Milestones in the history of machine translation

John Hutchins

South Ural State University, Chelyabinsk

"Topical Problems of Theoretical and Applied Linguistics"

1933: Artsrouni and Troyanskii

- Patents presented by:
  - Georges Artsrouni (Georgian 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)

1949: Weaver

- Warren Weaver (Rockefeller Foundation) writes memorandum on MT
- Currently collaborating with Claude Shannon on 'information theory'
- Discussed ideas with Andrew Booth in 1947
- Four main suggestions:
  - Disambiguation by examining adjacent words
  - Brain networks similar to computers
  - Use of cryptographic methods
  - Universal language

1952: first conference

- Bar-Hillel appointed May 1951, surveyed MT
- Convened first MT conference at MIT
- Topics covered:
  - Pre-editing, post-editing
  - 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

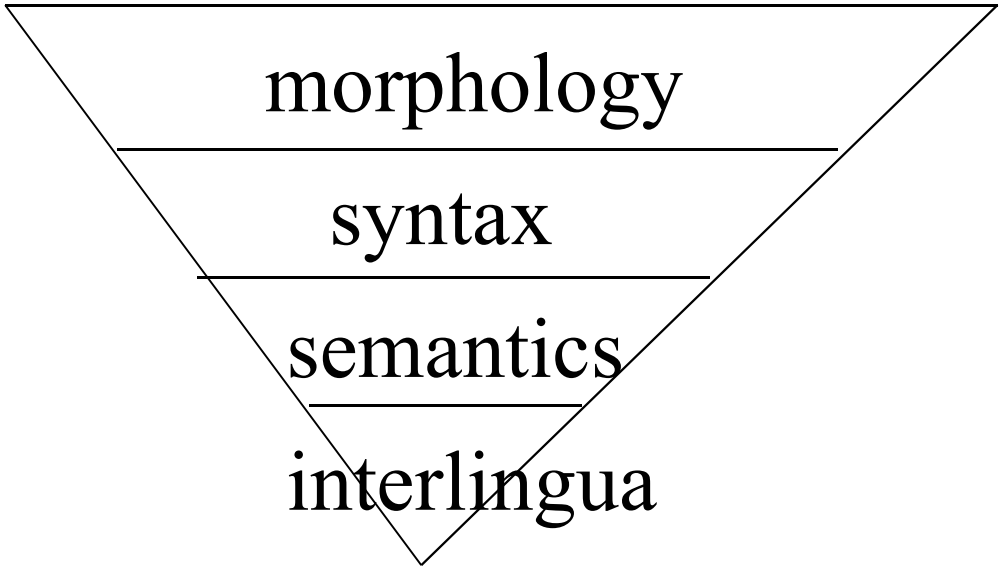
- Leon Dostert determined to show 'technical feasibility' of MT
- Collaboration of Georgetown University and IBM
- Public demonstration in New York, 1st January 1954 of Russian-English system
- Linguistic foundations by Paul Garvin of Georgetown U.
- Programming by Peter Sheridan of IBM
- Reported widely, worldwide interest
- Beginning of government funding - in both US and Soviet Union

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

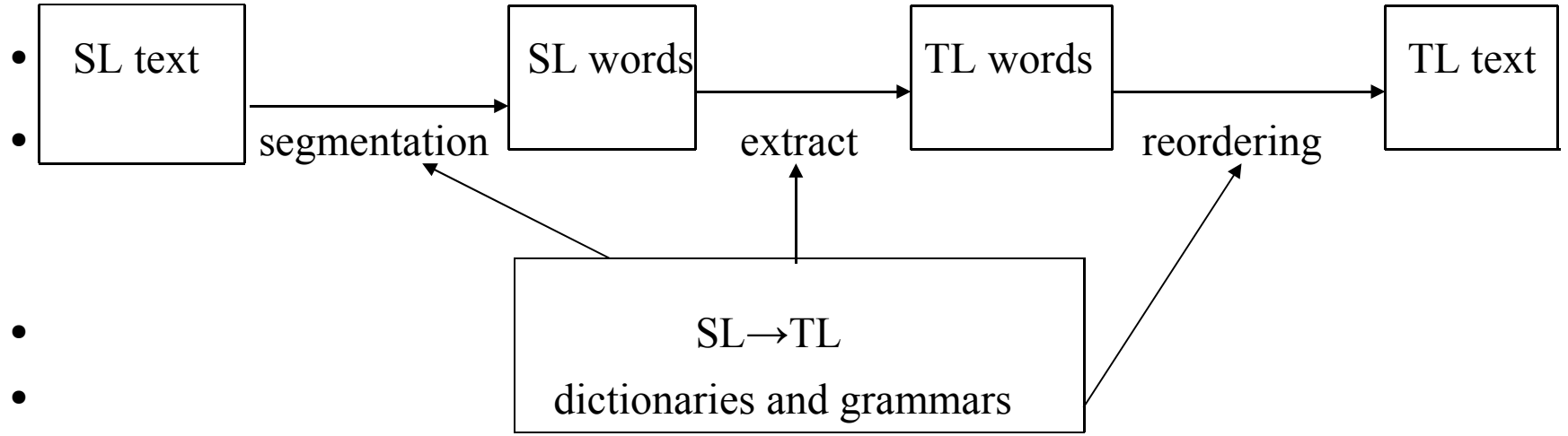
Main MT system types

- 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

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]
- 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]

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

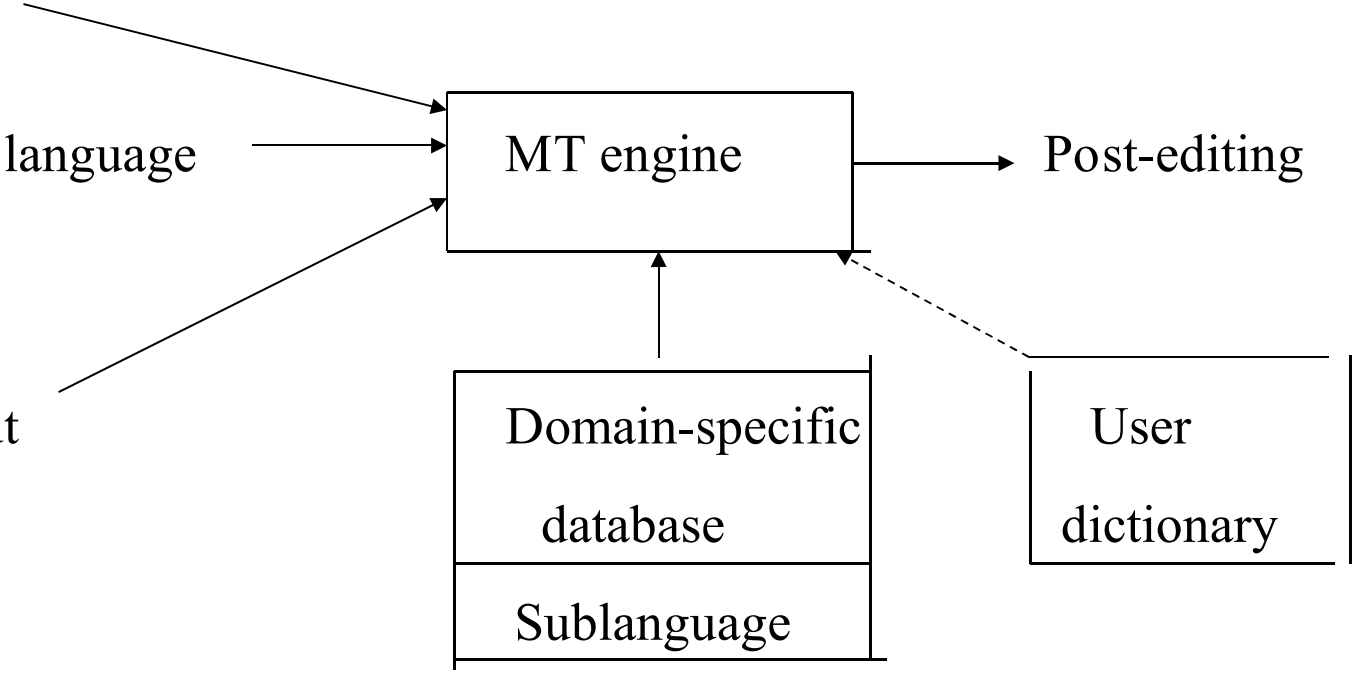
- identification of actual needs
  - assimilation vs. dissemination
- recognition that ‘perfectionism’ had neglected:
  - operational factors and requirements
  - expertise of translators
  - machine aids for translators
- henceforth three strands of MT:
  - translation tools
  - operational systems (post-editing, controlled languages, domain-specific systems)
  - research (new approaches, new methods)

From 1967 to 1976

- Continuation of research in US (Texas, Wayne State), Soviet Union, UK, Canada, France
- rule-based approaches: interlingua and transfer
- 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

Operational MT typical configurations

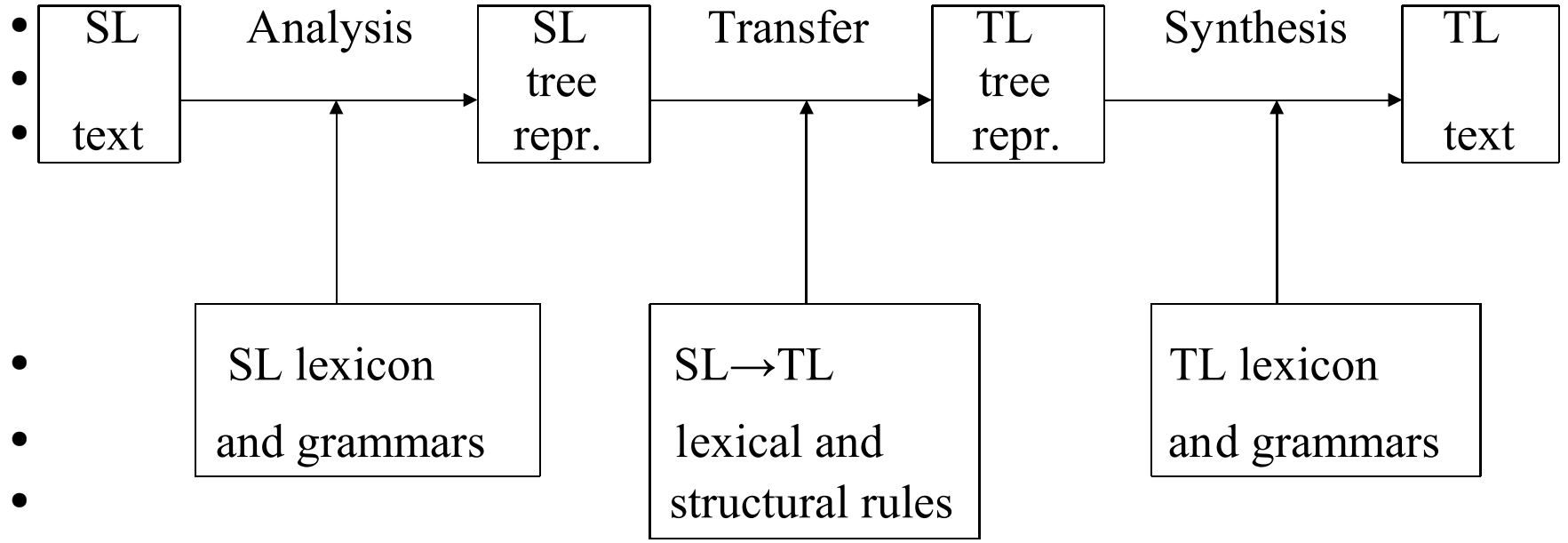
- Pre-edited input
- Controlled language input
- ‘Raw’ input



1978: transfer-based MT

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

Transfer-based MT model



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

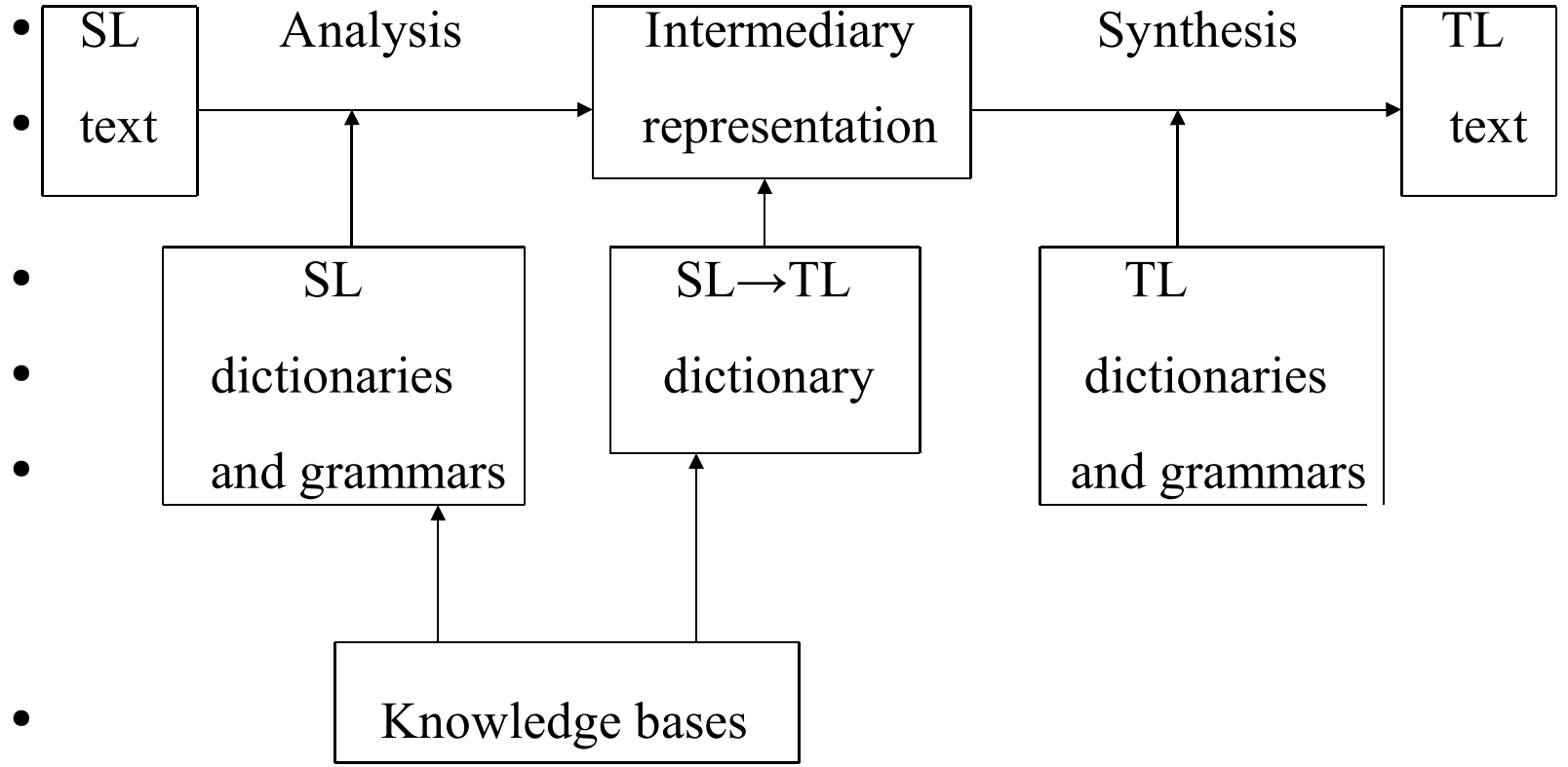
1981: MT for personal computers

- 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

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

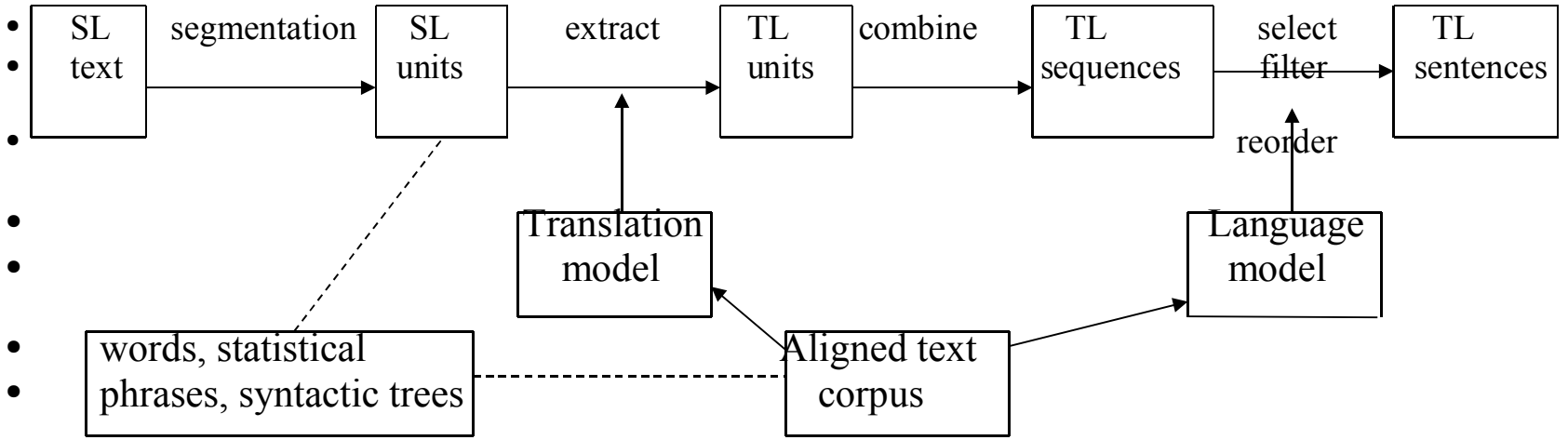
- ATR in Japan, JANUS at Carnegie-Mellon, Verbmobil (at various German universities)
- speech recognition, speech synthesis
- highly context dependent, use of ‘knowledge databases’
- discourse semantics, ‘ill-formed’ utterances
- ellipsis, use of stress, intonation, modality markers
- colloquial usage 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
- First article on Statistical MT, 1988 (research at IBM)
  - revival of Warren Weaver's idea ('decoding' SL as TL)
  - 'classic' SMT text in 1993
- Beginning of Example-based MT research, 1988-89
  - First proposed in 1981 by Makoto Nagao

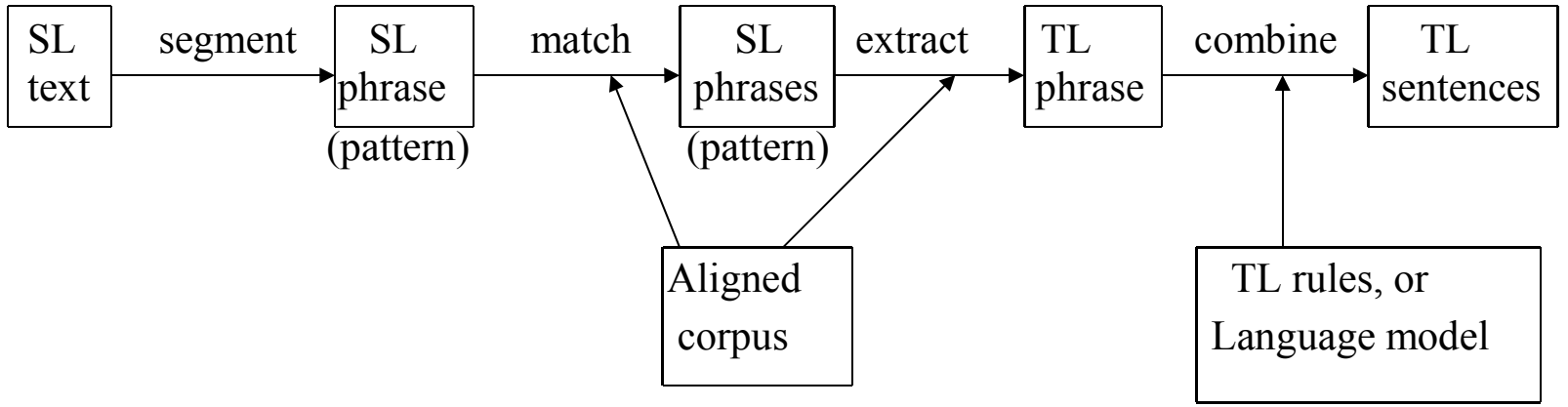
Statistics-based MT model

- 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



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

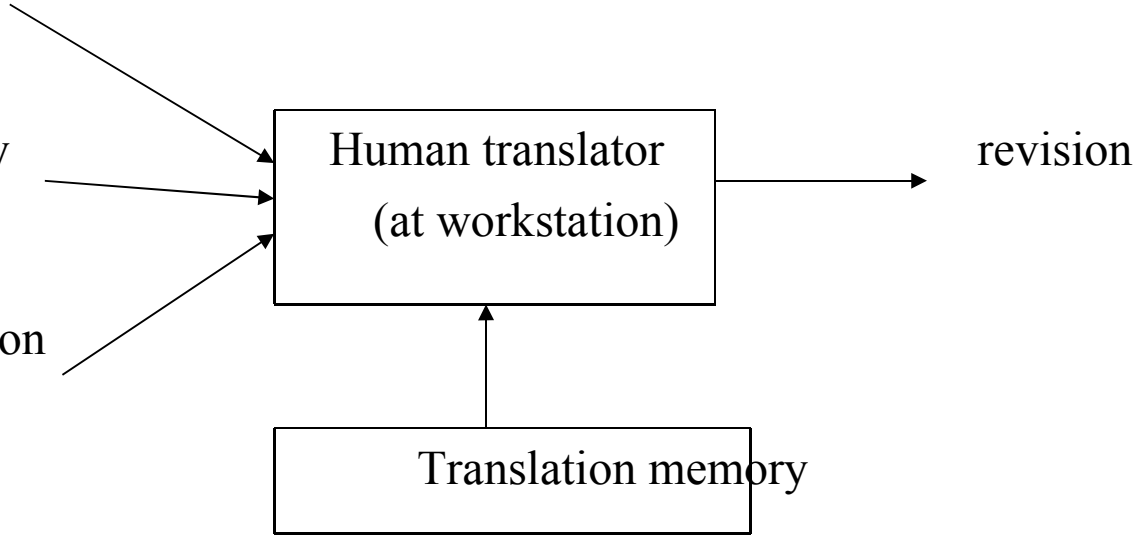


1993: translation memory

- 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

Machine-aided human translation

- Terminology
- database
- text glossary
-
- pre-translation
- (MT)
-



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)

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
- 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)

Resources

- Machine Translation Archive
  - http://www.mt-archive.info
- My website for history of MT
  - http://www.hutchinsweb.me.uk