# TC-Star, Review Workshop, Luxembourg, May 29, 2007

## TC-Star: Statistical MT of Text and Speech

**Hermann Ney**

**Human Language Technology and Pattern Recognition**
**Lehrstuhl für Informatik VI**
**Computer Science Department**
**RWTH Aachen University**
**D-52056 Aachen, Germany**

## Contents

# TC-Star: Objectives

**central topic: translation in a speech-to-speech translation task**

- **spoken language translation (SLT):**
  - **to significantly push the technology using statistical and data-driven techniques**
  - **to study the speech related problems in SLT and the tight integration of recognition and translation**
- **automatic speech recognition (ASR)**
- **text-to-speech synthesis (TTS)**

# 1 Approaches and Projects for MT

**area: machine translation of written language**

- **knowledge-based approaches**
  - **explicit rules (lexical, syntactic, semantic)**
  - **human effort to write down these rules**
- **memory-based translation:**
  - **goal: control of terminology**
  - **table of (source,target) phrase pairs**
- **example-based translation**
  - **(huge) database of (source,target) phrases**
  - **add generalization components**
- **statistical translation**

- **Systran:**
  - **(pragmatic) rules, in combination with dictionnary**
  - **optimized over several decades**

# Statistical MT

**IBM (1989-1994):**
**– design and implementation of a statistical approach to MT**
**– based on positive experience in speech recognition**

**task:**

- **input: written language (unlimited domain, large vocabulary)**
- **Canadian Hansards: French $\longrightarrow$ English**

**experimental evaluation:**

- **performance criterion: human evaluation (fluency + adequacy)**
- **result: slightly worse (?) than Systran**

# Projects 1992-2003

**tasks:**

- **speech input**
- **limited domains:**
  - **travelling, tourism, appointment scheduling**
  - **vocabulary size: 10000 words**
  - **language pairs: Chinese, Japanese, German, ... ↔ English**

**examples:**

- **C-Star consortium**
- **German BMBF: Verbmobil**
- **European: Eutrans, Nespole!, PF-Star, LC-Star, ...**

**automatic evaluation measures:**

- **WER/PER and BLEU/NIST are widely accepted**
- **allow competitive evaluations (helpful for progress!)**

# Challenges for TC-Star

- **work on a real-life task:**
  - **– unlimited domain**
  - **– large vocabulary**
- **speech input:**
  - **– cope with disfluencies**
  - **– handle recognition errors**
- **sentence segmentation**
- **reasonable performance**

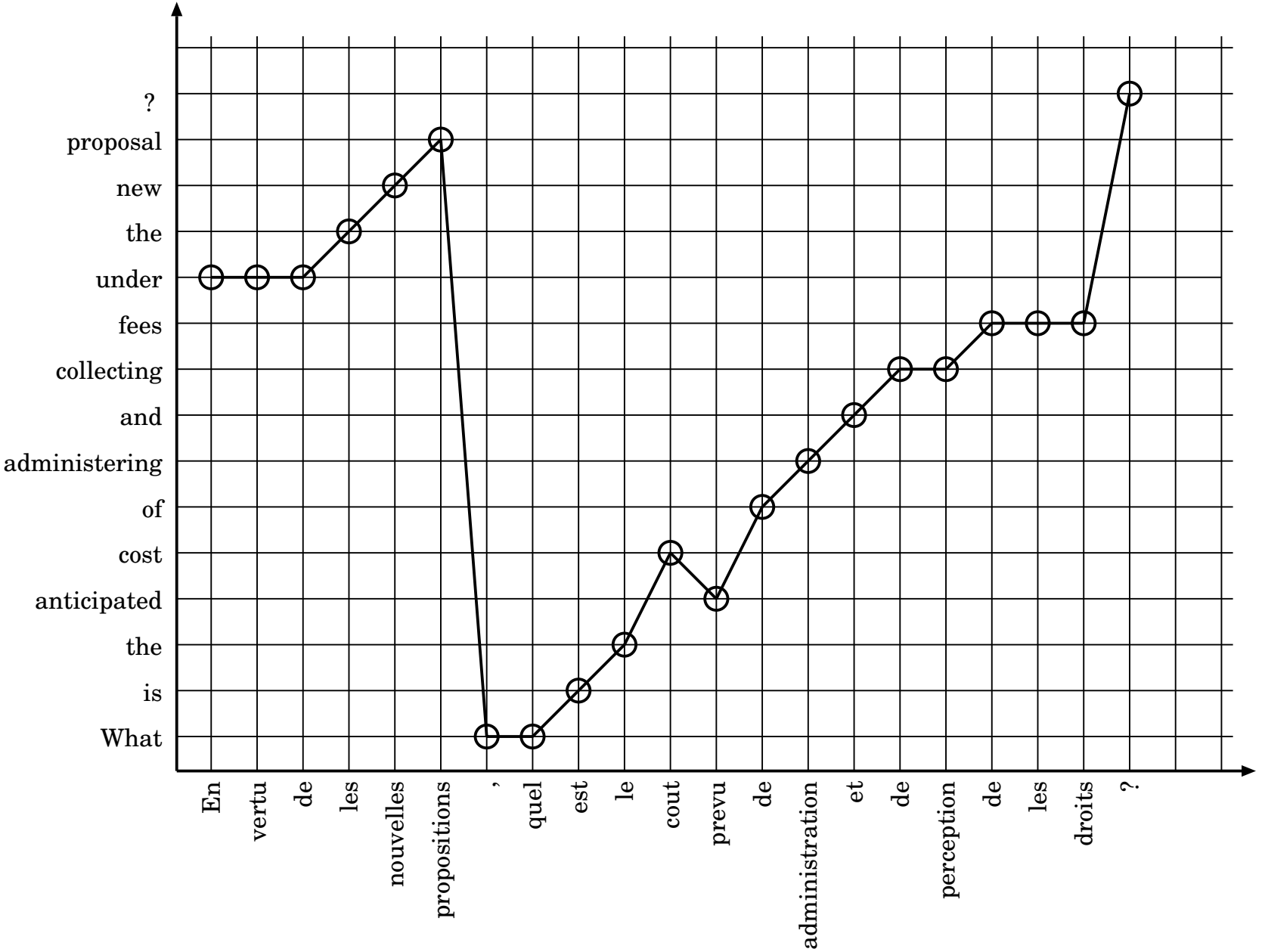
# 2 Statistical MT and TC-Star

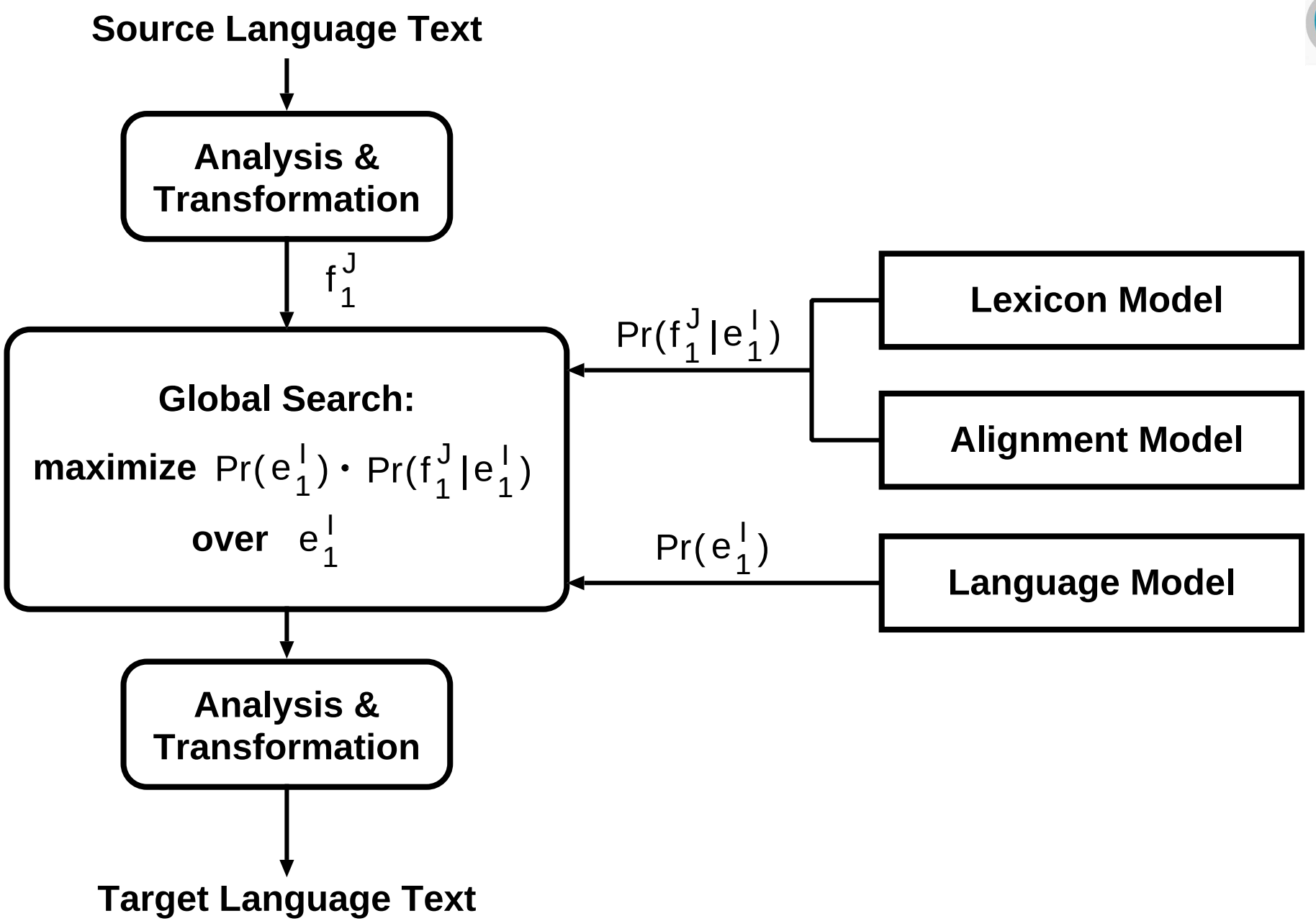
**Bayes decision rule:**

$$F \rightarrow \hat{E} = \arg\max_{E} \left\{ p(E) \cdot p(F|E) \right\}$$

- **distributions $p(E)$ and $p(F|E)$:**
  **– are unknown and must be learned**
  **– complex: distribution over strings of symbols**
  **– using them directly not possible (sparse data problem)!**
- **therefore: introduce (simple) structures by decomposition into smaller 'units'**
  **– that are easier to learn**
  **– and hopefully capture some true dependencies in the data**
- **example: ALIGNMENTS of words and positions: bilingual correspondences between words (rather than sentences) (counteracts sparse data and supports generalization capabilities)**

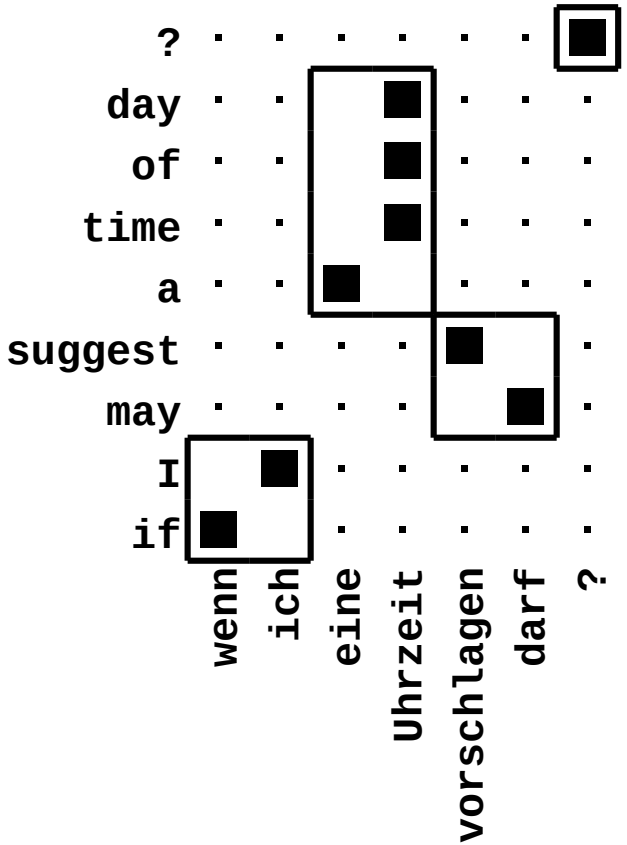
# Example of Alignment (Canadian Hansards)



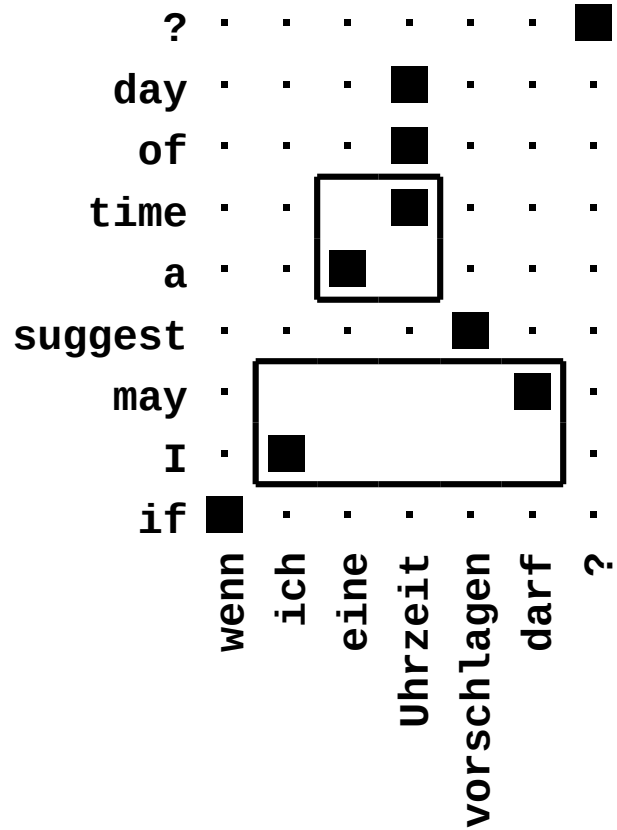

Source Language Text
Analysis & Transformation
$f_1^J$
Global Search:
maximize $Pr(e_1^I) \cdot Pr(f_1^J|e_1^I)$
over $e_1^I$
$Pr(f_1^J|e_1^I)$
Lexicon Model
Alignment Model
$Pr(e_1^I)$
Language Model
Analysis & Transformation
Target Language Text

# Phrase Extraction: Example

## possible phrase pairs:

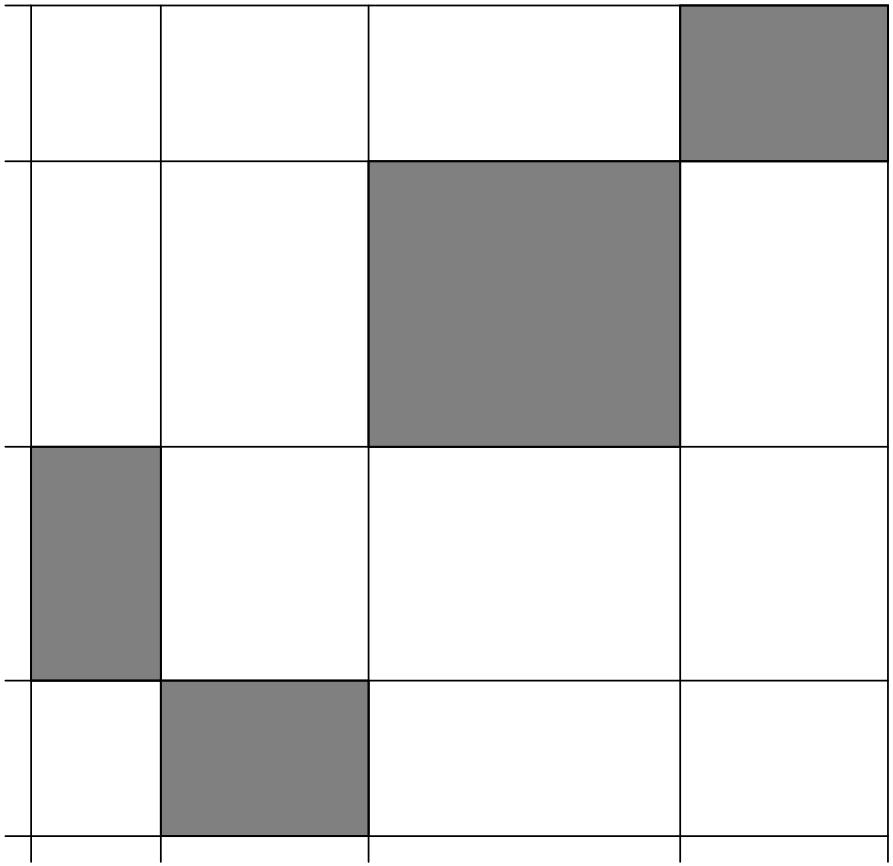



## impossible phrase pair:



# Translation Using Bilingual Phrases

**segmentation into two-dim. ’blocks’**
**with constraints:**
**no empty phrases, no gaps**
**and no overlaps**



**operations with interdependencies:**
**– find segment boundaries**
**– allow re-ordering in target language**
**– find most ’plausible’ sentence**

**similar to: memory-based and**
**example-based translation**

- **phrase-based approaches and extensions**
  **– extraction of phrase pairs, weighted FST, ...**
  **– estimation of phrase table probabilities**
- **improved alignment methods**
- **log-linear combination of models**
  **(scoring of competing hypotheses)**
- **use of morphosyntax**
  **(verb forms, numerus, noun/adjective,...)**
- **language modelling**
  **(neural net, sentence level, ...)**
- **word and phrase re-ordering**
  **(local re-ordering, shallow parsing, MaxEnt for phrases)**
- **generation (search):**
  **efficiency is crucial**

- **system combination for SLT**
  - **generate improved output from several MT engines**
  - **problem: word re-ordering**

- **interface ASR-SLT:**
  - **effect of word recognition errors**
  - **pass on ambiguities of ASR**
- **sentence segmentation**

- **public software**
- **software infrastructure for evaluation (UIMA)**
- **steady progress measured in yearly evaluations**
- **participation in international campaigns**

# 3 TC-Star: Results

**domain: EPPS = European Parliament Plenary Sessions**

**Training data:**

- **Sentence-aligned speeches and their translations**
- **Final text editions:**
  **from April 1996 to May 2006:**
  - **April 1996 – September 2004**
  - **December 2004 – May 2005**
  - **December 2005 – May 2006**
- **Verbatim transcriptions:**
  **from May 2004 to January 2006**

**Evaluation data 2007:**

**3 hours for each task (June-July 2006)**

# Evaluation: Participating Sites

**TC-Star participants:**

- **IBM: IBM Research Yorktown Heights**
- **IRST: ITC-IRST Trento**
- **LIMSI: CNRS Paris**
- **UKA: University of Karlsruhe (jointly with CMU)**
- **UPC: Universidad Politecnica de Catalunya**
- **RWTH: RWTH Aachen University**

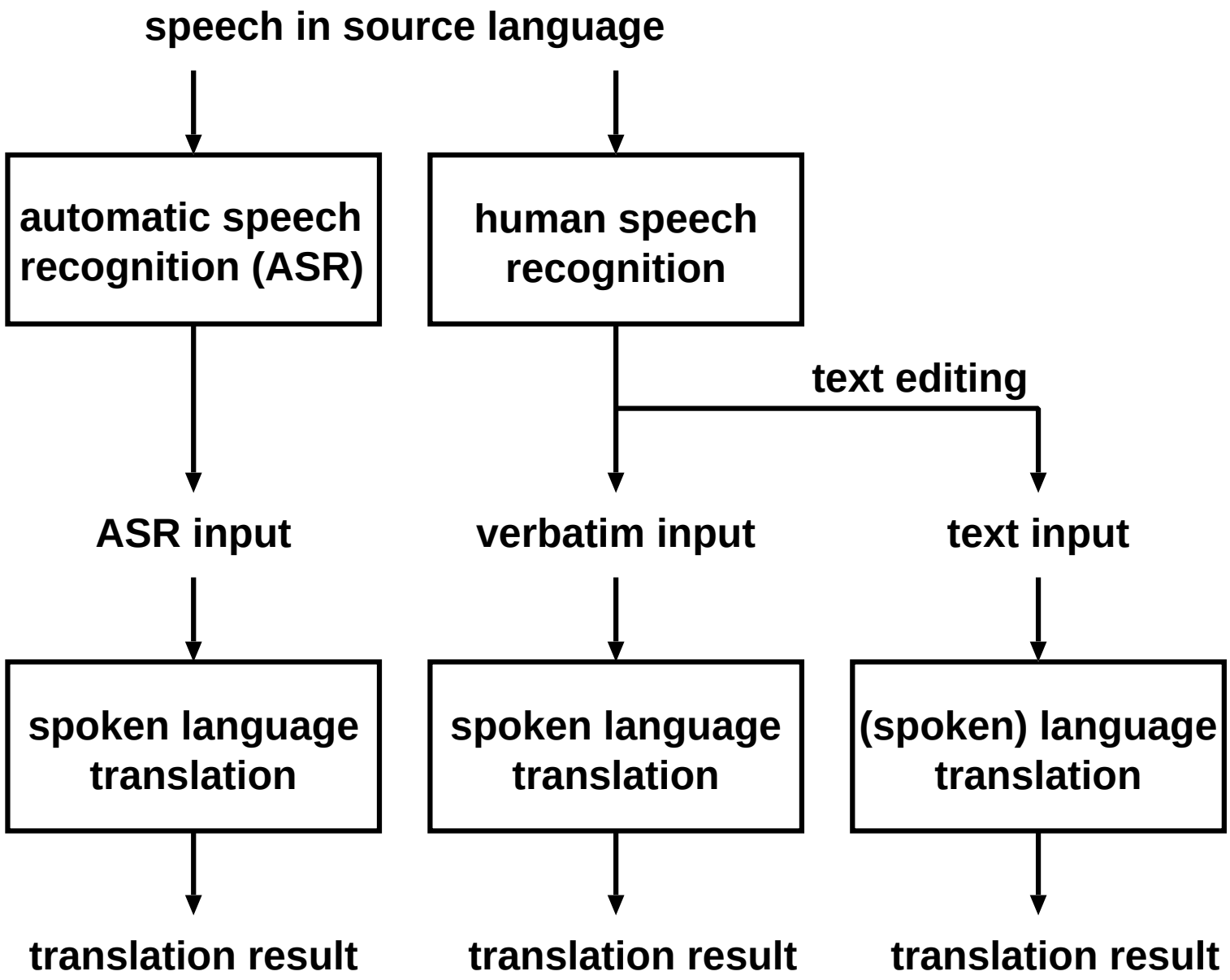
**external participants**

# Evaluation Measures

**automatic measures**
**based on single or multiple reference translations:**

- **WER = Word Error Rate (as in speech recognition):**
  **Levenshtein (edit) distance**
- **PER = Position independent word Error Rate (RWTH):**
  **ignore word order and count word errors**
- **BLEU = 'Bilingual Evaluation Understudy' (IBM)**
  **accuracy measure: geometric mean of n-gram precision + brevity penalty**
- **NIST = NIST variant of BLEU**
  **accuracy measure: arithmetic mean of n-gram precision + brevity penalty**

**remark: these automatic measures**
**correlate with human judgement (= adequacy + fluency)**


speech in source language
automatic speech recognition (ASR)
human speech recognition
text editing
ASR input
verbatim input
text input
spoken language translation
spoken language translation
(spoken) language translation
translation result
translation result
translation result

# Evaluation 2007: Spanish → English

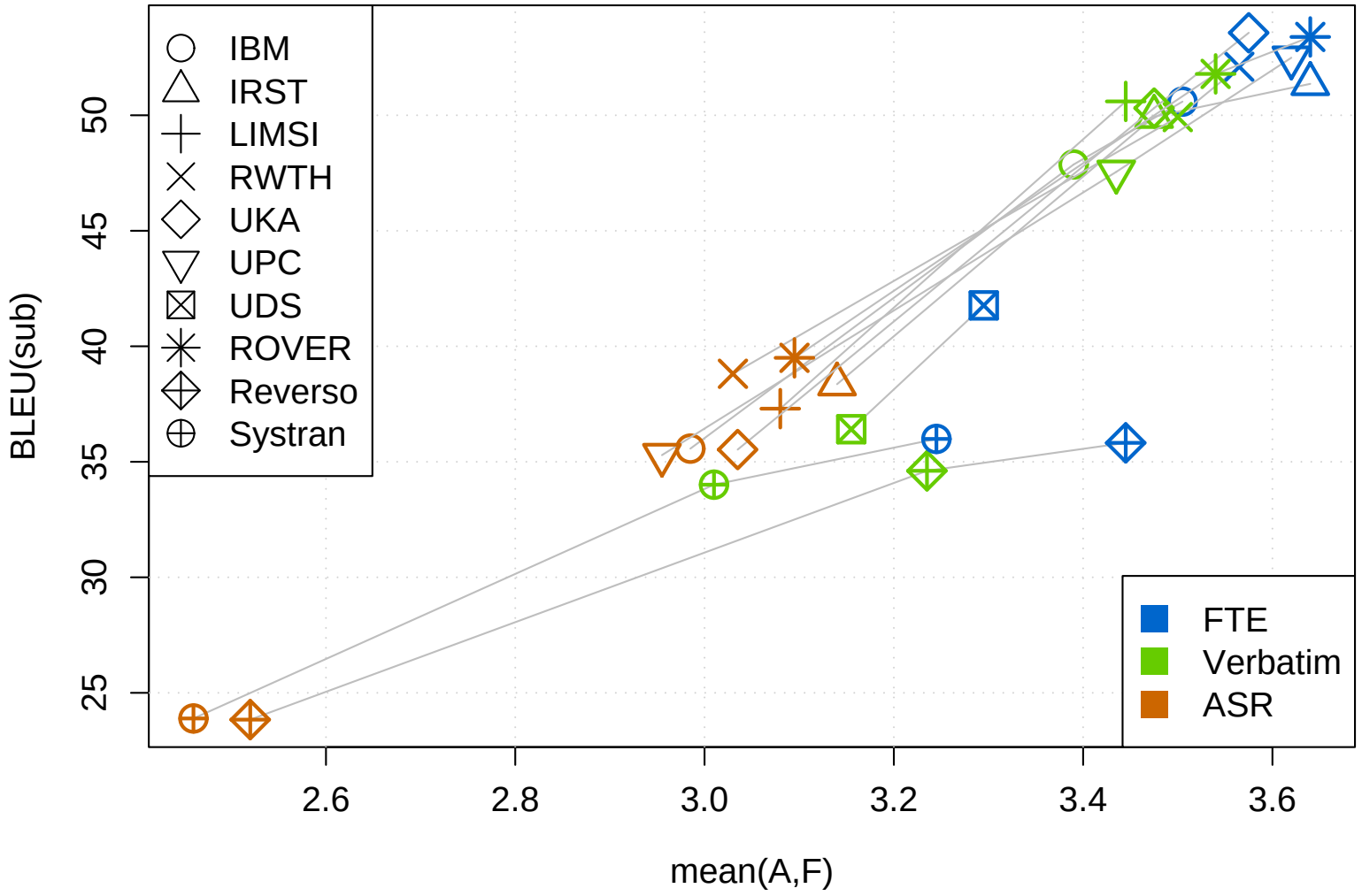
three types of input to translation:

- ASR: (erroneous) recognizer output
- verbatim: correct transcription
- text: final text edition
  (after removing effects of spoken language: false starts, hesitations, ...)

best results (system combination) of evaluation 2007:

| Input | BLEU [%] | PER [%] | WER [%] |
|---|---|---|---|
| ASR (WER= 5.9%) | 44.8 | 30.4 | 43.1 |
| Verbatim | 53.5 | 25.8 | 35.5 |
| Text | 53.6 | 26.7 | 37.2 |

# E → S (Text) 2007: Human vs. Automatic Evaluation



# English → Spanish: Human vs. Automatic Evaluation

**observations:**

- **good performance:**
  - **BLEU: close to 50%**
  - **PER: close to 70%**
- **fairly good correlation**
  **between adequacy/fluency (human) and BLEU (automatic)**
- **degradation:**
  **from text to verbatim: no or small**
  **from verbatim to ASR: ΔPER corresponds to ASR errors**

# Three-Year Improvements

**measure improvements over time on the 2007 eval data:**

- **2004: initial system**
- **2005: first evaluation**
- **2006: second evaluation**
- **2007: third evaluation**
- **2007: system combination**

**experiments:**

**relative improvement in BLEU by 40-60%**

# RWTH System over Time

**translation performance: BLEU[%] and relative improvment [%]**

| | S to E | | | | E to S | | C to E | |
|---|---|---|---|---|---|---|---|---|
| | EPPS | | CORTES | | | | | |
| | BLEU | impr. | BLEU | impr. | BLEU | impr. | BLEU | impr. |
| 2004 | 38.6 | – | 33.3 | – | 33.3 | – | 15.1 | – |
| 2005 | 47.8 | 23.8 | 40.6 | 21.9 | 46.1 | 38.4 | – | – |
| 2006 | 48.6 | 25.9 | 41.2 | 23.7 | 49.8 | 49.5 | – | – |
| 2007 | 51.2 | 32.6 | 44.6 | 33.9 | 52.5 | 57.7 | 25.3 | 67.5 |
| 2007 SysComb | 53.6 | 38.8 | 47.2 | 41.7 | 55.2 | 65.8 | – | – |

| | |
|---|---|
| **Source** | **Yo me ciño al texto que usted ha presentado, especial-mente . . .** |
| **2005** | **I am the text that you have tabled, especially . . .** |
| **2006** | **I am the text that you presented, in particular . . .** |
| **2007** | **I am holding on to the text that you have presented, especially . . .** |
| **SysComb** | **I am limiting myself to the text that you have presented, especially . . .** |
| **Reference** | **I am limiting myself to the text you submitted, especially . . .** |

| | |
|---|---|
| **Source** | **. . . han suscrito acuerdos con el Gobierno español para un uso de las lenguas cooficiales de España en sus actividades.** |
| **2005** | **. . . have signed agreements with the Spanish Government to use of languages citizens of Spain in its activities.** |
| **2006** | **. . . have signed agreements with the Spanish Government for a use of the languages means of Spain in its activities.** |
| **2007** | **. . . have signed agreements with the Spanish Government for a use of Spain's co-official languages in their activities.** |
| **SysComb** | **. . . have signed agreements with the Spanish Government for a use of the co-official languages of Spain in their activities .** |
| **Reference** | **. . . signed agreements with the Spanish Government concerning the use of Spain's co-official languages in their activities.** |

# 4 Other Projects and Evaluation Campaigns

**IWSLT: Int. Workshop on Spoken Language Translation**
**(organized by C-Star consortium)**

**task:**

- **input: SPOKEN language**
- **travelling and tourism: vocab.size = 10000 words**
- **language pairs: Chinese, Japanese, Arabic $\leftrightarrow$ English**
- **performance criterion: WER/PER, BLEU/NIST and human evaluation**

**experimental results (IWSLT 2006):**
**best performance by TC-Star systems**

# US-DARPA project GALE

**GALE: global autonomous language exploitation**

- **tasks: speech recognition, language translation, information extraction**

- **input: speech and text**
- **languages: Arabic, Chinese and English**
- **speech: broadcast news and conversations**
  **text: newswire and newsgroup**
- **output: result of distillation (= information extraction)**

**approach:**

- **three teams: IBM, BBN, SRI**
- **data and corpora: LDC**
- **each team: fully-fledged system with all components**
- **regular evaluations**

# NIST MT Evaluation

**NIST MT evaluation:**

- **written text**
- **Arabic and Chinese**
- **public evaluation**
- **measure: BLEU**

| Rank | Chinese NIST-06 | BLEU[%] Average | NIST (65%) | GALE (35%) |
|---|---|---|---|---|
| 1 | Information Sciences Institute | 27.0 | 33.9 | 14.1 |
| 2 | Google | 26.7 | 33.2 | 14.7 |
| 3 | Language Weaver | 25.9 | 32.8 | 13.0 |
| 4 | RWTH Aachen University | 23.8 | 30.2 | 11.9 |
| 5 | Institute of Computing Technology, CAS, Beijing | 23.1 | 29.1 | 11.9 |
| 6 # | University of Edinburgh | 22.6 | 28.3 | 12.0 |
| 7 | BBN Technologies | 22.2 | 27.8 | 11.7 |
| 8 | National Research Council Canada | 22.2 | 27.6 | 11.9 |
| 9 | ITC-irst | 22.1 | 27.5 | 11.9 |
| 10 | UMD-JHU | 21.6 | 27.0 | 11.4 |
| 11 | NTT, Japan | 20.8 | 26.0 | 11.2 |
| 12 | NICT, Japan | 19.8 | 24.5 | 11.1 |
| 13 | Carnegie Mellon University | 19.3 | 23.5 | 11.4 |
| 14 | Microsoft Research | 18.5 | 23.1 | 9.7 |
| 15 | Queen Mary University of London | 18.1 | 22.8 | 9.4 |
| 16 | HKUST | 17.0 | 20.8 | 9.8 |
| 17 | Universitat Politecnica de Catalunya | 16.8 | 20.7 | 9.3 |
| 18 | University of Pennsylvania | 16.0 | 19.6 | 9.2 |
| 19 * | Institute of Automation, CAS, Beijing | 15.5 | 18.9 | 9.0 |
| 20 | Institute of Software, CAS, Beijing | 14.8 | 18.2 | 8.6 |
| 21 | Language Computer | 14.7 | 18.1 | 8.1 |
| 22 | Xiamen University, Fujian, China | 12.9 | 15.8 | 7.5 |
| 23 # | Lingua Technologies Inc., Canada | 11.1 | 13.4 | 6.6 |
| 24 # | KCSL Inc., Canada | 4.0 | 5.1 | 2.0 |
| 25 | Kansas State University | 3.4 | 4.0 | 2.2 |

| Rank | Arabic NIST-06 | BLEU[%] Average | NIST (65%) | GALE (35%) |
|---|---|---|---|---|
| 1 | Google | 34.3 | 42.8 | 18.3 |
| 2 # | Applications Technology Inc. | 31.9 | 38.7 | 19.2 |
| 3 | IBM | 31.6 | 39.5 | 16.7 |
| 4 | Information Sciences Institute | 31.4 | 39.1 | 17.1 |
| 5 | RWTH Aachen University | 31.2 | 39.1 | 16.4 |
| 6 * | SRI | 30.0 | 37.4 | 16.1 |
| 7 | Language Weaver | 29.9 | 37.4 | 15.9 |
| 8 * | National Research Council Canada | 29.7 | 37.5 | 15.2 |
| 9 | NTT, Japan | 29.3 | 36.8 | 15.3 |
| 10 | BBN Technologies | 29.1 | 36.9 | 14.6 |
| 11 | ITC-irst | 27.7 | 34.7 | 14.8 |
| 12 | Sakhr Software Co. | 27.2 | 33.0 | 16.5 |
| 13 | Carnegie Mellon University | 26.8 | 33.7 | 13.9 |
| 14 | UMD-JHU | 26.5 | 33.3 | 13.7 |
| 15 # | University of Edinburgh | 26.1 | 33.0 | 13.1 |
| 16 | Queen Mary University of London | 23.6 | 29.0 | 13.5 |
| 17 | NICT, Japan | 23.3 | 29.3 | 11.9 |
| 18 | Language Computer | 22.1 | 27.8 | 11.3 |
| 19 | Universitat Politecnica de Catalunya | 21.9 | 27.4 | 11.5 |
| 20 | Columbia University | 19.4 | 24.7 | 9.6 |
| 21 | University of California Berkeley | 15.4 | 19.8 | 7.3 |
| 22 | The American University in Cairo | 12.2 | 15.3 | 6.4 |
| 23 | Dublin City University | 7.3 | 9.5 | 3.2 |
| 24 # | KCSL Inc., Canada | 4.0 | 5.2 | 1.8 |

# 5 Summary

- task: real-life, unrestricted domain
- fully automatic systems for spoken language translation
- complete chain: ASR, SLT, TTS
  - automatic segmentation of speech
  - interfaces to ASR and TTS
- progress monitoring by regular evaluations
  - steady improvement over time
  - external participants
- other projects and evaluation campaigns:
  - IWSLT
  - DARPA GALE
  - NIST-MT
- TC-systems: state of the art
  no superior technology around

# THE END