

# Verbmobil: The Development and Integration of a Large Speech-to-Speech Translation System

Norbert Reithinger  
DFKI GmbH  
Stuhlsatzenhausweg 3  
D-66123 Saarbrücken  
bert@dfki.de

Andreas Klüter  
DFKI GmbH  
Erwin-Schrödinger Straße Geb. 57  
D-67663 Kaiserslautern  
klueter@dfki.de

<http://verbmobil.dfki.de>



- **Overview of Verbmobil**
- **A walk through the live system**
  - Acoustic Processing
  - Dialog Translation
  - Selection and Speech Synthesis
- **Technical issues**
- **Human Factors and Experiences**

# Overview of Verbmobil

## Challenges, Partners, and General Approaches



# What is Verbmobil?

Verbmobil

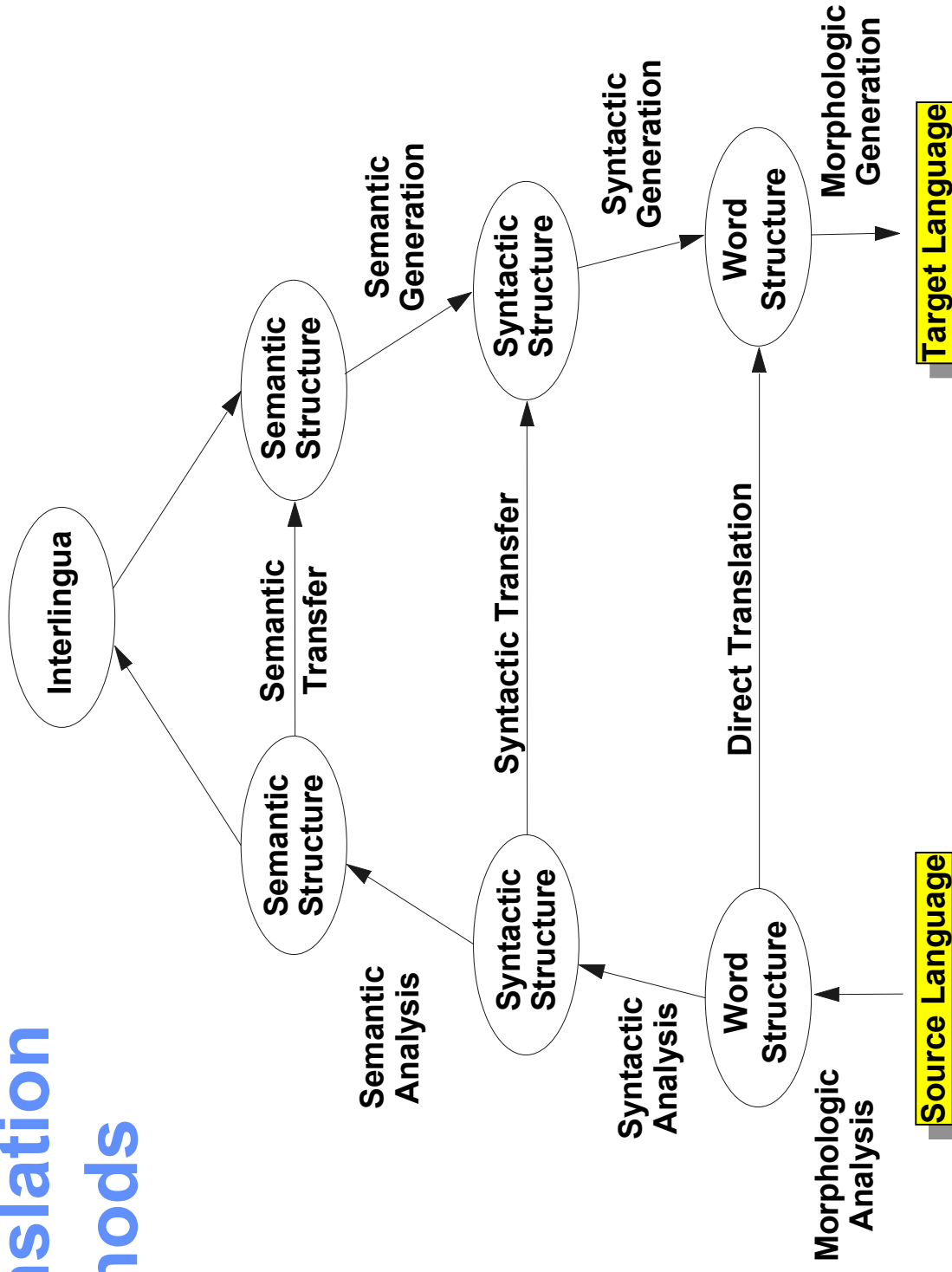
- **Speech-to-speech translation system**
- **Robust processing of spontaneous dialogs**
- **Speaker independent (adaptive)**
- **Languages: English, German, Japanese**
- **Domains: Appointment scheduling, travel planning and hotel reservation, remote PC maintenance**
- **The system **mediates** between two humans, it does not play an active role**
- **There is no control of the ongoing dialog by the system**

**Increasing Complexity** 

<b>Input Conditions</b>	<b>Naturalness</b>	<b>Adaptability</b>	<b>Dialog Capabilities</b>
Close-Speaking Microphone/Headset Push-to-talk	Isolated Words	Speaker Dependent	Monolog Dictation
Telephone, Pause-based Segmentation	Read Continuous Speech	Speaker Independent	Information- seeking Dialog
<b>Open Microphone, GSM Quality</b>	<b>Spontaneous Speech</b>	<b>Speaker Adaptive</b>	<b>Multiparty Negotiation</b>

# Verbmobil

# Classification of Machine Translation Methods





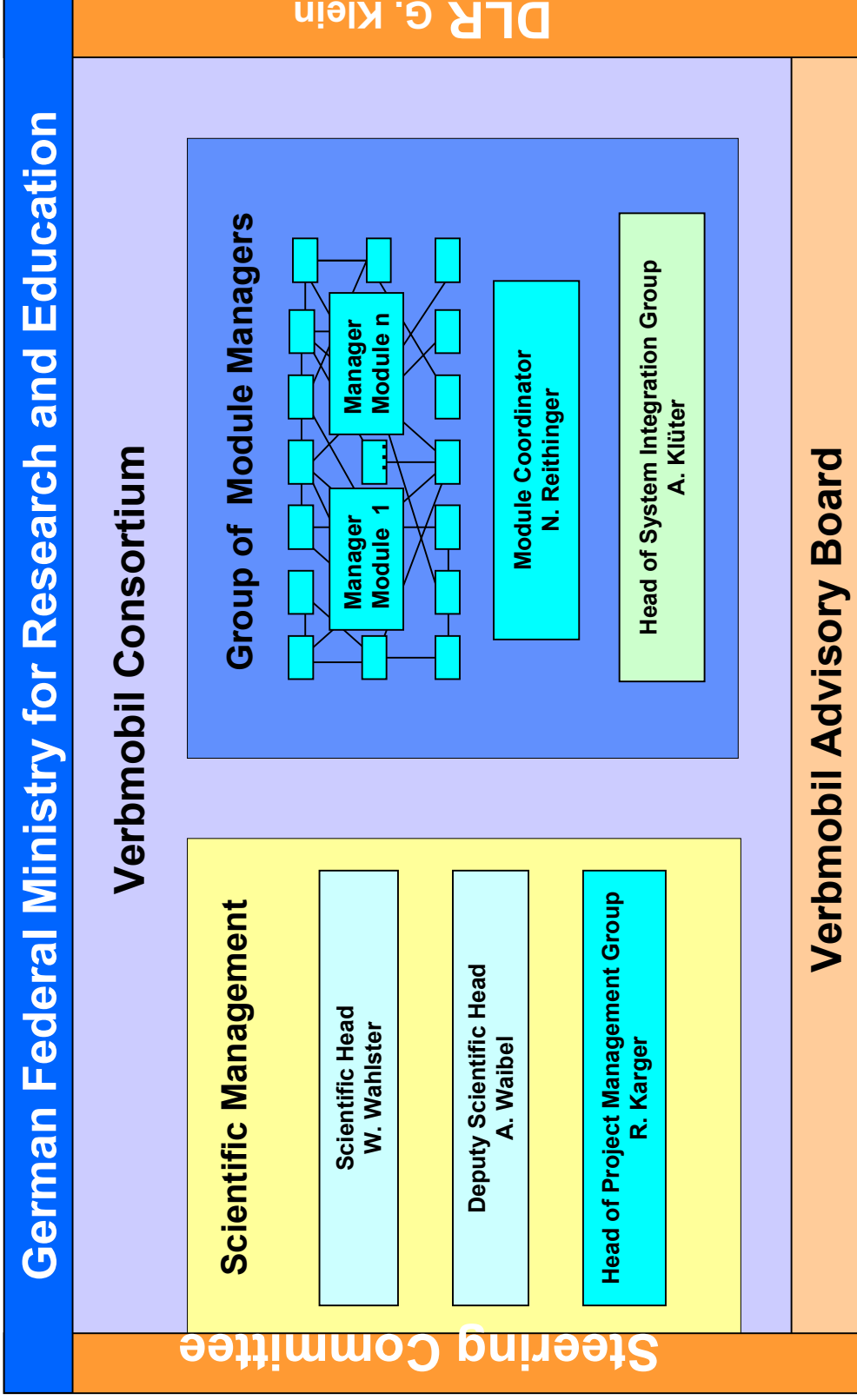


# Facts About the Project

- **23 participating institutions (in Verbmobil II), from Germany and the USA**
- **Over 900 full-time employees and students involved over the whole duration**
- **Funded by the German Ministry for Education and Science and the participating companies:**

BMBF-Funding Phase I, 1.01.93 – 31.12.96	62.7 Mio. DM	31.6 Mio €
BMBF-Funding Phase II, 1.01.97 - 30.9.2000	53.3 Mio. DM	27 Mio €
Industrial investment I+II	32.6 Mio. DM	16.5 Mio €
Related industrial R & D activities	ca. 20 Mio. DM	ca. 10 Mio €
<b>Total</b>	<b>168.6 Mio. DM</b>	<b>85.1 Mio €</b>

# Project Organization



# Verbmobil – The Book

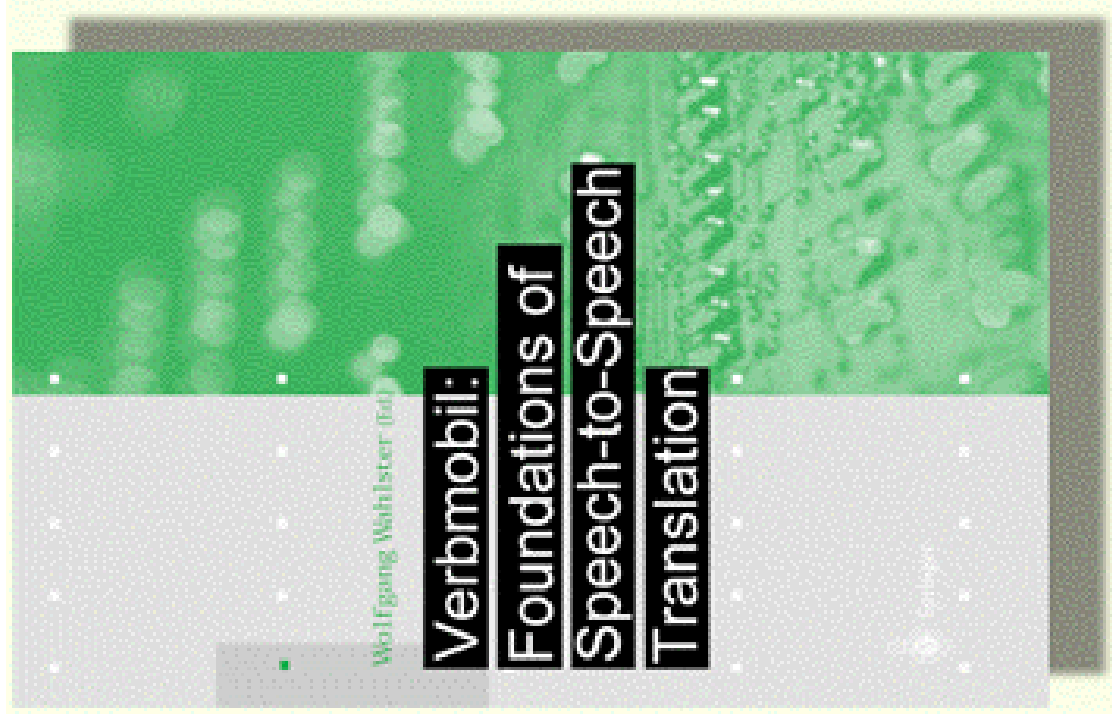
There are over 600 refereed papers on the various aspects of and achievements in Verbmobil.

**Wolfgang Wahlster (ed.):**

***"Verbmobil: Foundations of Speech-to-Speech Translation"***

**Springer-Verlag Berlin Heidelberg  
New York. 679 Pages**

**ISBN 3-540-67783-6**





# Verbmobil Hardware for This Tutorial

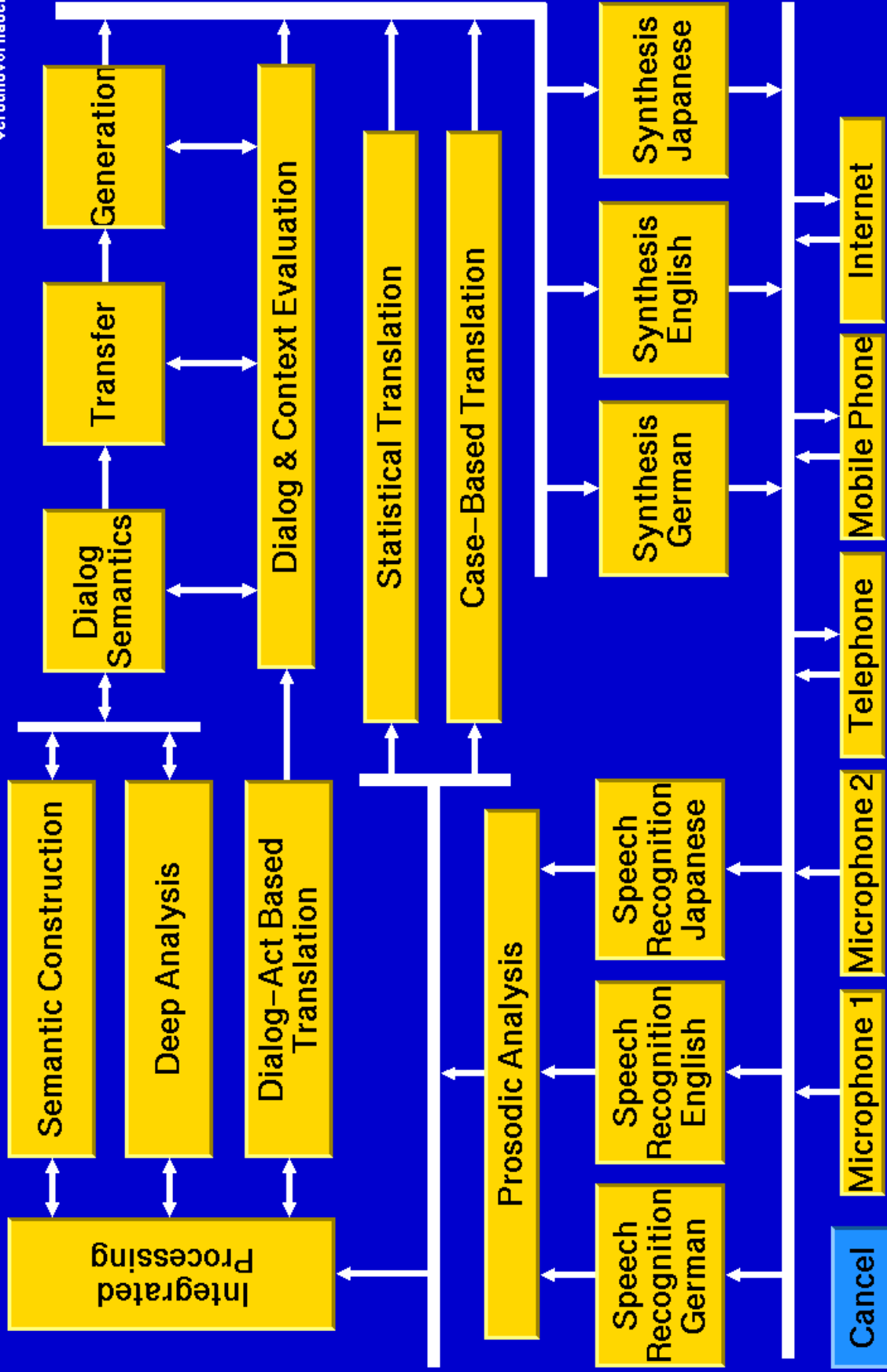
- SUN Ultra-Sparc 80
- 4 processors (450 MHz)
- 2 GB main memory
- 8 GB swap
- **no** special signal processing hardware
- Desklab Gradient A/D converter or Sun internal audio device
- close-speaking cordless microphones



bmb+f

Verbmobil

Verbundvorhaben



# Walk Through the Verbmobil System

## Detailed Module Presentation and Demonstration



# Acoustic Processing



# Recording, Synthesizing and Synchronization

- **Task:**  
Providing a uniform interface to varying audio hardware; synchronizing in- and output
- **Input:**  
Audio data and system states
- **Method:**  
Introducing audio modules; Finite State Machine for synchronizing
- **Result:**  
Audio Data and Synchronization
- **Benefit:**  
Encapsulating audio hardware, “open microphone”, preventing out-of-sync or overlapping system output
- **Responsible:**  
DFKI, Kaiserslautern

# Audio Configuration

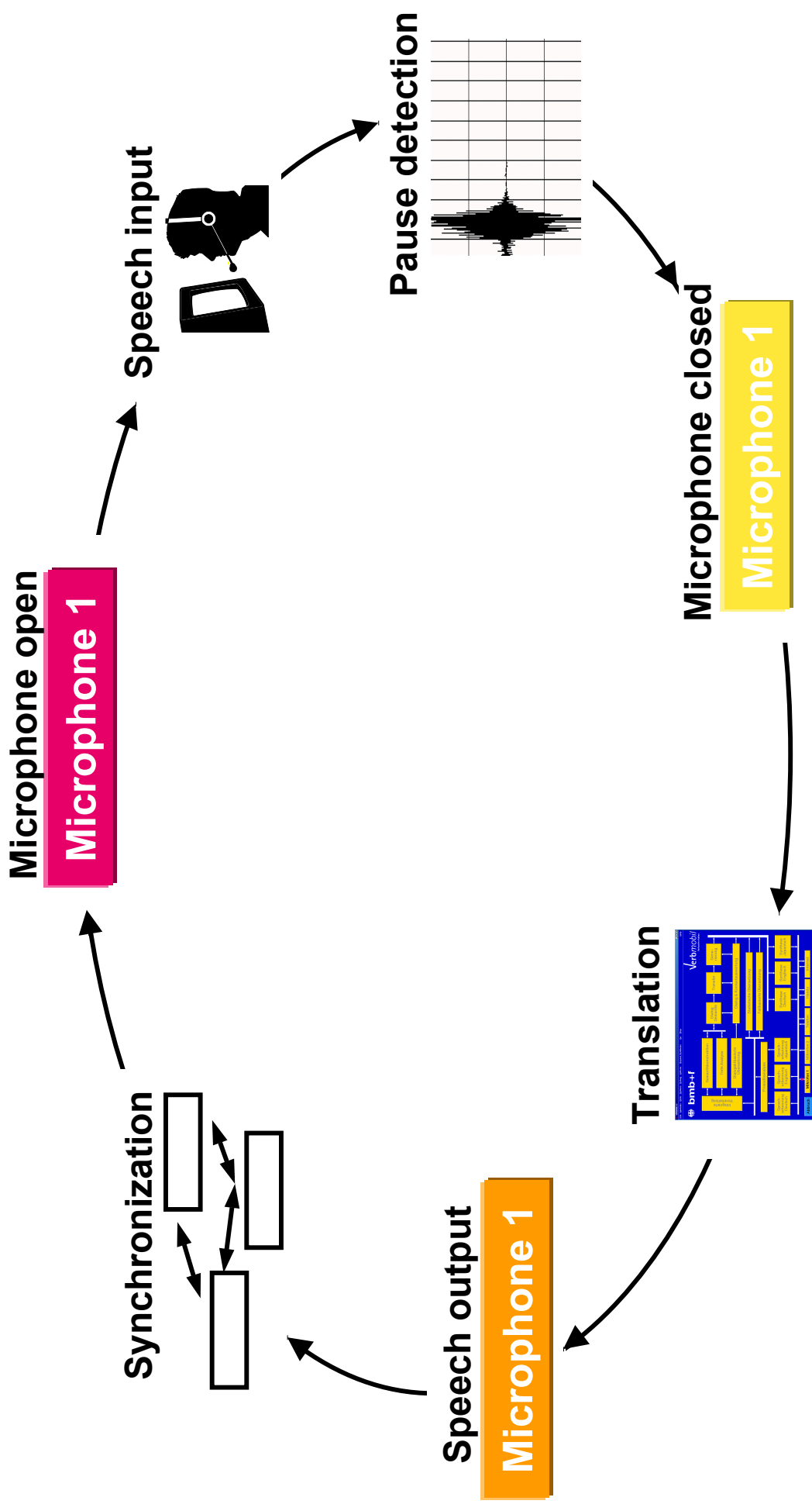
- **Configuration of the systems I/O behavior**
  - How many speakers?
  - For every (possible) speaker:
    - Input device (channel identification, speaker adaption)
    - Output device(s) (translation output, destination for man/machine dialogs)
    - Source language (or „unknown“)
  - Desired system output categories
- **Audio channel configuration**
  - Uniform configuration of heterogeneous audio hardware



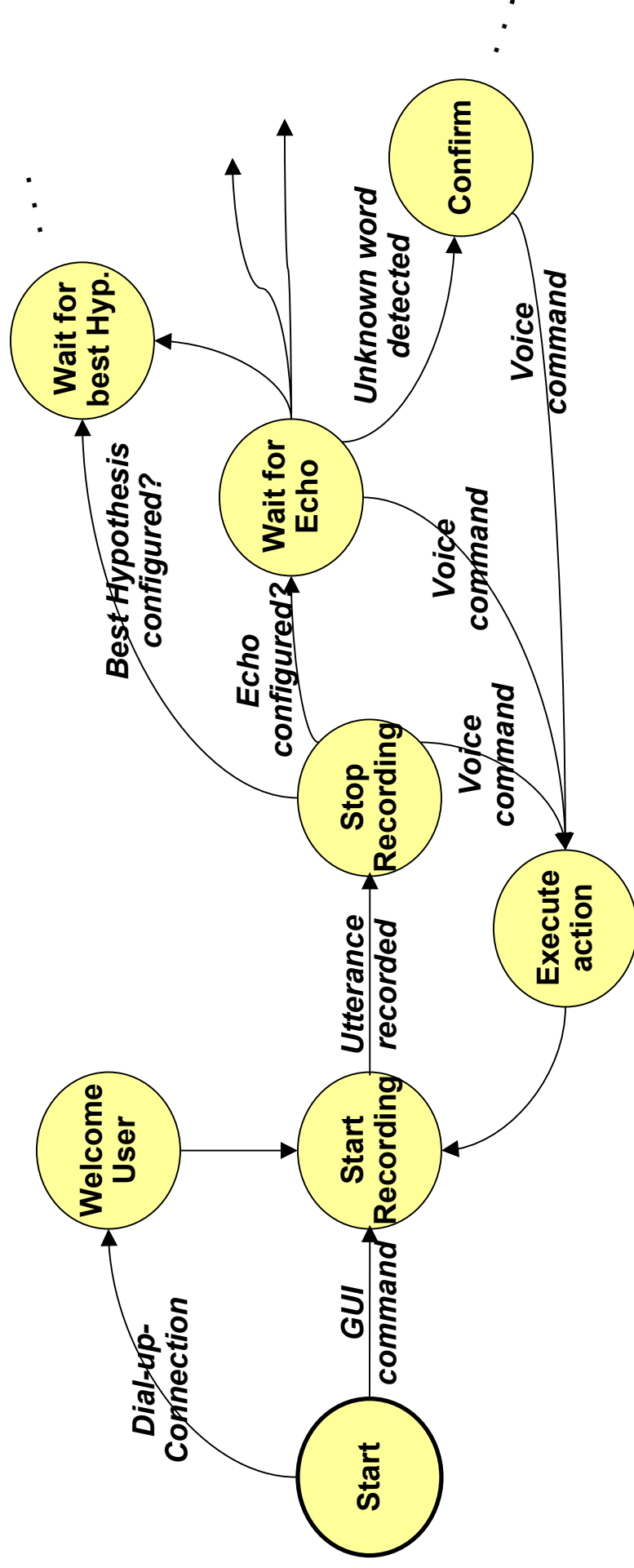
- **Turn-based processing, barge-in available for voice commands**
  - **Different audio quality:**
    - lab-quality close-speaking microphone (16kHz)
    - room microphone (16kHz)
    - telephone quality (8kHz)
    - GSM mobile (8kHz)
- ⇒ **Audio module concept**
- provides a uniform interface of different hardware devices to the system
  - # of channels is only limited by hardware
- **Open Microphone Approach (essential for telephone translation service!)**
  - **Input/output synchronization**
  - **No cross-talk allowed**

# Open Microphone Approach

Verbmobil



- Synchronization controls the high-level System behavior
- Realized via Finite State Machine



# Recognizing Speech

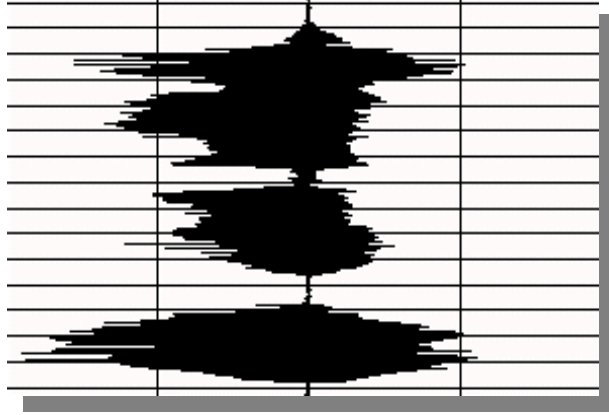
- **Task:**  
Analyzing continuous spontaneous speech signals
- **Input:**  
Audio data
- **Method:**  
HMMs, class based language models, etc.
- **Result:**  
Word Hypotheses Graphs (WHG) and speech commands
- **Benefit:**  
Compact representation of hypotheses of what has been said
- **Responsible:**  
DaimlerChrysler AG  
University of Karlsruhe  
RWTH Aachen  
Philips GmbH (Language Models)



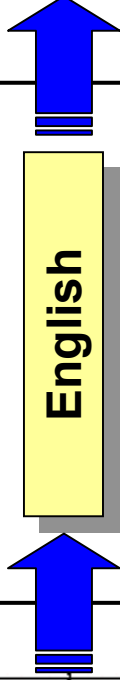
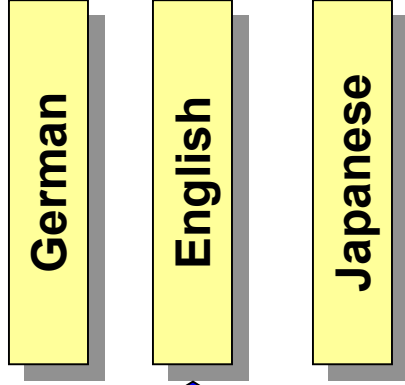
# General Speech Recognition Task



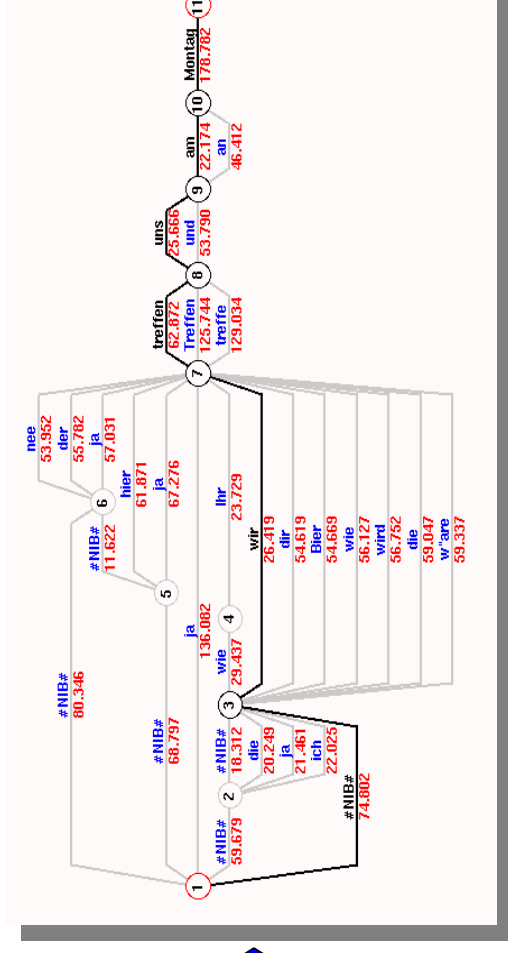
Audio Signal



Recognizers

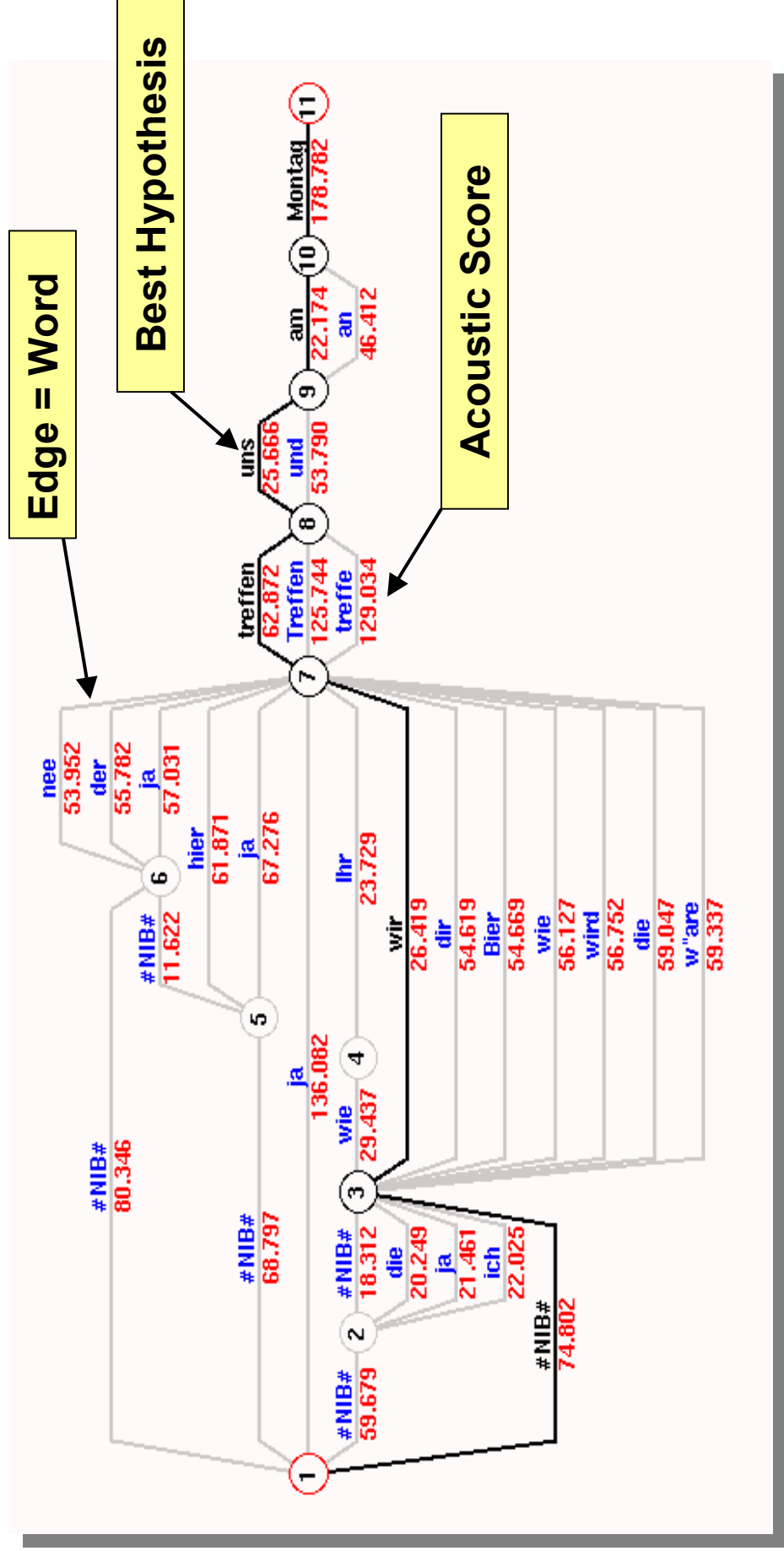


Word Hypotheses Graph



# Word Hypotheses Graphs (WHGs)

WHGs realize the interface between **acoustic** and **linguistic** processing

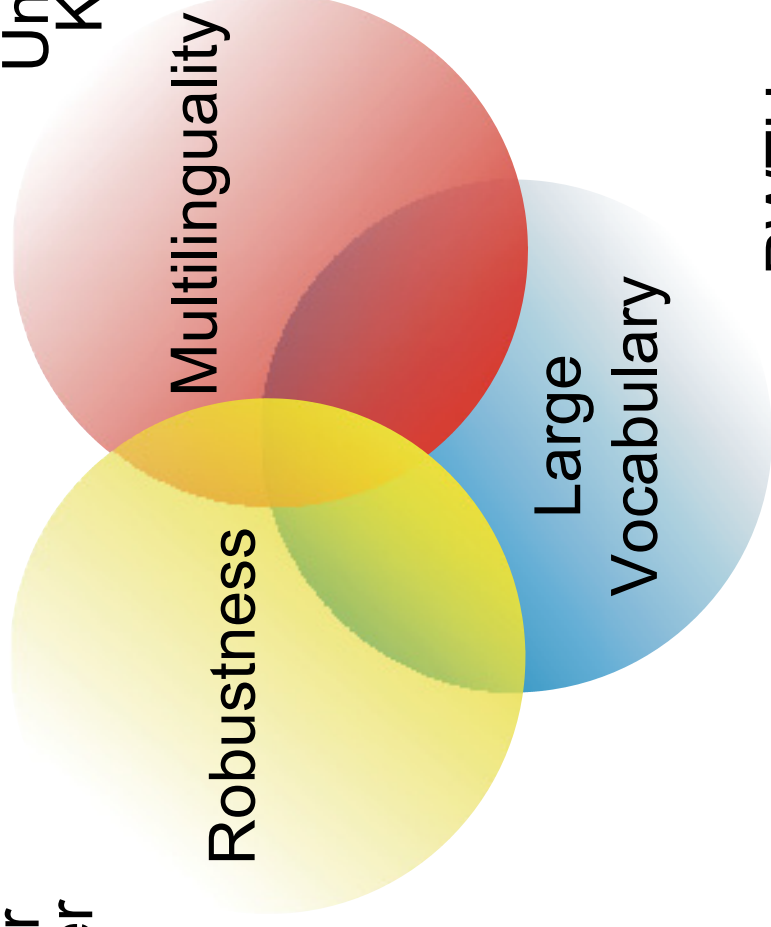


# Focuses of Speech Recognition in Verbmobil

Verbmobil

Daimler  
Chrysler

University of  
Karlsruhe



RWTH  
Aachen

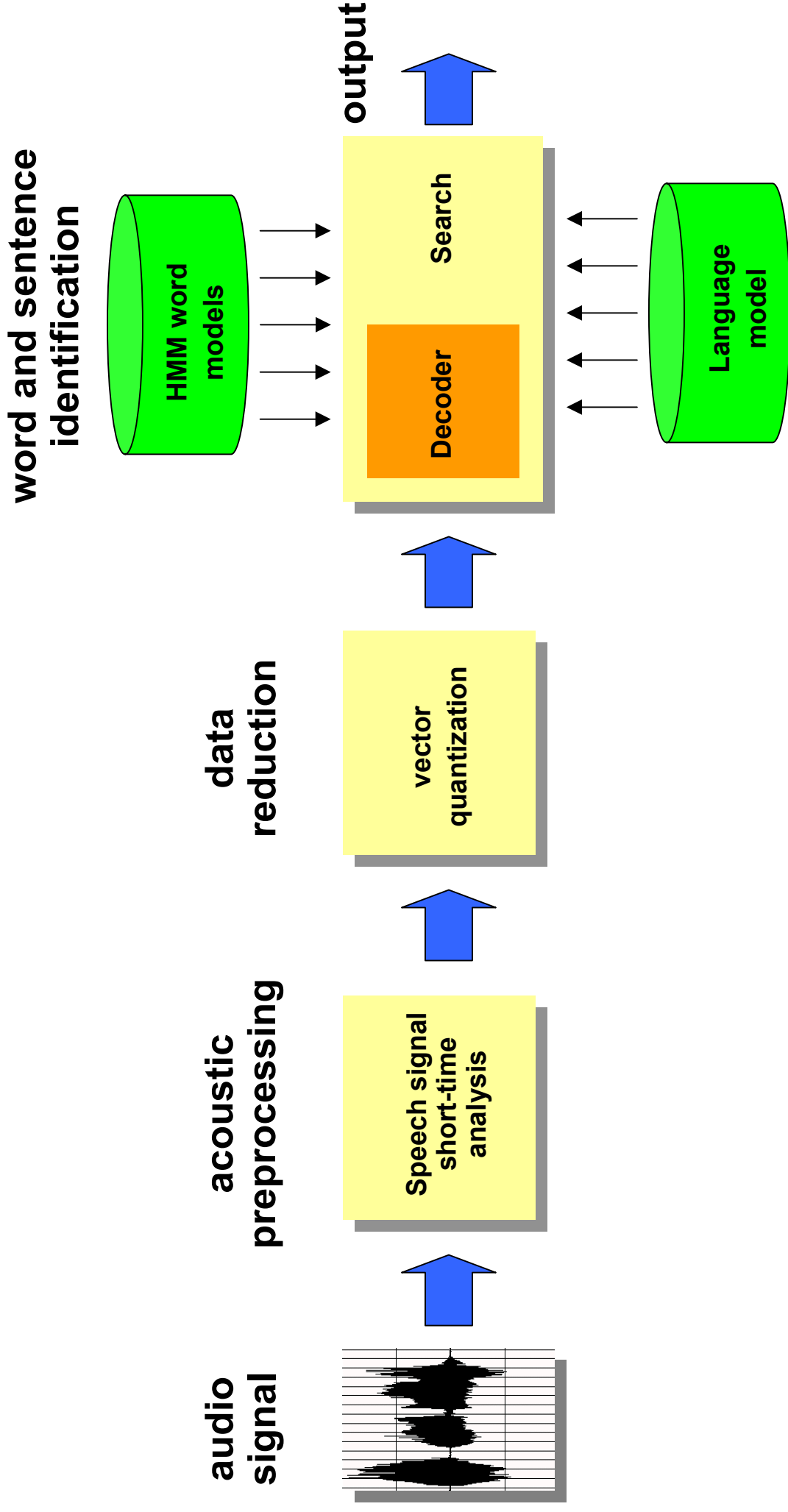
# Nine Available Recognizer Modules

Verbmobil

- **DaimlerChrysler**
  - German, 16 kHz, speaker adaptive, approx. 10000 words
  - German, 8 kHz, [telephone/GSM quality](#), speaker adaptive, approx. 10000 words
  - English, 8 kHz, [telephone/GSM quality](#), speaker adaptive, approx. 7000 words
- **University of Karlsruhe**
  - German, 16 kHz, speaker adaptive, approx. 10000 words
  - [English](#), 16 kHz, speaker adaptive, approx. 7000 words
  - [Japanese](#), 16 kHz, speaker adaptive, approx. 2600 words
  - [Language Identification Component](#) (German, English, Japanese)
- **RWTH Aachen**
  - German, 16 kHz, speaker adaptive, approx. 10000 words
  - German, 16 kHz, speaker dependent, approx. [30000 words](#)

# Principal Recognizer Architecture

Verbmobil



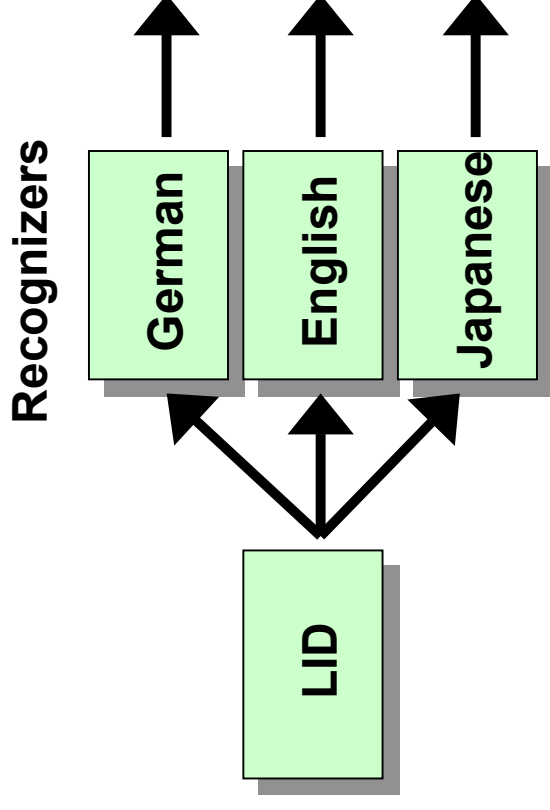
# The Speech Recognition Task

- **Some Highlights of the Verbmobil Recognizers:**
  - **Speaker adaptive** recognition:
    - Start speaker independent
    - Recognition results enhance during the dialog
  - Capable of **dividing speech and noise** input using garbage models
  - **Segmentation** of speech input allows incremental processing
  - **Word class based** language models and recognition allow flexible vocabulary extension
  - **Online vocabulary extension** through unknown word detection (names, towns, street names, ...)
  - Integrated continuous und **speech command** recognition

... **and many more**

- **Features**
  - ID on 3 seconds speech signal (maximum)
  - Real time factor 0.5
  - Speaker independent
  - Unknown audio channel
  - Using language model know-how

- **Flexible Architecture:  
LID can be combined with any  
speech recognizer**

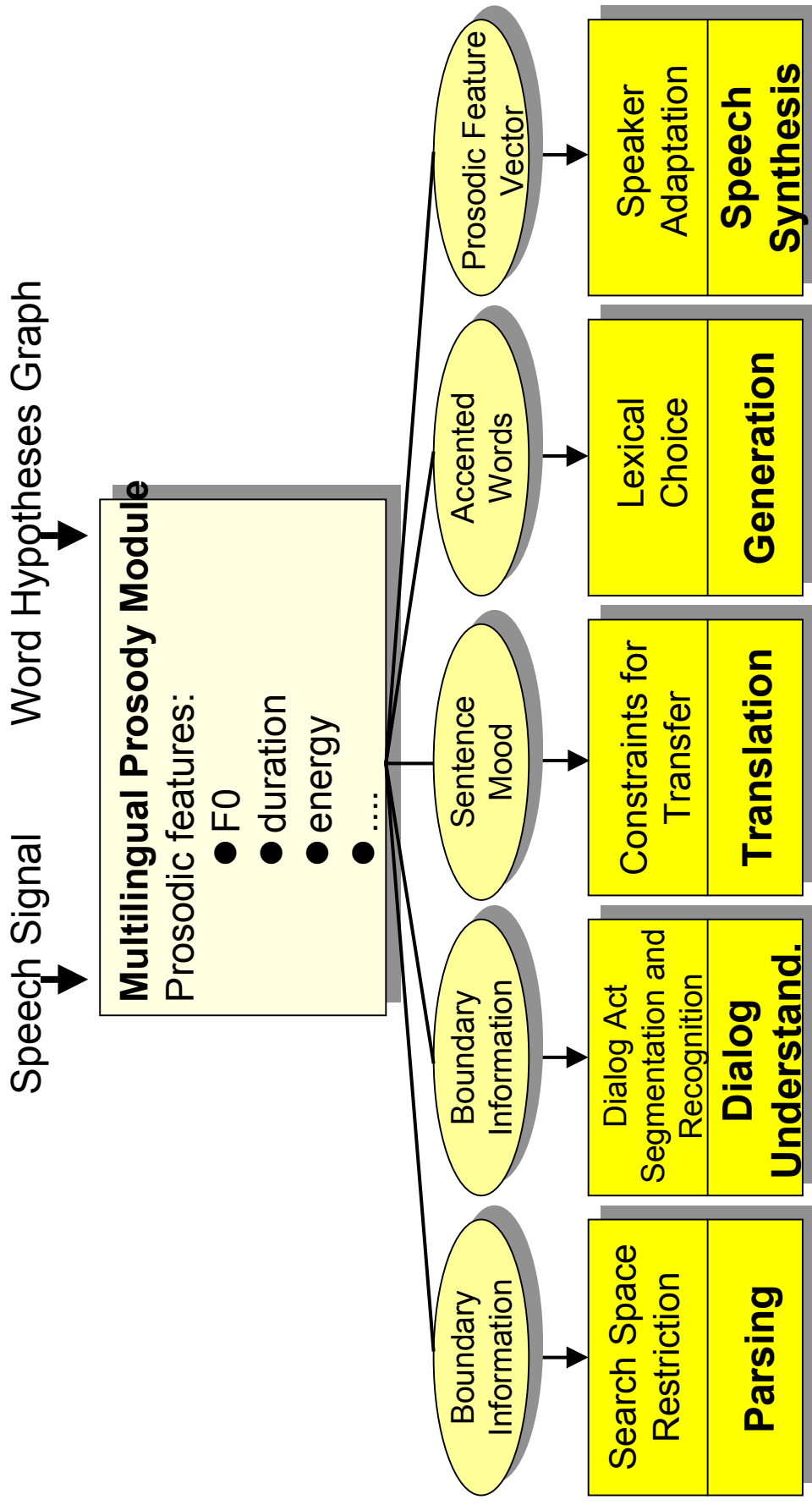


- **Task:**  
Recognizing prosodic phenomena (accents, sentence mood) and boundaries
- **Input:**  
WHG and speech signal
- **Method:**  
Neural networks and statistical classifiers
- **Result:**  
WHG annotated with accent and boundary information
- **Benefit:**  
Provides prosodic information needed for correct translation of spontaneous speech
- **Responsible:**  
Universität Erlangen-Nürnberg



**Prosody can help to disambiguate Parameters represented by Features**

- **lexical and phrasal accent**
- **phrasing (chunks of speech)**
- **sentence mood**
- **emotion, attitude, foreign accent**
- **F0 (fundamental frequency)**
- **Energy**
- **Duration**
- **Speech tempo**
- **Pause**



- Syntactic Boundaries  
*He saw ? the man ? with the telescope*      Prosody cannot help

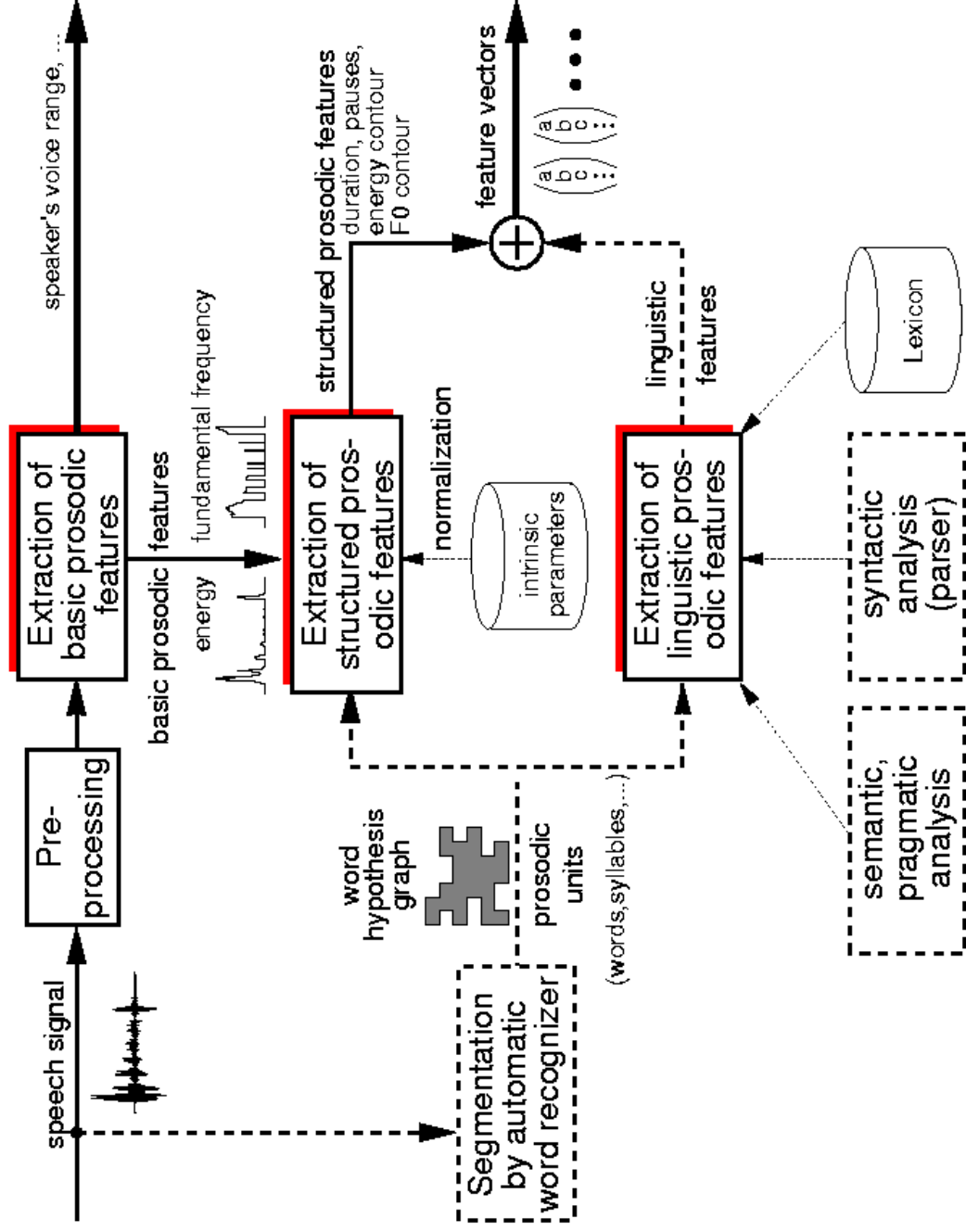
- Dialog Act Boundaries  
*No, I have no time at all on Thursday. D*  
*But how about on Friday?*

Dialog acts are pragmatic units that chunk the input into units which can be processed alone.

- Prosodic Syntactic Boundaries  
*Of course ? not ? on Saturday*  
Syntactic boundaries that correlate to the acoustic-phonetic reality; help during analysis within one chunk/dialog act.  
Important in spontaneous speech with elliptical utterances.

- **computed for each word**
- **from basic prosodic features and segmental information**
- **over different time contexts**
- **modeling of FO:**
  - linear regression coefficient, regression error, mean, median, minimum, maximum, onset, offset and their temporal locations**
- **modeling of energy--contour**
  - mean, median, maximum, max-pos, regression coefficient, ...**
  - and phoneme intrinsic normalizations**

# Extraction of Prosodic Features



# Prosodic Classification in Verbmobil

Verbmobil

- **five classes of boundaries: default, particles, phrases, clauses, sentences**
- **sentence mood: question vs. non-questions**
- **phrase accent: disambiguation of particles**
- **Computed by NN-classifiers and Language Models**
- **Language Models trained on a corpus annotated with syntactic prosodic boundaries and dialog act boundaries**

## I am calling about the trip to Hanover on the seventh and eighth of March

2	...	3	I	50.284023	34	46	(ID r3485)	(PR (S 1.00 0.00 0.00 0.00 0.00 0.00))	(A 0.82 0.18)	(F 0.92 0.00)
3	...	9	am	24.803406	47	52	(ID r3489)	(PR (S 1.00 0.00 0.00 0.00 0.00 0.00))	(A 0.84 0.16)	(F 0.81 0.00)
3	...	10	am	32.151409	47	54	(ID r3490)	(PR (S 1.00 0.00 0.00 0.00 0.00 0.00))	(A 0.88 0.12)	(F 0.37 0.00)
9	...	11	going	142.015503	53	91	(ID r3504)	(PR (S 0.94 0.00 0.05 0.00 0.00 0.00))	(A 0.14 0.86)	(F 0.10 0.00)
10		11	calling	131.019409	55	91	(ID r3505)	(PR (S 0.39 0.01 0.32 0.27 0.01 0.01))	(A 0.07 0.93)	(F 0.13 0.00)
11		12	about	125.144707	92	124	(ID r3506)	(PR (S 1.00 0.00 0.00 0.00 0.00 0.00))	(A 0.22 0.78)	(F 0.92 0.00)
12		13	the	40.895718	125	136	(ID r3507)	(PR (S 1.00 0.00 0.00 0.00 0.00 0.00))	(A 0.90 0.10)	(F 1.00 0.00)
12		13	that	42.615807	125	136	(ID r3508)	(PR (S 0.80 0.00 0.07 0.00 0.12 0.00))	(A 0.84 0.16)	(F 1.00 0.00)
13		14	trip	106.785835	137	167	(ID r3509)	(PR (S 0.10 0.00 0.80 0.10 0.00 0.00))	(A 0.24 0.76)	(F 0.03 0.00)
14		15	to	69.326729	168	188	(ID r3510)	(PR (S 0.86 0.02 0.08 0.02 0.02 0.02))	(A 0.85 0.15)	(F 1.00 0.00)
15		16	Hanover	245.755707	189	261	(ID r3511)	(PR (S 0.02 0.14 0.43 0.01 0.40 0.00))	(A 0.01 0.99)	(F 0.04 0.00)
16	...	18	and	69.891464	266	284	(ID r3514)	(PR (S 0.57 0.08 0.11 0.23 0.02 0.02))	(A 0.87 0.13)	(F 0.95 0.00)
17		18	on	75.358749	264	280	(ID r3515)	(PR (S 0.92 0.03 0.01 0.03 0.00 0.00))	(A 0.87 0.13)	(F 0.62 0.00)
18		19	the	37.180725	285	295	(ID r3516)	(PR (S 1.00 0.00 0.00 0.00 0.00 0.00))	(A 0.94 0.06)	(F 0.98 0.00)
19		20	seventh	184.631897	296	350	(ID r3517)	(PR (S 0.06 0.10 0.31 0.00 0.53 0.00))	(A 0.07 0.93)	(F 0.11 0.00)
20		21	and	44.750828	356	369	(ID r3518)	(PR (S 0.99 0.00 0.01 0.00 0.00 0.00))	(A 0.85 0.15)	(F 0.15 0.00)
21		22	the	42.576515	370	376	(ID r3520)	(PR (S 1.00 0.00 0.00 0.00 0.00 0.00))	(A 0.95 0.05)	(F 1.00 0.00)
22		23	eighth	134.293030	381	420	(ID r3521)	(PR (S 0.00 0.00 0.99 0.00 0.01 0.00))	(A 0.24 0.76)	(F 0.38 0.00)
23		24	of	62.543167	425	443	(ID r3522)	(PR (S 1.00 0.00 0.00 0.00 0.00 0.00))	(A 0.74 0.26)	(F 1.00 0.00)
24		25	March	204.886185	444	497	(ID r3523)	(PR (S 0.02 0.63 0.03 0.02 0.30 0.00))	(A 0.04 0.96)	(F 0.03 0.00)

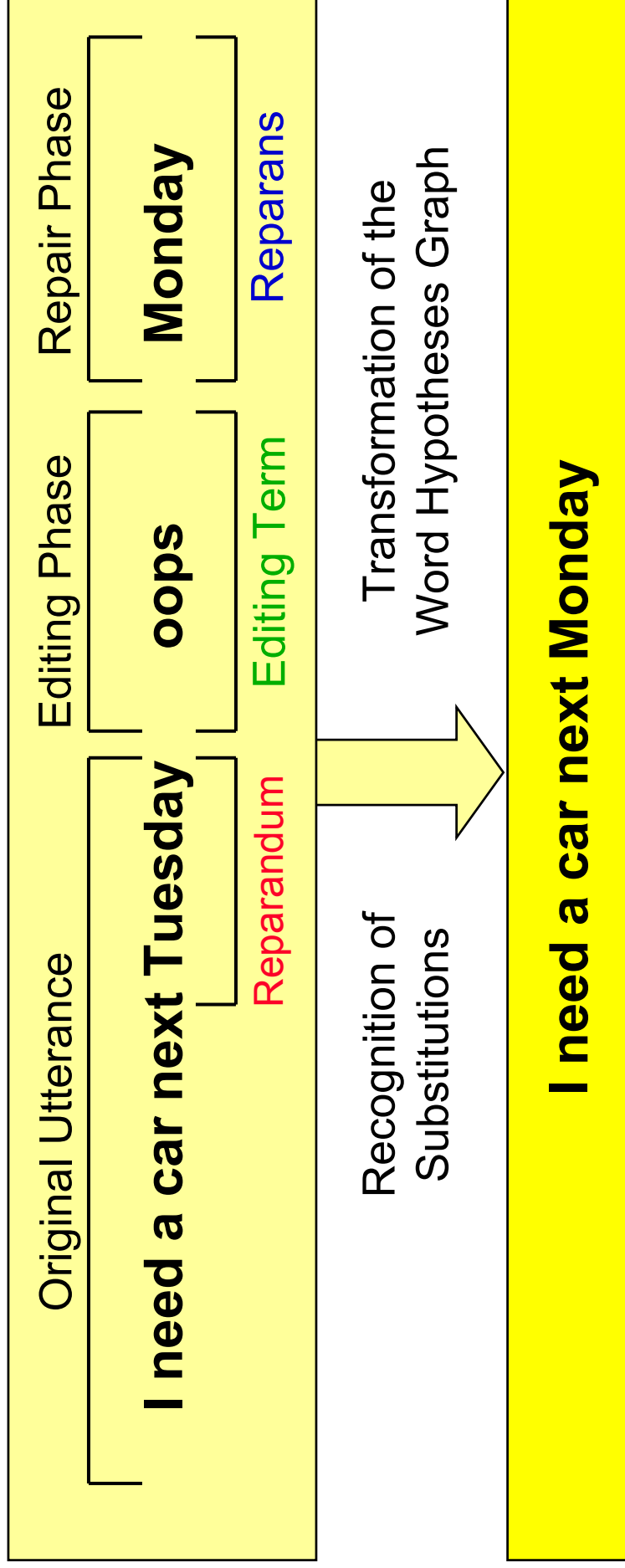
- **Task:**  
Detecting and repairing self-corrections
- **Input:**  
WHGs
- **Method:**  
Stochastic models
- **Result:**  
Enriched WHGs, including additional repaired hypotheses
- **Benefit:**  
Enabling Verbmobil to repair self-corrections of spontaneous speech input
- **Responsible:**

Universität Erlangen-Nürnberg



# The Understanding of Spontaneous Speech Repairs

Verbmobil

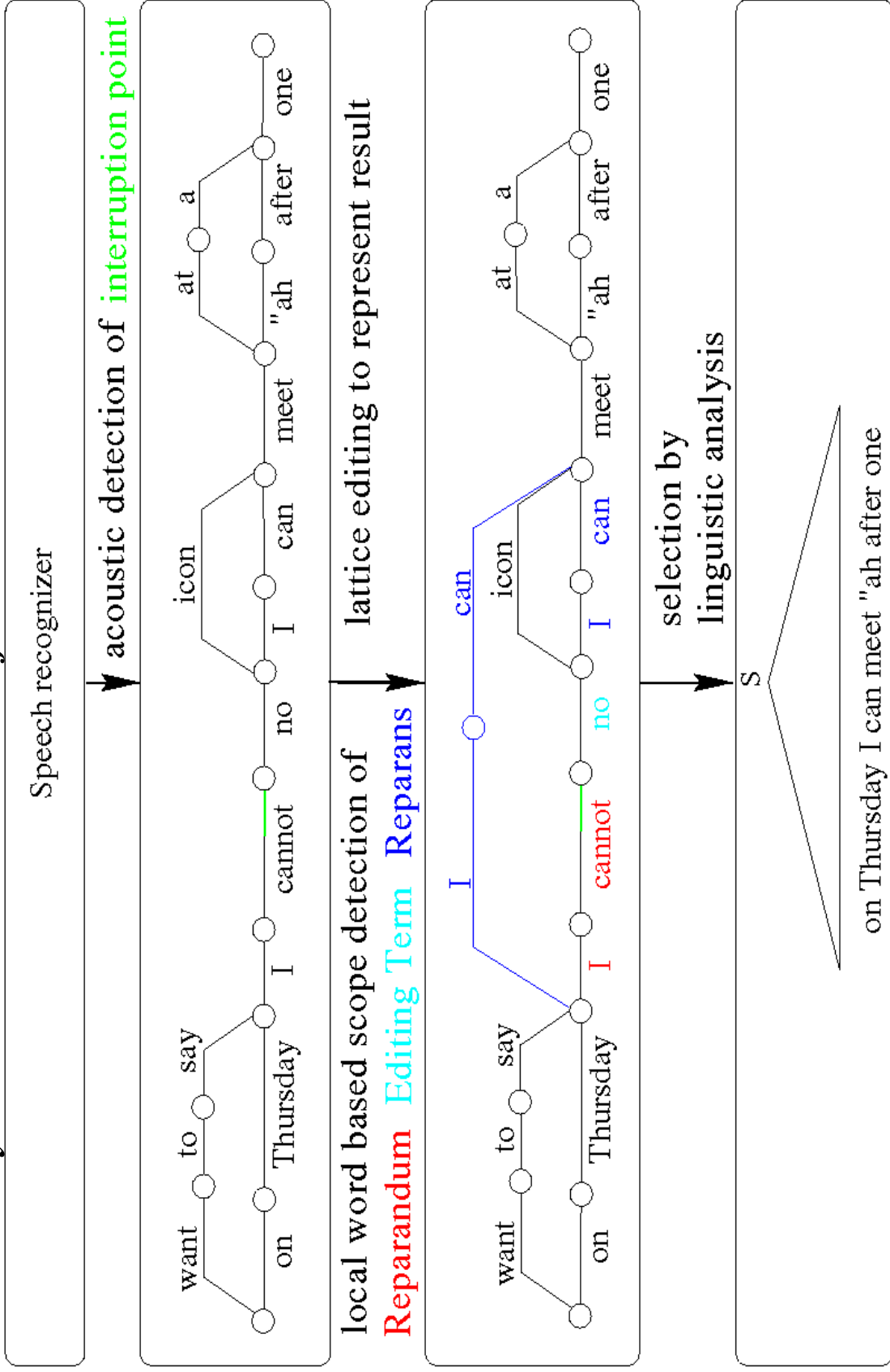


# Facts about Repairs in the Verbmobil Corpus

- **21% of all turns in the Verbmobil corpus (79 562 turns ) contain at least one self correction**
- **The syntactic category is preserved in most cases**  
(For example: Out of a sample of 266 verb replacements, 224 are again mapped to verbs)
- **Repairs take place in a restricted context**  
(in 98% the reparandum consists of less than 5 words)
- **Repair sequences underlie certain regularities**

# Architecture of Repair Processing

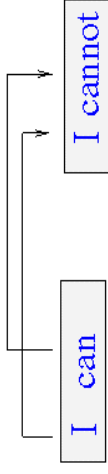
“On Thursday I cannot no I can meet äh after one”



- The editing term (ET) is given by the prosody
  - Wanted: Beginning (RB) and end (RE) of the Repair
  - Search the best replacement of a word order on the left hand side of ET through a word order on the right hand side of ET
- ⇒ rate the possible replacements
- search space is limited through looking at 4 words before and after ET
- ***Choose the best rated replacement over a certain threshold***

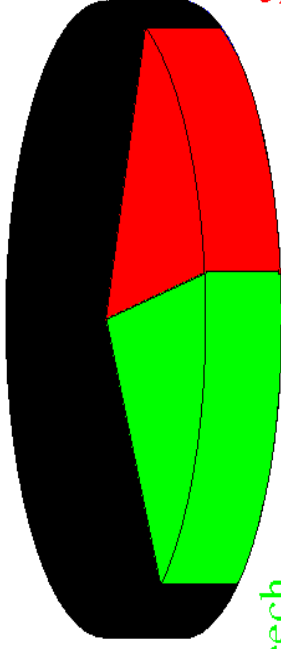
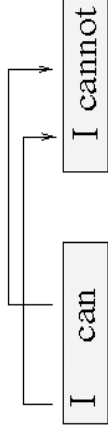
# Repair Detection and Word Smoothing

$$\begin{aligned}
 P_r(RD_j | RS_{a_j}) = & \\
 & \alpha * P(Word(RD_j) | Word(RS_{a_j})) \\
 & + \beta * P(SemClass(RD_j) | SemClass(RS_{a_j})) \\
 & + \gamma * P(POS(RD_j) | POS(RS_{a_j}))
 \end{aligned}$$



Linear Interpolation

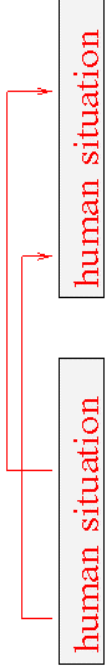
Word



Part of Speech



Semantic Class



# Dialog Translation



- **Mono-cultural approaches are dangerous**
  - humans vs. viruses ↓ diversity
  - Microsoft vs. ILOVEYOU and copycats ↓ alternative software solutions
- **Some sources of errors in a speech translation system**
  - external
    - spontaneous speech: not well formed, hesitations, repairs
    - bad acoustic conditions
    - human dialog behavior
  - internal
    - knowledge gaps in modules
    - software errors
    - probabilistic processing

□ **Use multiple engines, varying approaches on various stages of processing**

- **Exclusive alternatives: three different 16 kHz German speech recognizers with various capabilities**
- **Competing approaches:**
  - **three parsers: HPSG, Chunk, Statistical**
  - **five translation tracks: case-based, dialog-act based, statistical, substring-based, linguistic (deep) semantic translation**
- **Needed: selection and combination of results from competing tracks**
  - parsers: combination of partial analyses in the semantic processing modules
  - translation: preselection module

# Multiple Translation Tracks - Approaches and Advantages

- **Case-based:**
  - Approach: uses examples from the aligned bilingual Verbmobil corpus
  - Advantage: good translation if input matches example in corpus
- **Dialog-act based:**
  - Approach: extract core intention (dialog act) and content
  - Advantage: robust wrt. recognition errors
- **Statistical**
  - Approach: use statistical language and translation models
  - Advantage: guaranteed translation with high approximate correctness
- **Substring- based**
  - Approach: combines statistical word alignment with precomputation of translation "chunks" and contextual clustering
  - Advantage: guaranteed translation with high approximate correctness
- **Linguistic (deep) semantic translation**
  - Approach: "classic" approach using semantic transfer
  - Advantage: high quality translation in case of success

# Example Based Translation

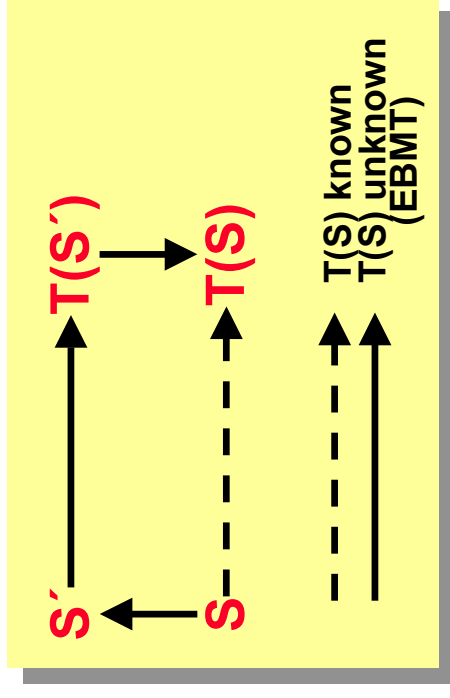
Verbmobil

- **Task:**  
Providing a translation based on translation templates and partial linguistic analysis
- **Input:**  
WHGs or best Hypothesis
- **Method:**  
Definite Clause Grammar (DCG), graph matching algorithms
- **Result:**  
Translation and a confidence value
- **Benefit:**  
Improving Verbmobil's translation capabilities through an additional translation path
- **Responsible:**  
DFKI, Kaiserslautern



- **Training is based on Verbmobil's bilingual corpus**
  - E:** I am on vacation, on the sixth and the seventh.
  - D:** ich bin am sechsten und siebten verreist.

- **Principle: Look up an example in the example storage that matches the input sentence best, use it's translation as output**



# Generalization in Example Based Machine Translation (EBMT)

- **Handicap of this naive approach: inadequate coverage**

S : I am not free on Friday.

S' : I am not free on Monday.

T(S') : am Montag habe ich keine Zeit.

- **Solution: partial generalization (analysis and generation)**

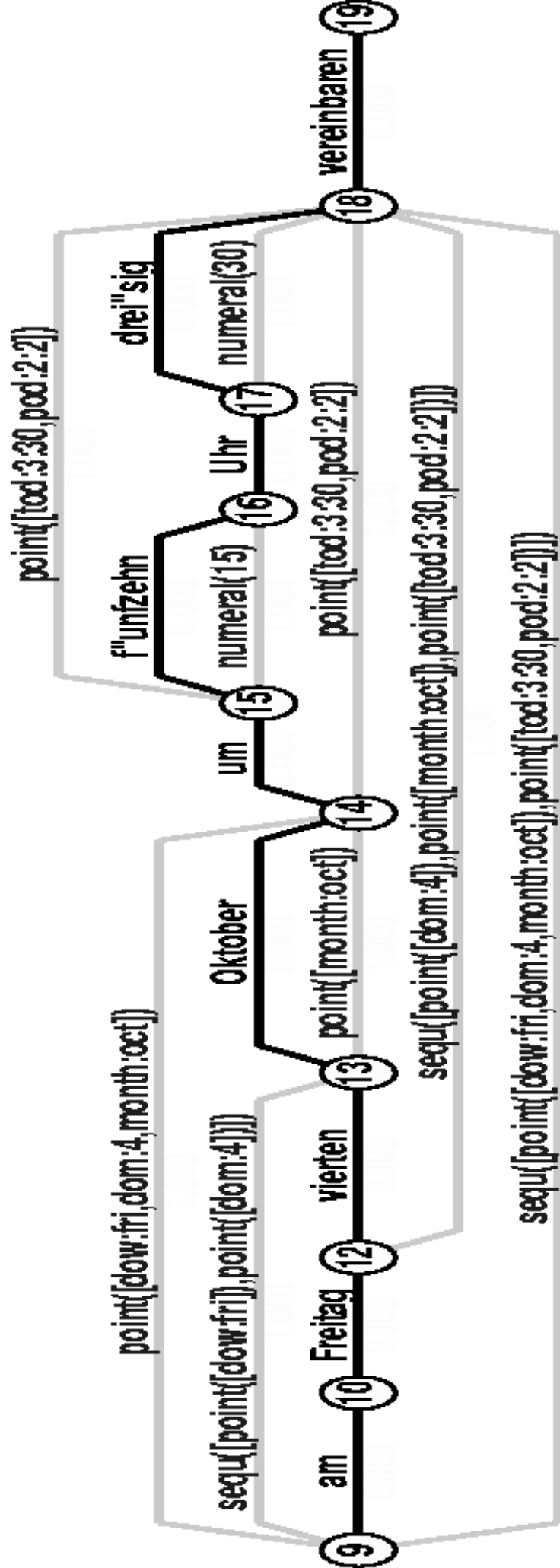
E: I am not free <Temp>.

D: <Temp> habe ich keine Zeit.

- **Automatic generalization approach:**
  - The grammar automatically generalizes the corpus (offline)
  - The runtime module generalizes incoming input (online)
  - Match generalized input sentence with generalized corpus example
  - Result: instantiated corpus translation

# Generalization of WHGs

## Verbobil



# Example Based Translation – Some More Features

- **Generalization grammar for temporals, names, locations (region, town, country), institutions**
- **Fast and robust WHG search:**
  - WHG packing
  - Optimal alignment for fast corpus search
  - Search space pruning
  - Search space caching
  - Any time capable
- **Adequate confidence value for selection**

- **Task:**  
Robustly provide a translation of core intentions and contents of the domain
- **Input:**  
Prosodically annotated best hypothesis (flat WHG)
- **Method:**  
Statistical dialog-act classifier and Finite State Transducers
- **Result:**  
Translation and a confidence value, additionally content descriptions for the dialog module
- **Benefit:**  
Robust translation and content extraction even when the recognition is erroneous
- **Responsible:**  
DFKI, Saarbrücken

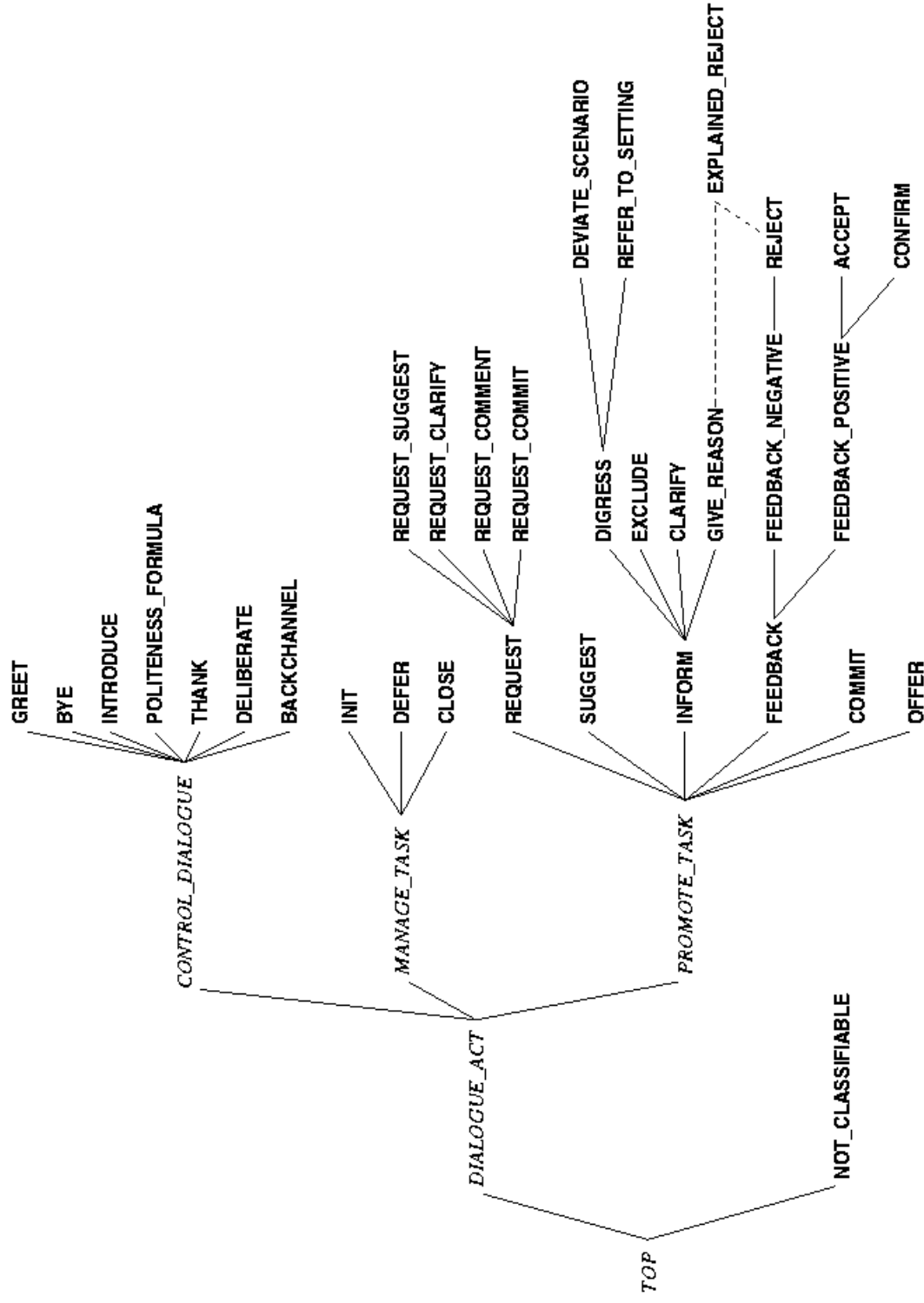


- Describe the core intention of an utterance
- 32 acts defined in a hierarchy, 19 used in processing
- 21 CD-ROMs with 1505 dialogs (German, English, Japanese) annotated with dialog acts for training and test purposes
- Computation uses bigram language models

$$D = \operatorname{argmax}_D P(w|D) \cdot P(D)$$

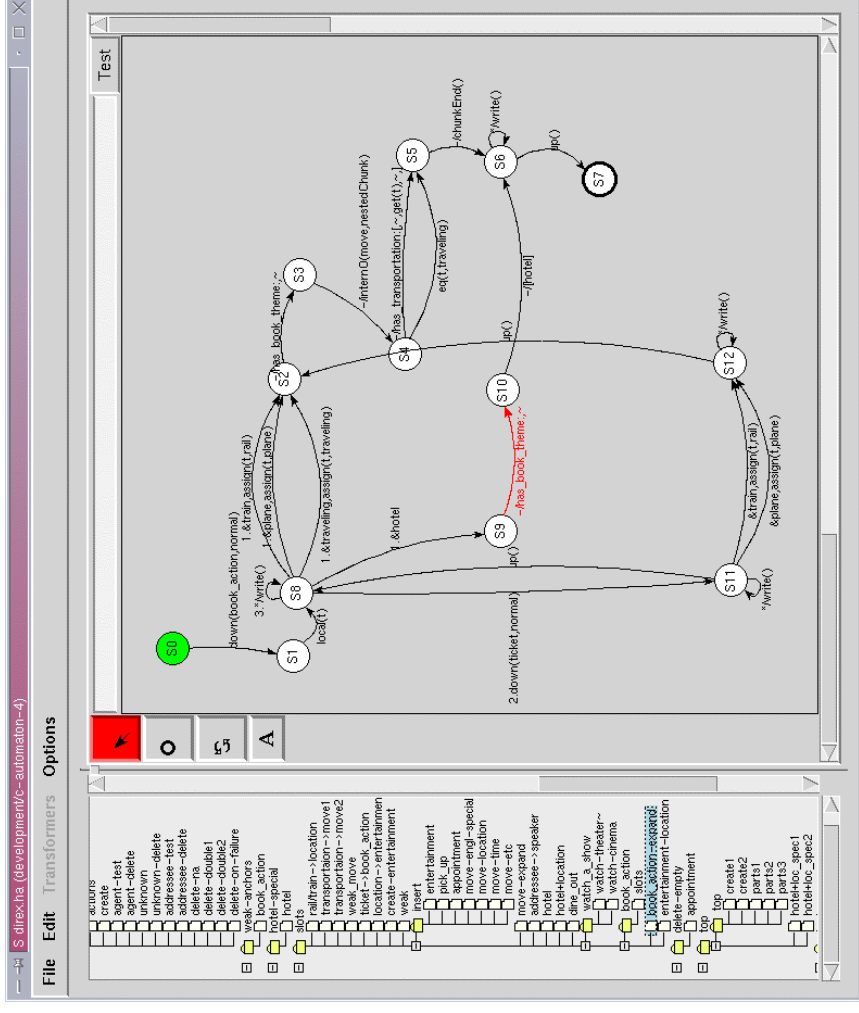
- Probabilities estimated from the annotated corpus
- Leave-One-Out test results for approx. 1000 German, English and Japanese dialogs: Recall 72.48 % (27185 of 37505), Precision 69.90 %

# Dialog Acts - The Hierarchy

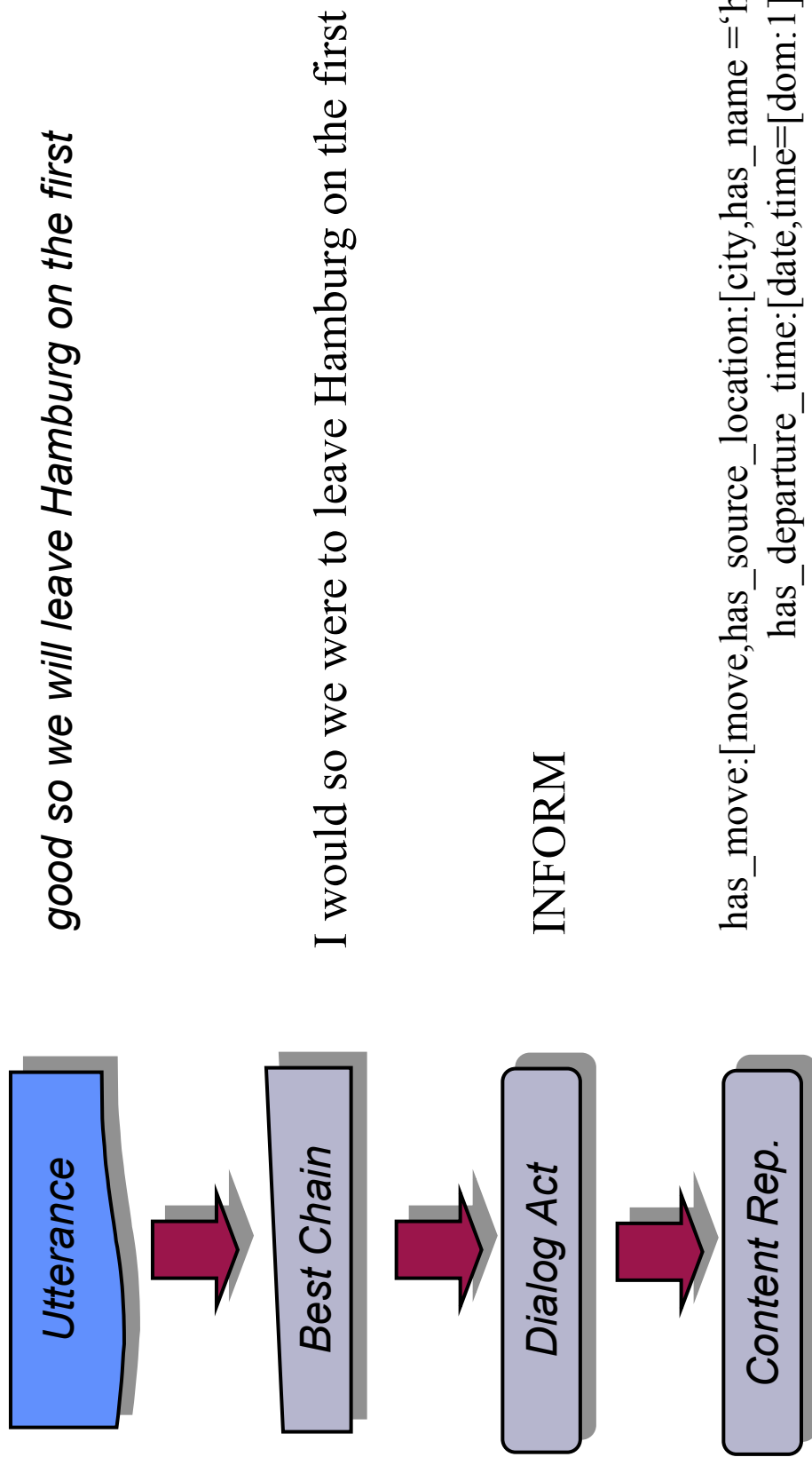


# Representation of Information and Extraction

- Semantic representation language, used also in the dialog and context modules
- Extraction using Finite State Transducers
- Semi-automatic creation exploiting semantic databases and lexica
- Comfortable development platform



# Processing Steps



- **Generation templates (>140)**, depending on dialog act, topic, content
- **Translated in Finite State Transducers**

- **Examples:**

suggest scheduling \$has\_date

g:ich w"urde \$\* vorschlagen &loc\_mode\_dat  
e:how about \$\*

suggest entertainment or(\$has\_location,\$has\_theme)

g:wir k"onnten \$\* gehen &loc\_mode\_acc  
e:we could go \$\*

request\_suggest

g:was schlagen Sie vor

e:what do you suggest

j:itsu ga yoroshii deshou ka

- **Result for our example:** *also wir fahren ab Hamburg am ersten*

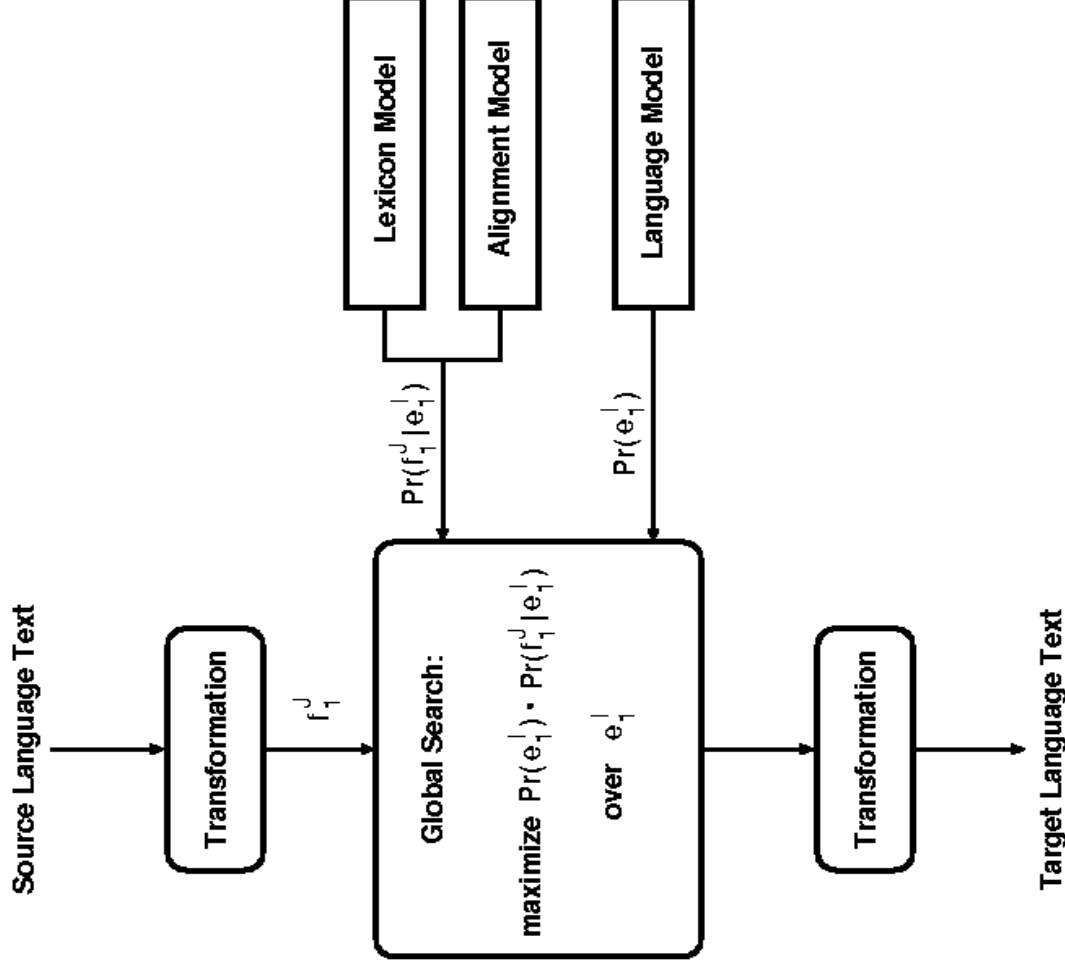
- **Task:**  
Provide approximative correct translations
- **Input:**  
Prosodically annotated best hypothesis (flat WHG)
- **Method:**  
Use statistical language and translation models
- **Result:**  
Translation and a confidence value
- **Benefit:**  
Approximative correct translation for spontaneous speech
- **Responsible:**  
RWTH Aachen

# The Statistical Translation Model

- **Task:** translate the source string  $f$  in the most probable target string  $e$ :

$$\begin{aligned}\hat{e}_1^I &= \arg \max_{e_1^I} \{p(e_1^I | f_1^J)\} \\ &= \arg \max_{e_1^I} \{p(e_1^I) \cdot p(f_1^J | e_1^I)\}\end{aligned}$$

- **Bayes' rule** needs language model of the target language, and lexicon and alignment models
- **Learned** from aligned corpus



- **Find corresponding words in source and target language sentences**
- **Difficult for language pairs with different word order**
- **Solution: alignment templates**
  - based on word classes (sparse data problem: approx. 40% of the words in the training corpus are singletons)
  - first step: statistically learn alignment of words for each translation direction
  - second step: combine the alignments of both directions
  - third step: statistically learn alignment of “phrases”, i.e. word sequences

## Word-to-Word

vs.

## Alignment Templates

days . . . . .  
 both . . . . .  
 on . . . . .  
 eight . . . . .  
 at . . . . .  
 it . . . . .  
 make . . . . .  
 can . . . . .  
 we . . . . .  
 if . . . . .  
 think . . . . .  
 I . . . . .  
 well . . . . .  
 ja . . . . .  
 ich . . . . .  
 denke . . . . .  
 wenn . . . . .  
 wir . . . . .  
 das . . . . .  
 hinkriegen . . . . .  
 an . . . . .  
 beiden . . . . .  
 Tagen . . . . .  
 acht . . . . .  
 Uhr . . . . .

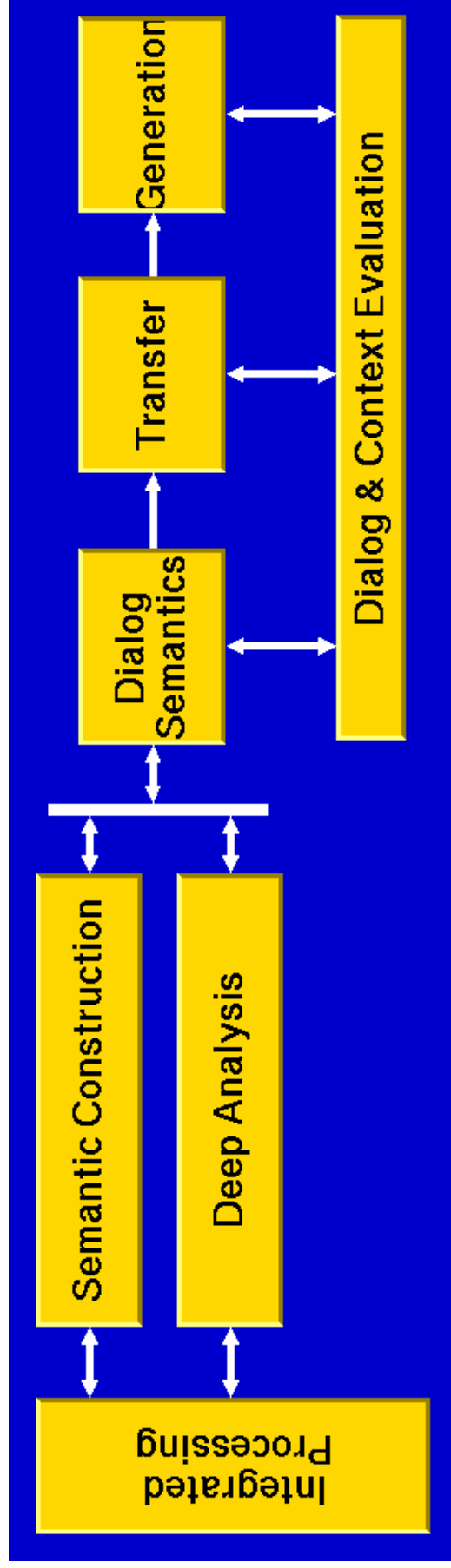
? . . . . .  
 afternoon . . . . .  
 the . . . . .  
 in . . . . .  
 o'clock . . . . .  
 two . . . . .  
 , . . . . .  
 maybe . . . . .  
 at . . . . .  
 nineteenth . . . . .  
 the . . . . .  
 about . . . . .  
 how . . . . .  
 , . . . . .  
 okay . . . . .  
 wie . . . . .  
 sieht . . . . .  
 es . . . . .  
 am . . . . .  
 neunzehnten . . . . .  
 aus . . . . .  
 , . . . . .  
 vielleicht . . . . .  
 um . . . . .  
 zwei . . . . .  
 Uhr . . . . .  
 nachmittags . . . . .  
 ? . . . . .



- **Task:**  
Provide high quality translations
- **Input:**  
Prosodically annotated WHG and contextual information
- **Method:**  
Use syntactic and semantic approaches to analysis, transfer, and generation

- **Result:**  
Translation containing content information, suited for high quality speech synthesis
- **Benefit:**  
Delivers the highest quality, but is sensitive to recognition errors and spontaneous speech phenomena
- **Responsible:**  
Siemens AG, DFKI Saarbrücken, Universität Tübingen, Universität des Saarlandes, Universität Stuttgart, TU Berlin, CSLI Stanford

# Modules Involved



- Integrated processing comprises
  - search through the WHG
  - statistic parser
  - chunk parser

- Semantic Construction provides VITs from statistic and chunk parser output

- Deep Analysis: HPSG Parser
- Dialog Semantics: combination of parsing results, and semantic resolution
- Transfer: VIT to VIT transfer
- Generation: TAG generation from VITs
- Dialog+Context: provides contextual information

- **Verbmobil uses three different syntactic parsers:**  
an HPSG parser, a chunk parser, and a probabilistic LR parser.
- **Every parser implements another level of parsing accuracy, depth of syntactic analysis, and robustness of the analyzing process.**
  - **Chunk parser:** Most robust but least accurate analysis
  - **HPSG parser:** Most accurate by least robust analysis
  - **Probabilistic parser:** Level of accuracy and robustness between HPSG and chunk parser

- Gets WHGs for the English, German, or Japanese speech input and dispatches WHG information to the three parsers
- Provides an A\* search algorithm that allows any connected parser to find the best scored path using
  - acoustic score of the speech recognizer
  - Verbmobil trigram language model
- Parsers analyze the same utterance simultaneously

- **Common syntactic-semantic interface**
- **Contains all linguistic information relevant for translation**
- **Record-like data structure: variable-free lists of non-recursive terms**
- **“Flat” set representations: semantic, scopal, sortal, morpho-syntactic, prosodic, and discourse information**
- **Labels relate different kinds of information**
- **Abstract Data Type implements construction, access, update, check, print, etc. facilities**

```
vit(vitID(sid(...)),
  []),
index(1250, 1234, i72),
[ start v(1248, i72) ],
  arg1(I248, i72, i75),
  nop(1240, h85),
  quest(1249, h84),
  time(1238, i73),
  abstr_vacation(1247, i75),
  pron(I242, i74),
  poss(1244, i75, i74),
  temp_loc(1239, i72, i73),
  def(I245, i75, h87, h86),
  whq(1235, i73, h83, h82) ],
[in_g(1235, 1237), ...],
leq(1234, h85), ...],
[ s_class(1240, mp), ...], i67, i66] ),
[ ana_ante(i74, [i75, i69, i67, i66]),
  prn_type(i74, third, std), ...],
[ gend(i75, masc), num(i75, sg) ],
[ ta_mood(i72, ind), ...],
[ ... ])
```

**%Segment ID**  
**%WHG-String**  
**%Index**  
**%Conditions**

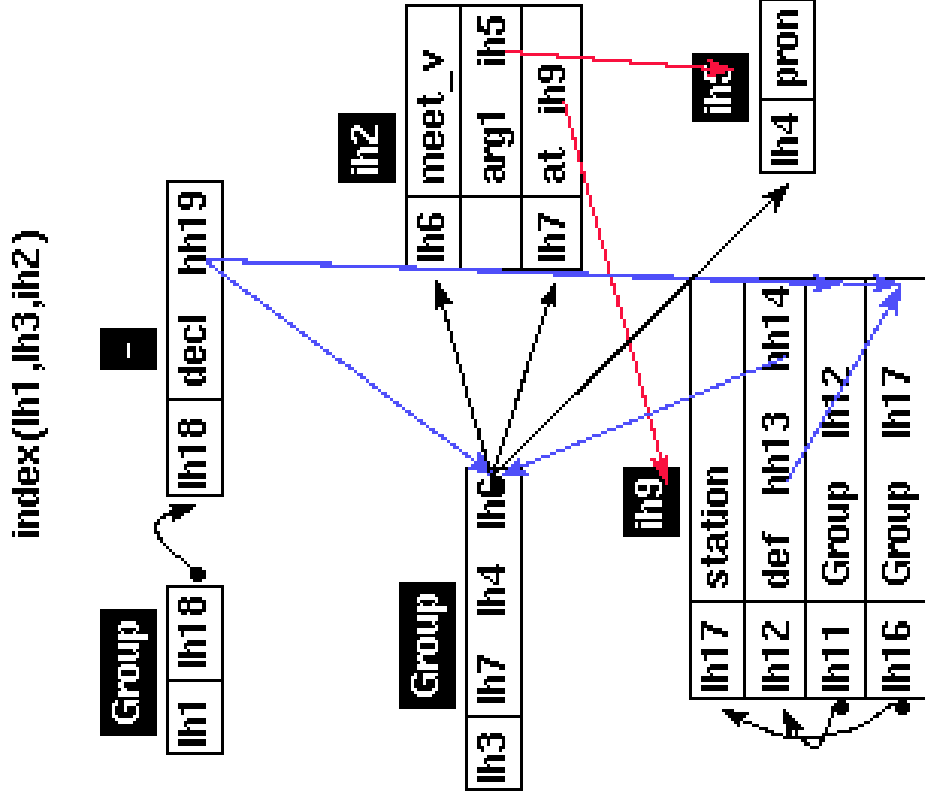
**%Constraints**

**%Sorts**  
**%Discourse**

**%Syntax**  
**%Tense and Aspect**  
**%Prosody**

*When do your vacations begin?*

# VIT: Verbmobil Interface Term



We meet at the station.



- **Task:**  
Thorough syntactic analysis
- **Input:**  
Word chains from integrated processing
- **Method:**  
Apply HPSG analysis
- **Result:**  
Source language VITs
- **Benefit:**  
Delivers the highest quality, but is sensitive to recognition errors and spontaneous speech phenomena
- **Responsible:**  
DFKI Saarbrücken, CSLI Stanford



# Head Driven Phrase Structure Grammar

- Well known advanced grammar theory in linguistics
- Based on the concept of a *sign* as integrated information structure for all types of linguistic information
- Inherently multilingual by distinguishing universal principles from language specific aspects
- Typed feature structures with inheritance
- Small number of rules, due to general principles
- Independent of specific processing strategies, usable for analysis and generation

- **Lexicalism:** Words carry all the important information about what they can be combined with, thus allowing to deal with regular and idiosyncratic properties in a uniform way
- **Heads:** Phrases contain a head which determines their combinatory potential, e.g. verbs as heads determine what complements must be present, and what modifiers they can combine with
- **Principles:** Few language independent general projection principles stating, e.g., how to combine a head with complements and modifiers
- **Unification:** Monotonically combines constraints from different sources

- **active chart parser allowing bidirectional and island parsing on word**
- **hypotheses graphs or strings**
- **fast processing by**
  - eliminating disjunctions, enabling fast conjunctive unification
  - precompiling type unifiability, avoiding runtime computations
  - quick checks on mostly relevant features, avoiding full unification
  - quick checks on possibly discontinuous constituents, e.g. separable verb prefixes in German, reducing the chart size
  - precompiling rule filters on possible rule sequences
  - scoring rule applications
- **anytime behavior**
- **robust: best partial analyses even for ungrammatical input**

- **Task:**  
Robust probabilistic parsing
- **Input:**  
n-best hypotheses
- **Method:**  
LR-Parser trained on Verbmobil's  
tree-bank
- **Result:**  
Syntactic tree representation of the  
input sentence
- **Benefit:**  
Increasing robustness in Verbmobil's  
multi-engine parser strategy
- **Responsible:**  
Siemens AG

- (Non-probabilistic) **LR-parsing** worked quite well for parsing speech in Verbmobil's first phase.
  - **LR-parsing** is well known to be able to parse huge amounts of input very efficiently.
  - Probabilistic **chart** parsing of spontaneous speech input had some problems i.e. the combinatorical explosion of edges in the chart on a word graph
- ⇒ try probabilistic **LR-Parser**

- **Training process: derivation of an LR table and the estimation of unknown probabilistic parameters from the Verbmobil tree bank**
  - Find the set of all context free rules (G) contained in the tree bank.
  - Construct an LR table from G using well known standard
  - Problems: sparse data, different annotation styles
    - ⇒ **eliminate rules that do occur less than N times**
- **Transformations:**
  - Needed **after parsing** to correct errors of the probabilistic context free parser
  - Rules are learned automatically from the training corpus

- **Task:**  
Robust and efficient partial parsing,  
even on ill-formed input
- **Input:**  
N-best hypotheses
- **Method:**  
Cascaded Finite State Transducers
- **Result:**  
Syntactic tree representation of the  
input sentence
- **Benefit:**  
Increasing robustness in Verbmobil's  
multi-engine parser strategy
- **Responsible:**  
Universität Tübingen

*1st Step: Chunk Parsing using Cascaded Finite State Transducers*

*“Chunks are non-recursive cores of ‘major’ phrases, i.e. NP, VP, PP, ...”*

*2nd Step:*

*Building a syntactic tree out of the parsing results*

**Benefit:** Robust and efficient parsing

**But:** Partial parsing: Often no spanning analysis

# Example for Chunks

*“Ich habe bei meinem letzten Besuch in Hannover so eine nette Kneipe entdeckt”*

### Chunks:

- [NX Ich] [VX habe] [PX bei [ NX meinem letzten Besuch]] in [NX Hannover]
- [PX so [NX eine nette Kneipe]] [VX entdeckt].

### where

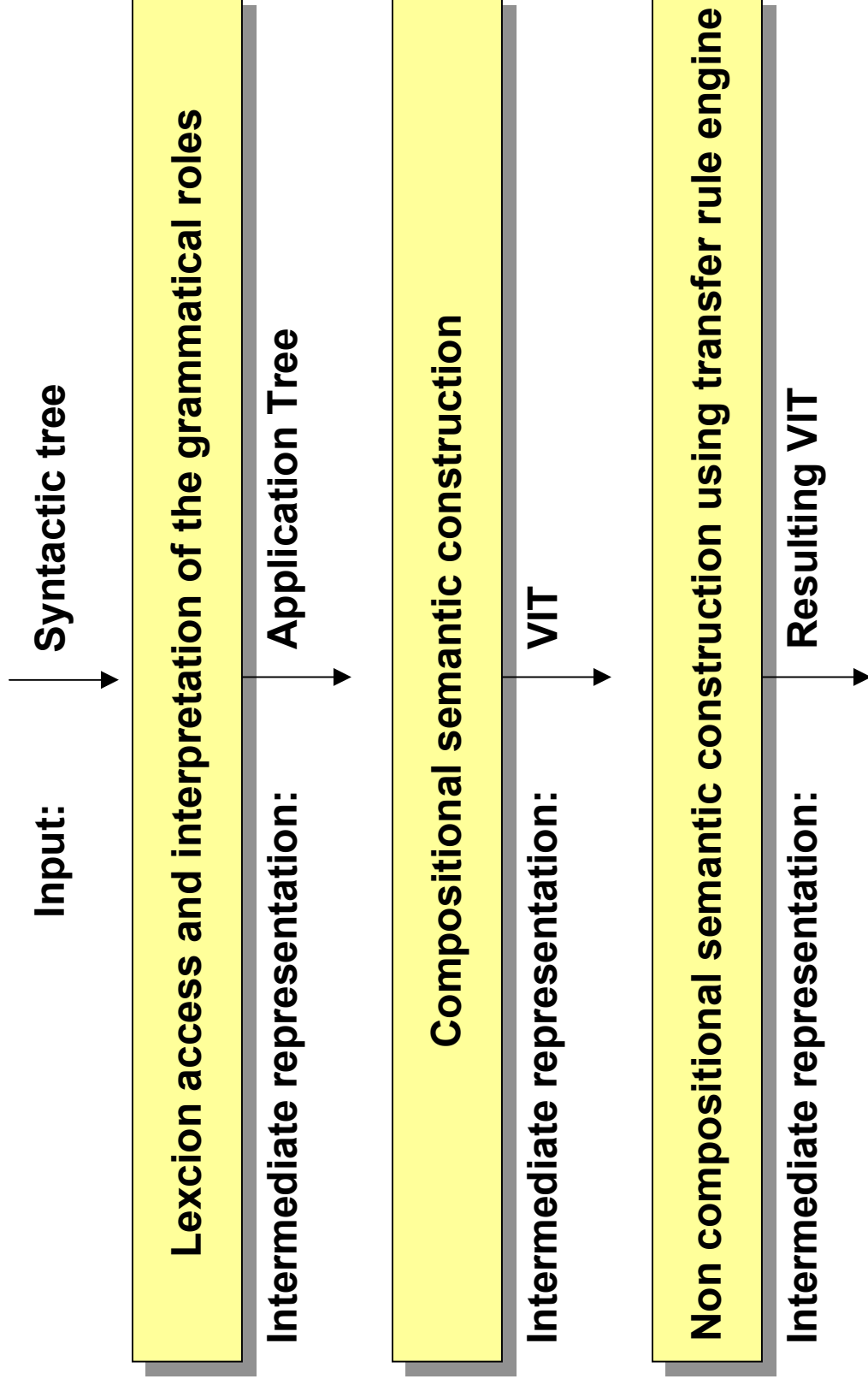
- **[NX]:** Extends from the beginning to the head of a NP
- **[VX]:** Includes all modals, auxiliary verbs and medial adverbs, but ends at the head verb or predicate adjective
- **[PX]:** Extends to the end of an [NX]

- **Determine the chunk position inside the syntactic tree**
- **Complete the internal chunk structure**
- **Determine functional categories and topological fields**
- **Rearrange chunks to obtain a complete syntactic tree**





- **Task:**  
Convert and extend syntax trees to VITs
- **Result:**  
VITs
- **Benefit:**  
Providing results of shallow parser to the deep analysis track
- **Responsible:**  
Universität Stuttgart (IMS)
- **Method:**  
Compositional construction using semantic lexicon



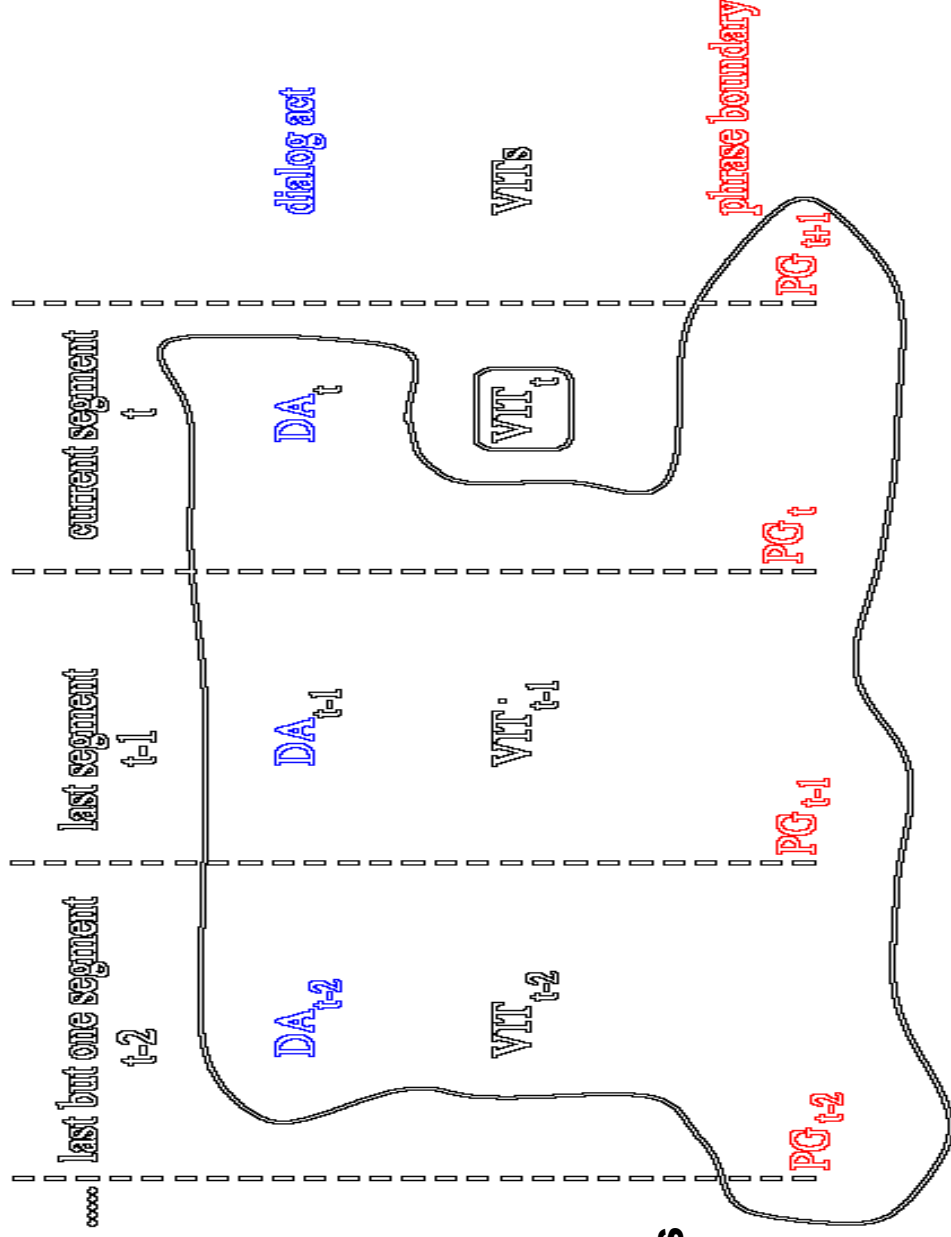
- **Task:**  
Combining results from various parsers, reinterpret and correct VITs, and resolve non-local ambiguities
- **Input:**  
VITs from different parsers
- **Method:**  
VIT models and rule based approaches
- **Result:**  
VIT ready for transfer
- **Benefit:**  
Enhances robustness of deep analysis and provides vital information for transfer
- **Responsible:**  
Universität des Saarlandes, Saarbrücken



# Combining Analyses from Various Parsers

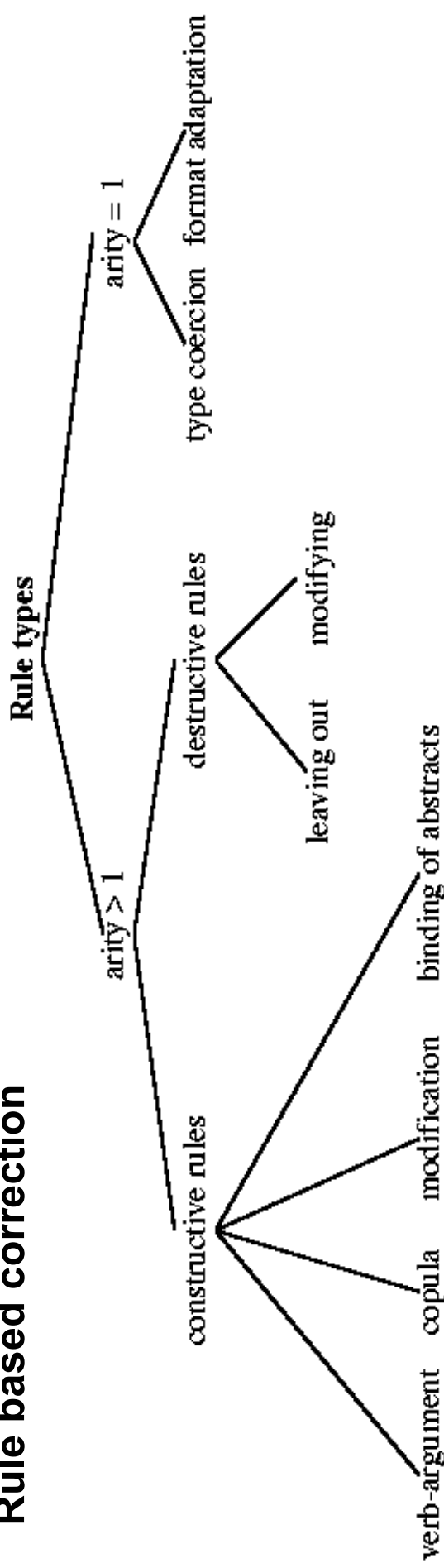
- **Parsers deliver VITs for segments of a turn**
- **May be spanning analyses or just partial fragments**
- **Combination necessary, both analyses of one parsers, but also analyses from various parsers**
- **Combination criteria**
  - HPSG is better than statistical parsers is better than chunk parser
  - Integrated results are better than fragments
  - Longer results are better than short ones

- **Parser internal scores not normalized  $\Rightarrow$  external scoring necessary**
- **Statistical model based on VIT content and dialog act (Tetragram language models)**
- **Search through Vit Hypotheses Graph VHG comparable to search through WHG**



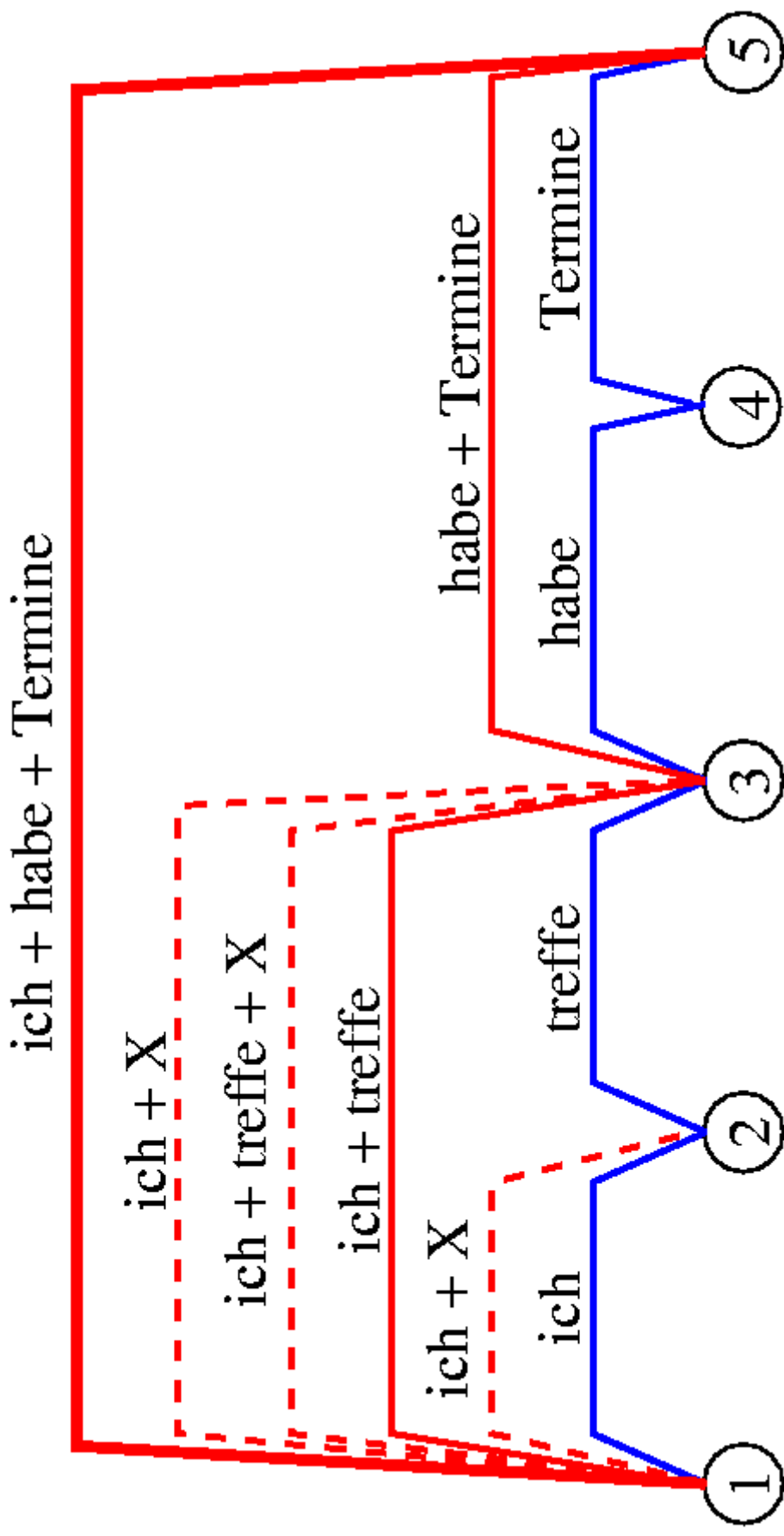
- **Partial results don't necessarily fit together**
  - phenomena of spontaneous speech
  - recognition errors
  - parsing errors

- **Rule based correction**



# Bridging Mechanism for False Starts

## Verbmobil



- **Based on prosody and dialog act information**
- **Ambiguities processed:**
  - Verb disambiguation:  
*Wir gehen in's Theater* (We go to the theater)  
*Montag geht bei mir nicht* (Monday does not suit me)
  - Sentence mood  
*Wir gehen in's Theater!* vs. *Wir gehen in's Theater?*
  - **Adverb disambiguation**  
*Wir gehen eher in's Theater* (We go to the theater earlier)  
*Montag geht bei mir eher nicht* (Monday does not really suit me)
  - **Anaphora and ellipsis resolution**
  - **Japanese: Definiteness, topic phrases, zero anaphora**

- **Task:**  
Transfer VITs from the source to the target language
- **Input:**  
VITs
- **Method:**  
Rule based transfer
- **Result:**  
VITs for generation
- **Benefit:**  
Translate VITs inside the deep translation path
- **Responsible:**  
Universität Stuttgart (IMS)

# The Transfer Approach: Rule Based Transfer

- VITs are mapped onto VITs: Transfer is a VIT rewriting system
- Rule based, context conditions restrict application
- Transfer rules remove matching source language expressions from the VIT
- Efficient implementation
- Examples:
  - Simple Rules: `adelig(L,I) -> noble(L,I)`
  - Simple Templates: `@mod(adelig, noble, L, I)`
  - Selectional restrictions: `#sort_check(I, human) -> true`  
`@mod(gross, tall, _, I)`  
`#sort_check(I, location) -> true`  
`@mod(gross, large, _, I)`

- **Structural changes:**
  - Adjective to PP: tagsüber -> during the day
  - Insertion: übernachtete -> spend the night
  - ...
- **Disambiguation:**

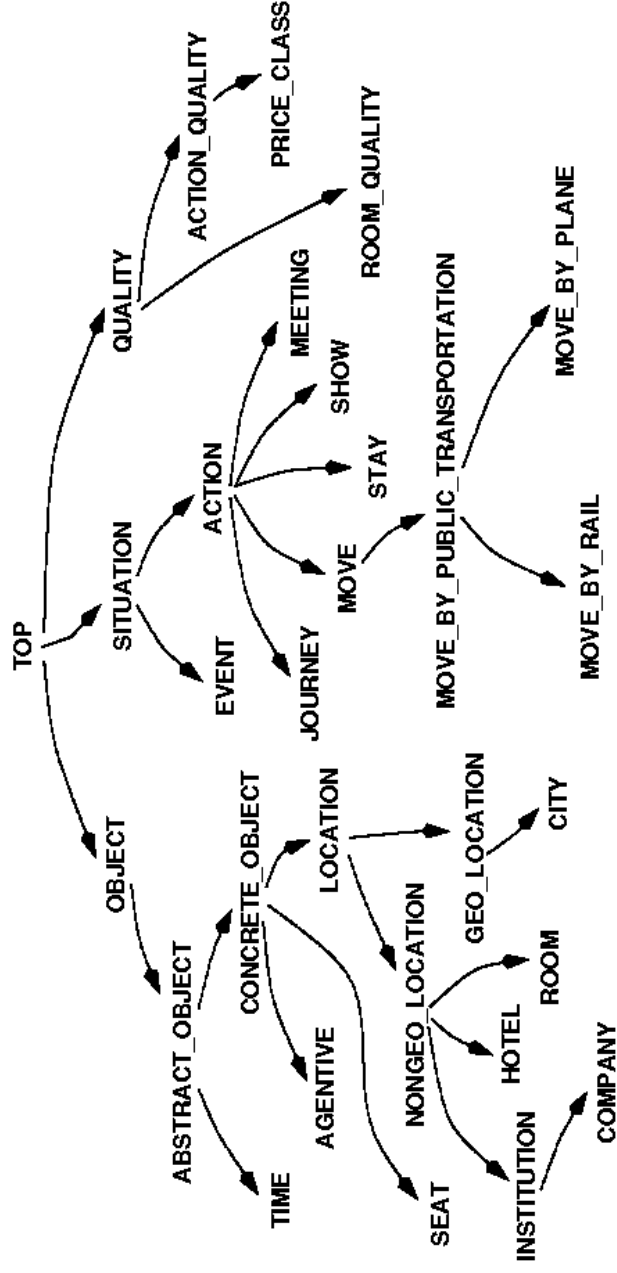
type of ambiguity	kinds of knowledge needed for disambiguation	modules that contribute to the resolution
lexical	syntactic, semantic, contrastive, domain, prosodic	parsers, semantic construction, discourse semantics, transfer, context
structural	syntactic, semantic, domain	parsers, semantic construction, transfer
anaphora and ellipsis	syntactic, semantic, domain	discourse semantics, context
semantic focus and operator scope	prosodic, syntactic, semantic, contrastive, domain	discourse semantics transfer

- Rules are compiled and packed
- 18088 rules German ↔ English
- 4694 rules German ↔ Japanese
- Mean runtime per sentence: 80 msec (Sun Ultra II, 300 MHz)

- **Task:**  
Resolving ambiguities in the dialog context during semantic transfer
- **Input:**  
Requests from transfer
- **Method:**  
Using world knowledge and rules
- **Result:**  
disambiguated transfer requests
- **Benefit:**  
Higher quality of transfer results
- **Responsible:**  
Technical University (TU) Berlin

# Context Evaluation - Tasks and Methods

- Supports semantic transfers and processes VITs
- Gets information from dialog module from shallow tracks
- Extends disambiguation of the dialog semantic module and uses ontological information



**Example: Platz → room / table / seat**

- 1 Nehmen wir dieses Hotel, ja. → Let us take this hotel.  
Ich reserviere einen **Platz**. → I will reserve a **room**.
- 2 Machen wir das Abendessen dort. → Let us have dinner there.  
Ich reserviere einen **Platz**. → I will reserve a **table**.
- 3 Gehen wir ins Theater. → Let us go to the theater.  
Ich möchte **Plätze** reservieren. → I would like to reserve **seats**.

- **Task:**  
Provides dialog context for all tracks  
and computes main information for  
dialog summaries
- **Input:**  
Data from a lot of modules
- **Method:**  
Frame-like topic structuring and rules
- **Result:**  
context information and dialog  
summaries and minutes
- **Benefit:**  
Verbmobil knows what happens  
throughout the dialog and can  
present it
- **Responsible:**  
DFKI, Saarbrücken

- **Dialog Memory:**
  - Stores information from each track
  - Only dialog act based and semantic transfer provide abstract representations:

Discourse Representation Language DRL:

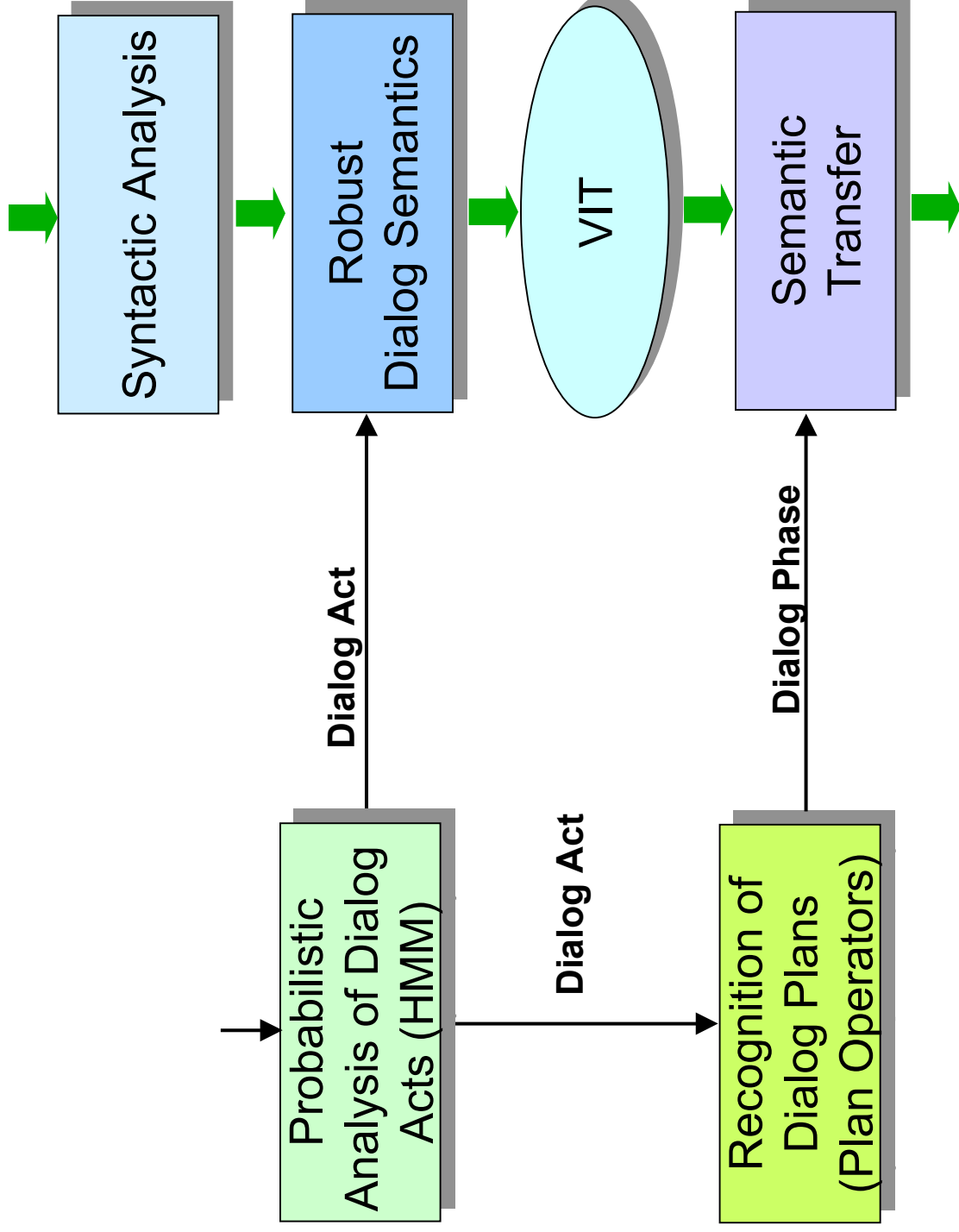
*I would so we were to leave Hamburg on the first*

```
[INFORM, has_move: [move, has_source_location: [city, has_name='hamburg' ,  
has_departure_time: [date, time='day:1'
```

- **Discourse Interpretation:**
  - Groups information into topics
  - Completes information
  - Keeps tracks of negotiation structure

# Dialog Information in Semantic Transfer

Verbmobil



# Collaboration for a New Functionality: Result Summaries

- **Provide the users with a summary of the topics that were agreed**
- **Two benefits**
  - have a piece of information to use in calendars etc.
  - control the translation
- **Approach: exploit already existing modules for**
  - content extraction
  - dialog interpretation
  - planning the summary
  - generation
  - transfer

**RESULT SUMMARY**

**Participants:** Speaker B, Speaker A  
**Date:** 22.2.2000  
**Time:** 8:57 AM to 09:37 AM  
**Theme:** Appointment schedule with trip and accommodation

**DIALOGUE RESULTS:**

**Scheduling:**  
Speaker B and speaker A will meet in the train station on the 1. of march 2000 at a quarter to 10 in the morning .

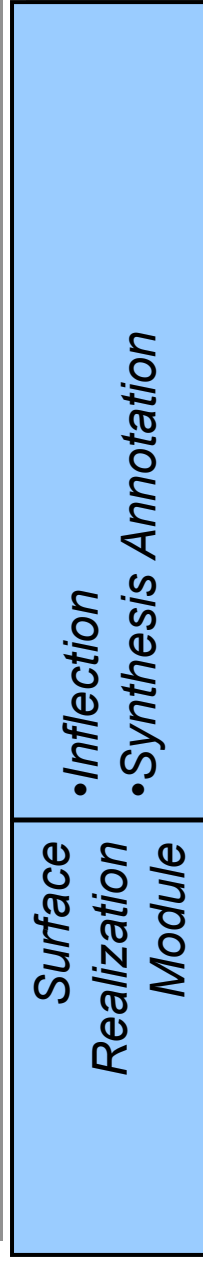
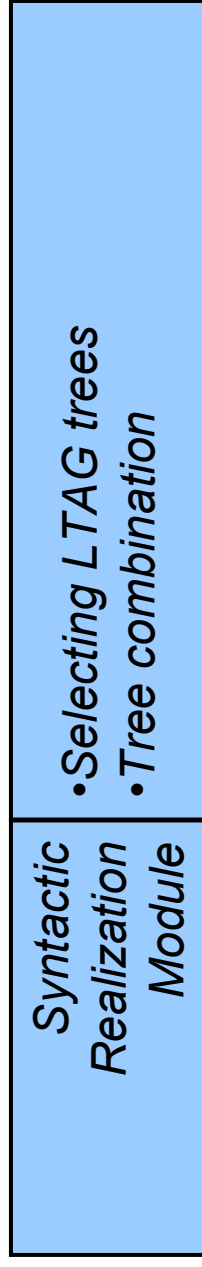
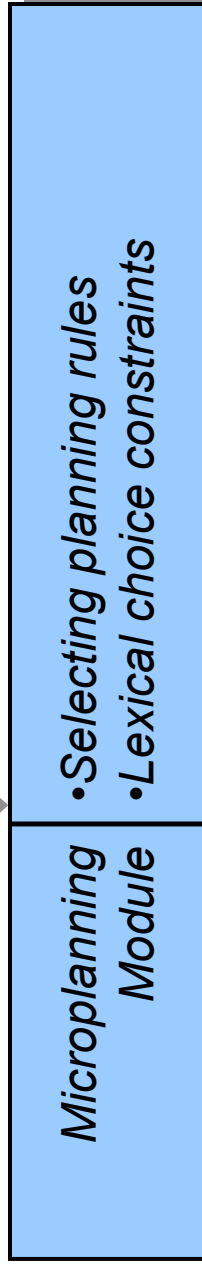
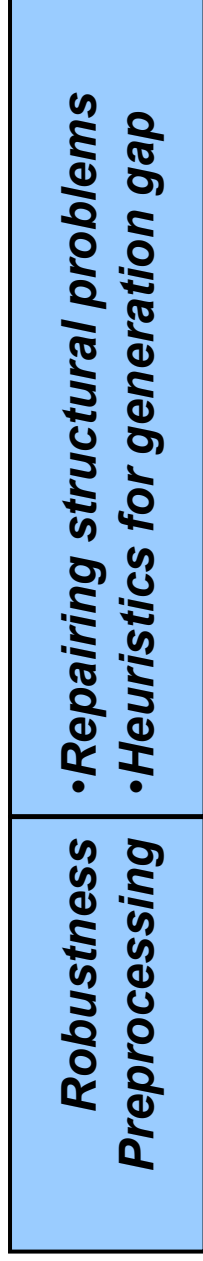
**Travelling:**  
There the trip from Hamburg to Hanover by train will start on the 2. of march at 10 o'clock in the morning . The way back by train will start on the 2. of march at half past 6 in the evening .

**Accommodation:**  
The hotel Luisenhof in Hanover was agreed on. Speaker A is taking care of the hotel reservation.

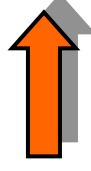
- **Task:**  
Robustly generate the output of the semantic transfer in German, English, or Japanese
- **Input:**  
VITs from transfer
- **Method:**  
Constraint system for micro-planning, TAG grammar (reusing HPSG grammars) for syntactic realization
- **Result:**  
Strings, enriched with content-to-speech (CTS) information to support synthesis
- **Benefit:**  
Output from the semantic transfer track
- **Responsible:**  
DFKI, Saarbrücken



VIT (Verbmobil Interface Term)



**Annotated String**



## Why *pre*-processing:

- Check and repair inconsistencies as early as possible
- Keep robustness and standard modules separate
- Alternative: relax constraints

## Preprocessing for robustness means:

- Executing a set of solution submodules in sequence
- For each problem found, the preprocessor lowers a *confidence value* for the generation output which measures the reliability of our result

# How much robustness?

- **PRO:**  
In a dialog system, a poor translation might still be *better than none* at all,
- **CON:**  
one of the shallow modules can be selected when deep processing fails,  
so respect the *inherent limitations of robustness*.  
⇒ **Generation** knows its limits and sometimes decides not to produce a string
- **Selection module:** uses training corpus and confidence values to select from the different translation paths

# Microplanning: Create Syntactic Building Blocks

Method: Mapping of dependency structures

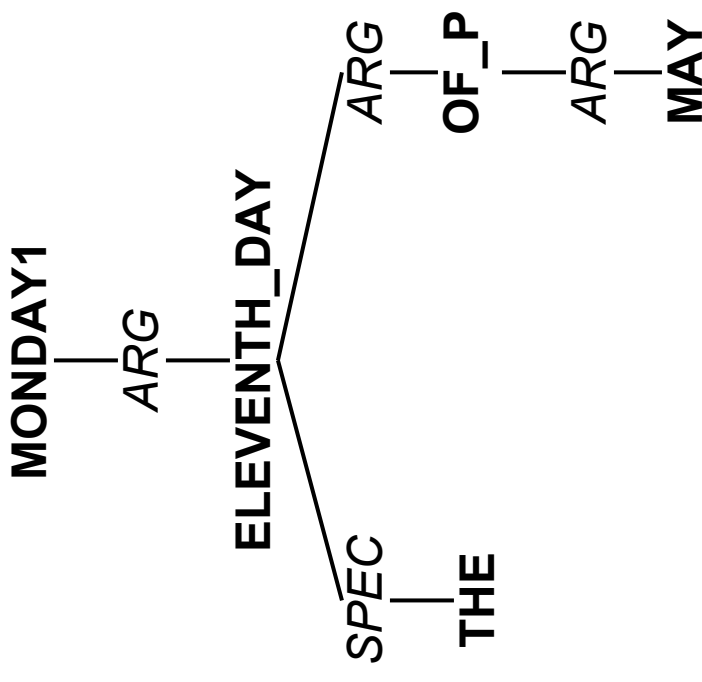
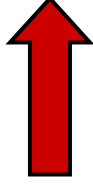
Example: Time Expressions

DEF (L,I,G,H)

DOWF (L1,I,I,mo)

ORD (L2,I,I,11)

MOFY (L3,I,I,may)



Semantical dependency: VIT

Syntactical dependency: TAG

- **Output annotated with information like speech act, syntactic grouping, word classes, prominence, ...**
- **Enhances synthesis quality**
- **Example:**

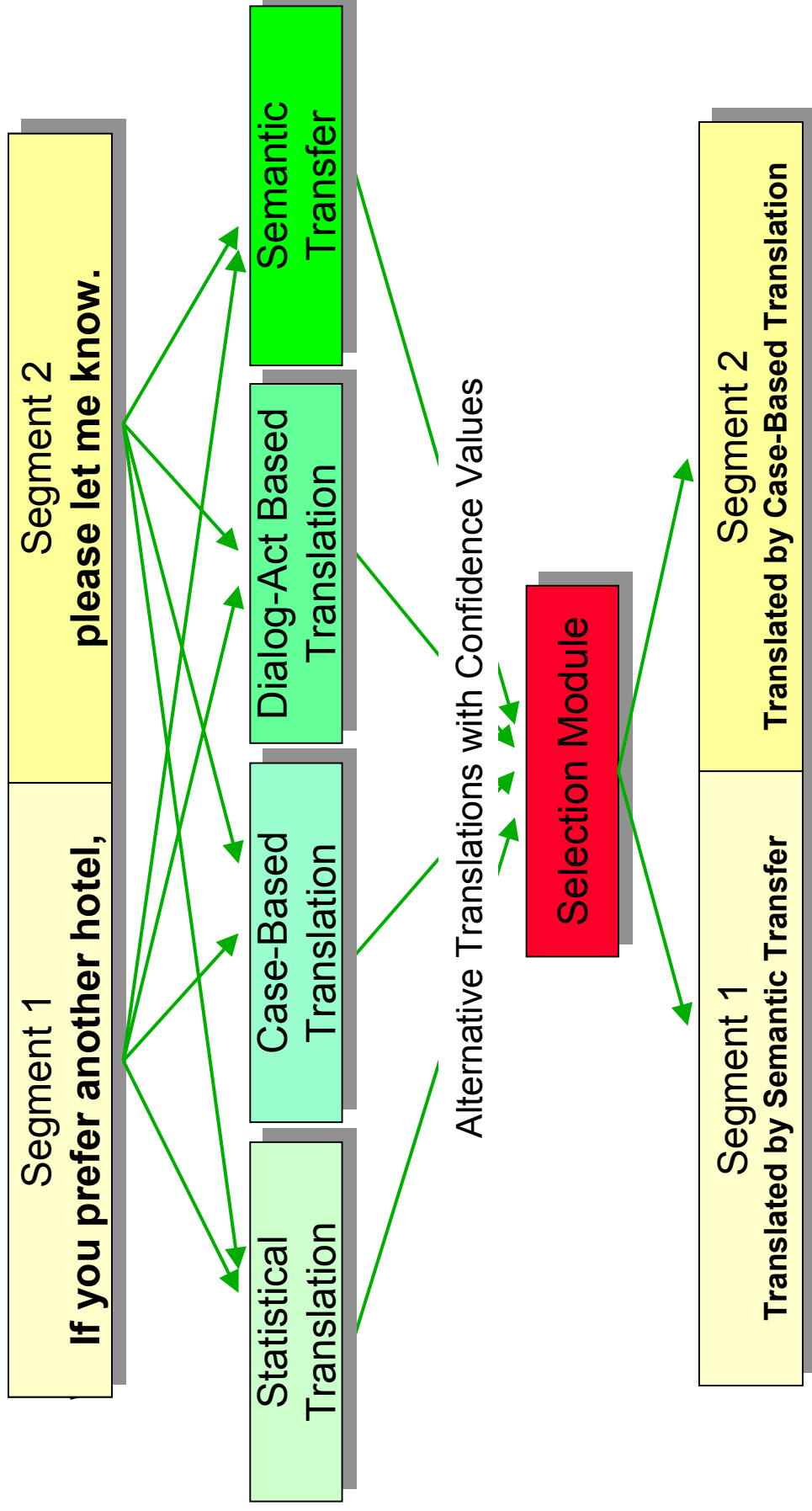
```
{SpeechAct:begin}{SpeechActType: Inform}{Language:English}{Utterance:begin}  
{SentenceType:Aussagesatz}{WordClass:N}Verbmobil{WordClass:AUX}is {WordClass: DET-ART}  
a{Prominence:18,2,3,0,15,0} {WordClass:ADJ}speaker_independent{WordClass:N}  
system{BorderProminence:5} {WordClass:CONJ-SYN}that {Prominence:15}{WordClass:V}offers  
{Prominence:6,18,4,0,13,4}{WordClass:N}translation_assistance{BorderProminence:2}  
{WordClass:PREP-SYN}in {Prominence:18,6,4,0,16,0}{WordClass:N}dialog {WordClass:N}situations  
{Utterance:end}
```

# Selection and Speech Synthesis



- **Task:**  
Select the “best” translation out of all deep and shallow translation paths
- **Input:**  
Translations (text or content)
- **Method:**  
Learning inequalities
- **Result:**  
Selected Translation (text or content)
- **Benefit:**  
Use the expertise of all translation paths for a particular utterance
- **Responsible:**  
TU Berlin

# Integrating Deep and Shallow Processing



**Selection is a difficult business:**

- **confidence values are difficult to compare**
  - probabilistic vs. knowledge based approaches
  - no bird's eyes view possible
- **re-training necessary after changes in the engines**
- **training data must be produced**

- **Task:**  
Synthesize the translation
- **Input:**  
text or content
- **Method:**  
Multilevel selection and concatenation  
of speech units from large speech  
corpora

- **Result:**  
Audio signal
- **Benefit:**  
“End of the chain” of the speech-to-  
speech system
- **Responsible:**  
Universität Bonn  
TU Dresden  
Universität Bochum  
Daimler Chrysler

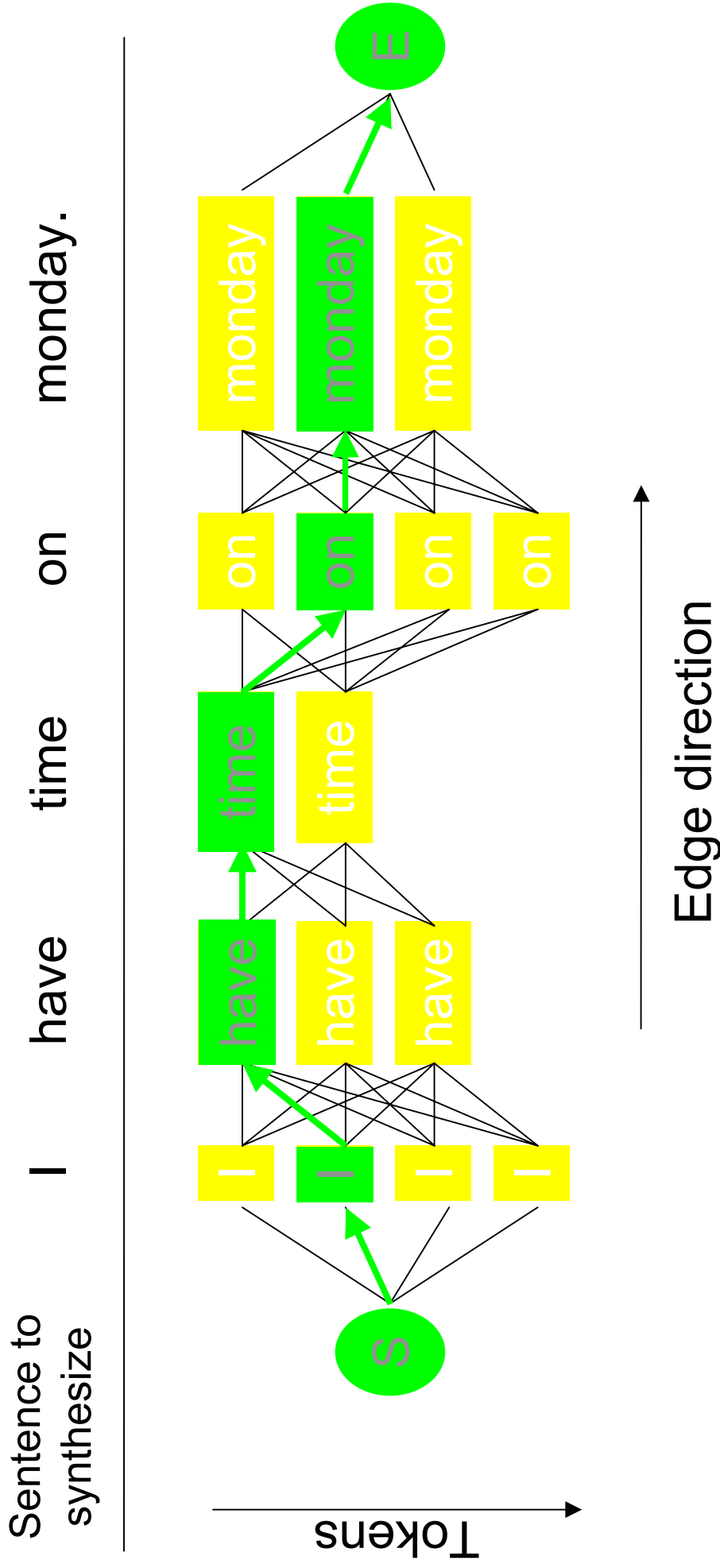
# Different Types of Synthesis

- *Text-to-Speech (TTS)*: reading machine from arbitrary text in orthographic form. Unlimited domain. The machine does not know what it is saying.
- *Concept-to-Speech [or content-to-speech] (CTS)*: spoken out-put from a database inquiry or from a dialog system. The input of the synthesizer comes from a semantic representation via a generation module. The machine should have full knowledge of what it is saying.
- *Reproductive Speech Synthesis*: spoken output from pre-recorded samples. For strictly limited domains.

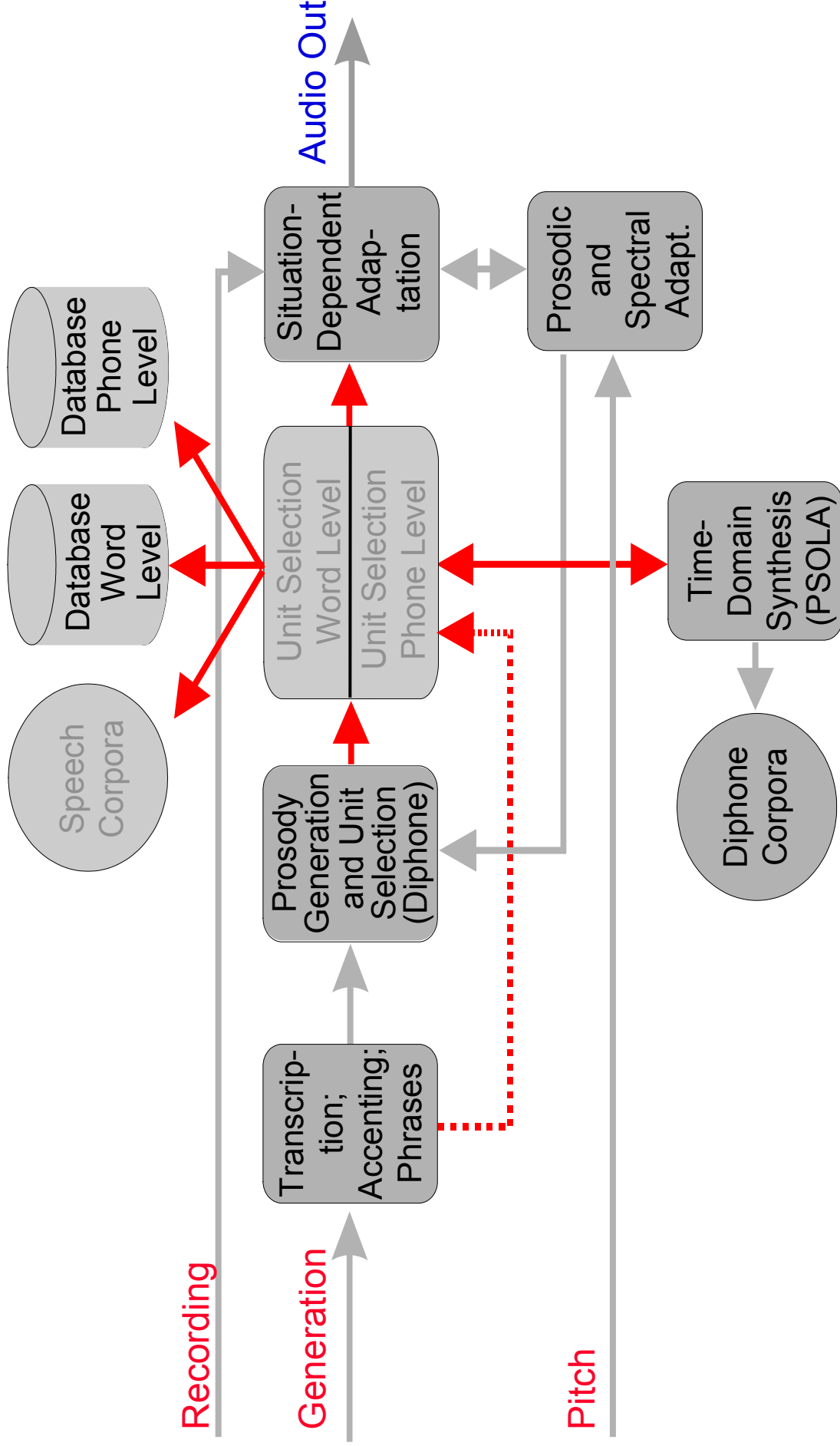
- Target utterances are synthesized from a corpus of utterances from within the domain.
- All units – whatever they are – have multiple instances in the corpus.
- No predefined units: the unit selection algorithm selects contiguous chunks of speech from the data base – the longer, the better.
- When units of word size and above are applied, much of the natural prosody is preserved.
- Problem: **coverage**. Words not in the database cannot be synthesized in this way.

# Unit Selection Algorithm

Verbmobil



- **Word** is the central unit and the starting point for all processing.
- Only if no suitable instance of a word is available in the database, an algorithm is invoked that composes a word from subword units which are currently phones.
- The principal strategy on both the word and the sub-word levels is to concatenate chunks that are as long as possible (up to a whole sentence).
- Like in CHATR, no prosodic manipulation is performed in this synthesis.
- In principle each word is needed in up to three positions (initial/medial, final declarative, final interrogative) and in both accented and unaccented mode.
- For Verbmobil this would mean that we need about 80000 word tokens to be recorded (which is prohibitive).
- Good coverage is reached by a selection of typical phrases from within the domain (dialogs from the Verbmobil dialog database).
- Additional utterances realize frequent words in relevant contexts (e.g., opening phrase, names of big cities).



# Verbmobil From a Software Engineering Point of View

**System Design and Software Integration**



## The goal

- Build an integrated system

## The situation

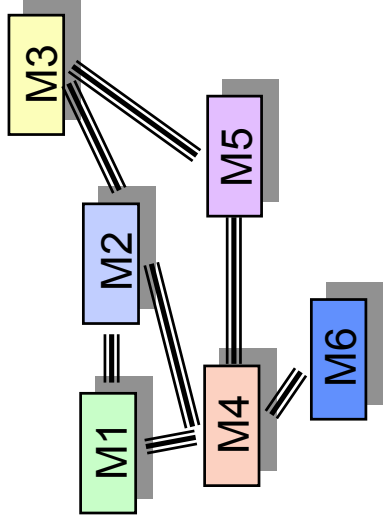
- Researchers do research
- Using different programming languages
- Researchers don't want to be bothered with technical details

## The solution

- Introducing: the **System Group**
- Maximal technical support for the researchers/developers

## Verbmobil I

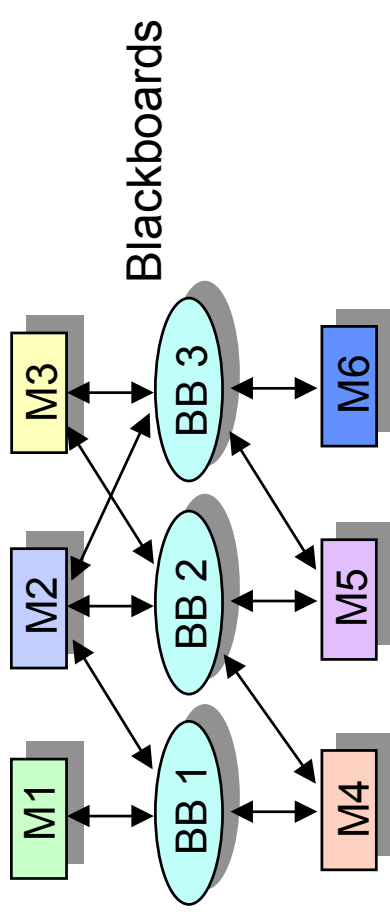
### Multi-Agent Architecture



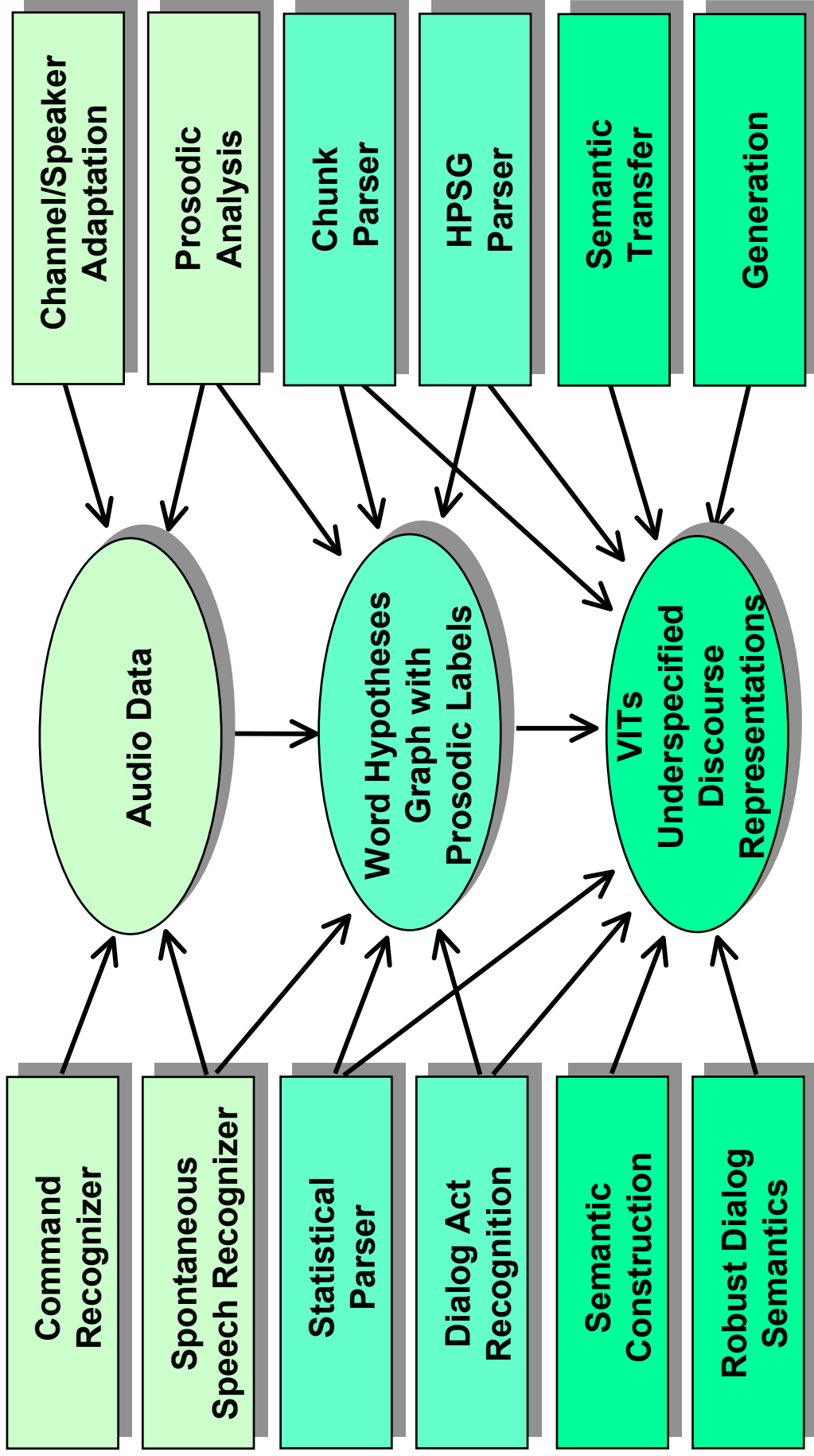
- Modules know all communication partners
- Direct communication between modules
- Reconfiguration difficult
- Software: ICE and ICE Master
- Basic Platform: PVM

## Verbmobil II

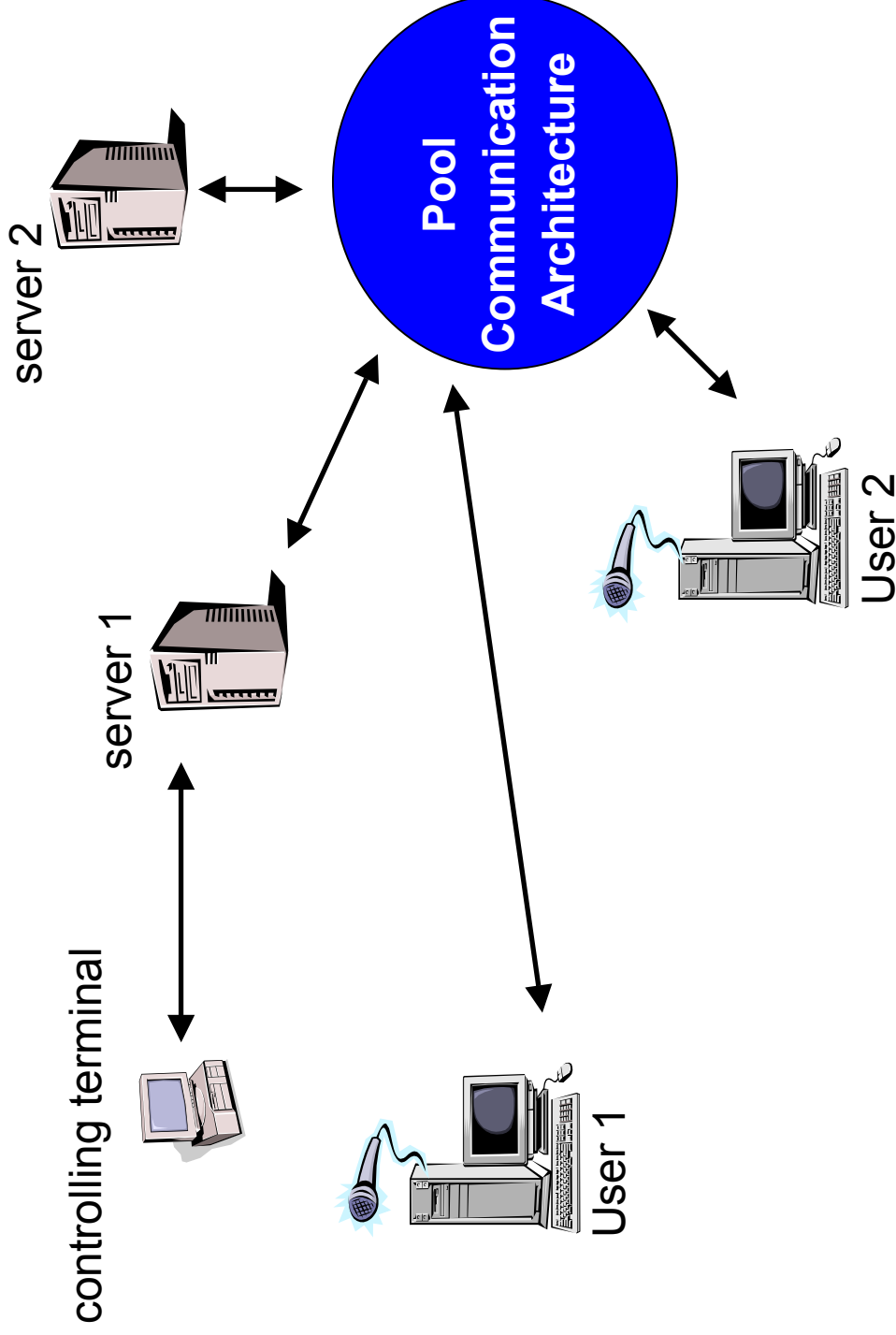
### Multi-Blackboard Architecture



- Modules know their I/O data pools
- No direct communication between modules
- **198 blackboards vs. 2380 direct comm. paths**
- Reconfiguration easy
- Several instances of one module/functionality
- Software: PCA and Module Manager
- Basic Platform: PVM



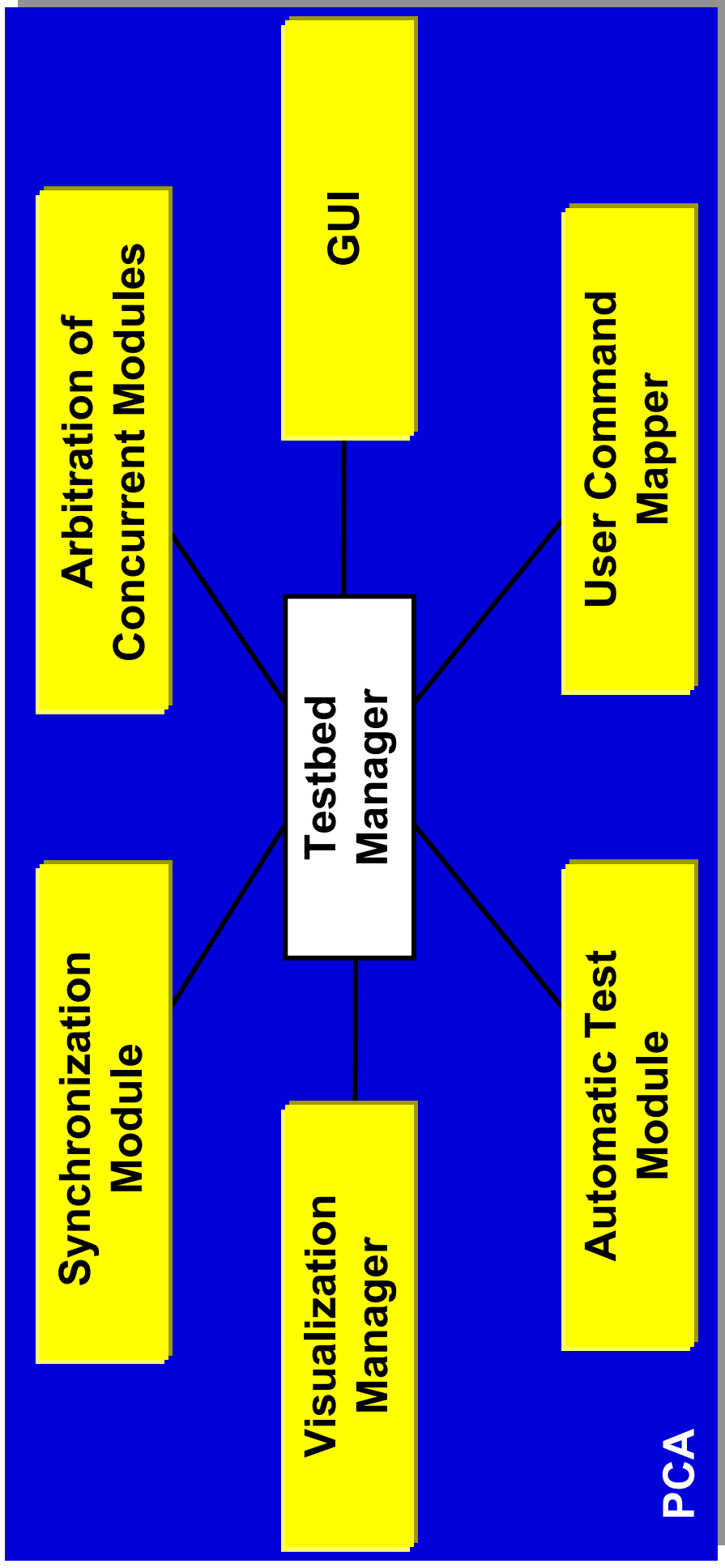
# Distributed Execution Supports Distributed Development



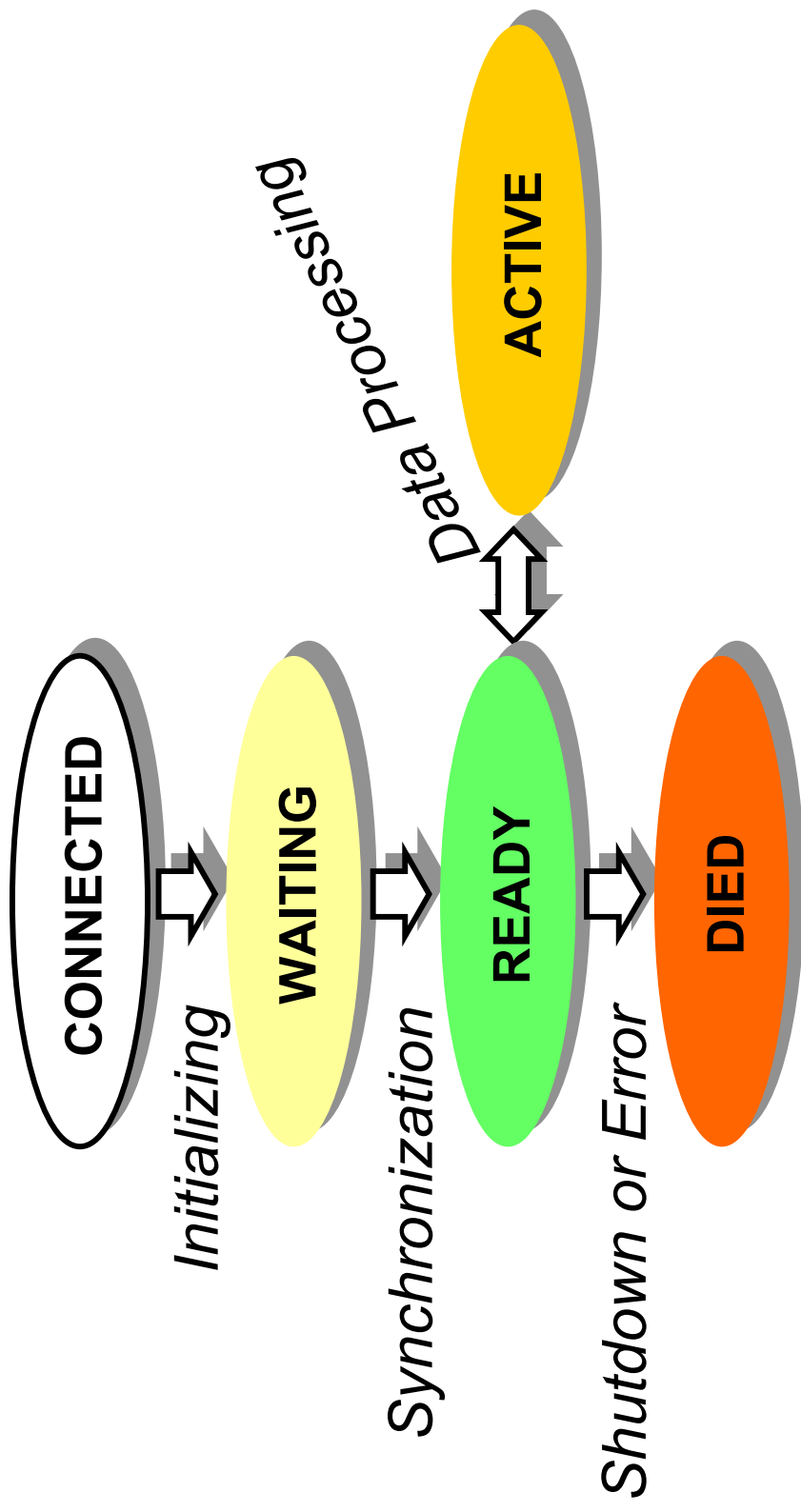
## Integration framework (**Testbed**) with

- **common communication mechanism for all used programming languages (C, C++, Lisp, Prolog, Java, Fortran, Tcl/Tk)**
- **Narrow interface for all used programming languages**
- **Overall system control infrastructure**
- **Standards on various levels**
  - Installation
  - Compilation
  - Communication formats between modules
  - ...
- **Toolbox for recording, replaying, testing, inspecting data exchanged between modules, ...**

# The **Testbed** is the Integration Framework for the Verbmobil System



# The Testbed controls the System: Module States



# The GUI- Visualization and Debug Tool

Verbmobil

The screenshot displays the Verbmobil GUI, which is divided into several functional areas:

- Top Panel:** Contains the 'bmb+f' logo and a menu bar with 'File', 'Modules', 'Options', 'Debug', 'Actions', 'Repeat\_Synthesis', 'Go', and 'Stop'.
- Workflow Diagram:** A central flowchart with yellow boxes labeled 'Dialog Semantics', 'Transfer', and 'Generation' connected by arrows. Below this, a vertical stack of boxes includes 'Dial', 'Stati', 'Case-', and 'Telephone'.
- Control Panel (Left):** Features a 'Pool Selection Filter' with options for 'word lattice', 'vit', and 'audio data'. It also includes language selection buttons for 'english', 'german', and 'japanese', and a 'Pool Selection' list with entries like 'recognized.Command.English' and 'recognized.Hypothesis.English.Lattice'.
- Control Panel (Right):** Includes fields for 'Turn Nr.', 'End & Request new', 'Ac. Chan.', 'Input Lang.', 'Start', and 'End'. It also has a 'Reading' section with 'Cur. Lang.', 'End of Turn', and 'Sender' fields, and a 'Textual SID' field.
- Visual Module Control (Bottom):** A window titled 'visual module control' showing a table of 'available modules' and 'selected modules'.
 

Module	Status
synthger_adapt	ready
synthger_prosgen	ready
synthger_timesynth	ready
synthger_transcription	ready
synthger_unitsel	ready
toptrans	not started
transfer	ready
user_command_mapper	ready
vim	ready
wismoc	ready
ltrans	ready

 Below the table are settings for 'transfer' (Host: serv-101, Startup: /V1\_2000\_exp/bin/transfer) and 'toptrans' (Host: serv-102, Startup: default).
- Bottom Bar:** A navigation bar with buttons for 'Cancel', 'Microphone 1', 'Microphone 2', and 'Telephone'.

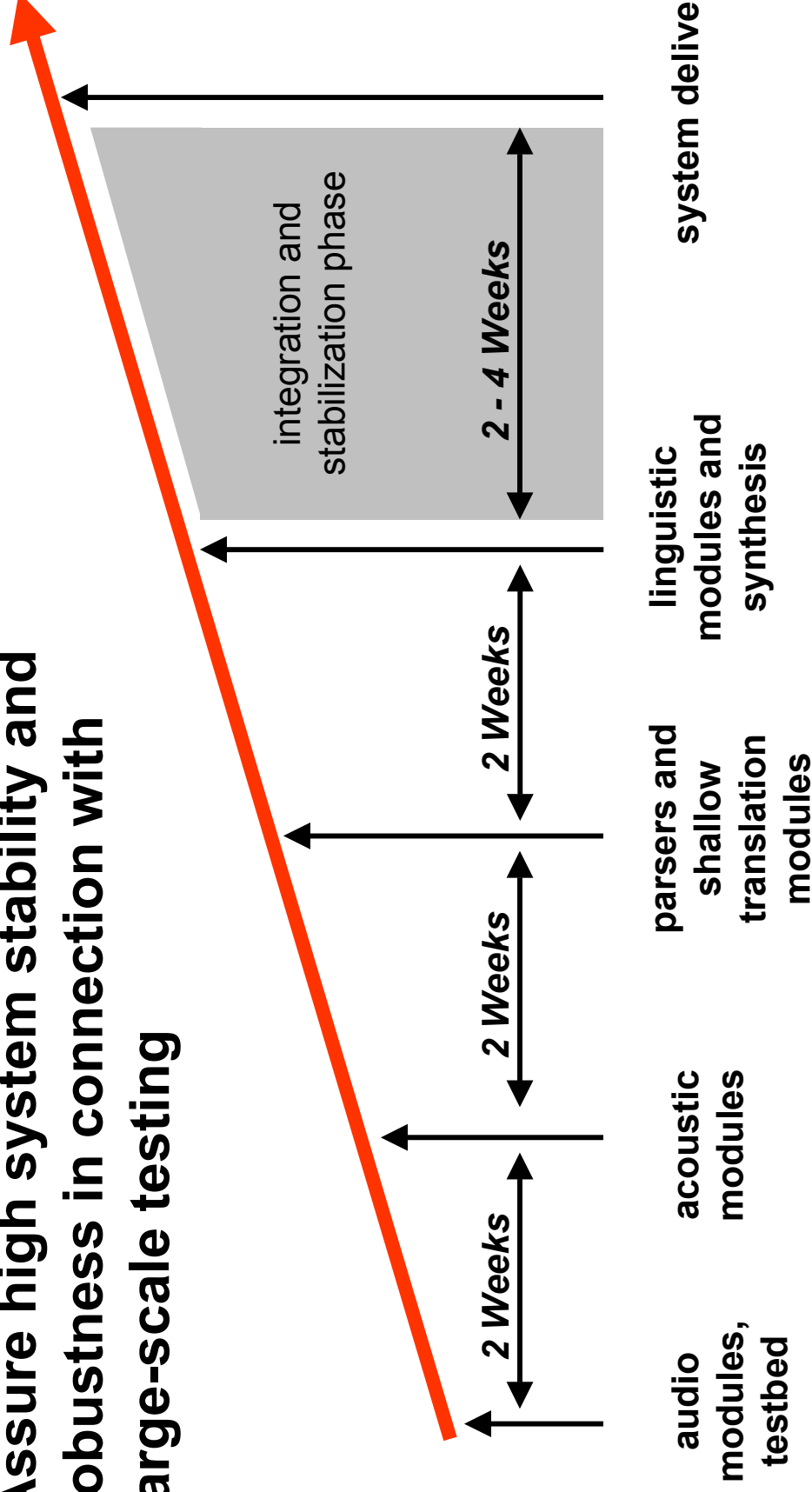
.... and much more



# Support from the System Group (2): Regular Integration Cycles

Verbmobil

**Assure high system stability and  
robustness in connection with  
large-scale testing**



# Human Factors



**8 years is a long time, especially since the invention of Internet time**

**1993**

- “You will need special hardware!”
- “1500 words speaker independent is impossible!”
- “Aren’t your goals unrealistic?”

**2000**

- “Does it run on my notebook?”
- “Only 10 000 words?”
- “Why can’t it also translate in the domains X, Y, and Z?”

**but**

**it is a unique chance for**

- **large scale, continuous research and development**
- **training people, collaborating, gaining experience**
- **collecting and annotating data**

# Management Challenges

## The goal

- Build an integrated system

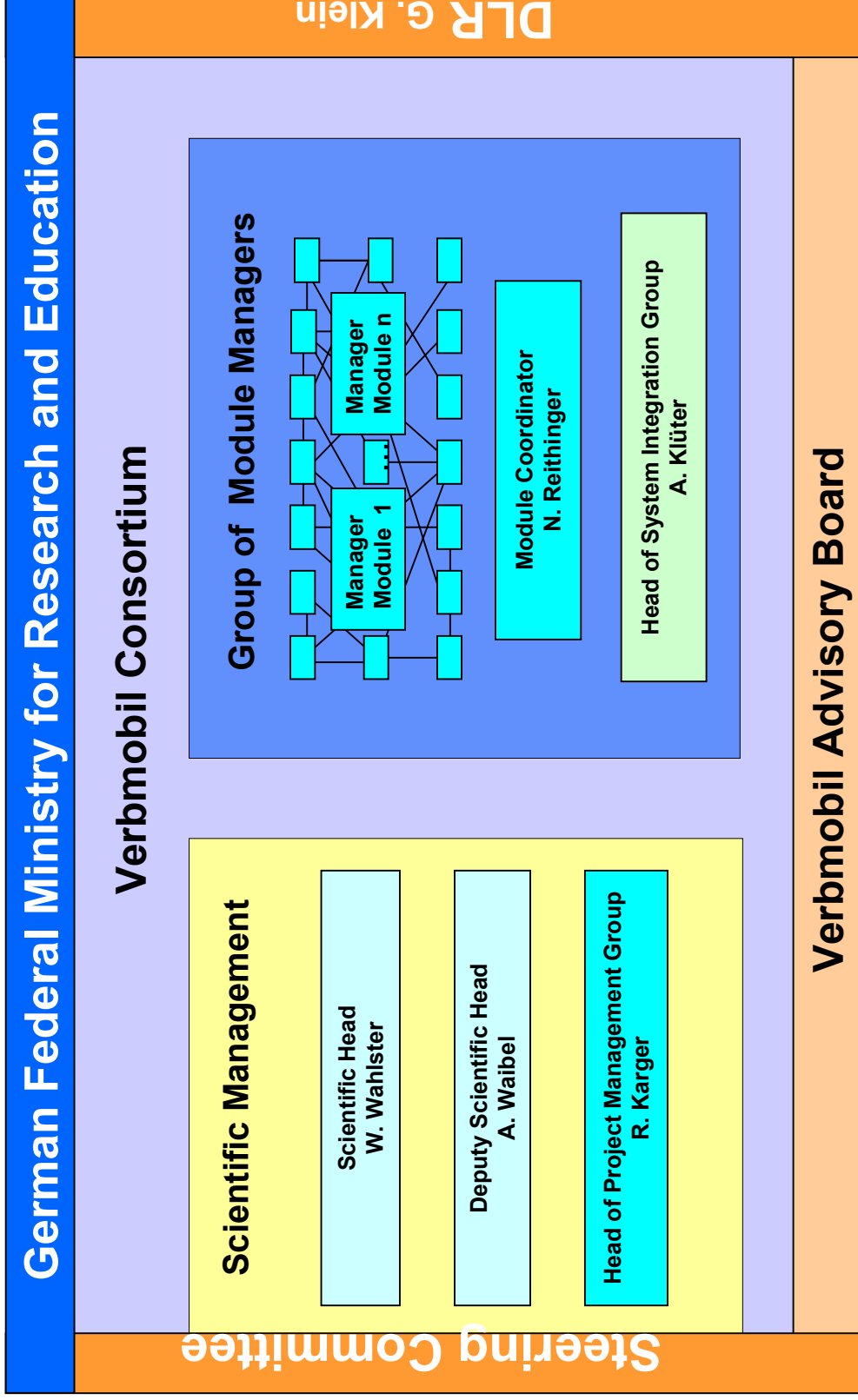
## The situation

- Partners distributed and pretty independent
- Great variation in project and background experience
- Adjustment of project plan and goals over time needed

## The solution

- Define a flat management structure
- Create a group spirit

# Project Organization



- **Have technical hands on experience**
- **Responsible for one module, even if it is developed at different sites**
- **Volunteers (sort of ...)**
- **Meet regularly, despite e-mail, phone and other devices**
- **Define next milestones**
- **Define data and software integration plans**

**Module coordinator coordinates the efforts and is the link to the scientific management**

- **21.02. Delivery of CeBit system** • **09.05. Delivery Verbmobil System 1.0**
- **21.02. - 30.04. Optimization phase** • **starting 09.05**
  - **15.03. - 28.04. End-To-End**
    - **speech recognizer evaluation**
    - **turn evaluation**
- **27.03. - 07.04. Workshop Deep Processing**

- **The group of module managers is a Good Thing™**
- **Common goals motivate**
- **Friendly peer pressure works most of the time**
- **Early problem detection and resolution in most cases**
- **Regular integration cycles focus and motivate**

## **Proactive consensus management (PCM)**

**You are invited to get more information at the ...**

# Verbmobil-Symposium

**30.7.2000, 10:30-18:00**

**Saarbrücken, Kongresshalle**

## Programm

(Keine Teilnahmegebühr)

Zeitraaster für das Verbmobil-Abschlusssymposium

Datum: 30.07.2000

Ort: Neue Congresshalle Saarbrücken

- 10:30 - 10:35 Eröffnung
- 10:35 - 10:45 Grußworte des BMBF (B. Reuse, BMBF)
- 10:45 - 11:30 Verbmobil (W. Wahlster)
- 11:30 - 12:00 Präsentation des Verbmobil-Systems (R. Karger)
- 12:00 - 12:45 Spracherkennung und Prosodieanalyse  
(A. Waibel, E. Nöth)
- 12:45 - 13:30 Imbiss
- 13:30 - 14:15 Multilinguale Analyse (U. Block, H. Uszkoreit)
- 14:15 - 15:00 Symbolische und Statistische Übersetzung  
(C. Rohrer, H.Ney)
- 15:00 - 15:30 Kaffee
- 15:30 - 16:15 Generierung und Synthese (T. Becker, W. Hess)
- 16:15 - 16:45 Evaluierung der End-to-End-Übersetzungsleistung  
(W. v.Hahn)
- 16:45 - 17:00 Verlesen des schriftlichen Abschlussgutachtens
- 17:00 - 18:00 Podiumsdiskussion: Sprachtechnologie  
as



**Thank you for your**

**interest and attention !!**

