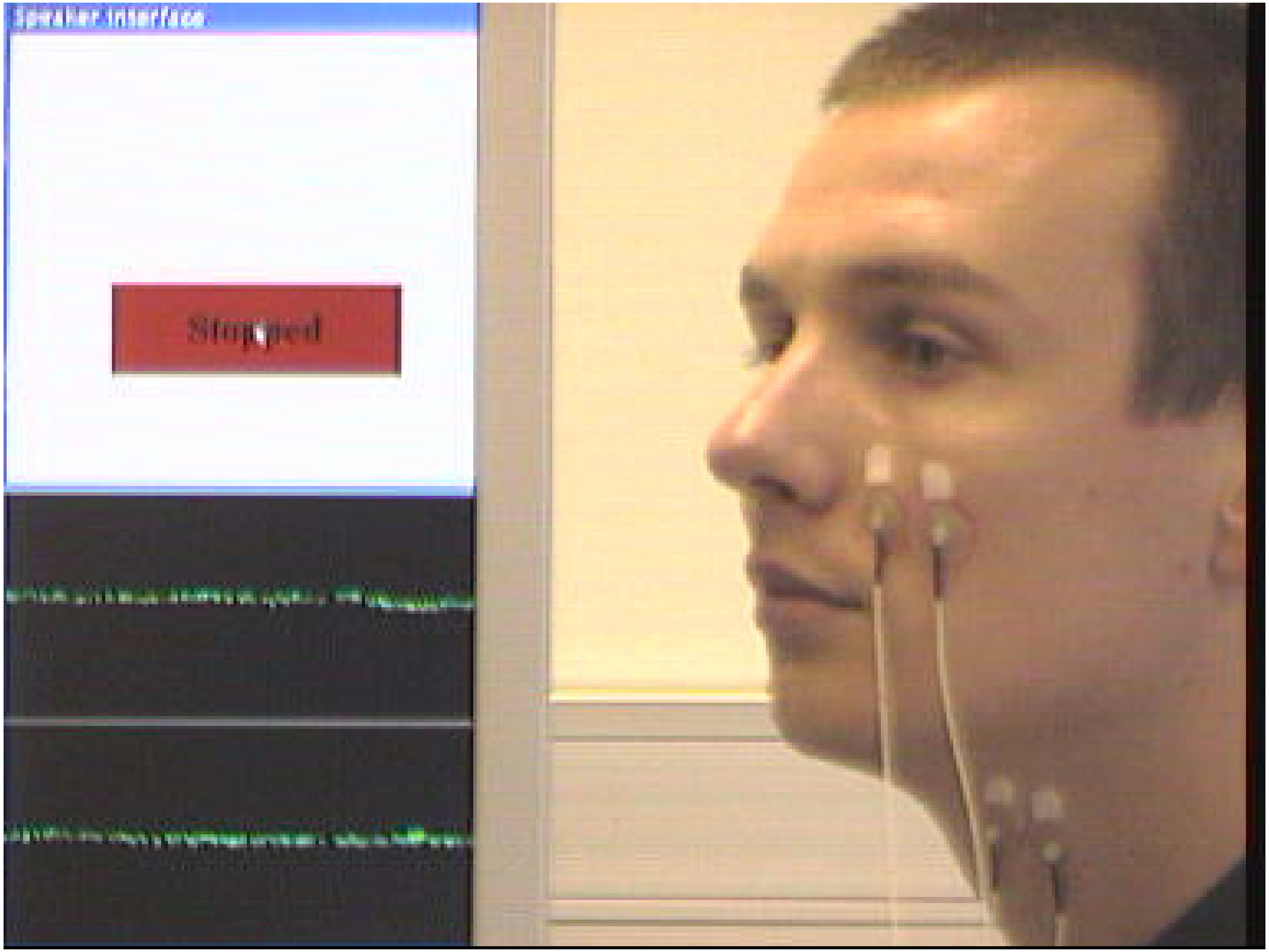
Implicit Information Delivery

Private and Public Information Delivery

- CHIL phone
- Steerable Camera Projector
- Targeted Audio
- Retinal and Heads-Up Displays



Silent Speech based on EMG Signals

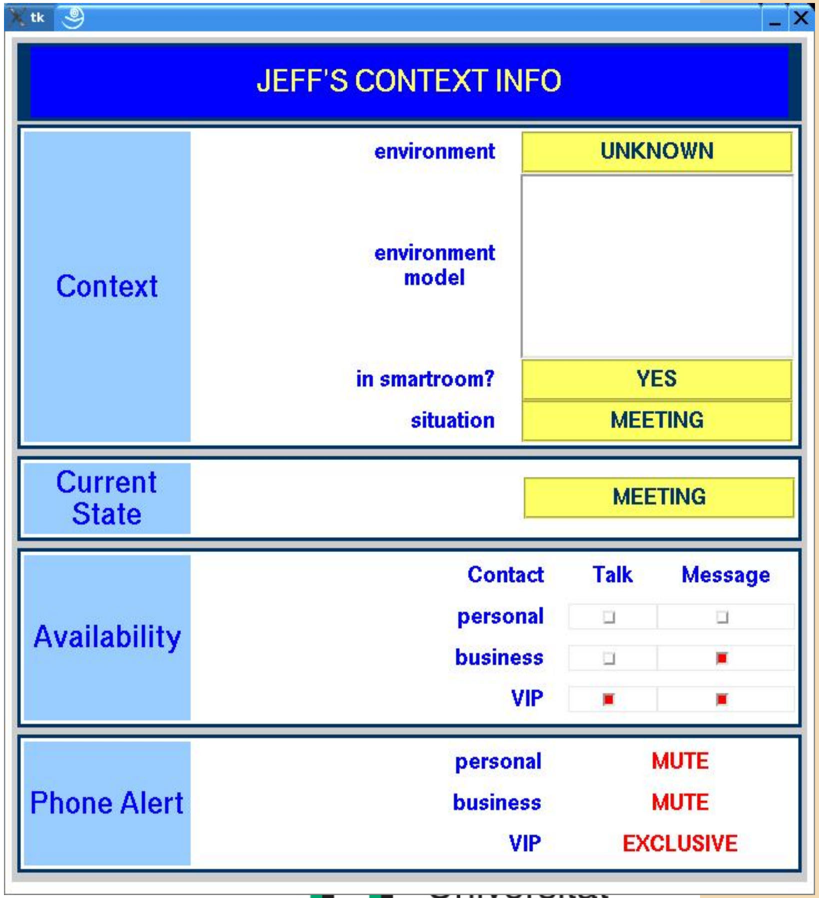


Human-Human Support Services

- **Connector**
  - Connects people through the right device at the right moment
- **Meeting Browser**
  - Create Corporate Memory of Events
- **Memory Jog**
  - Unobtrusive service. Helps meeting attendees with information
  - Provides pertinent information at the right time (proactive/reactive)
  - Lecture Tracking and Memory
- **Relational Report**
  - Informs the current speaker about interest/boredom of audience
  - Coaches Meetings to be More Effective
- **Socially Supportive Workspaces**
  - Physically shared infrastructure aimed at fostering collaboration
- ***Cross-Lingual Communication Services***
  - *Detect Language Need and Deliver Services Inobtrusively*
- … *(and more)*

The Connector

- Socially Appropriate Connection
  - Connect People when Appropriate by Appropriate Media
- Connecting People depends on:
  - Social Relationship of Parties
  - Space / Environment
  - Activity, User State
  - Urgency of Matter



JEFF'S CONTEXT INFO

Context

environment	UNKNOWN
environment model	
in smartroom?	YES
situation	MEETING

Current State: MEETING

Availability

Contact	Talk	Message
personal	□	□
business	□	■
VIP	■	■

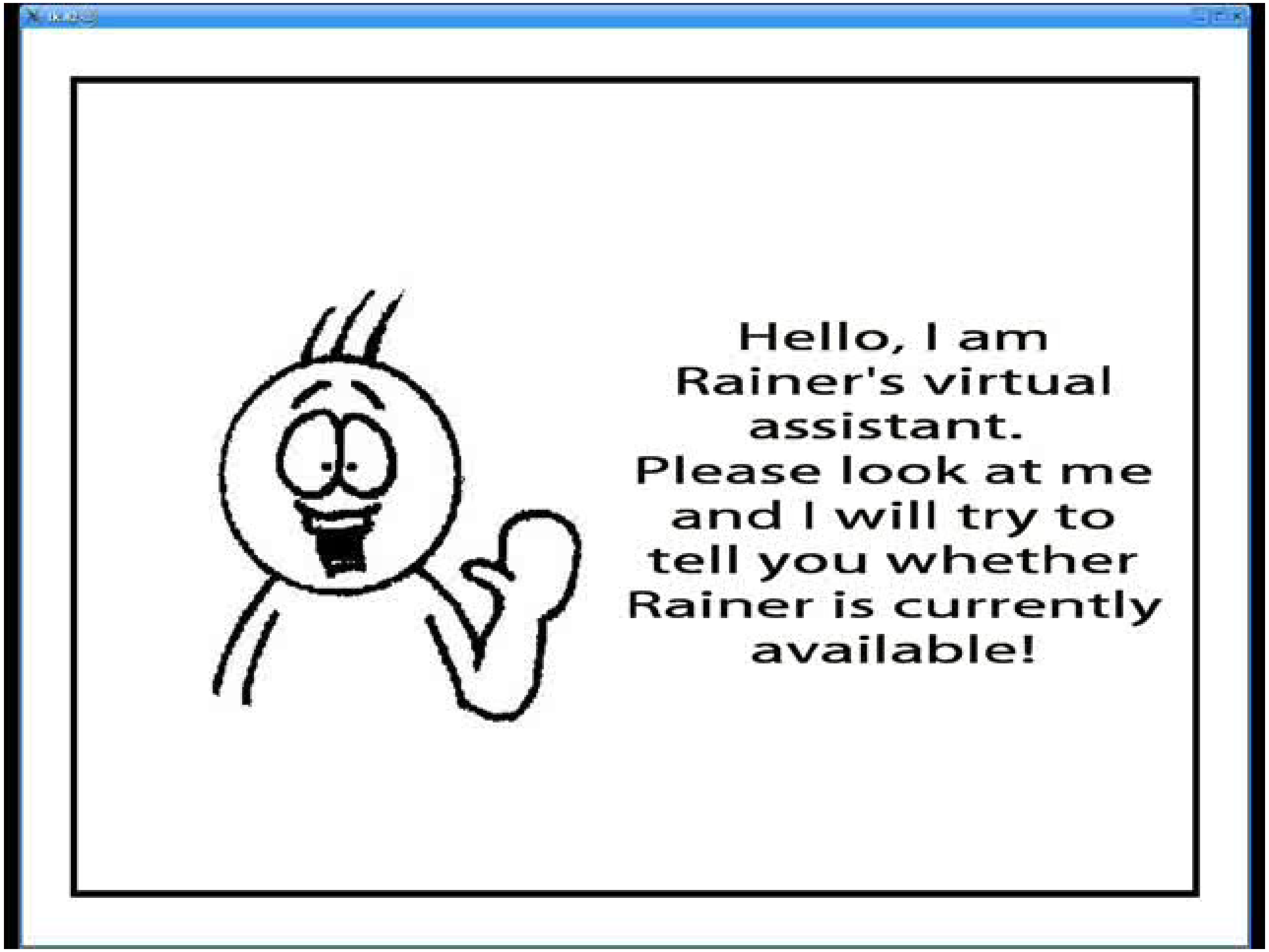
Phone Alert

personal	MUTE
business	MUTE
VIP	EXCLUSIVE

CHIL Connector – UKA

Connector - KTH

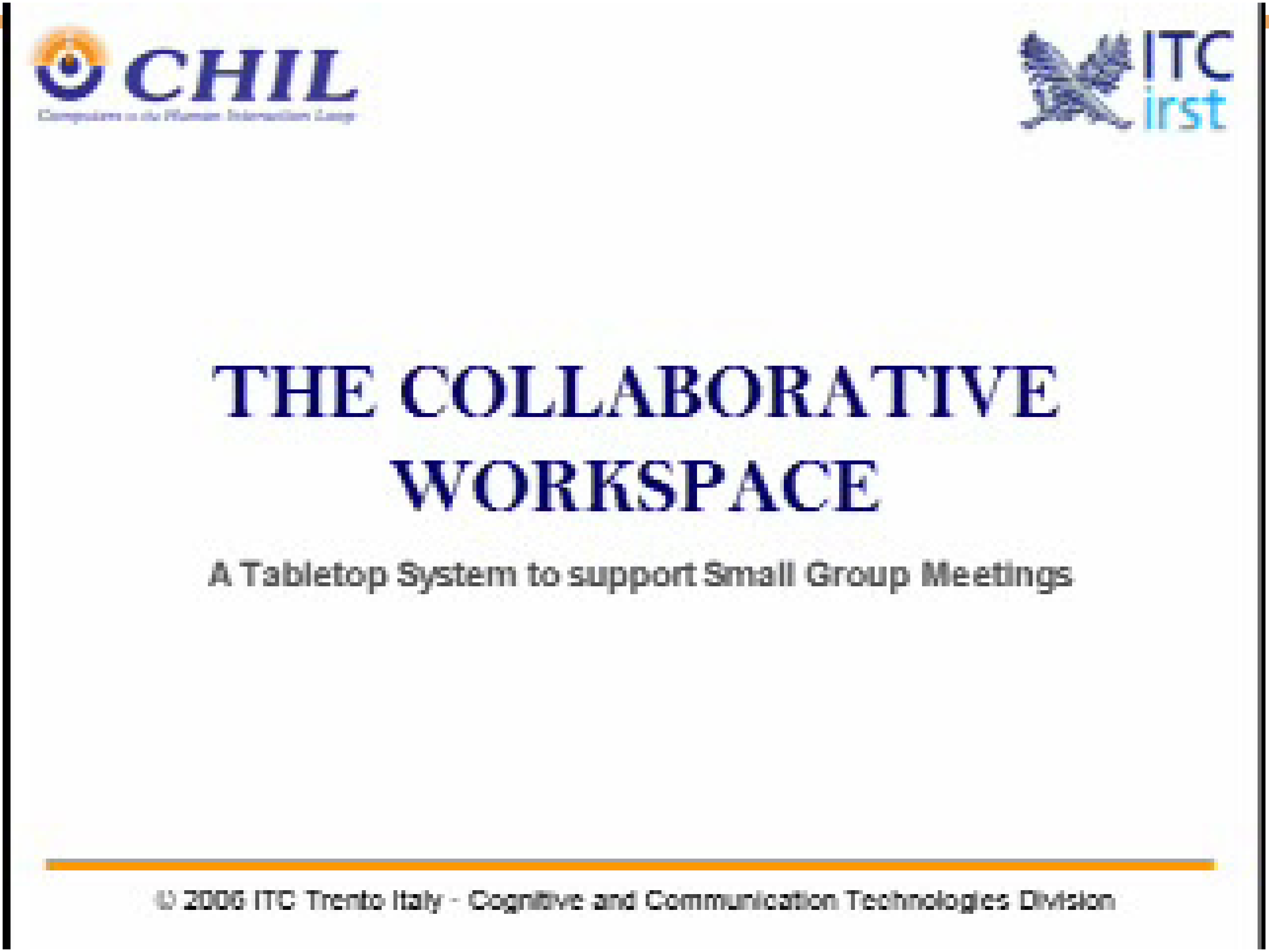
Office Receptionist



Hello, I am Rainer's virtual assistant.
Please look at me and I will try to tell you whether Rainer is currently available!

Memory Jog – UPC

interACT



CHIL

Computers in the Human Interaction Loop

ITC irst

THE COLLABORATIVE WORKSPACE

A Tabletop System to support Small Group Meetings

Carnegie Mellon

Universität
Karlsruhe (TH)

Interface Type 3

Human ←→ Machine ←→Human

Machine Mediation

Speech Translation

Motivation

- Dilemma:
  - Living in the Global Village
    - Globalization, Global Markets
    - Increased Exchange and Communication
    - European Integration
  - Cultural Diversity:
    - Beauty, Identity, Language, Culture, Customs
    - Pride and Individualism
  - Challenge:
    - Providing Access to Global Markets and Opportunities ←→ Maintaining Cultural Diversity
- Can Technology Provide Solutions?

Bridges Across the Linguistic Divide

Why is this so Hard ?

- Language is Ambiguous at All Levels:
  - Semantics:
    - The Spirit is Willing but the Flesh is Weak
    - → The Vodka is Good but the Meat is Rotten
  - Syntax:
    - Time Flies Like an Arrow → 6 Different Parses
  - Phonetics:
    - Give me a New Display → Give me a Nudist Play
- Problem:
  - A Sequence of Processing Modules will Compound Errors
- Solution:
  - Model Uncertainty Probabilistically
  - Maintain List or Lattice of Near Miss Working Hypotheses
  - Use Subsequent Knowledge Sources to Resolve Ambiguity

Conversational Speech

ja(2) guten Tag mein Name ist von Sudniz #AEHM# #ATMEN# #SCHMATZEN# #AEH# von #AEH# Frau oh denn also Sie sehen ich bin adelig #NICHT_ARTIKULATORISCH# #ATMEN# und Sudniz es oh #ATMEN# die denn die sechs #MIKROFON# #ATMEN# wenn wir das #AEH# #SCHMATZEN# #ATMEN# auch registriert haben da''s ich adelig bin und von Sudniz hei''se dann #ATMEN# w''urd' ich Sie doch #AEH# fragen wir m''ussen dringend #ATMEN# #SCHMATZEN# noch mal uns zusammensetzen #ATMEN# und ''uber unsere Reise kommende Woche #ATMEN# beziehungsweise ne nicht kommende Woche #ARTIKULATORISCH# was ich dann #ATMEN# unsre #ATMEN# #ARTIKULATORISCH# #AEH# Reise die wir vor hatten letzte Woche #ATMEN# und dann #ATMEN# an der Bar getroffen hatten und nach Kenia zusammen fliegen wollten und da wollten uns noch dar''uber unterhalten #ARTIKULATORISCH# #SCHMATZEN# #ATMEN# #AEHM# ja die Formalit''aten oder wie auch immer und ich w''urde dann vorschlagen da''s wir #ATMEN# uns m''oglichst demn''achst zusammensetzen und #ATMEN# #AEH# dann uns ''uberlegen #ATMEN# #AEH# wann wir nach Kenia fliegen und ob wir meine Safarib''uchse mitnehmen oder was wir da auch immer machen also #ATMEN# am #ATMEN# #SCHMATZEN# k''onnen Sie sich vielleicht schon vorstellen wann Sie da Zeit haben mal ''uber unsre Keniareise zu sprechen.

Text Based Translation...

...failed without Punctuation

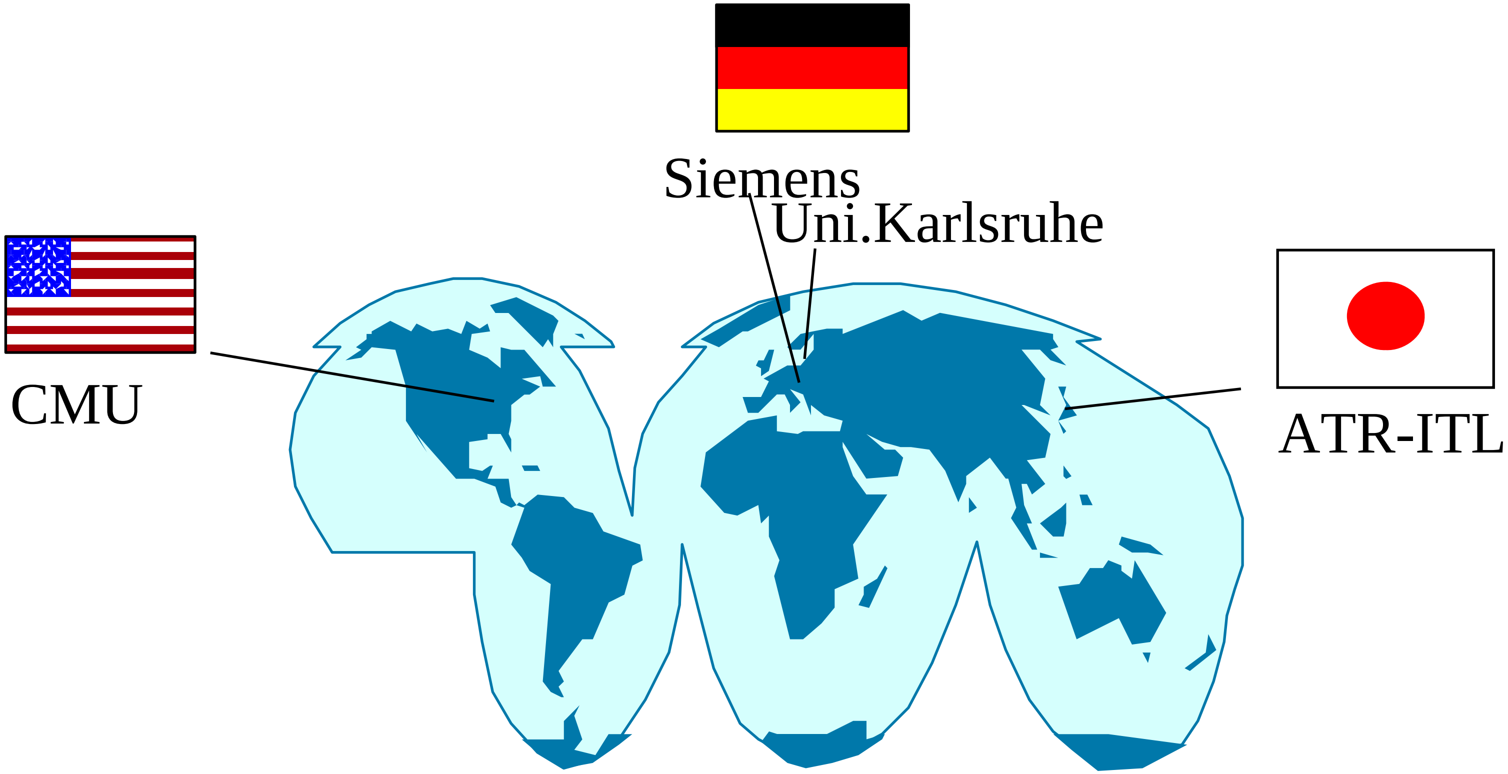
...after manually adding punctuation we get:

yes Hello, my name is of [Sudniz]. of woman. ah because therefore. You I see is titled, and [Sudniz]. it. ah, the because, the six, if we have noted the also, that I am titled, and of [Sudniz] hot, then I will ask however you: we must urgent again we compose and over our trip coming week respectively- [ne], not coming [Woche-] what I then. our trip, that we before had last week and then at the bar had met, and to [Kenia] together wants to fly; and there we want to talk still over it, the formalities, or as well as always, and I will propose then, that we compose ourselves preferably soon. and then we think, when we to [Kenia] fly. and if we my [Safaribüchse] brings. or what we there also always do. therefore at the could you introduce maybe already yourself, when you there does time have, to speak time over our [Keniareise]?

What the Human Wanted to Say:

"We should really schedule a meeting."

C-STAR



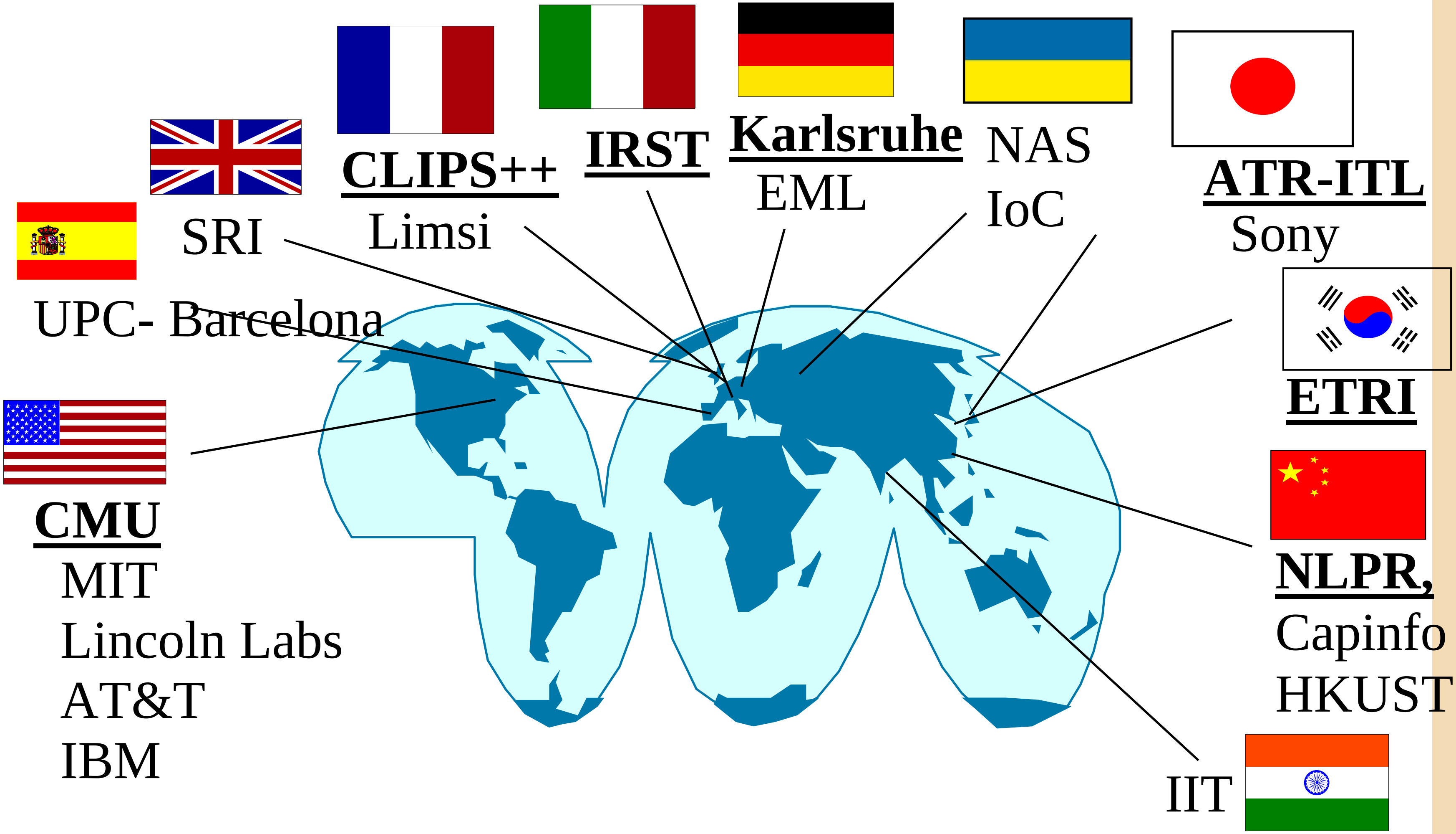
Siemens

Uni.Karlsruhe

CMU

ATR-ITL

C-STAR, now



UPC- Barcelona

SRI

CLIPS++
Limsi

IRST

Karlsruhe
EML

NAS
IoC

ATR-ITL
Sony

ETRI

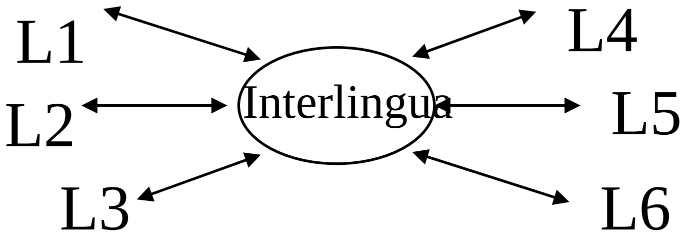
NLPR,
Capinfo
HKUST

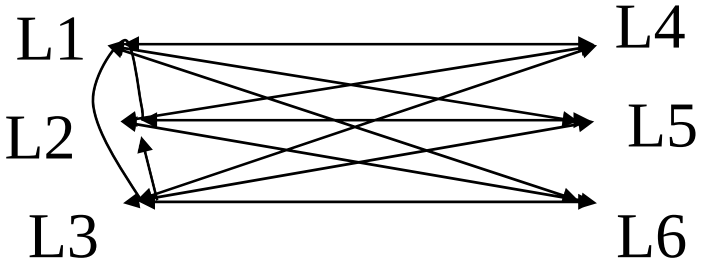
IIT

CMU
MIT
Lincoln Labs
AT&T
IBM

Interlingua Approach

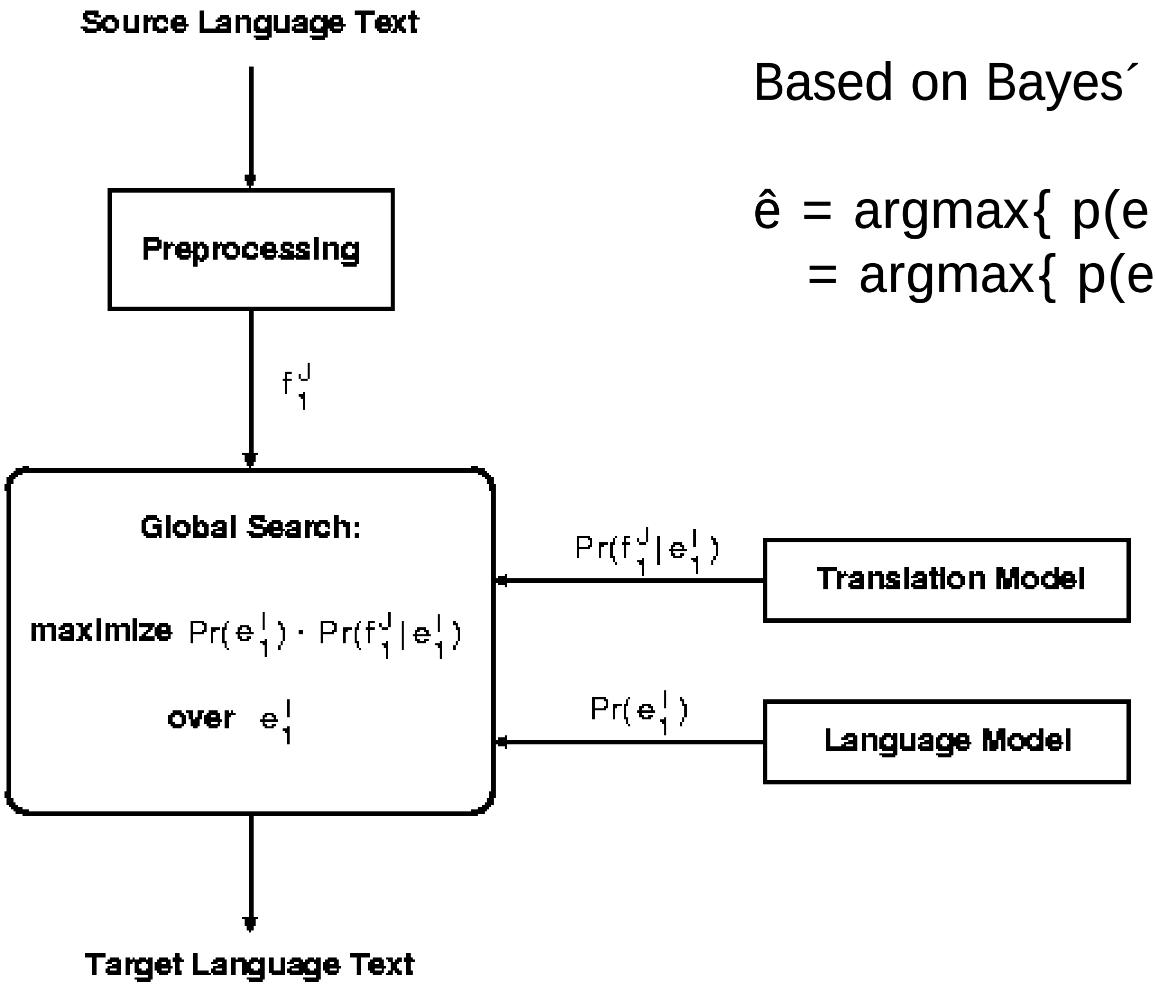
- C-STAR Partners Developed Common Interlingua – 6 languages
- Need only N parser/generators instead of N^2





- Rapid Addition of New Output Language
- Can generate culturally / contextually appropriate interpretation
- Eliminate Disfluencies, Clean-Up Language
- Generate Paraphrase in Own Language for Verification

Statistical Translation Approach



Source Language Text

Preprocessing

f_1^J

Global Search:

maximize $\Pr(e_1^I) \cdot \Pr(f_1^J | e_1^I)$

over e_1^I

$\Pr(f_1^J | e_1^I)$

Translation Model

$\Pr(e_1^I)$

Language Model

Target Language Text

Based on Bayes´ Decision Rule:

$$\begin{aligned} \hat{e} &= \operatorname{argmax}\{\ p(e \mid f)\ \} \\ &= \operatorname{argmax}\{\ p(e)\ p(f \mid e)\ \} \end{aligned}$$

The Grand Challenge

- **A World without Linguistic Borders**
- **Dimensions of the Problem:**
  - Overcoming Performance Limitations
    - Noise, Errors, Disfluencies
  - Expanding Domains and Scope
    - Hotel Reservation → Broadcast News, Lectures
  - Providing Suitable Access and Delivery
    - Mobile or Stationary Use
    - Modality → Speech, Image,
    - Natural Interaction → Human Factors/Devices
  - The Portability Problem
    - DARPA: 3 Languages
    - InterACT: 20 Languages
    - Speech and Language Companies: <40 Languages
    - Total World Languages: ~6,000

Speech Translation

History:

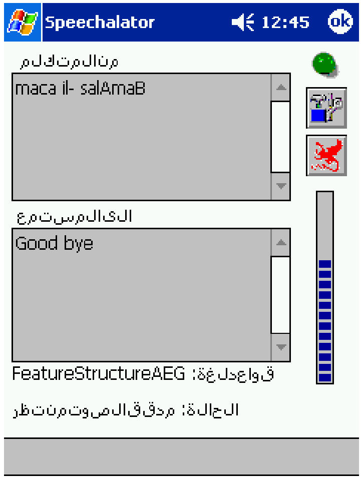
- Domain Limited, Clear Speaking Style (late 80’s-91)
  - Janus, ATT, NEC, ATR
- Domain Limited, Spontaneous (‘91-’00)
  - Janus II/III (work on 20 languages), Verbmobil, Nespole, Enthusiast, C-STAR, ATR, ETRI, NLPR,…
- Fieldable, Domain Limited, Spontaneous (current)
  - Transtac, Babylon, Phraselator, Thailator, ….

Fieldeable Domain Limited Speech Translation

Fieldable Systems:

PDA Speech Translators

- Tourism
  - Conferences
  - Business
  - Olympics
- Humanitarian
  - Refugee Registration
  - First Responder
  - Healthcare
    - USA, Latino Population
    - Europe, Expansion
    - Third World
- Government
  - Peace Keeping, Police



من المتكلم

maca il- salAmaB

الى المستمع

Good bye

قواعد لغة: FeatureStructureAEG

الحالة: مدقق الصوت منتظر

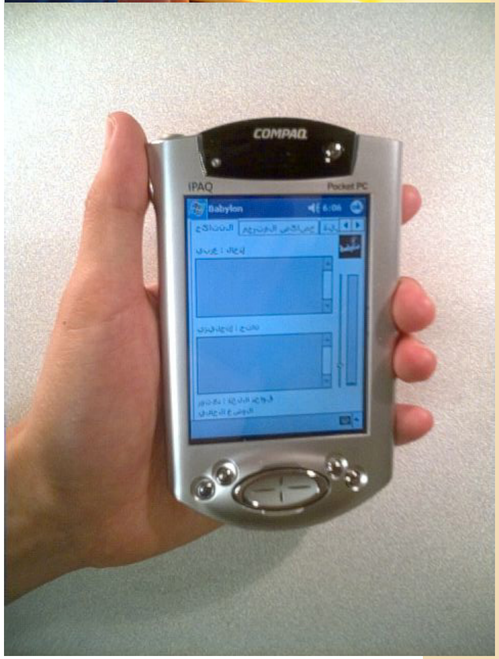
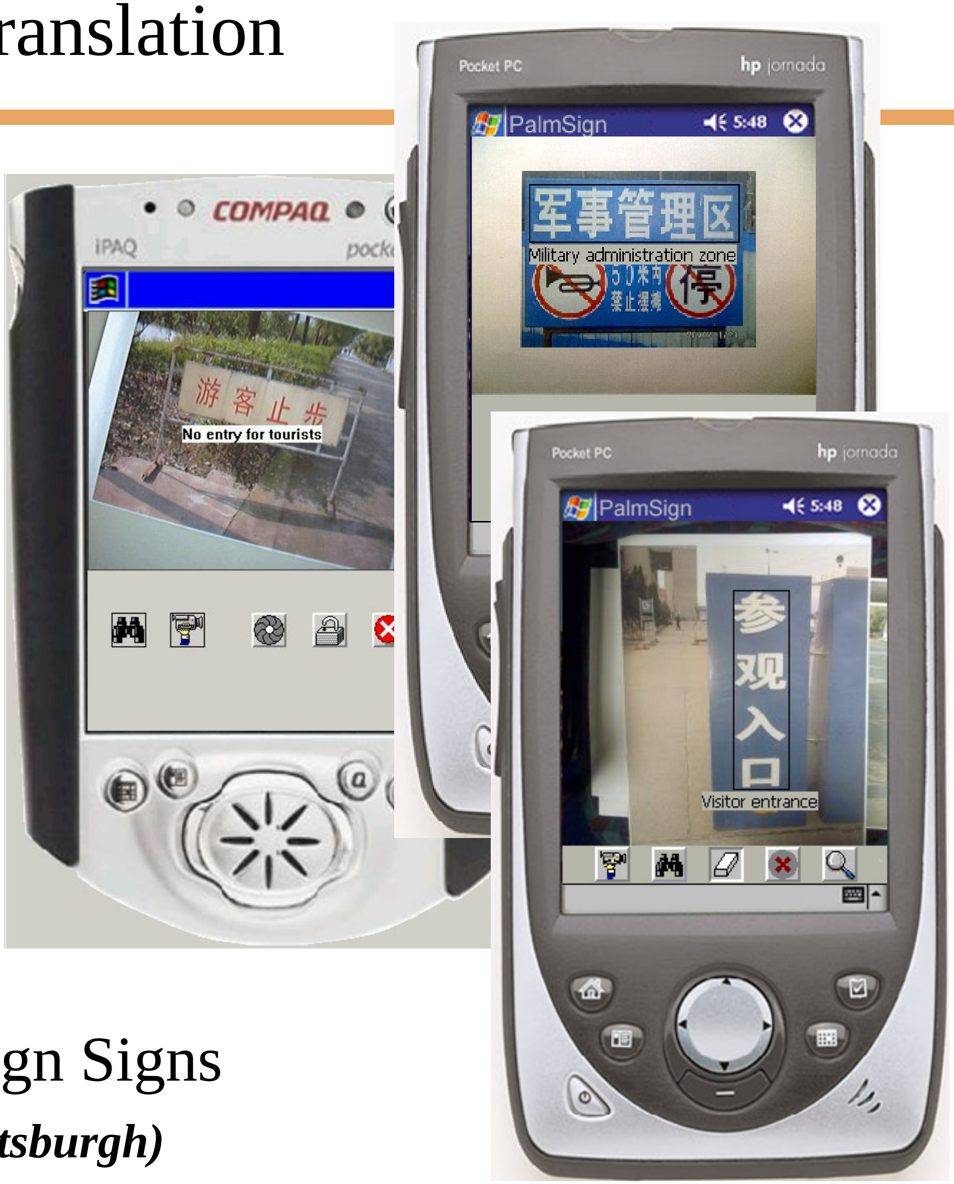


Image Translation

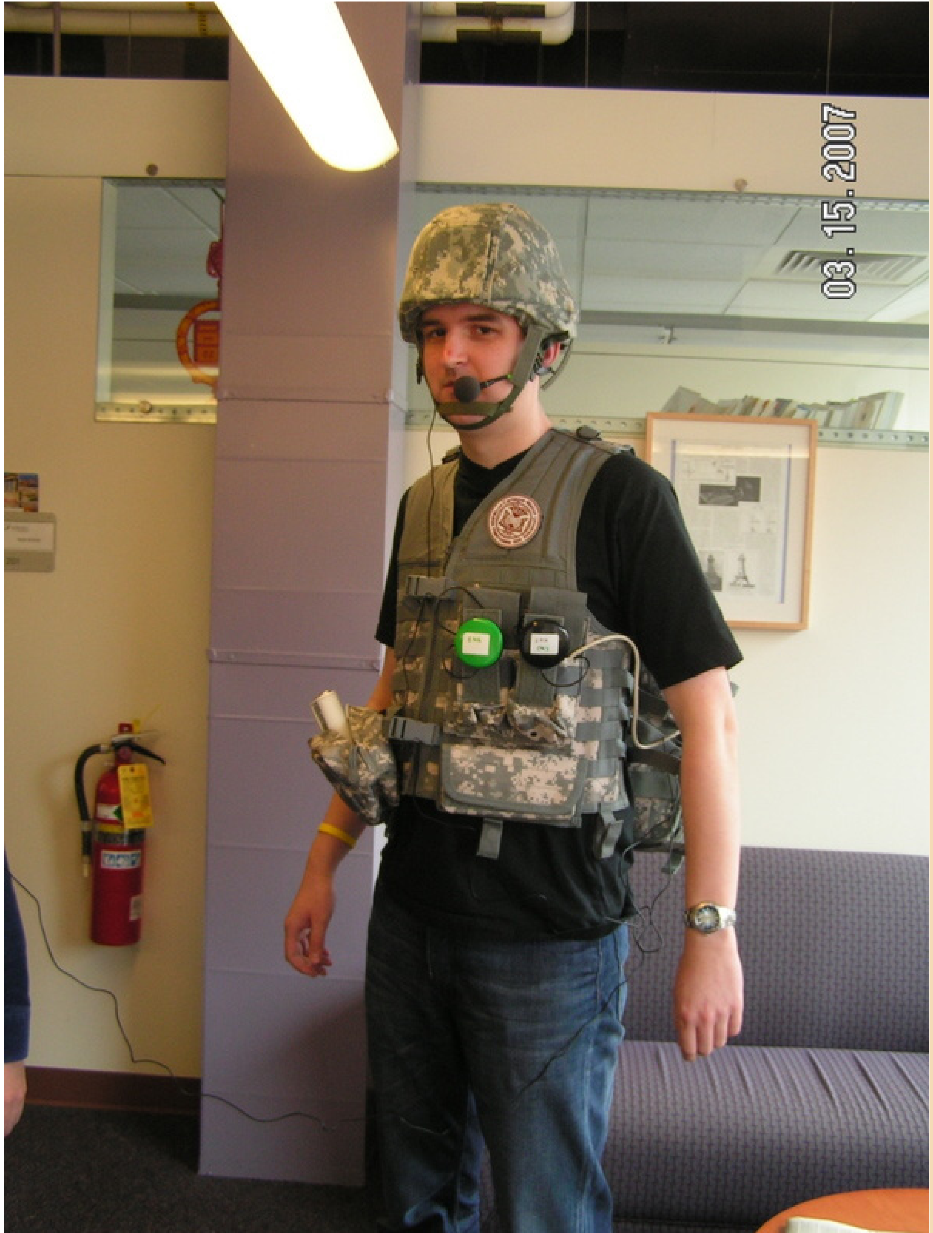


Pocket Translator of Foreign Signs

(Mobile Technologies, LLC Pittsburgh)

Mobility

- Hands-/Eye- Free Ops
- Integrated in Vest
- Close Speaking Mic
- Domain Limited
- Two-Way Device



03. 15. 2007

Demo

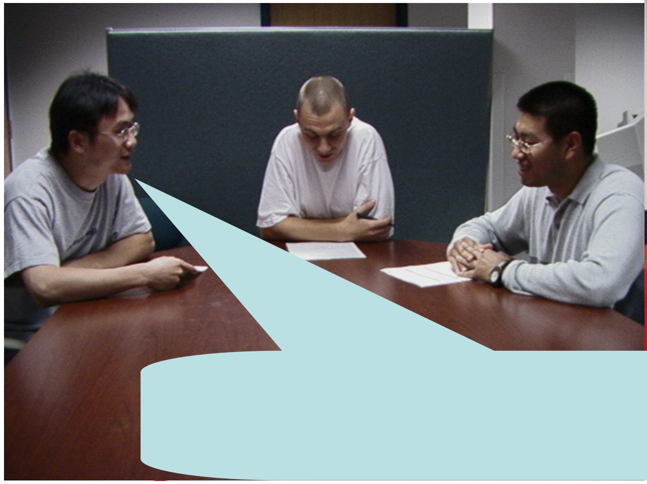
Language Support

….what is he saying?

Missing Science

Problem 1: Domain Limitation
cannot handle:

- TV/Radio Broadcast Translation
- Translation of Lectures and Speeches
- Parliamentary Speeches (UN, EU,..)
- Telephone Conversations
- Meeting Translation



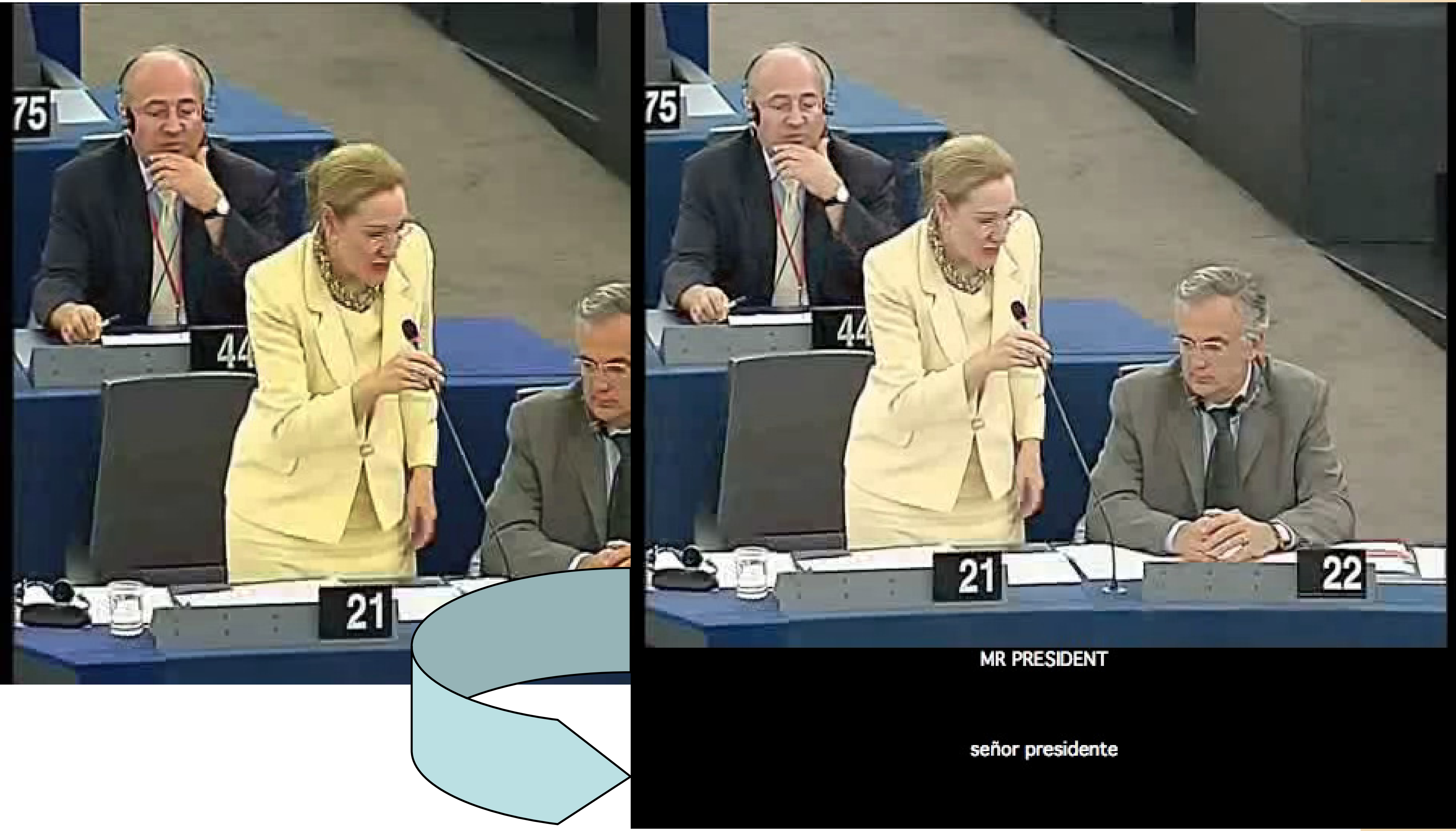
Language Support

Speech Translation

Progress:

- Domain Limited, Clear Speaking Style (late 80’s-91)
  - Janus (first European&US speech-to-speech system)
  - ATT, NEC, ATR
- Domain Limited, Spontaneous (‘91-’00)
  - Janus II/III (work on 20 languages),
    Verbmobil, Nespole, Enthusiast,
    C-STAR, ATR, ETRI, NLPR,…
- Fieldable, Domain Limited, Spontaneous (current)
  - Transtac, Babylon, Phraselator, Thailator, ….
- Domain Unlimited Speech Translation
  - Parliamentary Speeches (TC-STAR)
  - Broadcast News (GALE)
  - Lectures, Seminars (InterACT, STAR-DUST, TC-STAR)

Translation of Speeches



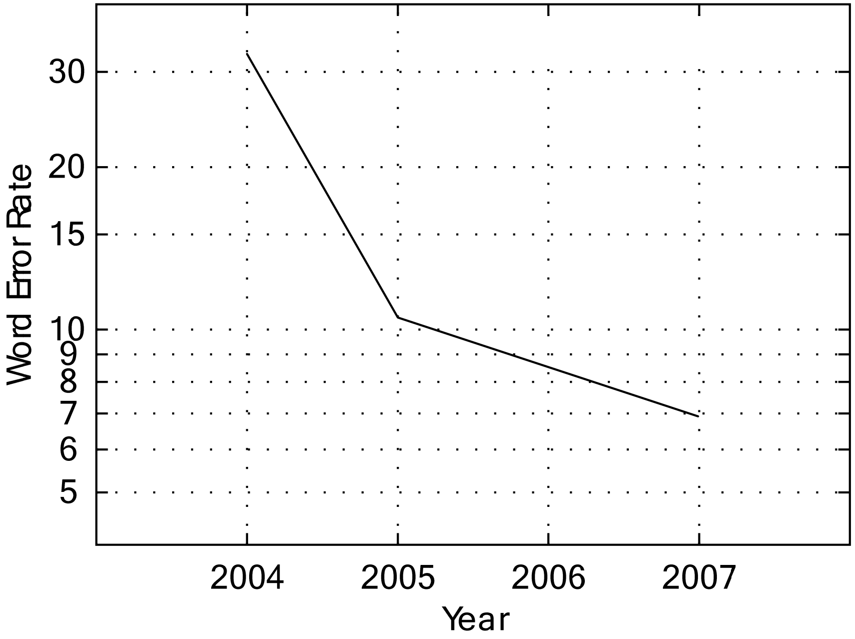
MR PRESIDENT

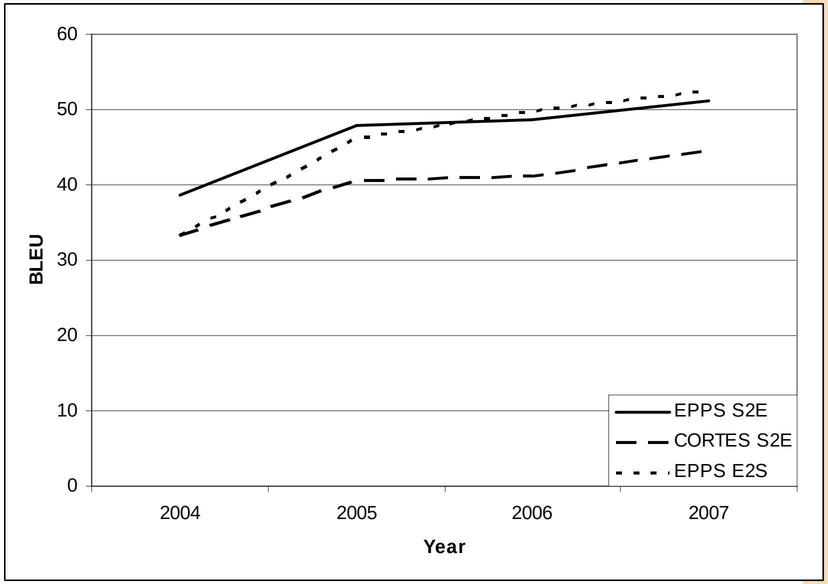
señor presidente

Translation of Speeches

- Technical Challenges:
  - Open Domain, Open Vocab, Open Speaking Style
  - No Sentence Markers/Boundaries
  - Too Complex to Program Rules
  - Reasonable Speaking Style, Prepared Speeches, Reasonable Acoustics
- How it is Done:
  - Statistical Learning Algorithms
  - Learn Speech and Translation Mappings from Large Example Corpora

Progress TC-STAR





Speech Recognition [WER]

Machine Translation [Bleue]

Translation of Lectures

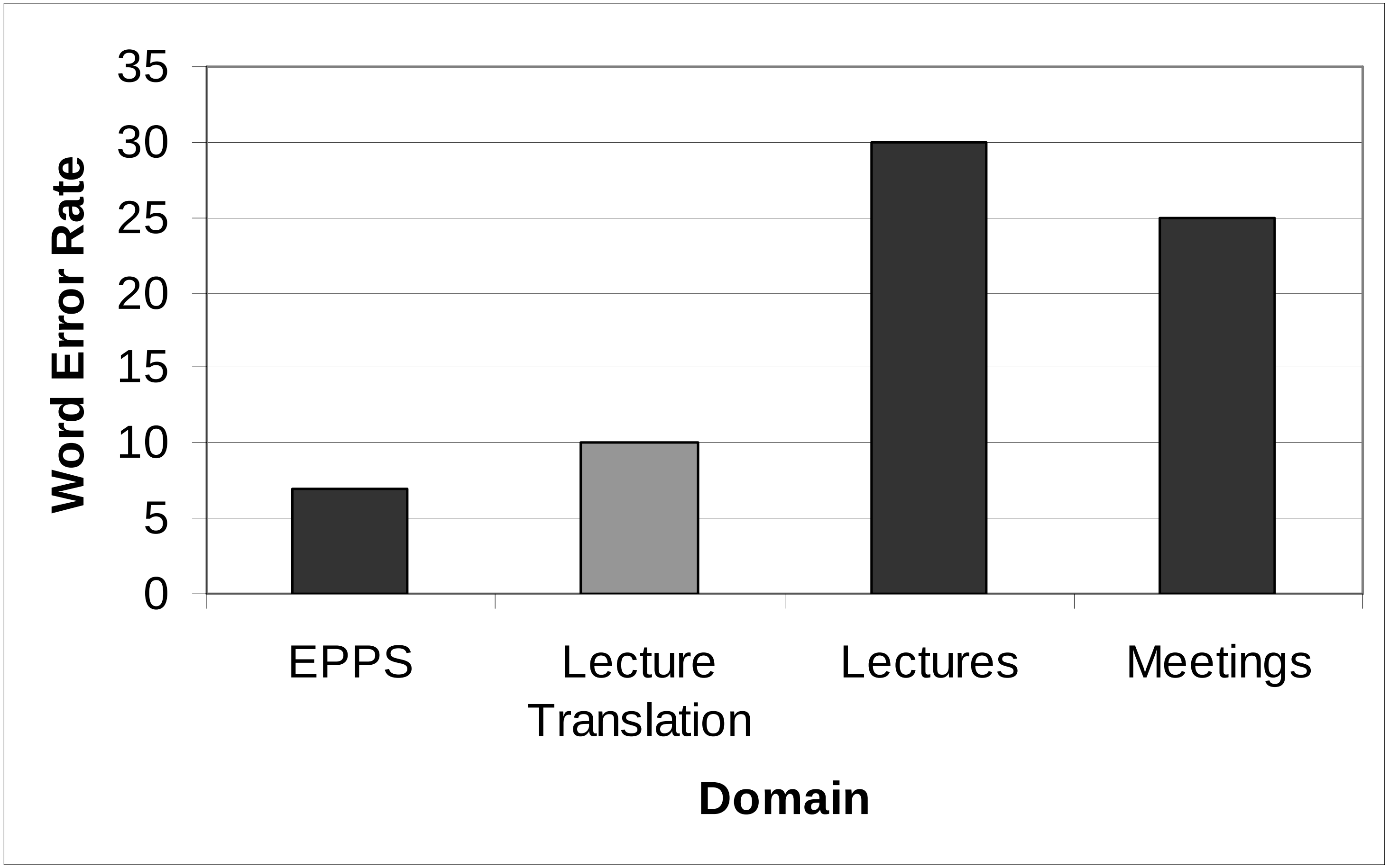
Formas de Entender la
Conciencia en Psicofisiologia
1. Percepción del mundo interno y externo
2. Conciencia como "darse cuenta"
3. Conciencia como "despertar"
4. Conciencia entendida como "atención focal" o "memoria a corto plazo"

interACT

Lecture Translator

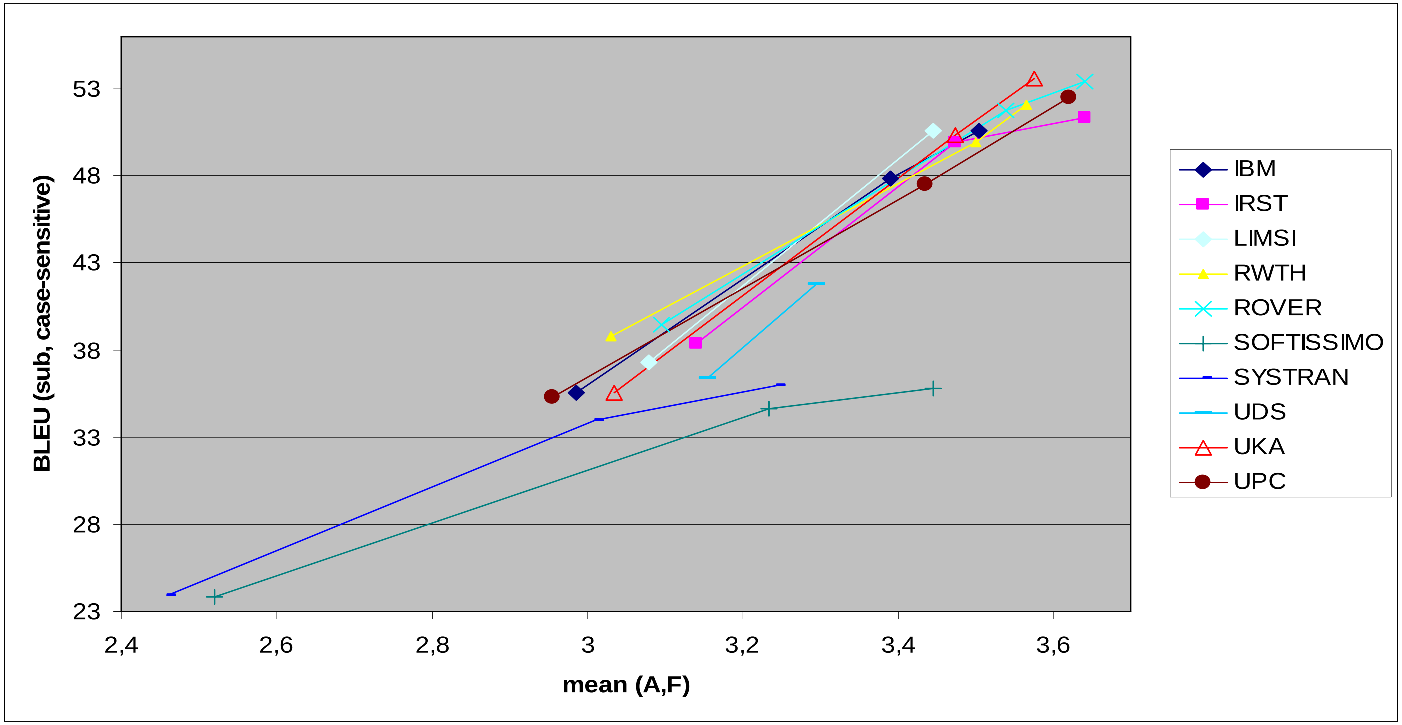
- Additional Technical Challenges:
  - Open Domain, Open Vocabulary, Open Speaking Style
  - Spontaneous Speech, Disfluencies, Ill-Formed Sentences
  - Suitable Chunking into Sentence Like Fragments for Translation
  - Specialty Topics, Dictionary, LM
  - Real-Time Requirement
- How it is Done:
  - Statistical Learning Algorithms
  - Adaptation: Voice, Specialty Dictionaries and LM’s from Speaker Info
  - Attention to Speed and Segmentation Issues

Speech Reco for Different Genres

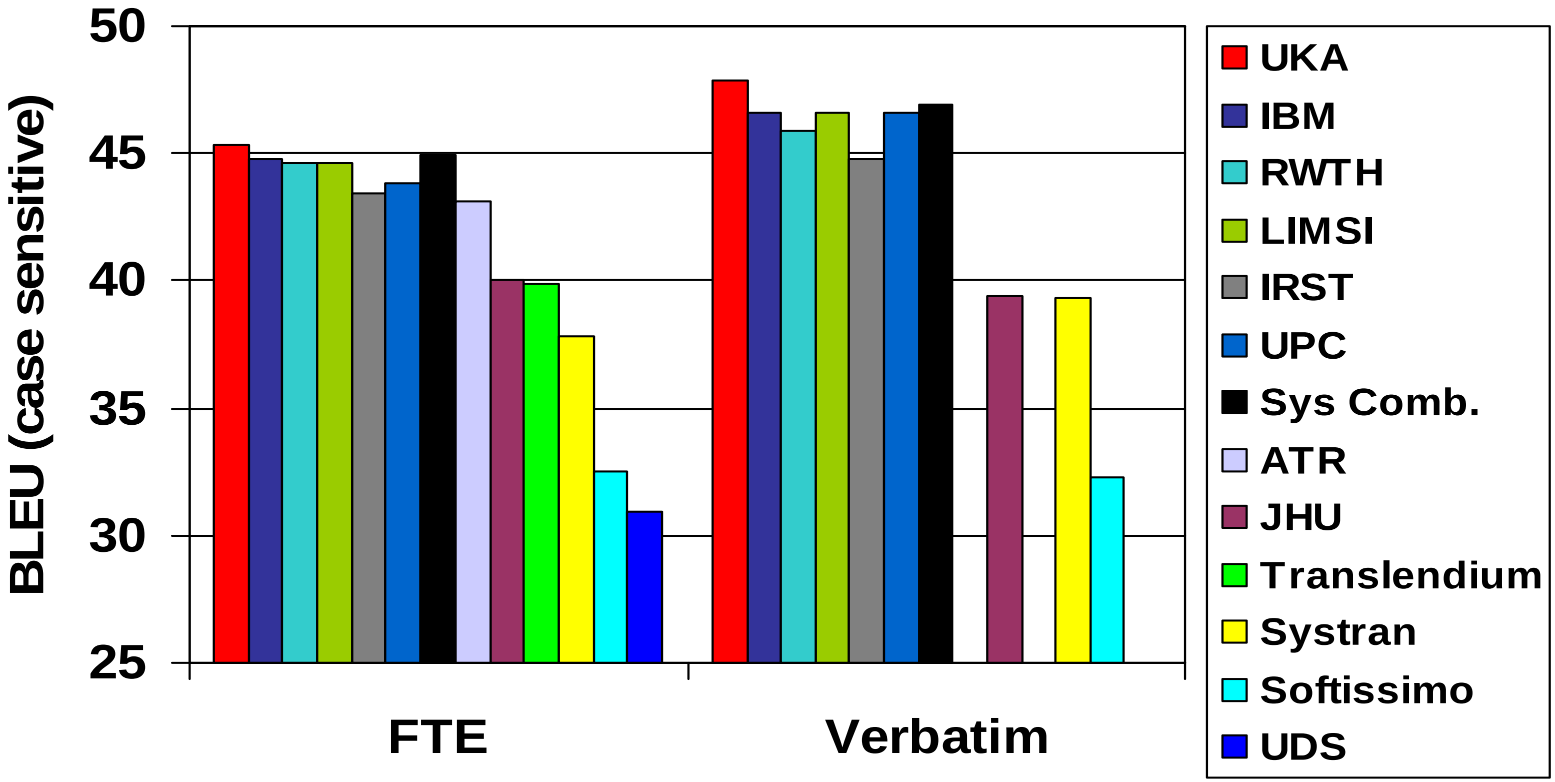


Human vs. Automatic Evaluation

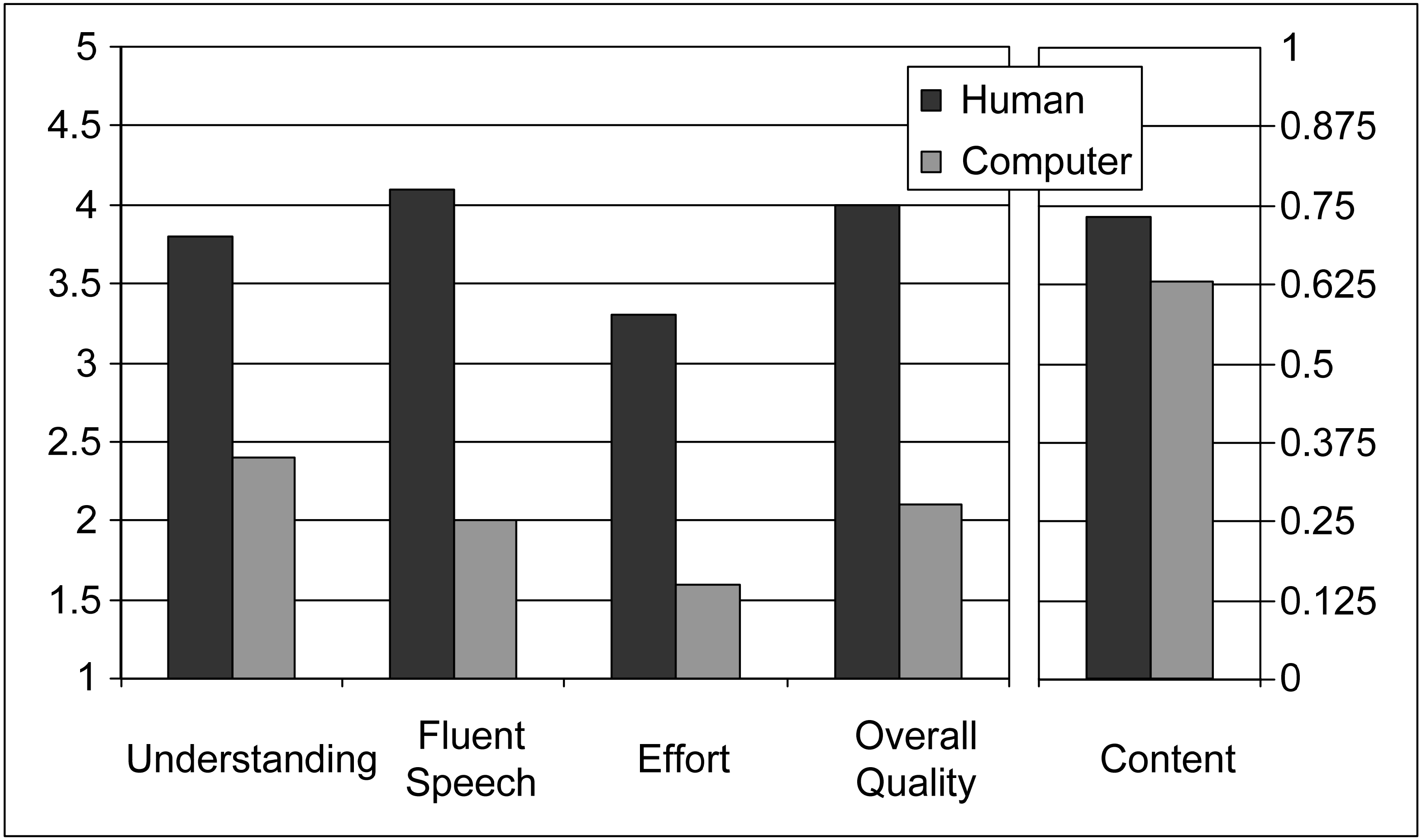
- TC-STAR SLT Eval ´07, English-Spanish
- Three data points: ASR, Verbatim, FTE task



TC-STAR SLT Eval ´07, Cortes Task



Human vs. Machine Performance



Missing Science

Problem 2: Translation Delivery

Has to be Appropriate for the Situation

- Should Allow for Fluent Communication
- Should Keep up with Input Speech
- Should Minimize Delay
- Should not Interfer with Human Tasks
- Should not Disturb Others
- Should Make Language Barrier Transparent

Integrating Emotional Cues into a Framework for Dialogue Management

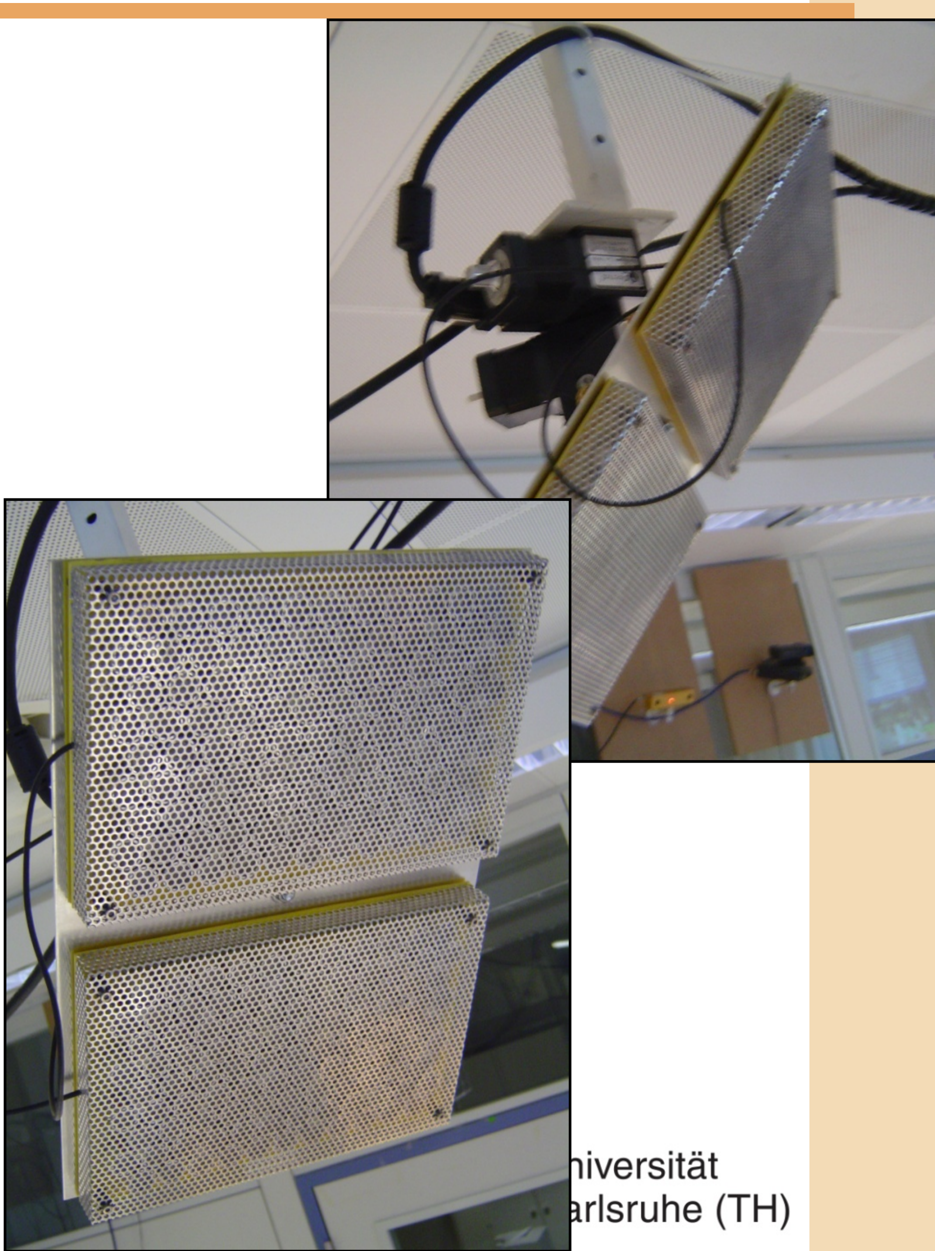
Delivery

Delivering Translation Output:

- Mobile Speech Translators
    - PDA's
    - In Vests or Clothing
- Hearing Personal Translations
    - Listen to Personal Simultaneous Translation Without Headsets and Without Disturbance
    - Targeted Audio Speakers
- Seeing Personal Translations
    - Reading Captions during Lecture
    - Heads-Up Display "Translation Goggles"
- Speaking in Foreign Languages
    - Producing Foreign Speech Without Knowing the Language
    - EMG Translation

Hearing Personal Translations

- Technology: Targeted Audio
  - Research under EC Project CHIL (Build Inobtrusive Computer Services)
  - Project Partner, Daimler-Chrysler
  - Array of Ultra-Sound Speakers
- Result: Narrow Sound Beam
  - Audible by one Individual Only
  - Others not Disturbed
  - Multiple Arrays Could Provide Multiple Languages
  - Steerable
  - Recognize/Track Individual Listener and Keep Language Beam on Target

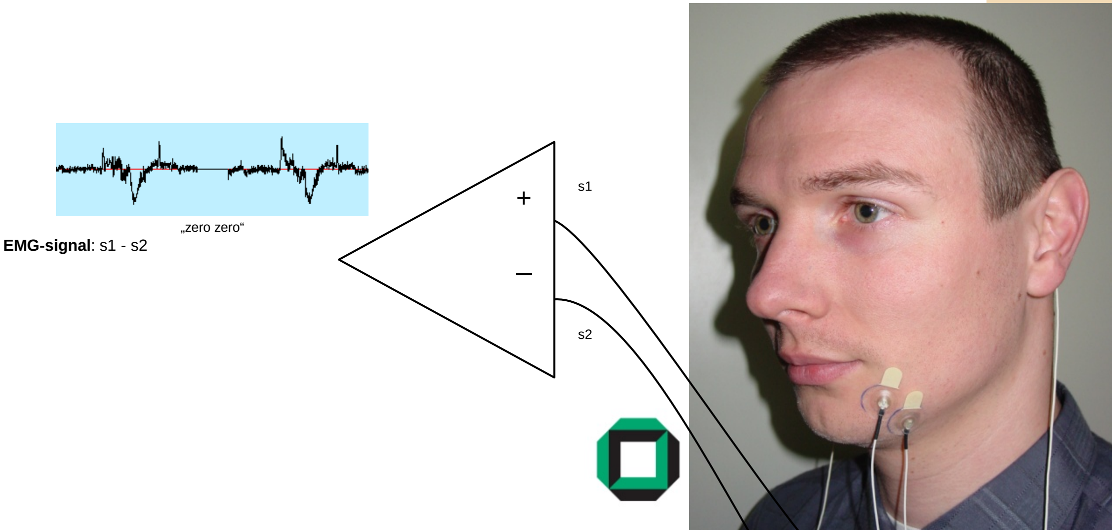


Seeing Personal Translations

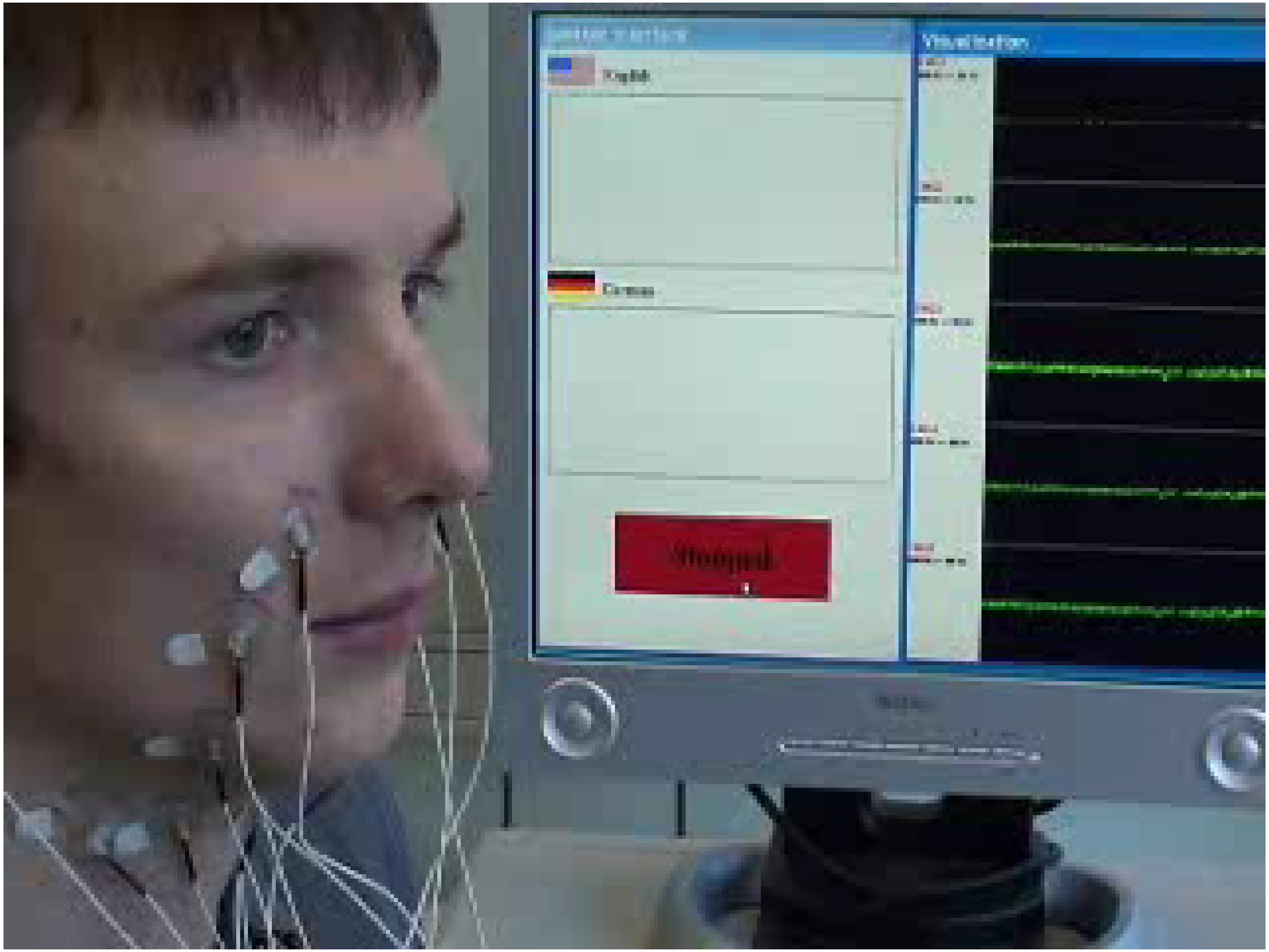
- Technology: Heads-up Display Goggles
  - Create Translation Goggles
  - Run Real-Time Simultaneous Translation of Speech
  - Text is Projected into Field of View of Listener
  - Translations are Seen as Text Captions Under Speaker
  - Output: Spanish, German,…

Speaking in Foreign Languages

- Technology: Silent Speech
  - Silently Motion Lips and Articulators in one Language (here: Chinese)
  - Capture Electrical Signals from Muscle Movement (Electromyography)
  - Recognition Engine Trained with EMG signals
  - Spoken Phrases are Recognized as Words and Translated
  - Synthetic Speech in Any Language and Any Voice is Produced
- First Prototype
  - Limited Set of Phrases, Positioning of Electrodes
  - Ongoing Work:
    - Robustness,
    - Large Vocabulary
    - Language Implants??



EMG Translator



CNN – Speech Translation in the News

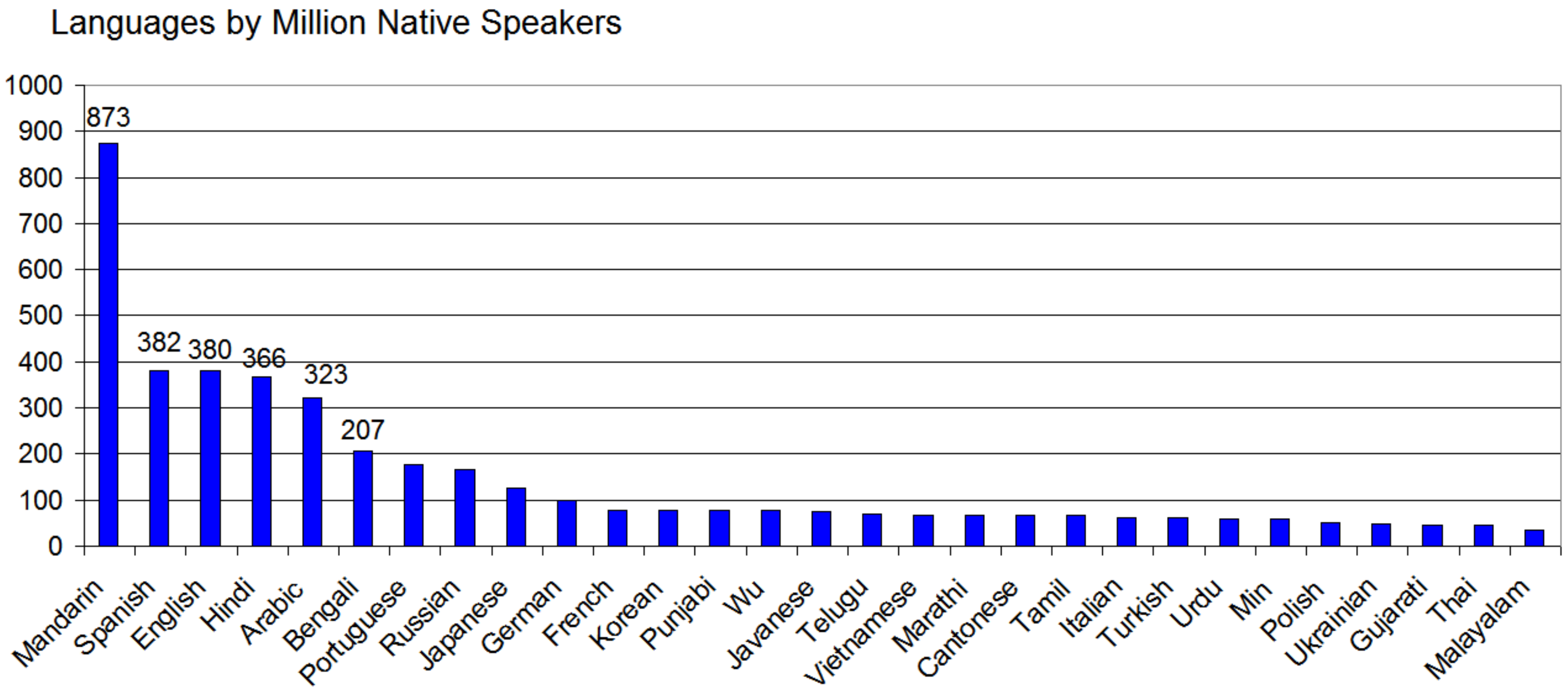
Discovery Channel – S2S in the News

Carnegie Mellon

Universität Karlsruhe (TH)

Discovery – S2S in the News

The Long Tail of Language – Portability



Reaching Out to a Larger World

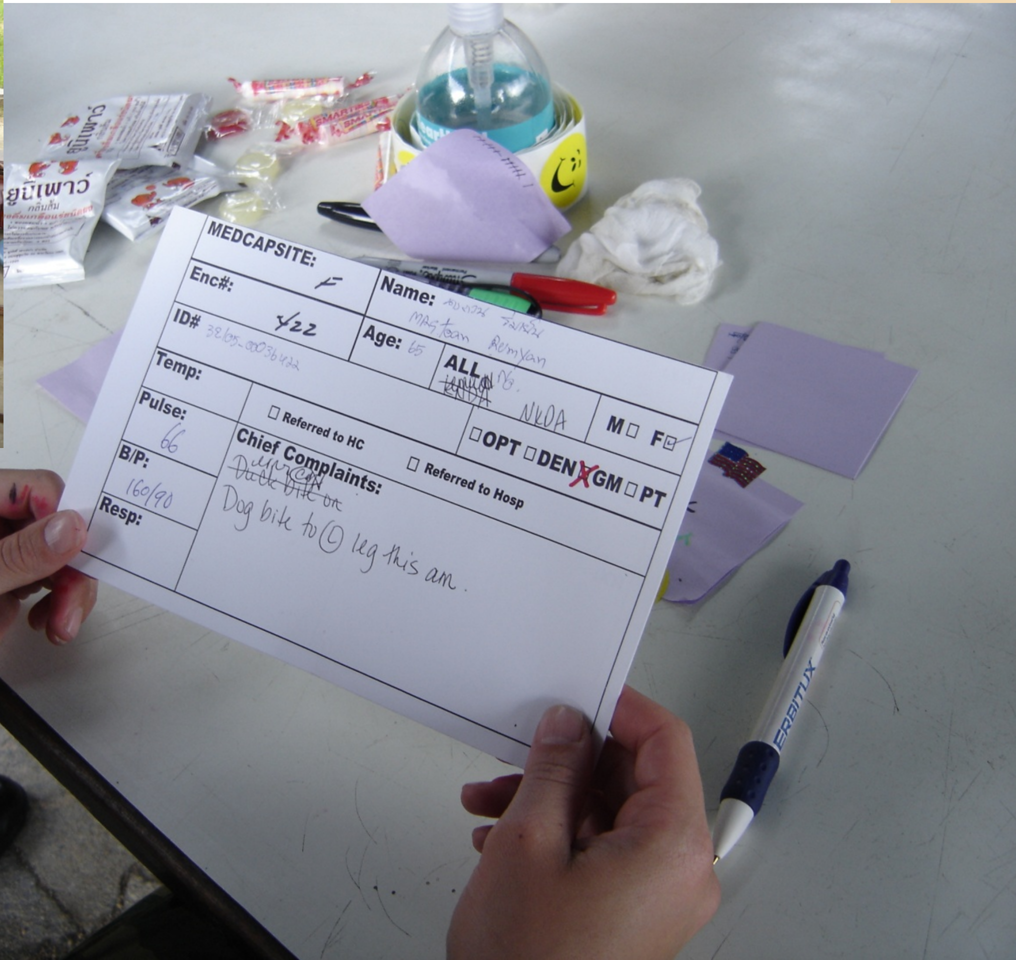
KL
og

Cobra Gold

interACT

DX THAI:
RX THAI:
RX:

Communication



MEDCAPSITE:

Enc#: F

ID# 422 32/05-00036422

Name: [illegible] MRS.Toan Remyan

Age: 65

ALL ~~[illegible]~~ NG. NKDA

M □ F ☑

Temp:

Pulse: 66

B/P: 160/90

Resp:

□ Referred to HC

□ Referred to Hosp

□ OPT □ DEN ☒ GM □ PT

Chief Complaints: ~~Duck bite on~~ Dog bite to Ⓛ leg this am.

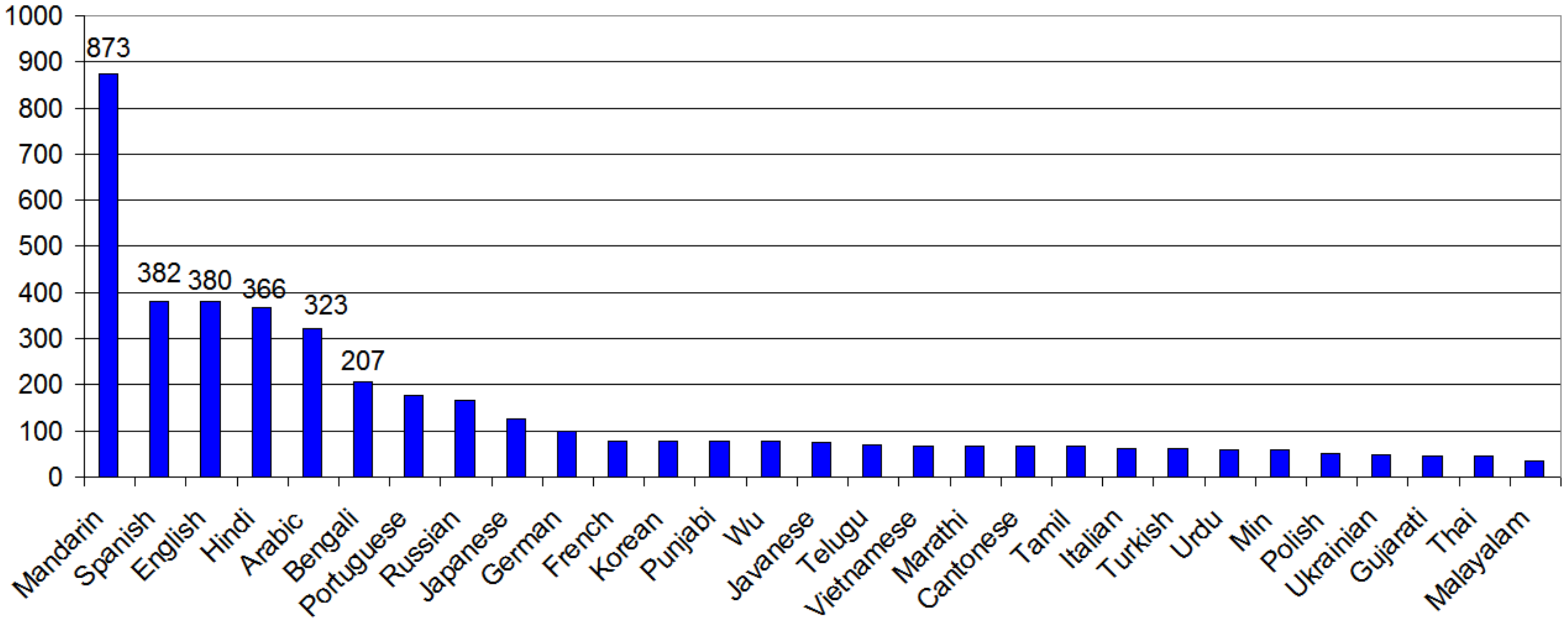
interACT

Communication by Machine

The Long Tail of Language – Portability

-

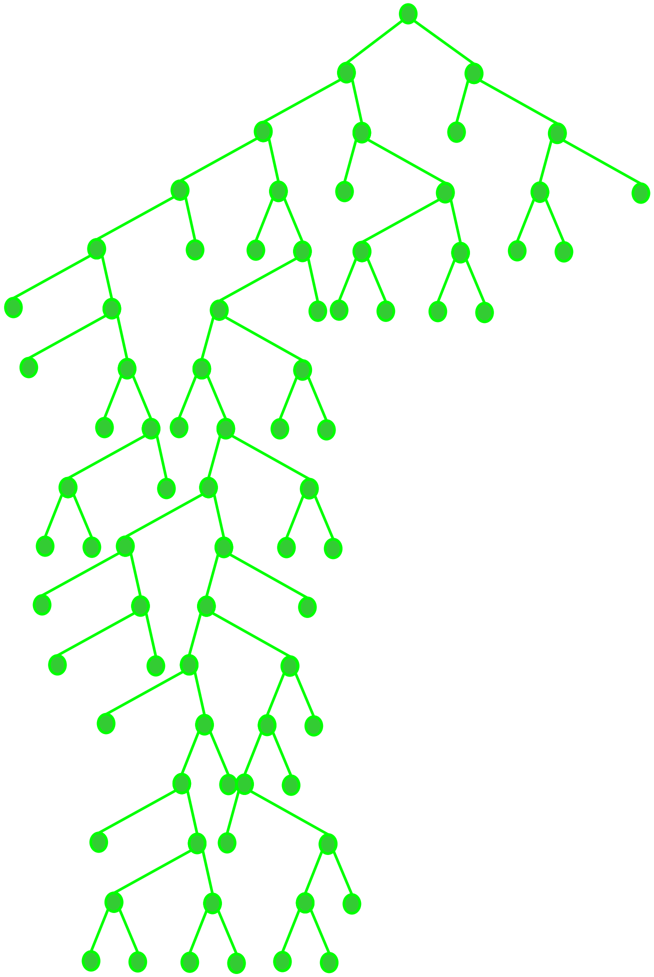
Languages by Million Native Speakers



How to Achieve it?

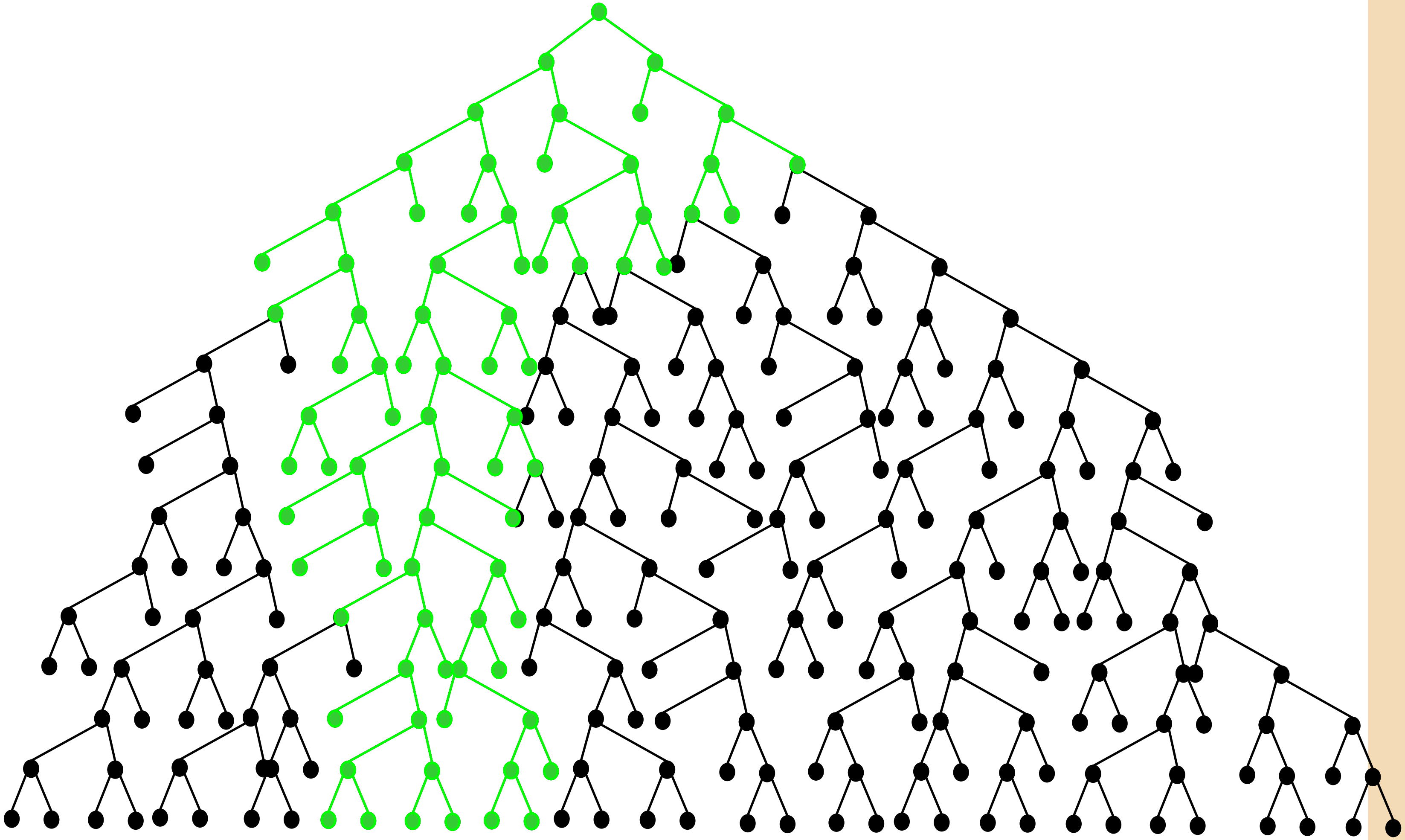
- Do Massive Data Collection Effort
- Make Process Cheaper
- Make Modules Language Independent/Adaptive
- Use Interlingua or Pivot Languages
- Improve Performance with Less Data
- Select Data more Carefully
- Acquire Data Interactively
  - Without people knowing ??

interACT Language Independent/Adaptive Acoustics



English Polyphone Tree

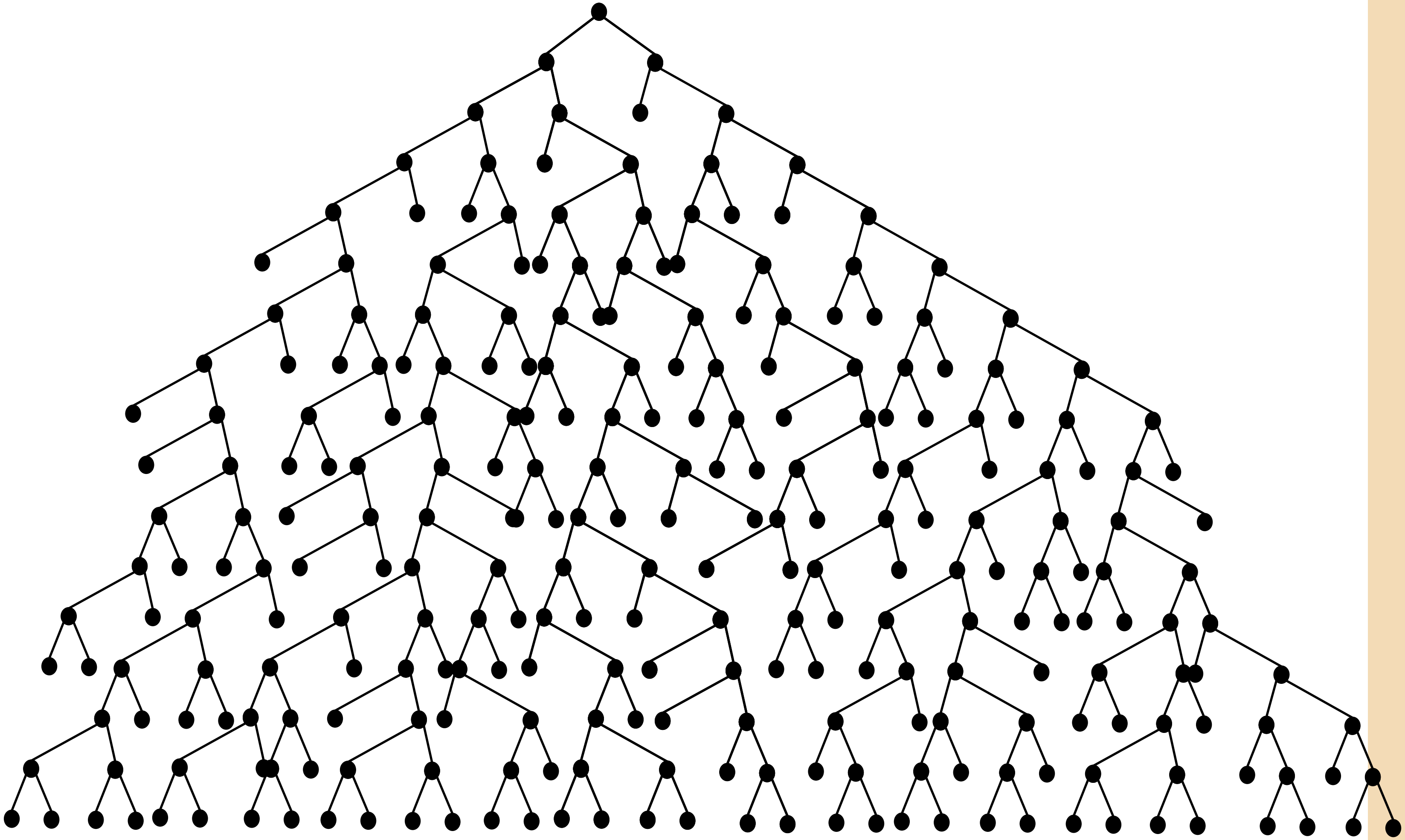
Language Independent/Adaptive Acoustics



English

Other languages

Language Independent/Adaptive Acoustics

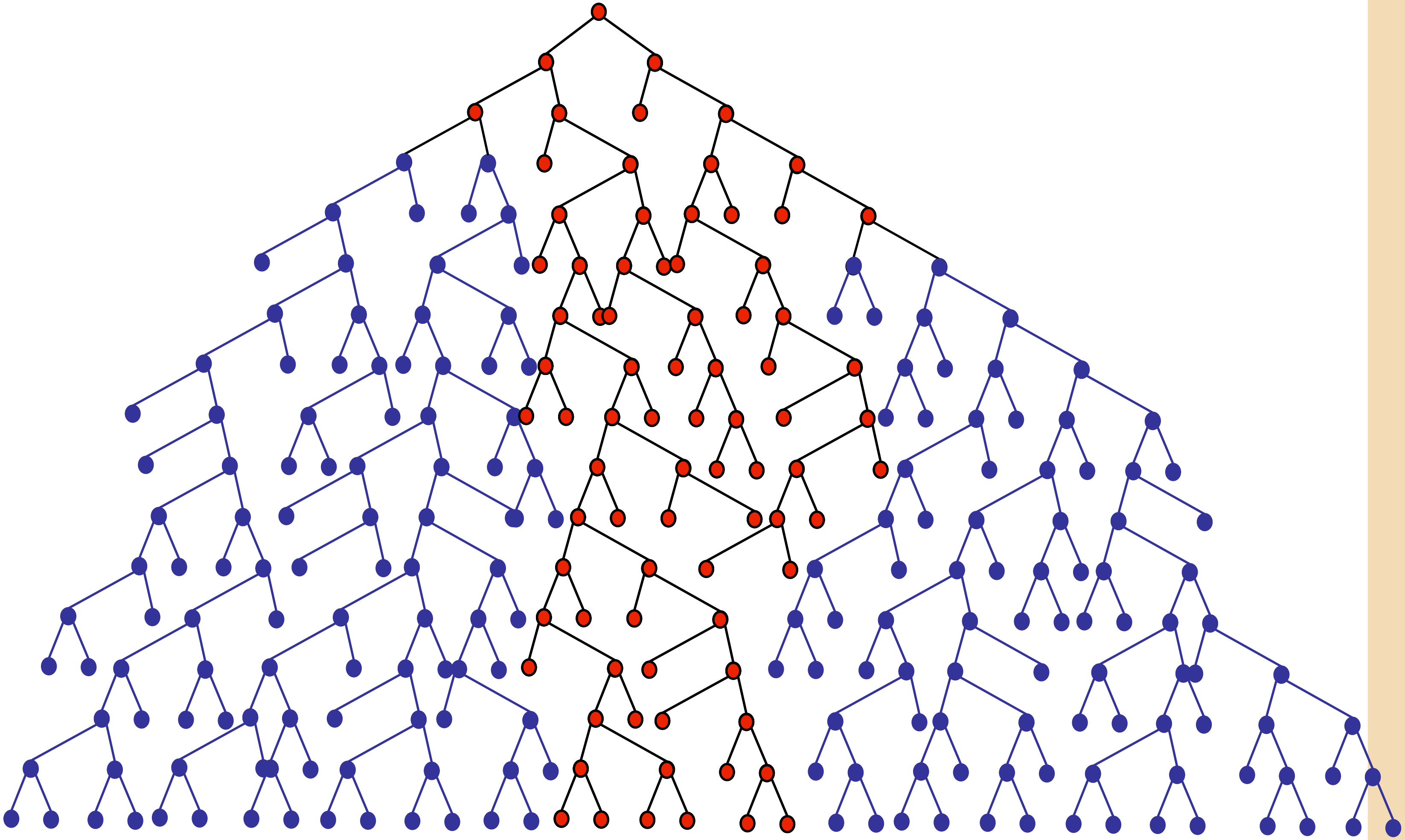


Multilingual Polyphone Tree

Carnegie Mellon

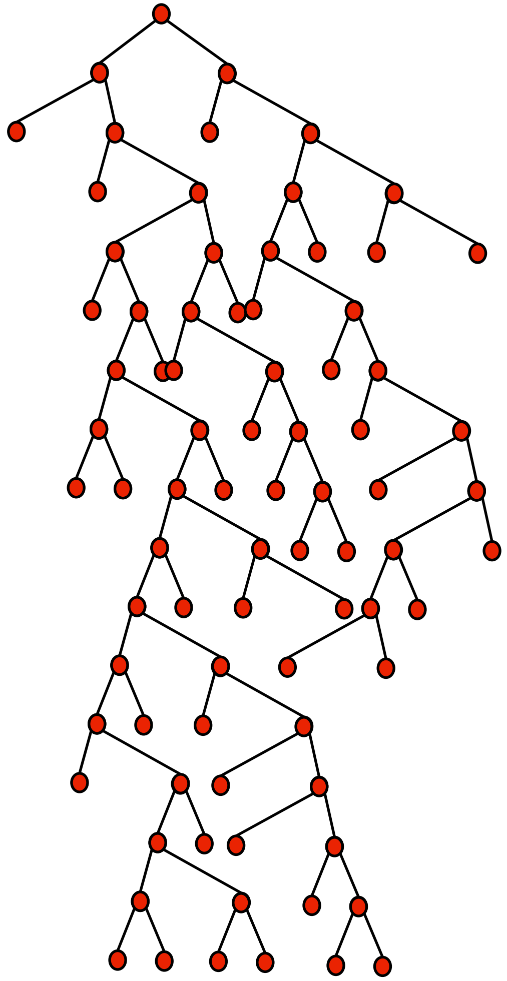
Universität
Karlsruhe (TH)

Language Independent/Adaptive Acoustics



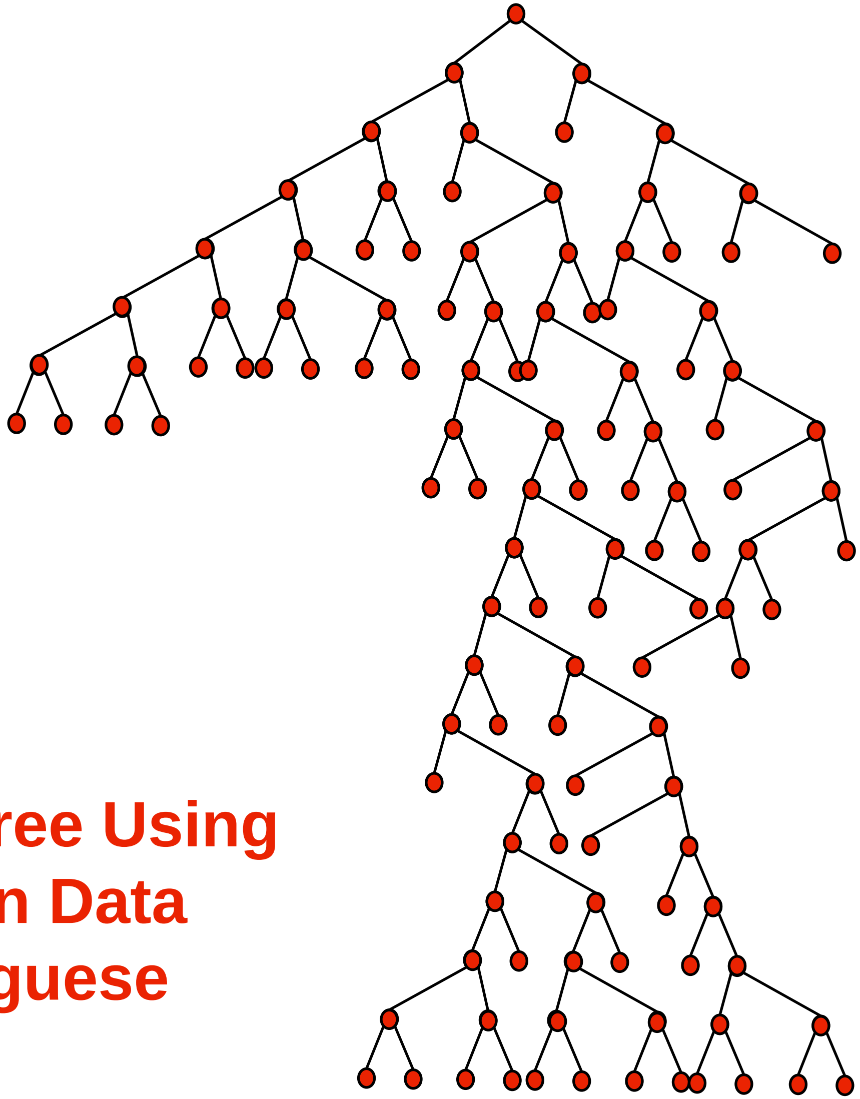
Polyphones found in Portuguese

Language Independent/Adaptive Acoustics



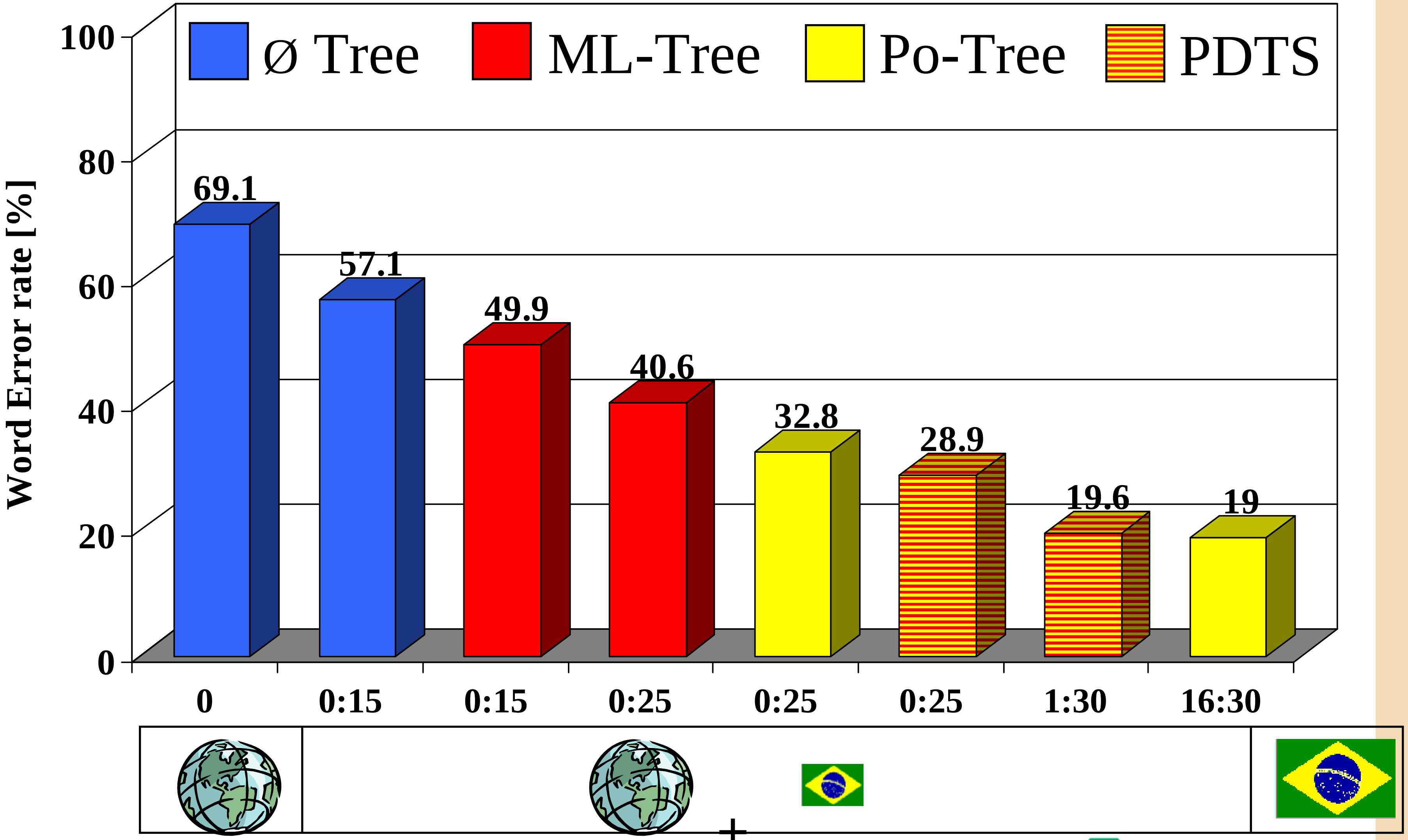
Pruning the Tree to Portuguese

Language Independent/Adaptive Acoustics



Regrow Tree Using Adaptation Data for Portuguese

Language Adaptation Experiments



On-line Grammar Acquisition

> do i have any mail

I understand "do i have any mail"

> arrange by recency

I don't understand right away what you mean but let me guess...

"arrange by recency" is a way to express:

1. count mail, e.g. "count"
2. list mail, e.g. "list"
3. sort mail, e.g. "sort"
0. None of the above

> sort

"recency" is a way to express:

1. sort by size, e.g. "size"
2. sort by date, e.g. "date"
3. sort by sender, e.g. "sender"
0. None of the above

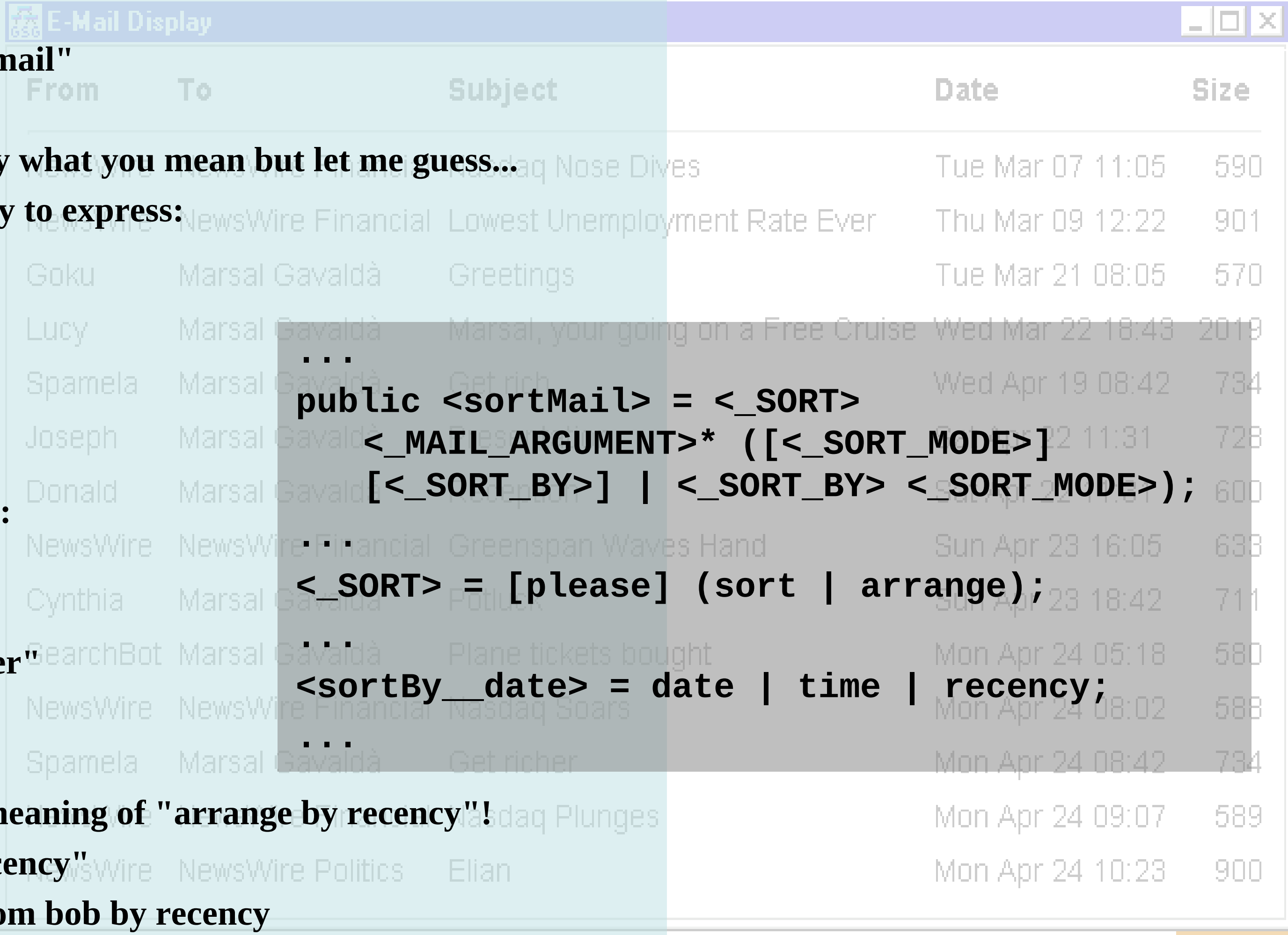
> by date

Thanks for teaching me the meaning of "arrange by recency"!

I understand "arrange by recency"

> please arrange messages from bob by recency

I understand "please arrange messages from bob by recency"



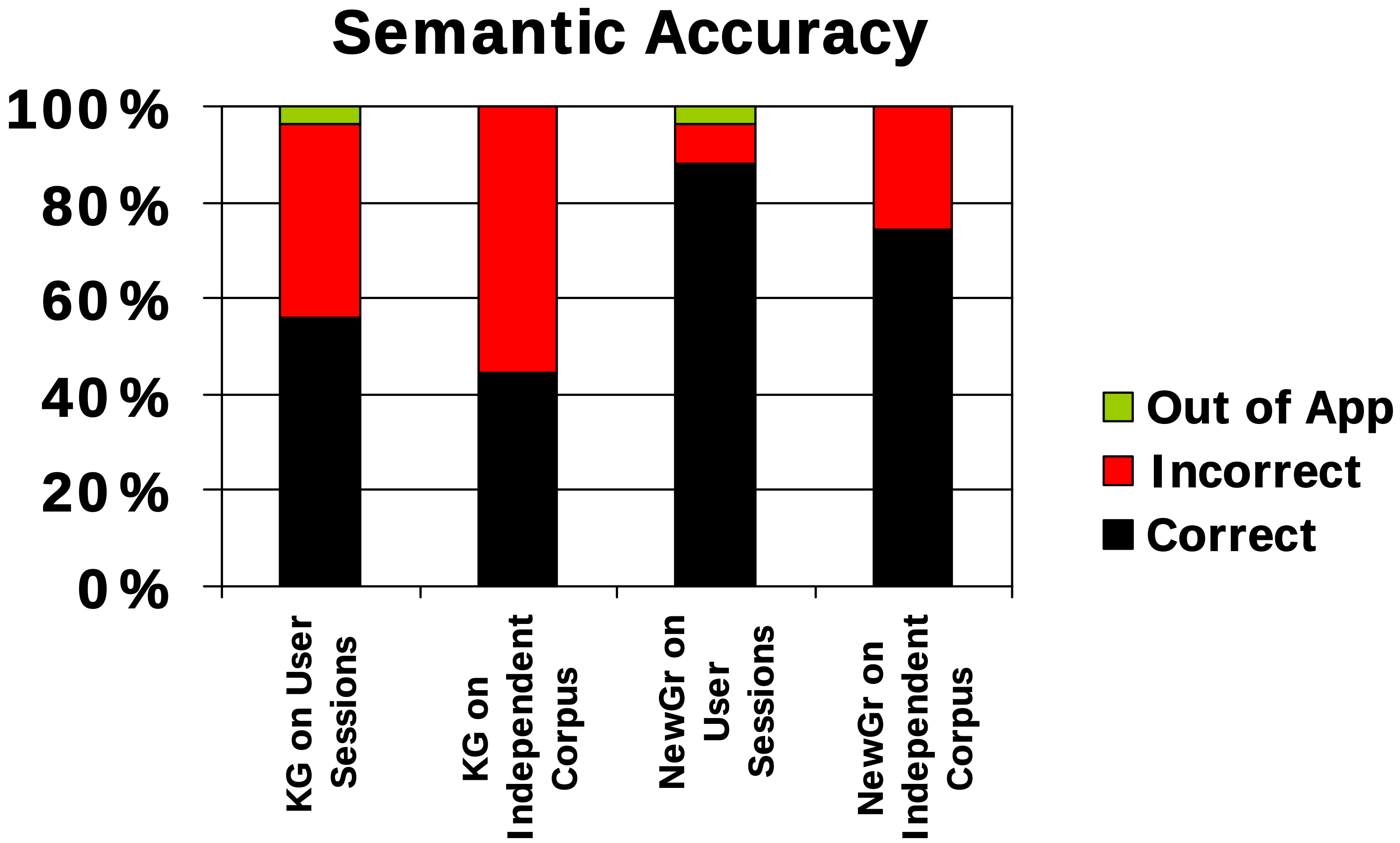
From	To	Subject	Date	Size
NewsWire	NewsWire Financial	Nasdaq Nose Dives	Tue Mar 07 11:05	590
NewsWire	NewsWire Financial	Lowest Unemployment Rate Ever	Thu Mar 09 12:22	901
Goku	Marsal Gavaldà	Greetings	Tue Mar 21 08:05	570
Lucy	Marsal Gavaldà	Marsal, your going on a Free Cruise	Wed Mar 22 18:43	2019
Spamela	Marsal Gavaldà	Get rich	Wed Apr 19 08:42	734
Joseph	Marsal Gavaldà	Presentation	Sat Apr 22 11:31	728
Donald	Marsal Gavaldà	Reception	Sat Apr 22 11:01	600
NewsWire	NewsWire Financial	Greenspan Waves Hand	Sun Apr 23 16:05	633
Cynthia	Marsal Gavaldà	Potluck	Sun Apr 23 18:42	711
SearchBot	Marsal Gavaldà	Plane tickets bought	Mon Apr 24 05:18	580
NewsWire	NewsWire Financial	Nasdaq Soars	Mon Apr 24 08:02	588
Spamela	Marsal Gavaldà	Get richer	Mon Apr 24 08:42	734
NewsWire	NewsWire Financial	Nasdaq Plunges	Mon Apr 24 09:07	589
NewsWire	NewsWire Politics	Elian	Mon Apr 24 10:23	900

```
...
public <sortMail> = <_SORT>
    <_MAIL_ARGUMENT>* ([<_SORT_MODE>]
    [<_SORT_BY>] | <_SORT_BY> <_SORT_MODE>);
...
<_SORT> = [please] (sort | arrange);
...
<sortBy__date> = date | time | recency;
...
```

Carnegie Mellon

Universität Karlsruhe (TH)

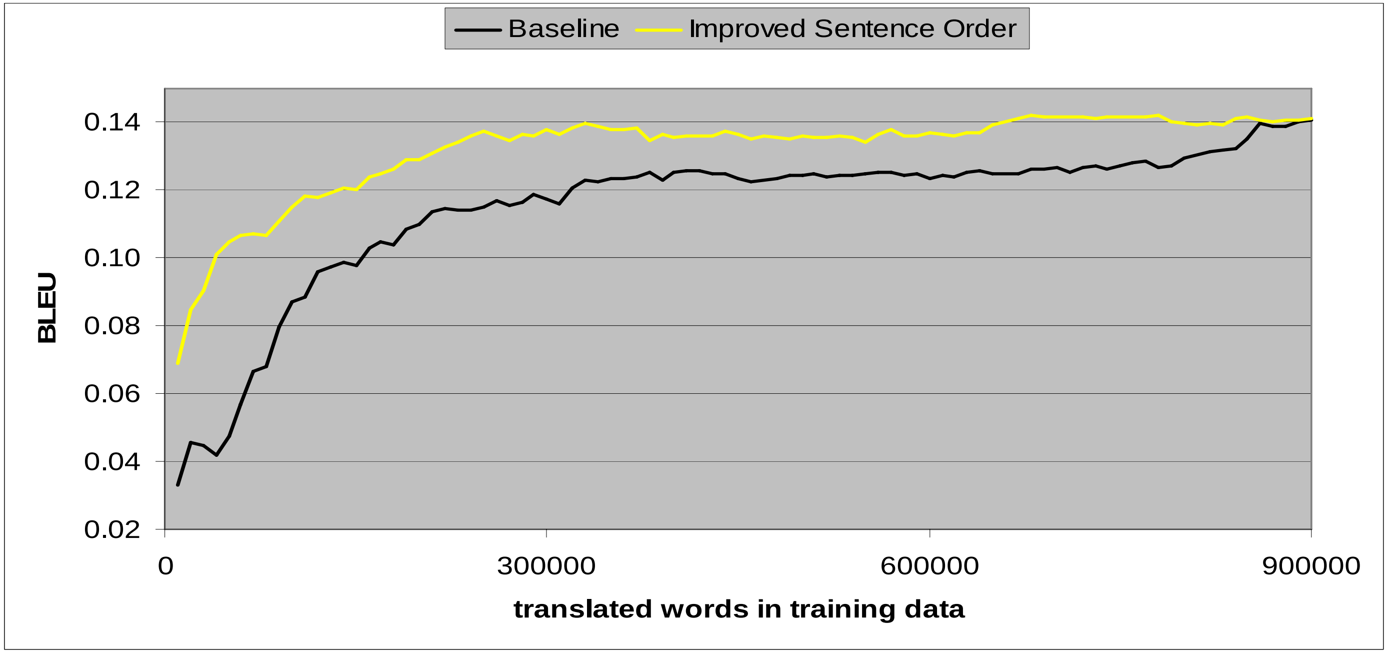
Results: Interactive Grammar Learning



Redundancies...

130: I'd like to make a hotel reservation.
131: Do you have a room for tonight?
132: How long do we stay here?
133: I'd like a shave, please.
134: I'd like a haircut.
...
173: Another one, please.
174: May I have another glass of water?
175: May I have another fork?
176: I'll show you to your room.
...
227: Overseas operator, please.
228: This is Mr. Sato in room one two three four.
229: I'd like to call Tokyo, Japan.
230: Miki Hayakawa.
231: Operator, please.

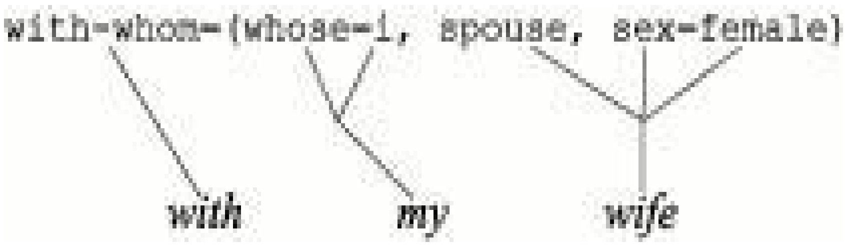
Make Do with Less Data



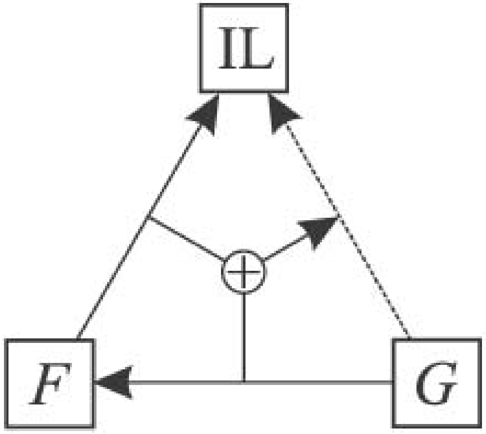
- If Parallel Corpus has to be Developed
  - Choose English Seed Sentences Opportunistically
- Sentences sorted according to:
  - Frequency of unseen uni-, bi and trigrams per sentence length

Statistical Interlingua MT

- Interlingua is a Language, too!
  - But:
    - Order Invariant
    - Tree Structured



- Reformulate Statistical Translation
- Train SIMT
  - Tagged Corpus
- ‘Grammar’ Projection to *New* Language
- *(Refence: Kauers et al., ICSLP’02)*

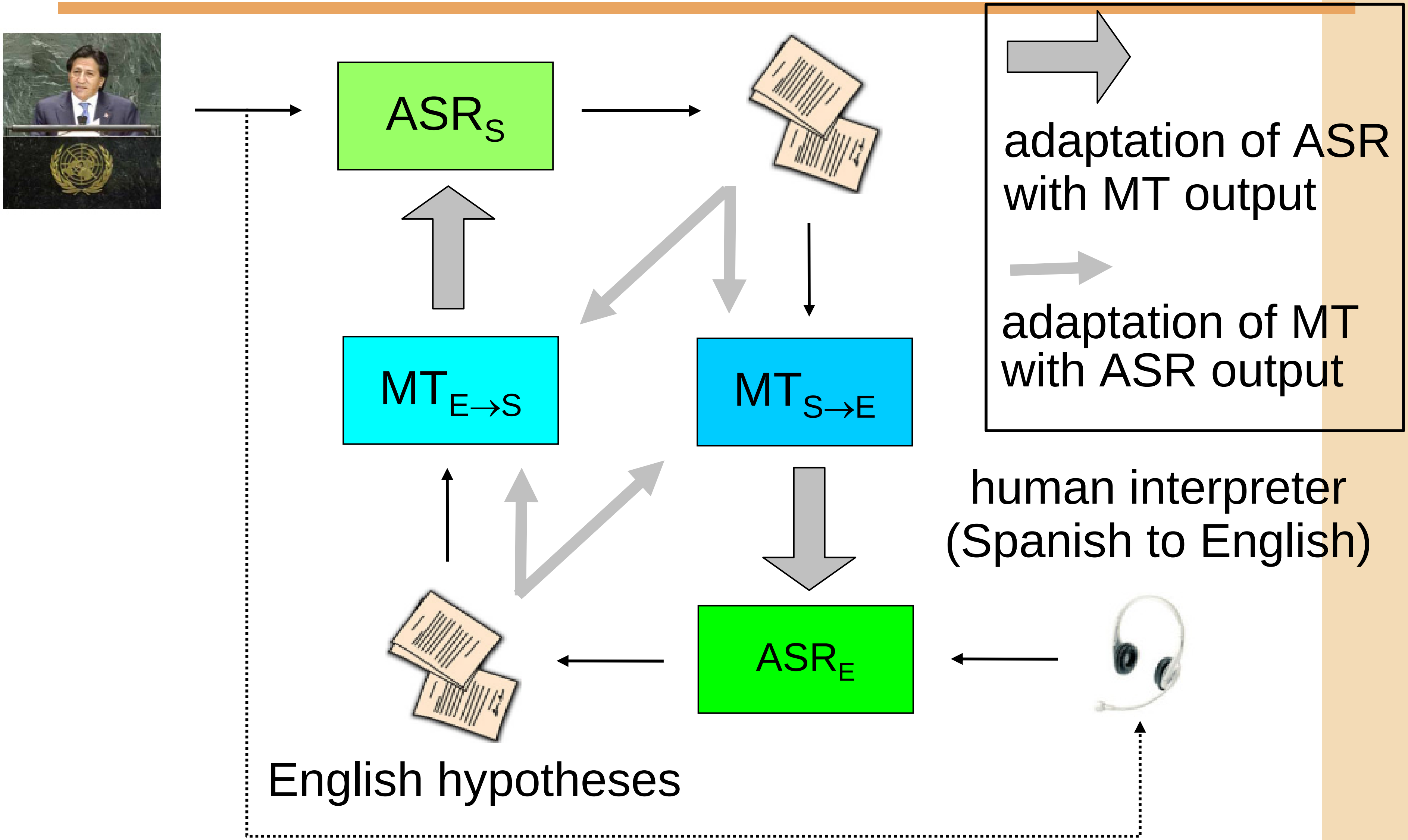


Exploiting Experts in the Field

- Is it possible to Train Speech Translators from Recording Simultaneous Translators?
  - …skip text altogether
  - Cheaper to do than transcription
  - Best for really low density languages
- First Results:
  - Existing Speech Translators Improve from parallel speech corpora, ASR and MT Modules adapted iteratively

interACT

Spanish hypotheses



ASR_S

$MT_{E \rightarrow S}$

$MT_{S \rightarrow E}$

ASR_E

adaptation of ASR with MT output

adaptation of MT with ASR output

human interpreter (Spanish to English)

English hypotheses

Carnegie Mellon

Universität
Karlsruhe (TH)

Conclusion

- Multimodal Human-Human Communication
  - New Class of Computer Interaction
  - Supported by Multimodal Perceptual User Interfaces
- Grand Challenge Problem
  - Crossing the Language Divide Anywhere, Anytime
  - Handling the Long Tail of Language

Questions