Milestones in the history of machine translation

John Hutchins

Presentation
15 November 2012
Science Museum, London

1933: Artsrouni and Troyanskii

- Pre-computer proposals
- Patents presented by:
  - Georges Artsrouni (Georgian based in France)
  - Petr Petrovich Troyanskii (Russian)
- Both no more than mechanized bi- or multilingual dictionaries
  - Origins traceable to 17th century
- Although Troyanskii included codes (Esperanto based)

1948: Richens and Booth: dictionary word for word

French input with segmentation:

Il n'est pas étonn*ant de constat*er que les hormone*s de croissance ag*issent sur certain*es espèce*s, alors qu'elles sont in*opér*antes sur d'autre*s, si l'on song*e à la grand*e spécificité de ces substance*s

English output:

v not is not/step astonish *v* of establish v that/which? *v* hormone *m* of growth act *m* on certain *m* species *m*, then that/which? *v* not operate *m* on of other *m* if *v* one dream/consider *z* to *v* great *v* specificity of those substance *m*.

(where *v* is untranslated French word, *m* multiple/plural, *z* unspecific)

1949: Weaver

- Warren Weaver (Rockefeller Foundation) writes memorandum on MT
- At the time, collaborating with Claude Shannon on 'information theory'
- Discussed ideas with Andrew Booth of Birkbeck College in 1946 and 1947
- Four main suggestions:
  - Disambiguation by examining adjacent words
  - Brain networks similar to computers
  - Use of cryptographic methods: decoding of source text – 'When I look at an article in Russian, I say: “This is really written in English, but it has been coded in some strange symbols.I will now proceed to decode.”'
  - Universal language

1952: first conference

- Bar-Hillel appointed at MIT in May 1951, surveyed MT
- Convened first MT conference at MIT
- Topics covered:
  - Pre-editing, post-editing (Reifler)
  - Controlled language (Dodd's Model English)
  - Domain restriction (Oswald's microglossaries)
  - Syntactic analysis (Bar-Hillel's categorial grammar)
  - Computer hardware, programming
  - Funding

1954: first public demonstration

- Leon Dostert determined to show 'technical feasibility' of MT
- Collaboration of Georgetown University and IBM
- Public demonstration in New York, 7th January 1954 of Russian-English system
- Linguistic foundations by Paul Garvin of Georgetown U.
  - 250 words, 6 rules
- Programming by Peter Sheridan of IBM
- Reported widely, worldwide interest
- Beginning of government funding - in both US and Soviet Union

Georgetown University-IBM program

Russian input	English equivalents		1st code	2nd code	3rd code	rule
	Eng_1	Eng_2	(PID)	(CDD_1)	(CDD_2)	
vyelyichyina	magnitude	---	***	***	**	6
ugl-	coal	angle	121	***	25	2
-a	of	---	131	222	25	3
opryedyelyayetsya	is determined	---	***	***	**	6
otnoshyenyi-	relation	the relation	151	***	**	5
-yem	by	---	131	***	**	3
dlyin-	length	---	***	***	**	6
-i	of	---	131	***	25	3
dug-	arc	---	***	***	**	6
-yi	of	---	131	***	25	3
k	to	for	121	***	23	2
radyius-	radius	---	***	221	**	6
-u	to	---	131	***	**	3

IBM 701 at New York headquarters

1955: beginnings of MT in Soviet Union

1953: death of Stalin, March 1953 opened access to science of the west: cybernetics, structural linguistics, and computers

1954: news of Georgetown-IBM demonstration

1955: first attempts using BESM

1956: foundation of groups: Inst Precision Mechnaics, Steklov Mathematical Institute, Inst Linguistics, Leningrad University

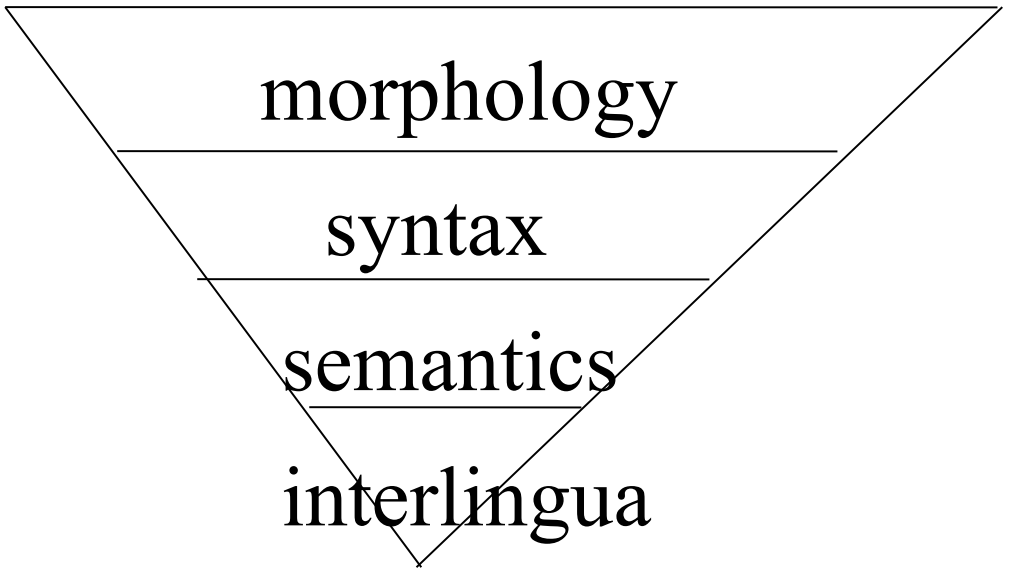
BESM program

Dealing with multiple translations of English *many*:

a(b,c)	Check preceding word (directly) for *how*
b(0,0)	сколько (numeral, not declined)
c(d,e)	Check preceding word (directly) for *as*
d(0,0)	столько (numeral declined)
e(g,i)	Check given word for *much*
f(0,0)	Not translated (adverb)
g(f,k)	Check preceding word (directly) for *very*
h(0,0)	многий (adjective, hard stem, with sibilant)
i(h,j)	Check preceding word for preposition and succeeding word for noun
j(0,0)	много (adverb)
k(l,j)	Check succeeding word for noun
1(0,0)	много (numeral, declined)

Main MT system types

- SL direct/lexical TL
- morphology
- syntax
- semantics
- interlingua



- [based on Vauquois' triangle]

Main groups in 1960s

- University of Washington, IBM ('direct translation') [Reifler, King]
- Harvard (massive dictionary, predictive syntax) [Oettinger]
- Massachusetts Institute of Technology (syntactic transfer) [Yngve]
- Georgetown (multiple levels of analysis) [Dostert, Zarechnak]
- Cambridge Language Research Unit (interlingua, lattices) [Masterman]
- Birkbeck College London [Booth]
- National Physical Laboratory, Teddington
- Milan University (interlinga) [Ceccato]
- Institute of Precision Mechanics and Computer Technology [Panov]
- Leningrad University (interlingua) [Andreev]
- Leningrad University (statistical) [Piotrowski]
- Institute of Linguistics, Moscow (syntax, semantics) [Kulagina, Mel'chuk]

Georgetown University

1954 founded by Leon Dostert

Largest MT group in USA(over 20 researchers), funded by CIA

Variety of methods for Russian-English system examined: code-matching, syntactic analysis (Paul Garvin), sentence-by-sentence (Antony Brown), general analysis (Michael Zarechnak)

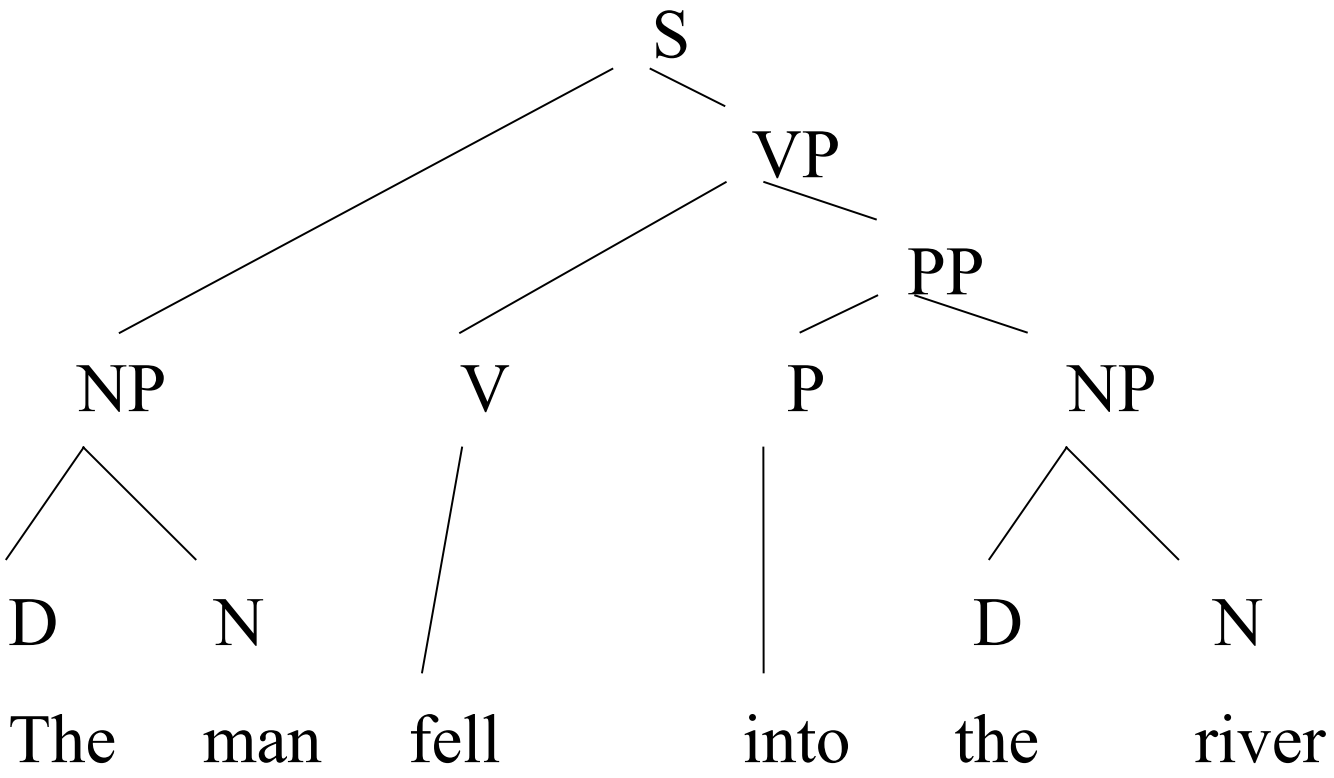
GAT eventually adopted: multiple levels of analysis: morphological, idiom identification, syntagmatic analysis (agreements, government), syntactic analysis (subject-predicates)

Implemented on SERNA (Peter Toma)

1961 demonstrated at Pentagon; 1963 installed at Euratom (Ispra, Italy); 1964 installed at Oak Ridge National Laboratory

Syntactic analysis

One major focus of MT research in 1960s



Massachusetts Institute of Technology

1951: appointed Bar-Hillel, convened first MT conference

1953-1965: directed by Victor Yngve

Fundamental research, not “short-cut methods” (other researchers included Chomsky)

Linguistic analysis: German-English

Programming language (COMIT): first non-numerical, string-processor

Syntactic transfer (SL tree representation to TL trees)

Sentence production (first generator system)

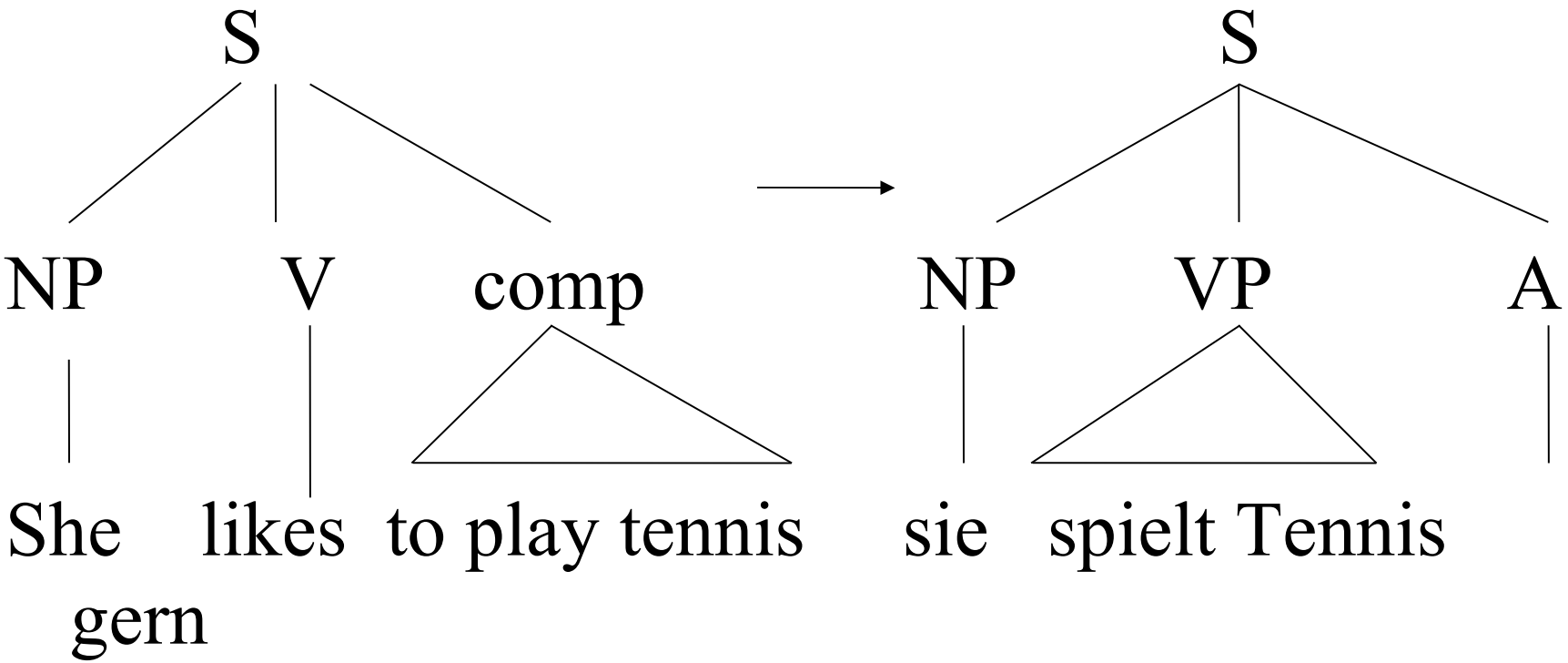
Other languages: Finnish, Arabic

1964: reached “semantic barrier”

Yngve editor of “Mechanical Translation”, co-founder of Association of Ccomputational Linguistics

Syntactic transfer

Model for MIT and later systems



Harvard University

1954 founded by Anthony Oettinger

Large-scale Russian-English dictionary, producing word-for-word translations and a research tool

1959 Syntactic analysis: based on predictive analyzer developed at National Bureau of Standards by Ida Rhodes

1963-1966: Further developments at Harvard: pushdown store, multiple-path predictive analyzer (Susumu Kuno, Warren Plath)

Birkbeck College

1948: Booth's collaboration with Richard Richens on morphology

1953: first research by Booth on dictionary: first test of procedure on APEXC built at Birkbeck

1955: fast dictionary lookup, 'binary division'

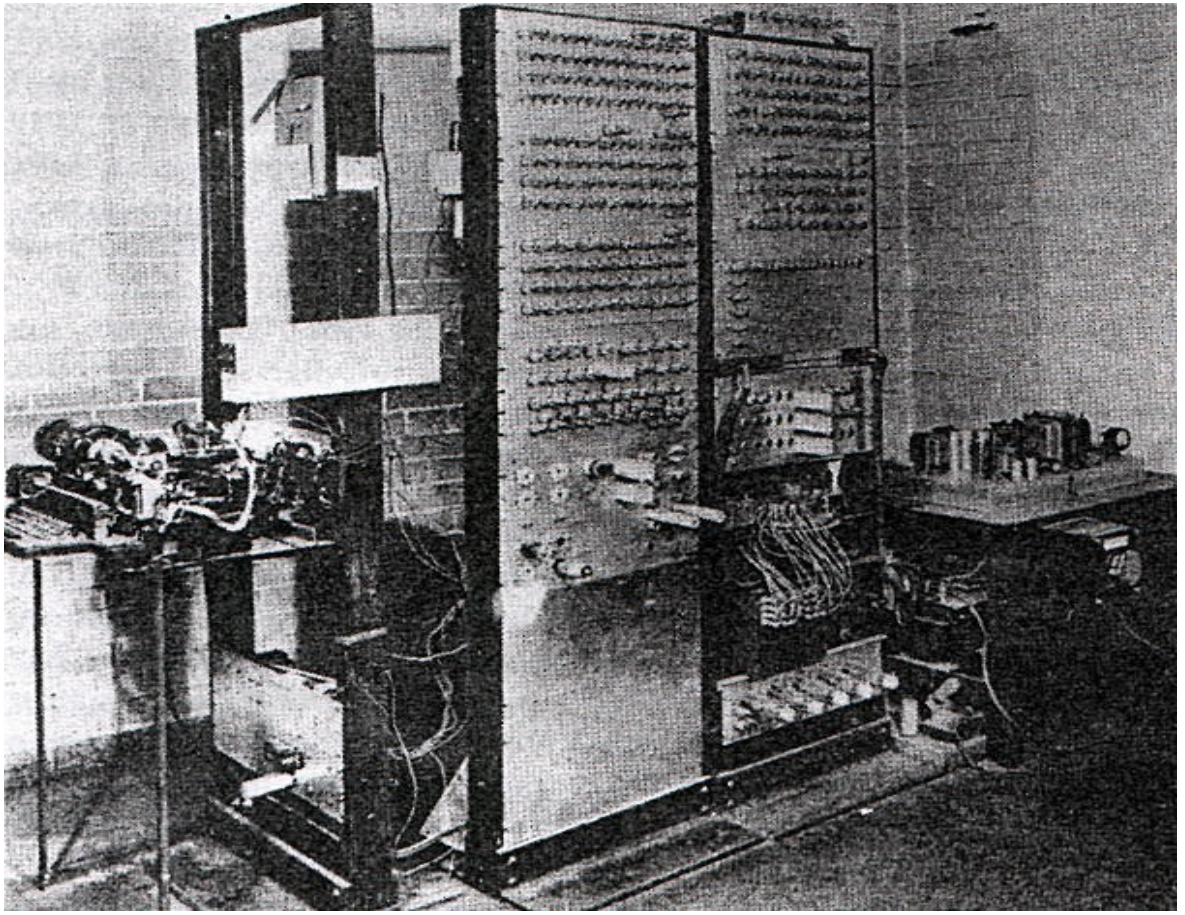
1955: funding by Nuffield Foundation for French-English system – tested on APEXC 1958

1958: syntactic research on German (not programmed because of computer storage limitations)

1962: Departure for University of Saskatchewan

1965: Trial for Canadian MT system

Booth's APE(R)C about 1952



Cambridge Language Research Unit

Founded 1956, director Margaret Masterman

Interlingua: continuation of Richens naked ideas (Nude)

- 51 elements (semantic classifiers): ASK BANG BE BEAST CAN CAUSE CHANGE COUNT DO … from which dictionary entries constructed, e.g. “to distress” (CAUSE/ SENSE: NOT PLEASE), “to join” (CAUSE/ (BE/PART)
- Influence on Artificial Intelligence (Schank, etc.)

Thesaurus approach: lexical items under 'heads', multiple meanings under different heads,e.g.'plant' under place, vegetable, agriculture, trick, tool

Pidgin translation: interim results of thesaurus analysis

Lattice theory

CLRU: pidgin translation

From Latin:

Among the+Swiss by+far noble-est was and rich-est the+chief+Orgetorix. He, during+the+consulate of+M.Messalla and +M.Piso kingdom desire-s induced conspiracy persuaded-s, that or limit-s own when(ith) all-s resources might+go+out-they: a+mere+nothing to+be when(ith) strength all-s excel-they+might, the+whole+of Gaul control to+gain+the mastery+of.

National Physical Laboratory

leader: John McDaniel

1959 began on ACE computer

Russian-English scientific texts

Use of Harvard Russian-English dictionary (18,000 words) – not operational until 1963

'international' words transliterated

syntactic analysis: noun groups, verb groups, however results little better than word-for-word translations

evaluation in 1966: “slightly less than good”

1961 organized international conference

1960: Bar-Hillel's survey

- Critical of most MT groups for unrealistic aims
- Demonstration of 'non-feasibility' of fully automatic high-quality MT – need encyclopedic knowledge (of ‘pens’ and ‘boxes’)
  - The pen was in the box
  - The box was in the pen
- Example was convincing for many at the time
  - although some contemporary researchers claimed to have methods for dealing with it (lexical adjacency, semantic analysis, etc.)
- Later artificial intelligence and statistical methods make it no longer compelling

1964/65: World's Fair

Based on research since 1950 at University of Washington by MT pioneer, Erwin Reifler – a dictionary-driven Russian-English system

Photoscopic disk developed at International Telemeter Corporation

Research effort transferred to IBM under Gilbert King

1959: system delivered to USAF

1965: second version delivered to USAF Foreign Technology Division, where remained in daily operation until 1970

World's Fair system: translation from Russian

Machine Output

All this page is machine translation of Russian text, which is printed on preceding page. This page of translation from Russian into English is not perfect due to unsolved problems of grammar.Before machine can translate from one language into another,linguists have to introduce in memory unit of machine large quantity of grammatical rules, which increase intelligibility of given translation. But because in languages exists significant variety and complexity, all grammatical rules of any language are not developed completely at present time for use by computers. System for automatic translation of languages must also consider problem of words, having identical spellings, but different meanings. (In English language, exists large quantity of such words: can, will, type, store, fair, through, content, rule, port, even, mean etc.) In such cases linguists have to find rules for resolution of conflicts in meaning on the basis of analysis of speech and other forms of context. However, if necessary grammatical rules is/are lacking, machine can print several meanings, divided by line. As can be seen from this page, many linguistic questions of machine translation still are not solved. But today translations from one language into another, performed by computers, help scientists to recognize immediately contents/allowance of foreign technical literature and to determine necessity of further study of this information.

Babelfish translation, 2011

This entire page is the machine transfer of the Russian text, which is printed on previous page. This page of transfer from the Russian to English not [sovershenna] as a result of the unresolved questions of grammar. Before machine it can translating from one language to another, linguists must introduce into the storage unit the large number of grammatical rules, which increase the clearness of this transfer. [N]o because in the languages there is a significant variety and a complexity, all grammatical rules of any language are not developed completely at present for the use by computers. System for machine translation of languages must also consider the problem of the words, which have identical writing, but different values. (In the English there is a large number of these words: can, will, type, store, fair, through, content, rule, port, even, mean and so forth) in such cases linguists must find rules for resolution of conflicts in the value of the words on the basis of the analysis of vocal and other forms of context. However, if there are no necessary grammatical rules, machine can print several values, divided by feature. As can be seen from this page, many linguistic problems of machine transfer are not yet solved. But today translations from one language into another, produced by computers, help scientific workers to learn immediately the content of foreign technical literature and to determine the need for further study of this information.

1966: the ALPAC report

- Set up by NSF for US sponsors of MT research
- Concluded: No effective MT despite massive funding, and none in prospect
- Poor quality output
- Criticised at time for short sightedness
- Brought to end US funding for many years
- Affected funding elsewhere

Consequences of ALPAC

identification of user needs:

- Dissemination vs. assimilation

recognition that 'perfectionism' had neglected:

- Operational factors and requirements
- Expertise of translators
- Machine aids for translators

henceforth three strands of research

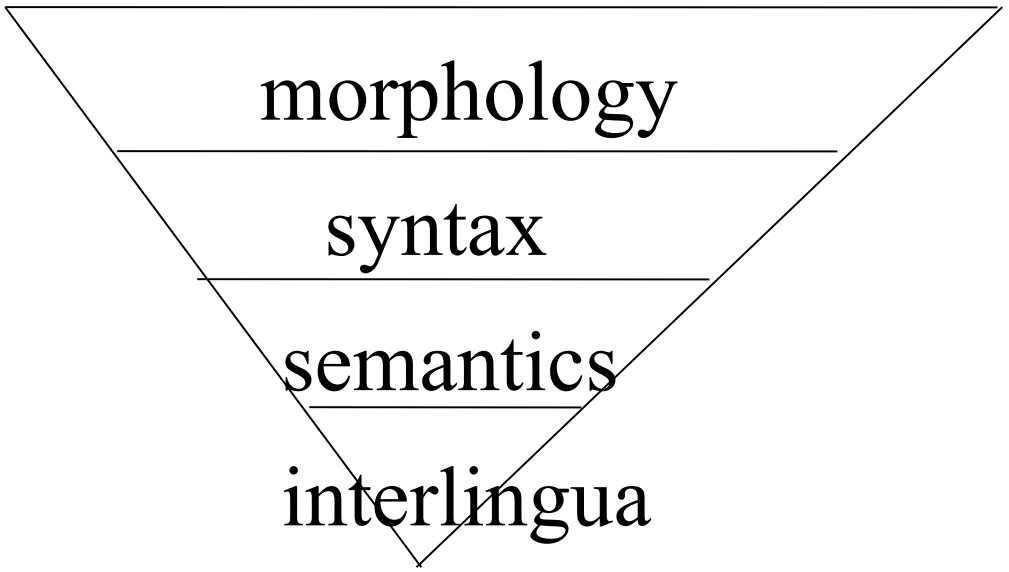
- Translation tools and aids
- Operational tools: post-editing, controlled languages, domain-specific systems
- New systems, new approaches, new methods

From 1967 to 1978

- Continuation of research in US (Texas, Wayne State), Soviet Union, UK, Canada, France
- 1970: Systran installed at USAF (Foreign Technology Division)
- 1970: TITUS installed (restricted language: textile industry abstracts)
- 1975: Météo ‘sublanguage’ English-French system (weather broadcasts)
- 1975: CULT Chinese-English (restricted language: mathematics)
- 1976: European Commission acquires Systran:
- 1978: Xerox Corporation uses Systran with controlled language (Caterpillar English)

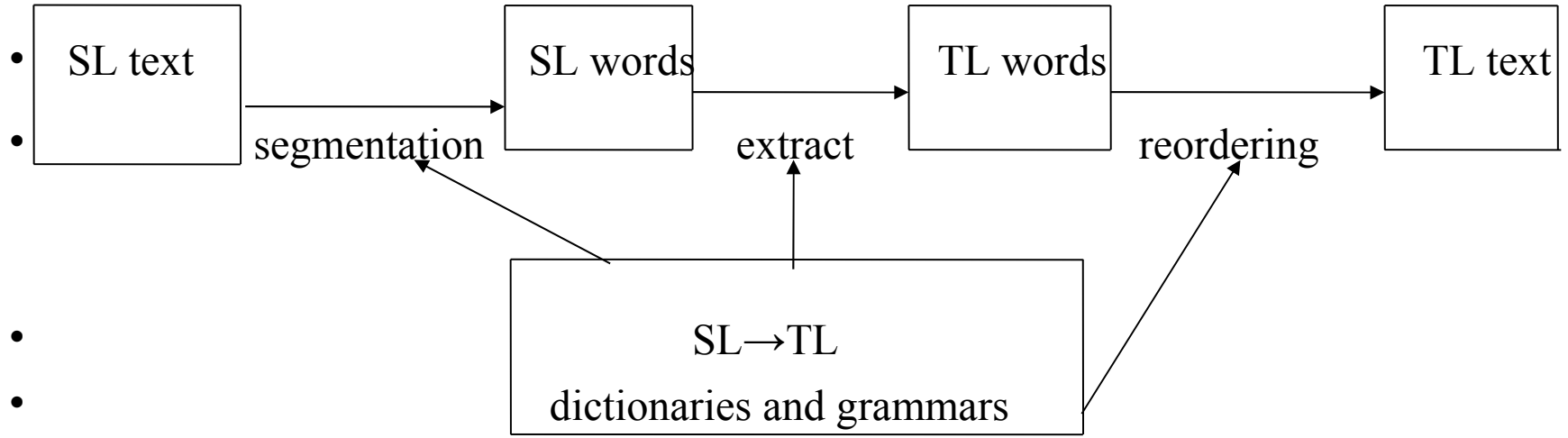
Main MT system types

- SL direct/lexical TL
- morphology
- syntax
- semantics
- interlingua



- [based on Vauquois' triangle]

Direct translation model

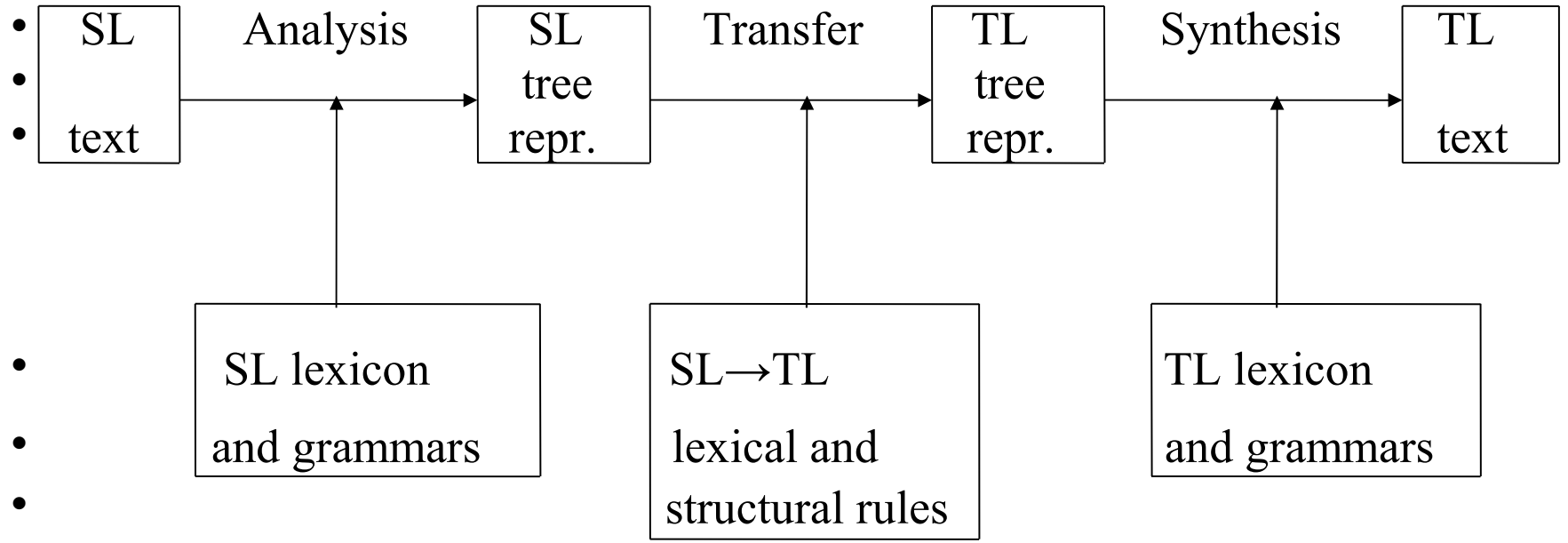


- model: segment, extract, combine/reorder
- word for word, some morphology, some reordering of TL

1978: transfer-based MT

- Beginning of research on :
- ARIANE system at Grenoble University (France) [Vauquois, Boitet] – Russian-French, English-French, German-French
- Eurotra system funded by European Commission
- Logos (USA) [Scott] - German-English
- Mu system, Kyoto University (Japan) [Nagao] - Japanese/English
- METAL, University of Texas (USA) [Lehmann] -German-English
- Meaning-Text Model (Moscow) [Mel'chuk]
- ETAP (Moscow) [Apres'yan]

Transfer-based MT model



- Multi-level representations (morphology, syntax, semantics), syntax-oriented, tree transduction

ARIANE

Founded 1960 in Paris and Grenoble as CETA (Centre d'Etudes de la Traduction Automatique)

Interlingua model 1960-1970: director: Bernard Vauquois

Failures: reduction to interlingua representations, destruction of useful SL information

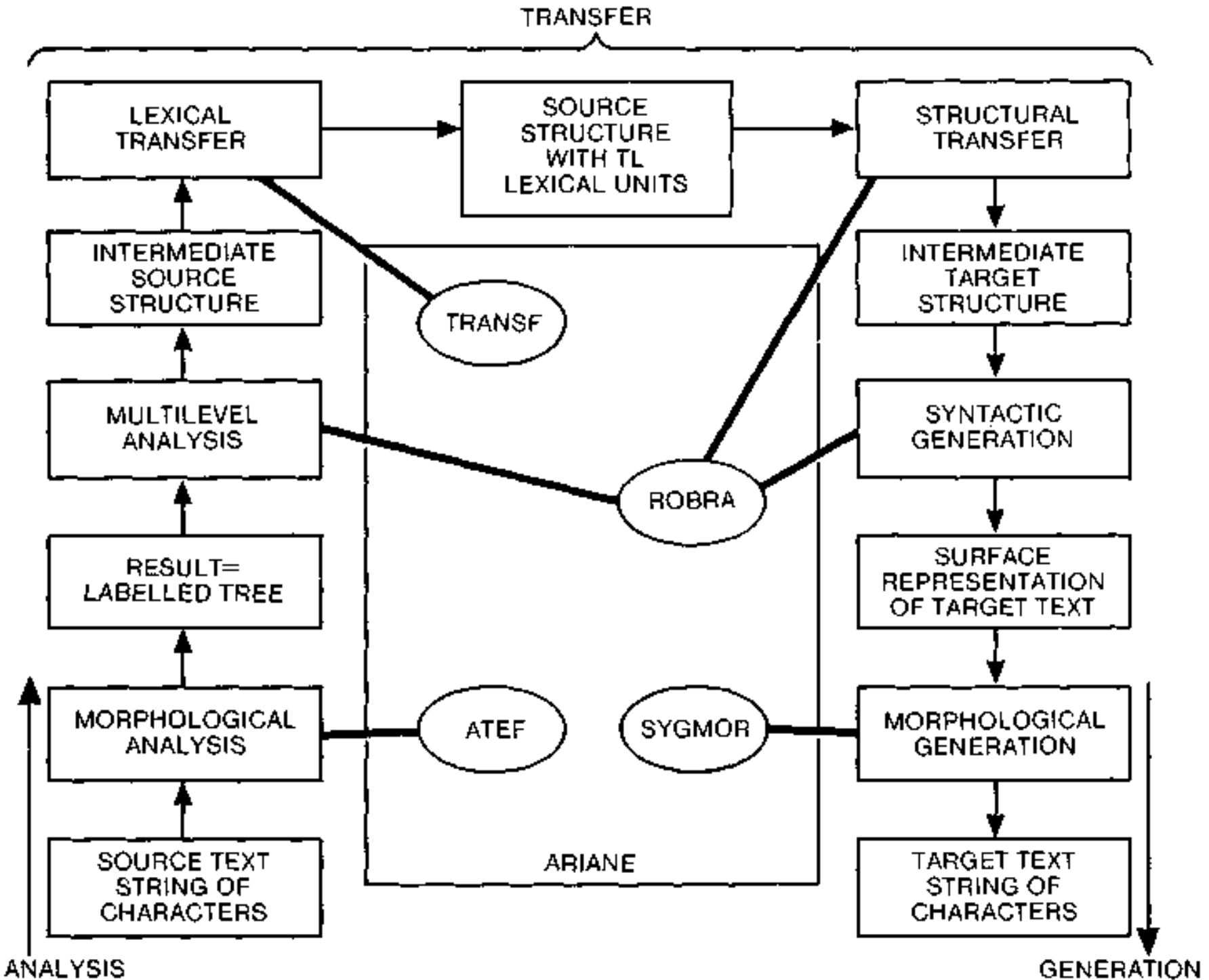
Group renamed GETA (Groupe d'Etudes pour la Traduction Automatique)

Ariane transfer model 1971-

Strict separation of linguistic data and programming

stages: morphological analysis, syntactic analysis, intermediate structure, lexical transfer, SL structure with TL units, structural transfer, syntactic generation, morphological generation

One of most influential MT system


TRANSFER
LEXICAL TRANSFER
SOURCE STRUCTURE WITH TL LEXICAL UNITS
STRUCTURAL TRANSFER
INTERMEDIATE SOURCE STRUCTURE
TRANSF
INTERMEDIATE TARGET STRUCTURE
MULTILEVEL ANALYSIS
ROBRA
SYNTACTIC GENERATION
RESULT= LABELLED TREE
SURFACE REPRESENTATION OF TARGET TEXT
MORPHOLOGICAL ANALYSIS
ATEF
SYGMOR
MORPHOLOGICAL GENERATION
SOURCE TEXT STRING OF CHARACTERS
ARIANE
TARGET TEXT STRING OF CHARACTERS
ANALYSIS
GENERATION

Eurotra

Funded by European Commission 1979-1992

Intended to replace Systran (adopted by EC in 1976)

80 researchers in eight member states (UK, France, Germany, Belgium, Denmark, Netherlands, Italy, Spain)

Multilingual transfer design, intended to be operational as soon as possible

Required precise specification of analysis, transfer and synthesis programs; ambiguities dealt with by monolingual analysis programs; transfer not interlingual by 'Euro-versals'

Excellent and influential linguistic research; but neglected dictionary construction, industrial prototype not delivered

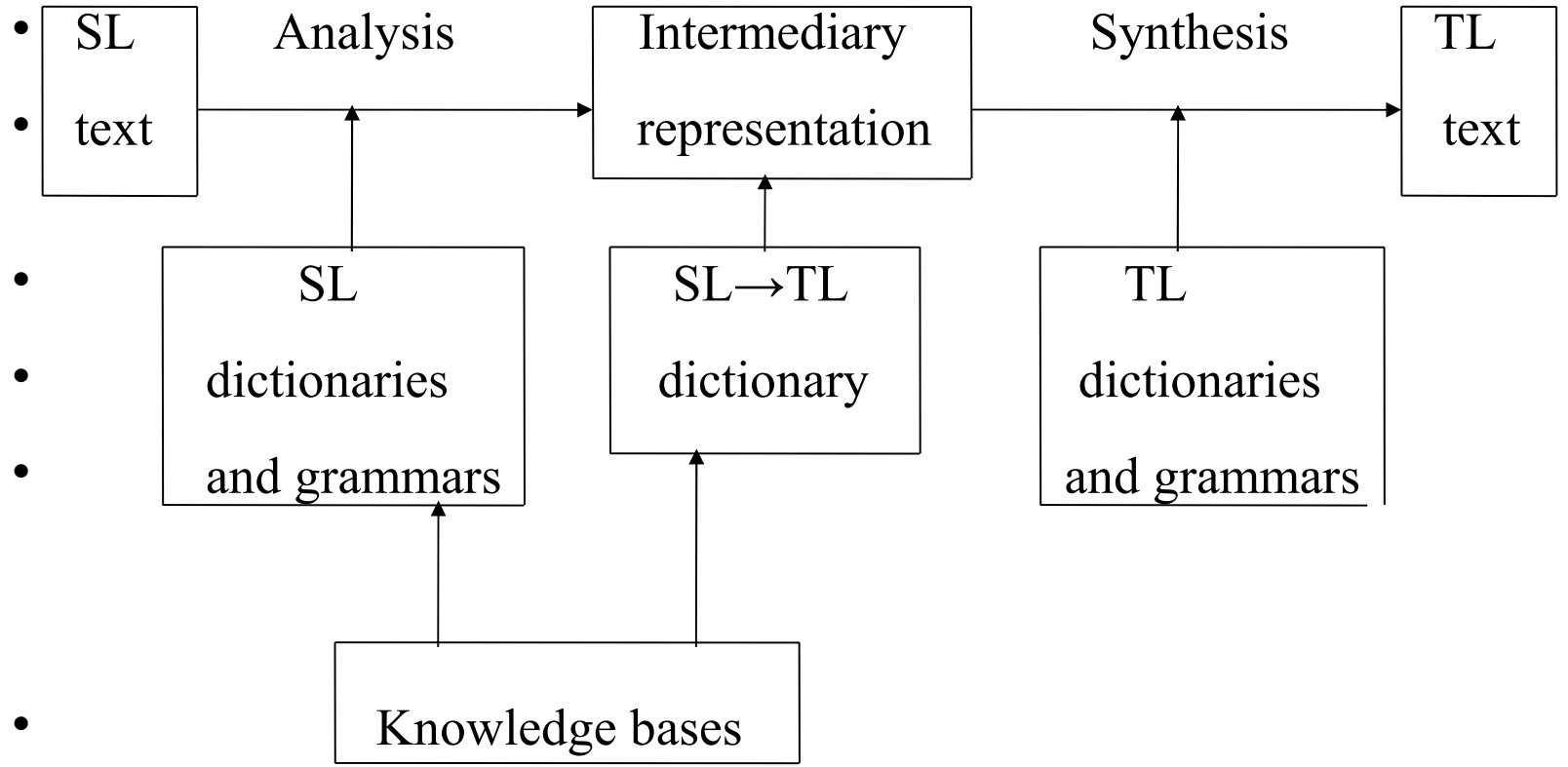
1981: MT for personal computers

- Previously all MT systems for mainframe computers
- ALPS
- Weidner/Bravis
- subsequently (in 1980s and 1990s):
  - ESI, Instant Spanish, LogoMedia, Personal Translator, PeTra, PROMT, Systran
- many Japanese systems
  - e.g. Crossroad, LogoVista

1982: AI and interlinguas

- Beginning of 'Fifth Generation' (AI) program in Japan; influence on US research
- Research on interlingua systems
  - At Philips (Rosetta) – implementing Montague grammar
  - At Utrecht (DLT) – modified Esperanto, bilingual knowledge bank
- Research on knowledge-based systems
  - At Colgate University, Carnegie-Mellon University, New Mexico State University (PANGLOSS)

Interlingua MT model



Theories and formalisms

- (For linguistics-based models of MT, up to late 1980s)
- Categorial grammar
- Transformational-generative grammar, Government-binding theory
- Case grammar
- Dependency grammar
- Stratificational grammar, Meaning-text model
- Montague grammar
- Lexical Functional Grammar

1986: speech translation

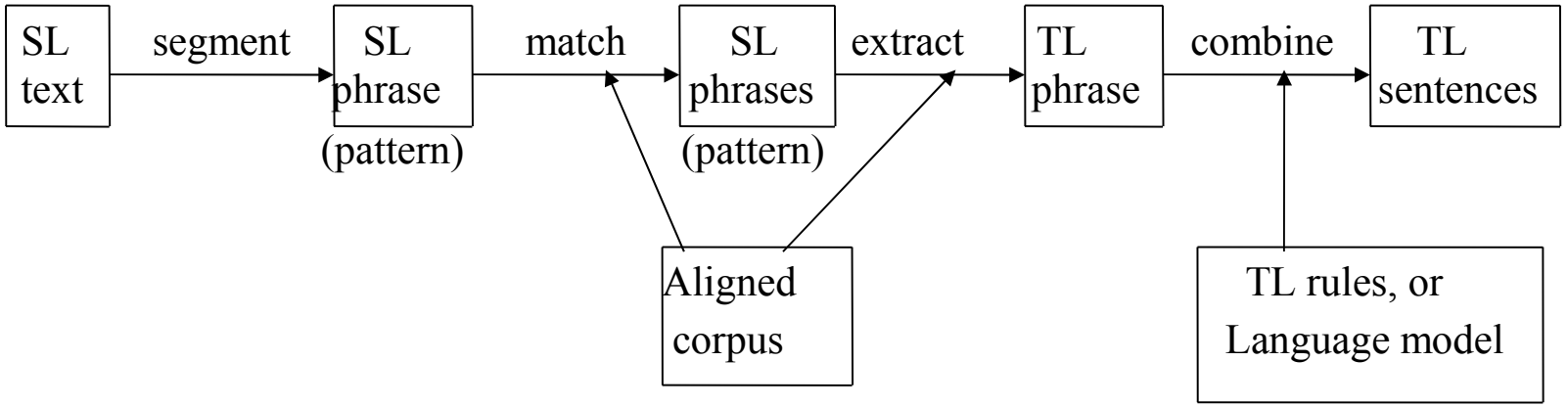
- ATR in Japan, JANUS at Carnegie-Mellon, Verbmobil (at various German universities)
- speech recognition, speech synthesis
- highly context dependent, use of ‘knowledge databases’
- discourse semantics, ‘ill-formed’ utterances
- ellipsis, use of stress, intonation, modality markers
- colloquial usage not yet investigated sufficiently (even in linguistics)
- Restricted fields (telephone booking of hotels and conferences)
- Still continuing

1988: corpus-based MT

- Availability of large bilingual corpora
- Beginning of Example-based MT research, 1988-89
  - First proposed in 1981 by Makoto Nagao
- First article on Statistical MT, 1988 (research at IBM)
  - revival of Warren Weaver's idea ('decoding' SL as TL)

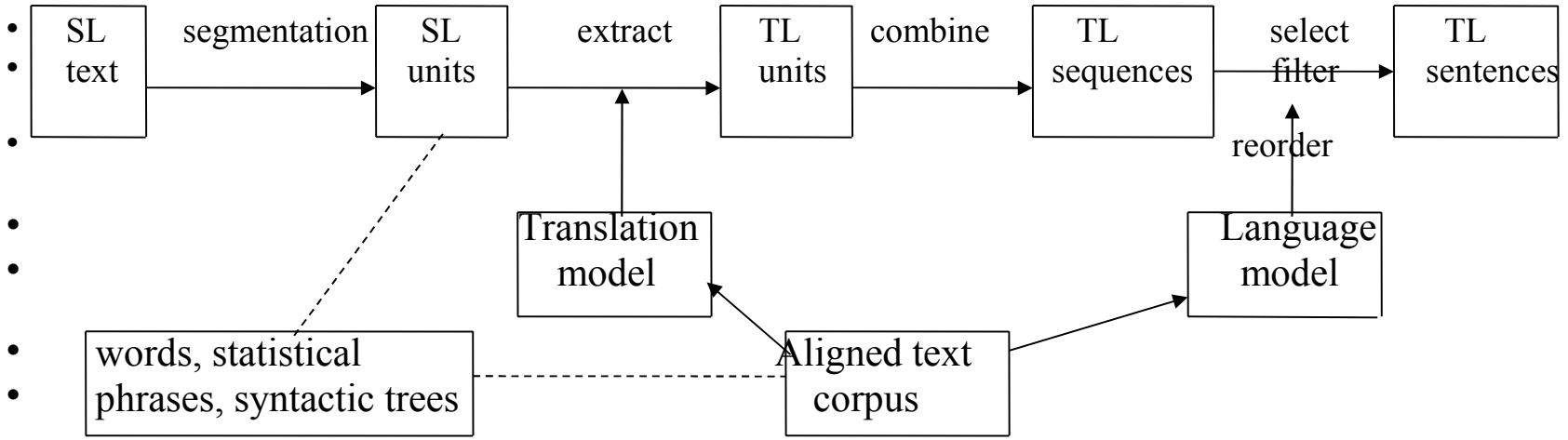
Example-based MT model

- Based on observation that translators try to find similar SL phrases and sentences and their TL equivalents in previously translated texts
  - seek sets of analogies and examples from bilingual corpora
  - in essence, continuation of 'transfer' model, with statistical methods



Statistics-based MT model

- TL words/phrases are chosen as those most likely to correspond with the SL words/phrases in specific context (probabilities, frequencies)
- TL words/phrases are combined in ways most appropriate for the TL in a specific context/domain and style/register etc. (maximizing probabilities)
  - minimal use of linguistic information (morphology, syntax) - but now growing
  - in essence, revival of 'direct translation' (segment, extract, combine/reorder) and Weaver's cryptographic and information-theoretic ideas



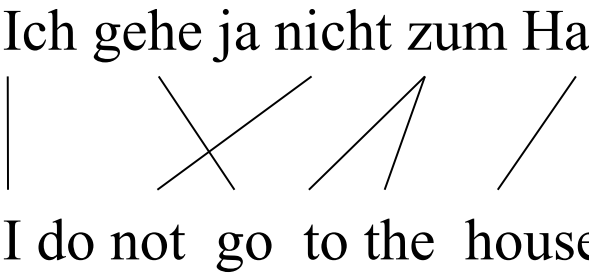
1993: translation memory

- Previous tools: dictionaries, termbanks, concordances
- in 1993 launch of first commercial system: Trados
  - later followed by Transit, Déjà Vu, ProMemoria, WordFast, ...
- using aligned bilingual corpora (of human translation), searchable by words and phrases
- Attractiveness for translators:
  - Components and facilities controlled by users
  - Terminology management
  - Facilities for building dictionaries (e.g. from Internet)
  - Compatible with authoring and publishing systems

Developments of statistical MT

Bi-lingual alignment and monolingual corpora (as 'Language model')

Word alignment



Phrase-based alignment

Features/tags on source trees (corpora) to aid reordering

Comparable corpora (not bilingual translations)

Syntax-aware SMT: syntactic pre-ordering, tree-string decoding

Crowd sourcing (for data, evaluations)

1992: automatic evaluation of MT

Previous evaluations by human judgments: FEMTI (Framework for the Evaluation of Machine Translation in ISLE)

In 1992/1994 DARPA investigates unedited US systems, comparing automatic measures and human judgments of adequacy, fluency, informativeness

Development of MT evaluation metrics (in parallel with SMT):

2001: BLEU (Bilingual Evaluation Understudy)

Statistical measures of similarity of SMT output and human translations (reference translations)

2001-2005: NIST (National Institute of Standards and Technology)

2005: METEOR (Carnegie Mellon), etc.

1997: free online MT

- Origins:
  - Minitel service from 1988 (22 lines of text per minute, charge of $1.20 per page; potentially accessible to 4.5 million users in France)
  - CompuServe started testing in 1992 (limited to subscribers of some forums)
  - Systran offered online translations of webpages since 1996

- Babel Fish launched on AltaVista on December 9, 1997
  - free for all Internet users
- subsequently: FreeTranslation, PROMT, Google, etc.
- usage: mainly short phrases, text not webpages, into native language
- rapid growth
  - Babelfish: 500,000 per day (May 1998) to 1.3 million (October 2000)
  - FreeTranslation: 50,000 per day (December 1999) to 3.4 million (September 2006)

Since 2004: open source toolkits

GIZA ++: tool for alignment in SMT

Moses: platform for building SMT systems

Joshua: decoder for syntax-based (hierarchical) SMT

Apertium: platform for building rule-based MT

META-SHARE: data for EU projects

LetsMT: cloud-based resource for supporting MT research

2006- : some current projects

Euromatrix, founded 2006

Goal: MT systems between all EU languages (over 500 language pairs), many centres involved, lead by Philipp Koehn at Edinburgh University

Statistical phrase-based MT, hybrid systems, statistics-based; tree-transfer (dependency)

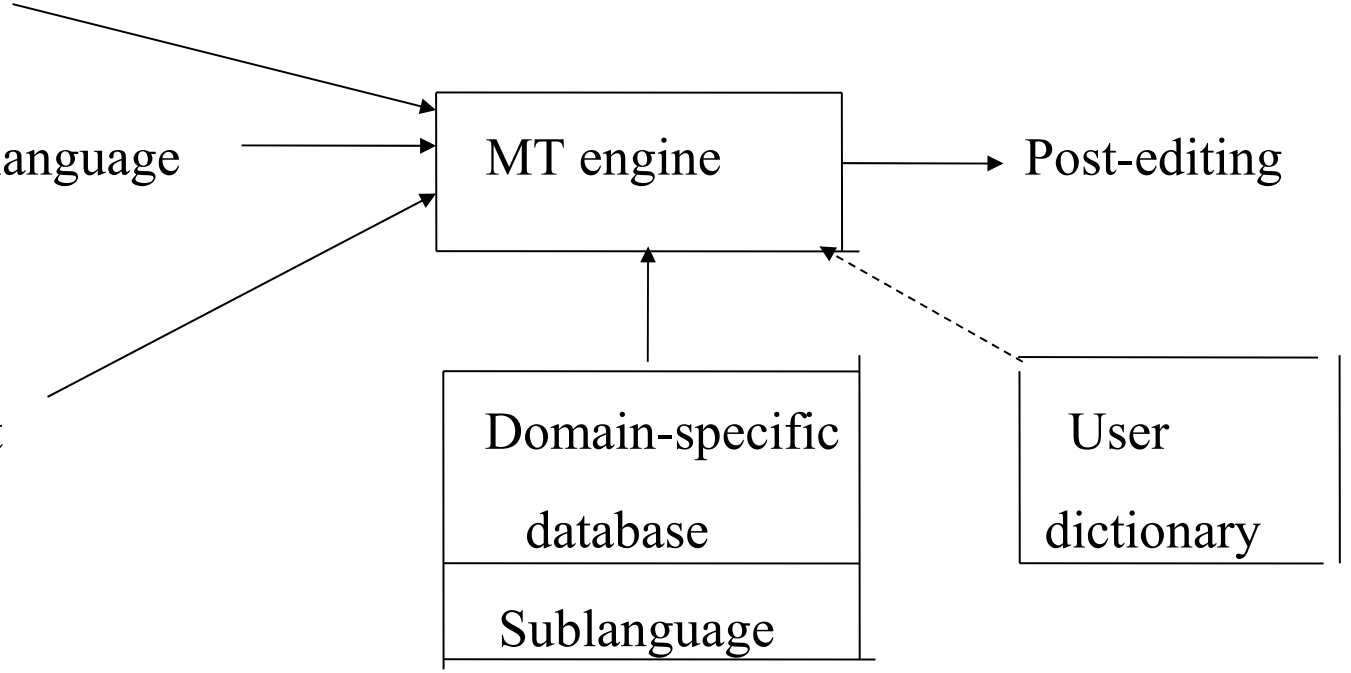
Open source (Moses), collaboration, shared resources; rapid development of systems, open continuous evaluation

Lets MT – fouded 2010 in Baltic states

Online platform for data sharing, and building SMT systems, with easy user interaction; cloud computing

Operational MT typical configurations

- Pre-edited input
- Controlled language
- input
- 'Raw' input
-
-



Machine translation (MT) and human translation (HT) in complementation

- HT for literature, and other ‘culturally-sensitive’ translation
- MT for technical, scientific, medical (etc.) texts which are culturally neutral
- HT (with translation aids) and human-aided MT for dissemination (publishable quality)
- MT for assimilation (rough ‘gist’)
- MT for real-time on-line translation (is this its ‘real’ niche?)
  - **the less the user knows of the source language, the more useful becomes fully automatic translation**
- HT for spoken language translation
- MT for integrating translation with other LT tasks

Summary: current situation

- Commercial MT systems
- Online MT, crowd sourcing; cloud computing
- Statistical machine translation as dominant framework
- Automatic evaluation of MT systems
- Hybrid research systems (rule-based and statistical)
- Speech translation
- Special-purpose MT (e.g. medical, patents, ...)
- MT for less resourced languages (e.g. African, Indian)
- Aids for translators, aids for non-translators (e.g. tourists, military personnel); aids for MT developers

Resources

Machine Translation Archive

- http://www.mt-archive.info

My website for history of MT

http://www.hutchinsweb.me.uk

Readings in machine translation, ed.Sergei Nirenburg, Harold Somers, Yorick Wilks (MIT Press, 2003)

Introduction to SMT

Philipp Koehn: *Statistical machine translation* (Cambridge University Press, 2010)