

**Latest Developments  
in  
Machine Translation  
Technology**

**John Hutchins**

**University of East Anglia, Norwich, England**

# Introduction

- not a comprehensive survey of all MT research in the past few years
- examples of projects only indicative of trends
- highlight new developments/approaches/methods
- no discussion of projects and systems well established in the late 1980s (e.g. Systran, Logos, EDR, ATR, Mu/JICST, METAL)

---

## Contents

MT in the 1980s

Rule-based MT (transfer, interlingua, formalisms)

Corpus-based MT (statistical, example-based)

Text corpora and alignment

Generation

Controlled, domain-specific, user-specific MT

Workstations

Eras of MT

Towards "third generation" systems

# **MT during the 1980s**

**at first MT Summit conference (1987):**

- **dominant frameworks**

  - indirect translation  
(transfer, interlingua)  
linguistics-based**

- **knowledge-based MT still innovative**

- **multinational-multilingual projects**

- **some well-established systems  
(Systran, Meteo, Weidner/Bravice,  
PAHO systems, Logos)**

- **many new systems/projects from  
Japanese companies  
(NEC, Toshiba, Fujitsu, Oki,  
Ricoh, Sharp, etc.)**

# **Rule-based MT**

## **Transfer-based (typical "second generation")**

- **three stages: analysis, transfer, synthesis**
- **abstract semantico-syntactic interfaces**
- **multiple levels/strata:**
  - morphology, syntax, semantics**
- **syntax-oriented, tree-transduction approach**
- **batch processing, post-edited**
- **little pragmatic/discourse information**

**projects: [Ariane, Eurotra], Eurolang (SITE, METAL),  
LMT 'Logic programming MT' (IBM centres)**

## **Interlingua-based**

- **two stages: analysis, synthesis**
- **abstract language-neutral representation**
- **multistratal: morphology, syntax, semantics**
- **semantics-oriented ('understanding')**
- **[knowledge bases]**

**projects: [DLT, Rosetta], PIVOT (NEC),  
Carnegie-Mellon University (KBMT89,  
KANT, CATALYST)  
ULTRA (New Mexico State University)  
MCC (Microelectronics and  
Computer Technology Corporation)  
UNITRAN**

# Formalisms

**Non-transformational**

**Constraint-based formalisms**

**Unification formalisms**

**Lexical-Functional Grammar**

**Logic programming (Q-systems, UNITRAN Prolog)**

**Definite Clause Grammar**

**Slot grammar (LMT)**

**Generalized Phrase Structure Grammar**

**Head-driven Phrase Structure Grammar**

**Categorial Grammar**

**Principles-based MT (Government-Binding Theory)**

**projects: ITS (Geneva), UNITRAN**

**Lexicalist approaches**

**[transfer rules = simple bilingual lexical equivalences]**

**projects: CRITTER, ACQUILEX,  
'Shake-and-bake' (Sharp)**

**Reversibility**

**projects: [Rosetta], ISSCO, CRITTER, UNITRAN**

# Tree transduction model (Eurotra)

Input text in L

Analysis

Grammar (1)  $\rightarrow$  Representation (1)

$\leftarrow$  Mapping rules T (1/2)

Grammar (2)  $\rightarrow$  Representation (2)

$\leftarrow$  Mapping rules T (2/3)

Transfer

Grammar (n)  $\rightarrow$  Representation (n)

$\leftarrow$  Mapping rules T (n/n')

Grammar (n')  $\rightarrow$  Representation (n')

$\leftarrow$  Mapping rules T (n'/n'-1)

Synthesis

Grammar (2')  $\rightarrow$  Representation (2')

$\leftarrow$  Mapping rules T (2'/1')

Grammar (1')  $\rightarrow$  Representation (1')

Output Text in L'

# Constraint-based formalism (LFG)

John likes Mary <--> Marie plait à Jean

*like, V:*

(↑PRED) = like <SUBJ, OBJ>

(τ↑PRED FN) = plaire <SUBJ, OBJ>

(τ↑AOBJ OBJ) = τ(SUBJ)

(τ↑SUBJ) = τ(↑OBJ)

*john, N:*

(↑PRED) = john

(τ↑PRED FN) = jean

*mary, N:*

(↑PRED) = mary

(τ↑PRED FN) = marie

TL f-structure:

PRED	plaire	
SUBJ	[PRED marie ]	
AOBJ	[OBJ [PRED jean ] ]	

Student is likely to work <--> Il est probable que l'étudiant travaillera

*likely, A:*

(↑PRED) = likely <XCOMP> SUBJ

(↑SUBJ) = (↑XCOMP SUBJ)

(τ↑PRED FN) = probable

(τ↑COMP) = τ(↑XCOMP)

*probable, A:*

(↑PRED) = probable <COMP>SUBJ

(↑SUBJ FORM) = il

(↑COMP COMPL) = que

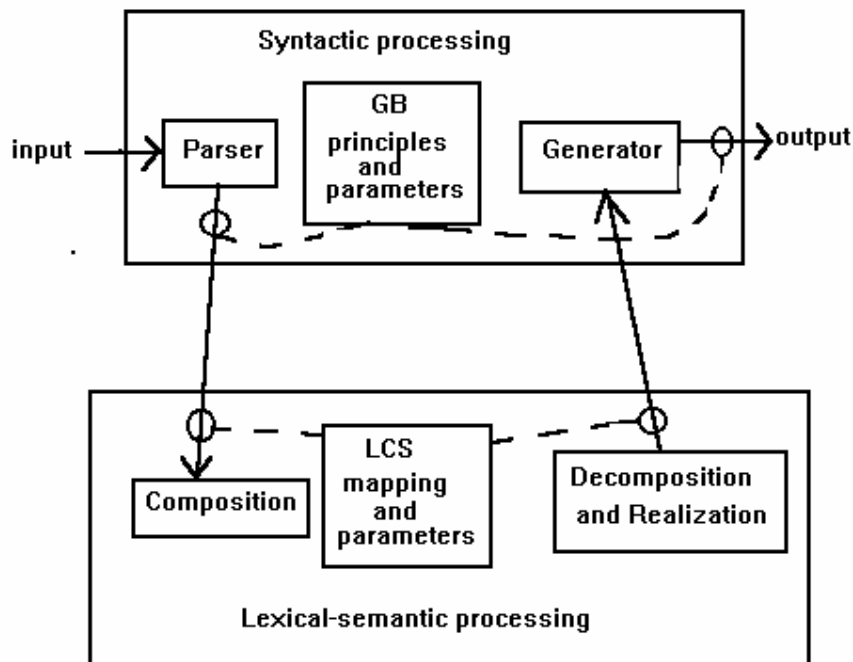
TL f-structure:

PRED	probable	
SUBJ	[FORM il ]	
	[PRED travailler ]	
COMP	[COMPL que ]	
	[SUBJ [...]]	

## Knowledge-based MT: lexical entry (CMU)

```
(find
  (make-frame
    +find-v1
    (CAT (value v))
    (STUFF
      (DEFN "to discover by chance, to come across")
      (EXAMPLES "found X in the bedroom"
        "found X sleeping upstairs")
        "found that X was sleeping on the porch")
      (MORPH
        (IRREG (*v+past* found) (*v+past-part* found))
      (SYN-STRUC
        (*OR* ((root $var0)
          (subj (root $var1)(cat N))
          (obj (root $var2)(cat N))
          ((root $var0)
            (subj (root $var1)(cat N))
            (xcomp(root $var2)(cat N)(form pres-part)))
          ((root $var0)
            (subj (root $var1)(cat N))
            (comp (root $var2)(cat V)(form fin))))))
      (SEM
        (LEX-MAP
          (%involuntary-perceptual-event
            (experiencer (value ^$var1))
            (theme (value ^$var2))))))
```

# UNITRAN system design



Example parameters:

Syntactic divergences: parameters

Constituent order *I have seen him* – *Ich habe ihn gesehen*  
 Null subject *I saw the book* – *(Yo) Vi el libro*

Lexical-semantic divergences: parameters

structural *X entered the house* – *X trat ins Haus hinein*  
 thematic *I like Mary* – *Mary gefällt mir*  
 categorial *I am hungry* – *Ich habe Hunger*  
 demotional *I like eating* – *Ich esse gern*

## Lexicalist approach ('Shake-and-bake')

monolingual lexical entry (English):

[12]	ORTHO    like		
	SEM        E1:	(like (E1), { role (E1, experiencer, X1), (role (E1, stimulus, Y1)	)   }   )
	ARG0      E1		
	ARG1      X1		
	ARG2      Y1		

monolingual lexical entry (Spanish):

[13]	ORTHO    gust-		
	SEM        E2:	{ gustar (E2) { role (E2, stimulus, X2) { role (E2, experiencer, Y2)	)   }   )
	ARG0      E2		
	ARG1      X2		
	ARG2      Y2		

bilingual lexical entry for *like-gustar*:

SPANISH [13]	SEM	[ ARG0 E ] ] ]
		[ ARG1 X ] ] ]
		[ ARG2 Y ] ] ]
ENGLISH [12]	SEM	[ ARG0 E ] ] ]
		[ ARG1 Y ] ] ]
		[ ARG2 X ] ] ]

### Reversibility (CRITTER)

eat <--> manger

miss (1: X, 2: Y) <--> manquer (1: Y', 2: X')

walk (inv-1: across (2: X)) <-->

traverser (2: X, inv-1: \$manner (2: à\_pied))

## **General-purpose NLP systems**

ELU

(Environnement Linguistique d'Unification) (ISSCO)

KIELIKONE (Finland)

Core Language Engine (SRI, Cambridge)

PLNLP

(Programming Language for Natural Language  
Processing) (IBM):

[SHALT-2, C-SHALT, KSHALT, PORTUGA]

## Summary of trends in rule-based approaches

### mid-1980s

syntax-orientation  
complex transfer rules  
stratificational/multi-level  
representation/monostratal  
tree transduction [filters]  
analysis/transfer  
understanding/disambiguation  
uni-directional  
linguistic information  
databanks  
lexicon compilation

### mid-1990s

lexicalist orientation  
simple lexical transfer  
single  
constraints/unification  
generation  
style/quality output  
reversible  
lexical/conceptual  
lexicon acquisition

## **Corpus-based MT**

**a) the direct use of information derived from corpora for the analysis, transfer and generation of translations**

**b) the indirect use of corpora**  
- as sources of information for deriving or compiling lexical, grammatical and knowledge databases,  
- as sources of statistical information about source and target languages.

# Statistics-based MT

Alignment: (IBM Candide, using corpus of the Canadian Hansard)

The proposal will not now be implemented

Les propositions ne seront pas mises en application maintenant

---

English: *not*

French	Probability	Fertility	Probability
<i>pas</i>	.469	2	.758
<i>ne</i>	.460	0	.133
<i>non</i>	.024	1	.106
<i>pas du tout</i>	.003		
<i>faux</i>	.003		
<i>plus</i>	.002		
etc.			

---

Later modifications: syntactic transformations

e.g.

English questions:

*Has the store any eggs? -> The store has any eggs QINV*

English 'adverbs':

*John does not like turnips -> John likes do\_not\_M1 turnips*

French negation:

*Je ne sais pas -> Je sais ne\_pas*

French pronouns:

*Je vous le donnerai -> Je donnerai le\_DPRO vous\_IPRO*

## Example-based MT

DLT: Bilingual Knowledge Bank

the main fields	les principaux domaines
the following fields	les domaines suivantes
these two fields	ces deux domaines
the specialized fields	les domaines spécialisés
the para-medical fields	activités paramédicales
the magnetic fields	les champs magnétiques
the coal fields	les bassins-houilliers
the corn fields	les champs de blé

le livre <i>de</i> mon père ->	my father's book
un verre <i>d'eau</i> - >	glass of water
il est certain <i>de</i> réussir ->	certain to succeed
il est capable <i>de</i> résister ->	capable of resisting
il vient <i>de</i> Paris ->	he comes from Paris
le train <i>de</i> Paris ->	the train to/from Paris
il partit <i>de</i> nuit ->	he left at night
il partit <i>de</i> bonne heure ->	he left in good time
je suis âgé <i>de</i> trente ans ->	I am thirty years old

English ...*have a/an effect on*... -> French

have a direct effect on ->	ont une influence directe à
have a direct effect on ->	intéressent directement
have a direct effect on ->	ont eu une répercussion directe sur
has had a marked effect on ->	a largement influencé
had a positive effect on ->	s'est avérée positive dans
had a highly negative effect on [X] ->	[X] en auraient été gravement affectés
will have a decisive effect on ->	influencera de façon déterminante
would have a detrimental effect on ->	aurait de fâcheuses répercussions sur

## Example-based method with probabilistic scores

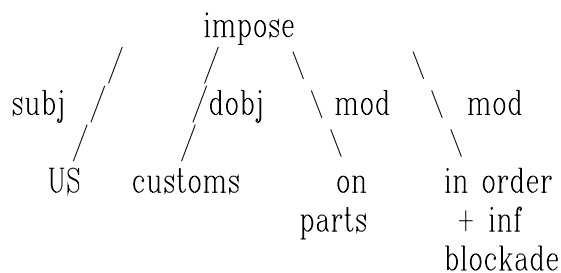
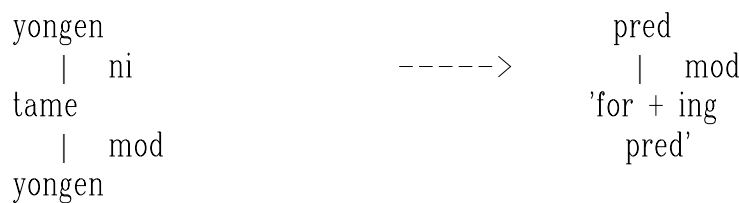
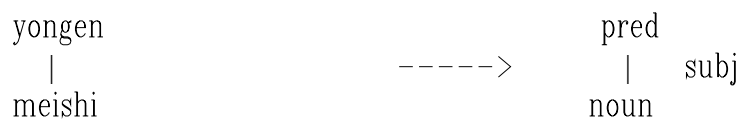
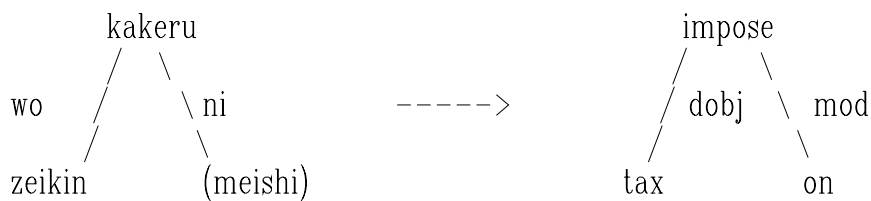
SL input (Japanese):

*US ga ... wo fusegu tame ni buhin ni kanzei wo kakeru*

			kakeru		
	ga	wo	ni	ni	
US		kanzei	buhin	tame	
					mod
					fusegu

similarities (with probabilities):

- |       |                               |                       |
|-------|-------------------------------|-----------------------|
| 5.988 | (meishi) ni zeikin no kakeru  | impose tax on (noun)  |
| 3.077 | (meishi) no saibin ni kakeru  | take (noun) to court  |
| 2.717 | (meishi) no made ni kakeru    | hang (noun) in window |
| 2.554 | (meishi) no sutoobu ni kakeru | put (noun) on stove   |



TL output (English):

*US imposes tax on parts in order to blockade...*

# Connectionist approaches

## Spreading activation: Modification Deciding Network (Matsushita)

example of positive (co-operative) links:

- 1) single phrase modifying a number of verb phrases simultaneously:

*Kare wa hon o kai, sore o yonda*

He book buy it read

'He bought a book and read it'

– *kare* modifies both *kai* and *yonda*

examples of negative (exclusive) links:

- 2) one word cannot have two different meanings in same sentence:

*Watashi wa 1-nensei no eigo o motte iru*

I freshmen English take charge of

'I teach freshmen English'

*motte* means either 'take charge of' or 'have'

here it can only be 'take charge of'

- 3) modifications should not intersect:

*Watashi wa kare ga kinoo sakyokushita kyoku o kiita*

I he yesterday compose song listened

'I listened to the song he composed yesterday'

the modification of *kiita* by *kare* and of *sakkyokushita*

by *watashi* would violate the non-intersection condition.

examples of control rules:

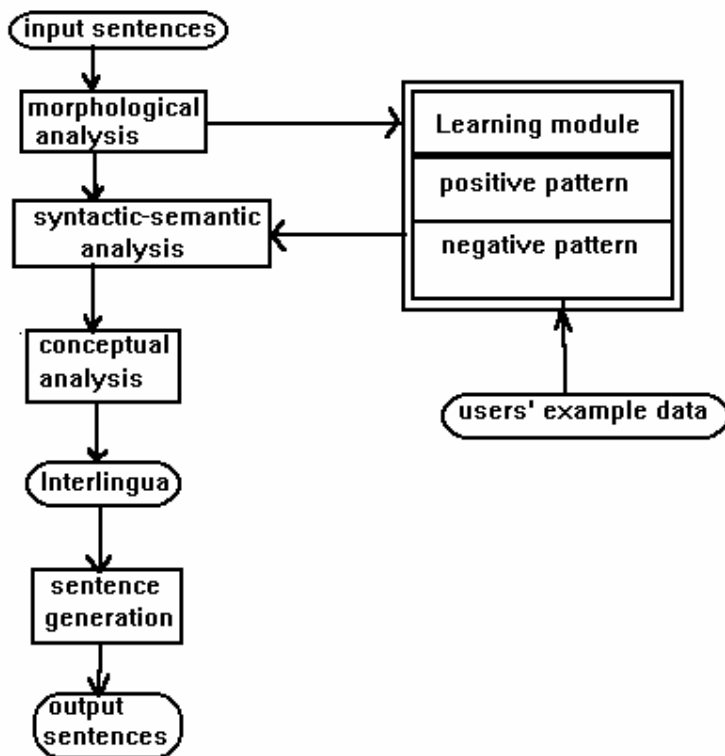
- 4) obligatory case is given precedence to optional case

- 5) modifications close to verbs are given precedence  
to modifications more distant

- 6) modifications with modifier closely related semantically  
to verb are preferred

# Learning systems:

## Model - NEC



# **Corpora as sources of information**

## **text corpora:**

**LDC (Linguistic Data Consortium)**  
**EAGLES (Expert Advisory Group on  
Language Engineering Standards)**

## **sublanguage or domain-specific systems**

## **lexical/knowledge databases:**

**UNITRAN**  
**Carnegie-Mellon systems**  
**ULTRA, Pangloss**  
**LMT**  
**Electronic Dictionary Research**

## **Bilingual text databases (aligned):**

**- for example-based systems**  
**- for translator's workstations**  
**AT&T**  
**ACQUILEX**  
**CWARC (Canadian Workplace Automation  
Research Center)**

## **Direct use of statistical information**

**PIVOT (NEC)**  
**ArchTran**

# Generation

- **stylistic improvement**
- **discourse features**
- **dialogue translation**
- **illocutionary acts**

**example-based methods**

**multilingual generation (e.g. RAREAS)**

**dialogue-based MT**

(translation by monolinguals not  
knowing target language)

**UMIST**

**Brussels (Babel-2)**

**Grenoble (LIDIA)**

**Kuala Lumpur (Malaysia)**

# **Controlled, domain- and user-specific MT**

## **Controlled input MT systems**

**Xerox (Systran)**

**Smart Corporation**

**Perkins Engines (Weidner)**

**CATALYST (Carnegie-Mellon for  
Caterpillar)**

## **Domain-specific and sublanguage MT**

**Meteo, CRITTER, ELU**

**Pangloss, CATALYST**

**ATR speech translation**

**VERBMOBIL**

## **User-specific and custom-built MT systems**

**Winger**

**Le Routier**

**Volmac Lingware Services**

**TRADEX**

**CSK (ARGO system)**

**HESS (Hangul-English Support System)**

## **Translator Work Stations**

- **multilingual word processing**
- **optical character recognition, electronic receipt and transmission of texts**
- **terminology management software**
- **automated access to dictionaries, terminology databanks and other information sources (on-line, remote access, CD-ROM, local network, etc.)**
- **concordance software**
- **storage of and access to existing translations (for later (partial) reuse or revision)**
- **access to example translations [aligned bilingual text corpora]**
- **access to automatic translation facilities (individual words; phrases; sentence by sentence; full text)**
- **'pre-translation' facilities**

---

**examples:**

**Canadian Workplace Automation Research Center  
Carnegie-Mellon Centre for Machine Translation  
TWB (Translator's Workbench): [METAL]  
IBM workstation TranslationManager/2: [LMT]  
English-Malay translation workstation: [JEMAH]**

## Pre-translation

### English original:

Momentary loads can occur repeatedly during the duty cycle but are of short duration, not exceeding 1 min. at any occurrence. When several momentary loads occur within the same 1 min. period and a discrete sequence cannot be established, the load shall be assumed to be the sum of all momentary loads occurring within that minute.

### French pre-translation:

charges\_momentanées peuvent occur de\_façon\_répétée pendant le cycle\_opératoire mais sont de courte durée, not exceeding 1 min at any occurrence. Lorsque several charges\_momentanées occur within le même 1 min période et une discrete séquence ne\_peut\_pas être établi, la charge doit être assumed to être la somme de all charges\_momentanées occurring within that minute.

# The global view

## Commercial systems since 1988:

- Globalink
- PC-Translator
- Tovna
- DP/Translator
- Toltran
- Translate (Finalsoft Corporation)
- XLT (Socatra)
- AppTek
- Hypertrans
- Lexitrans
- Language Assistant (Microtac)
  
- RMT/EJ (Ricoh)
- DuetQt (Sharp)
- STAR (Catena), LogoVista E to J  
(Language Engineering Corporation)
- EZ JapaneseWriter
- Meltran (Mitsubishi)

## Worldwide activity:

mainly:

United States, Canada

Western Europe

Japan

also:

Malaysia (e.g. JEMAH)

Thailand

China (mainland and Taiwan)

India

Korea (HESS, KSHALT, MATES, etc.)

Eastern Europe (AMPAR, ETAP-2, PARS)

## The five eras of MT history

### [1] 1947-1954:

- Weaver's memorandum (July 1949)
- Bar-Hillel at MIT (1951-52)
- word for word translation (Booth, Richens, Reifler)
- statistical methods (Kaplan, RAND)
- first MT conference (MIT, 1952)

### [2] 1954-1966:

- IBM-Georgetown demonstration (Jan 1954)
- direct translation ("first generation")
- syntactic analysis (MIT, ATN)
- conceptual interlingua (CLRU)
- Georgetown systems installed (1963, 1964)

### [3] 1966-76:

- ALPAC report (1966)
- indirect approach ('interlingua' CETA, LRC)
- rule-based formalisms
- sublanguage (Meteo 1976)
- Systran at USAF (1970) and at CEC (1976)

### [4] 1976-89:

- transfer "second generation" (Ariane, Eurotra, Mu)
- multinational, multilingual projects (Eurotra, CICC)
- linguistics-based interlingua (DLT, Rosetta)
- knowledge-based interlingua (CMU)
- operational systems (Systran, Meteo, PAHO, CSK)
- controlled input (Xerox, Smart)
- commercial systems (Logos, Weidner, ATLAS, PENSEE, PIVOT, HICATS, ASTRANSAC, DUET, METAL, Tovna, etc.)
- PC systems (PC-Translator, Globalink)

### [5] 1989-

## **Developments since late 1980s**

**MT Summit conferences (1987, 1989, 1991, 1993)  
International Association for Machine Translation (1991)  
US government hearings (1990), JTEC report (1992)**

**statistics-based IBM Candide (1989)  
example-based MT (1989)  
constraint-based/lexicalist tendencies in rule-based MT  
systems for monolingual use (Babel-R, LIDIA, UMIST)  
spoken language (ATR 1986, VERMOBIL 1992)  
lexicon acquisition (two conferences 1993)**

**end of European "second generation" projects (Ariane,  
Eurotra, DLT, Rosetta) and of Eastern European  
projects after 1989/90 political changes  
start of new multilingual European project (Eurolang)  
operational implementation of knowledge-based MT  
(Pangloss, CATALYST)  
integration of MT/MAT in documentation systems  
(e.g. Xerox DocuTran, Krupp Industrietechnik)  
user-specific custom-built systems  
(e.g. TRADEX, Volmac, Winger, etc.)  
generation from non-textual sources  
evaluation: methodology, benchmarks, etc.**

## MT chronology

	oper. systems	research	other
1945			
	46		
	47		Weaver letter
	48		
	49		Weaver memo
1950			
	51		
	52		
	53		
	54	IBM-GU demo	
1955			
	56		
	57		
	58		
	59		
1960	Mark II (IBM)		
	61		
	62		
	63		
	64	Georgetown	
1965			
	66	ALPAC	
	67		
	68		
	69		
1970			
	71	Systran	GETA, LRC, TAUM
	72		
	73		
	74		
1975	Meteo		
	76	CEC Systran	Stanford, KBMT
	77		
	78		Eurotra
	79		
1980			
	81	Weidner	
	82	LOGOS	DLT, Rosetta, CMU
	83		
	84	Fujitsu ATLAS	EBMT, Unification
1985	PC-Translator		
	86	NEC PIVOT	
	87	Hitachi HICATS	ATR
	88	METAL, Sharp Duet, Tovna, Globalink	
	89		IBM Candide
1990			
	91	Trados	Eurolang
	92		
	93		

## **"Third generation" systems ?**

- rule base less abstract than that of the 'indirect' models
  - syntactic analysis restricted to surface constituency and dependency relations (roles and cases); single monostratal representation (unification/constraint-based analysis)
  - semantic analysis limited mainly to identification of sentence and clause roles (agents, patients, etc.)
  - broad-brush disambiguation by simple semantic features (human, animal, etc.)
  - lexical information derived primarily from standard dictionary sources ('crude' syntactic categories and semantic features)
  - lexical/structural transfer rules (constraint-based) operating on monostratal ('shallow') representations
- 
- example translations (aligned bilingual corpora)
  - statistical data about lexical collocations and vocabulary frequencies (monolingual)
  - probabilities of lexical transfer (bilingual)
  - domain-specific knowledge bases (both linguistic and subject knowledge)
  - feedback ('learning') for grammar/lexicon improvement (connectionist ?)
  - greater emphasis on discourse and text stylistic aspects
  - integration into documentation processing and publishing systems (as in translator's workstations)

## Possible precursors of the "third generation"

SHALT (IBM Japan), with five types of knowledge sources:

- grammar rules
- concept definitions
- mapping rules
- conceptual paraphrasing rules
- corpus of example sentences

ArchTran (Taiwan)

- language model (inductively derived)
- corpus as data source
- statistical translation scores (lexical, syntactic, semantic)
- probabilistic transfer
- feedback to adjust parameters for specific users

Transfer-Driven MT (ATR, Japan)

- thesaurus coding for distance measurement
- string level transfer:  
*sochira* -> *this* (desu), *you* (okura), *it* (miru)  
e.g. *sochira ni tsutaeru* -> *you convey*  
[okura (send) <-> tsutaeru (convey)]
- pattern level transfer:  
*X onegaishimasu* -> *please.../may I...*
- grammar level transfer:  
[grammatical category sequence]
- example database:  
e.g. Japanese *N<sub>1</sub> no N<sub>2</sub>*

*fee for the conference*  
*conference in Tokyo*  
*week's holiday*  
*hotel reservation*

not: *fee of the conference*,  
not: *conference of Tokyo*  
not: *holiday of a week*  
not: *reservation of hotel*

# **MT and users**

**Multilingual/multinational corporations**

**Translation agencies/services**

- **post-edited, pre-edited, controlled input**
- **single source text and multiple target languages**
- **partial retranslation of revised texts**

**occasional non-professional translator.**

- **bilingual non-translators (PC-based systems)**
- **monolingual users knowing only the source language (UMIST, Babel-R, LIDIA)**
- **monolingual users knowing only the target language (mainframe/batch systems, PC-based systems)**

**research implications:**

**adoption of realistic and realizable aims**  
**recognition of usefulness of the less than perfect**

**definition of user environments**

**no attempt to 'mimic' the human translator**

**user/developer collaboration**

**user/researcher collaboration**

**evaluation of systems, benchmarks, standards of performance**