Machine translation and translation aids:

systems, problems, uses, prospects

John Hutchins

(Email: WJHutchins@compuserve.com)
[http://ourworld.compuserve.com/homepages/WJHutchins]

Università di Bologna, SSLMIT, Forlí

December 2002

Contents

- Types of linguistics rule-based systems
- Corpus-based systems
- Problems for MT systems
- Computer-based translation tools, translation workstations, translation memories
- use of MT systems in large organizations
- MT for professional translators
- quality and evaluation
- MT for ‘home use’ and online
- MT and other language systems
- future prospects

The development of MT: 1950s and 1960s

- Sponsored by government bodies in USA and USSR (also CIA and KGB)
  - assumed goal was fully automatic quality output (i.e. of publishable quality) [dissemination]
  - actual need was translation for information gathering [assimilation]
- Survey by Bar-Hillel of MT research:
  - criticised assumption of FAHQT as goal
  - demonstrated 'non-feasibility' of FAHQT (without 'unrealisable' encyclopedic knowledge bases)
  - advocated "man-machine symbiosis", i.e. HAMT and MAHT
- ALPAC 1966, set up by disillusioned funding agencies
  - compared latest systems with early unedited MT output (IBM-GU demo, 1954), criticised for still needing post-editing
  - advocated machine aids, and no further support of MT research
  - but failed to identify the actual needs of funders [assimilation]
  - therefore failed to see that output of IBM-USAF Translator and Georgetown systems were used and appreciated

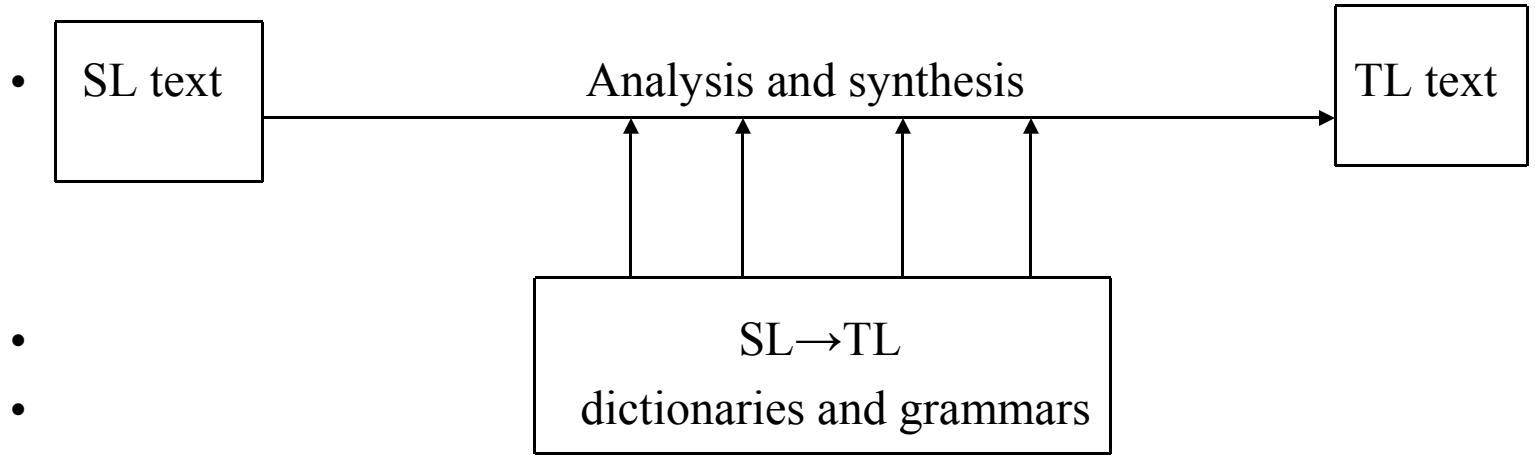
Consequences of ALPAC

- MT research virtually ended in US
- identification of actual needs
  - assimilation vs. dissemination
- full automation vs. HAMT and MAHT
- recognition that ‘perfectionism’ (FAHQT) 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)

System architectures and strategies

- Rule-based
  - Direct translation
  - Interlingua-based MT
  - Transfer-based MT
- Corpus-based MT
  - Statistics-based
  - Example-based
- Hybrid systems

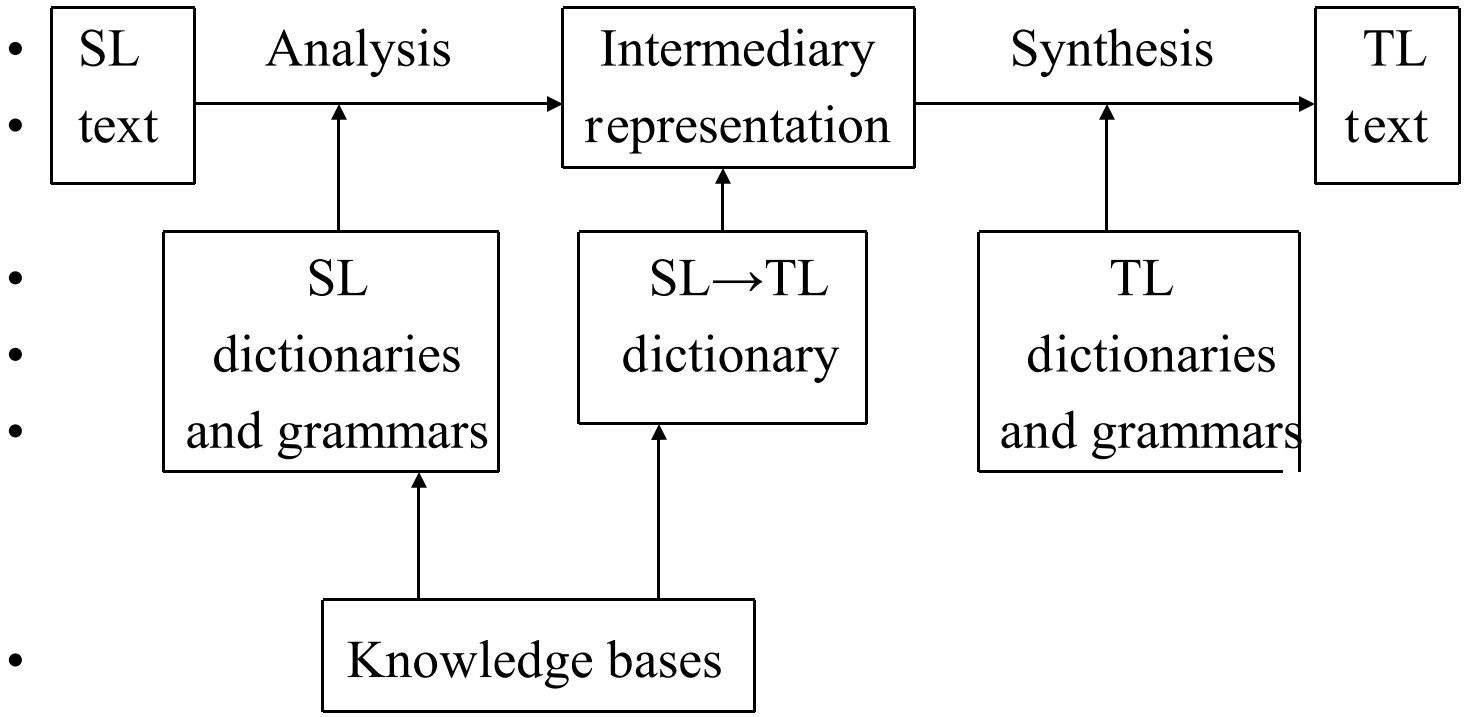
Direct translation



Direct translation

- Analysis of SL only as much as necessary for conversion into particular TL; dictionary lookup followed by TL word-for-word output, then TL rearrangement based on dictionary entries
- Use of ‘cover’ words (most frequent not most appropriate)
- no analysis of SL syntax or semantics
- output too close to SL structure
- example (Russian to English):
  - On dopisal stranitsu i otložil ručku v storonu.
  - It wrote a page and put off a knob to the side
  - (i.e.) “He finished writing the page and laid his pen aside”
- problems of direct translation systems:
  - too complex for modification and enhancement (not just computationally)
  - mixture of lexical rules and syntactic rules (no linguistic or translation ‘theory’)
- systems:
  - Univ.Washington, IBM (US), Georgetown University (US), Ramo-Wooldridge (US), Institute for Precision Mechanics and Computer Technology (USSR), National Physical Laboratory (UK)

‘Interlingual’ system



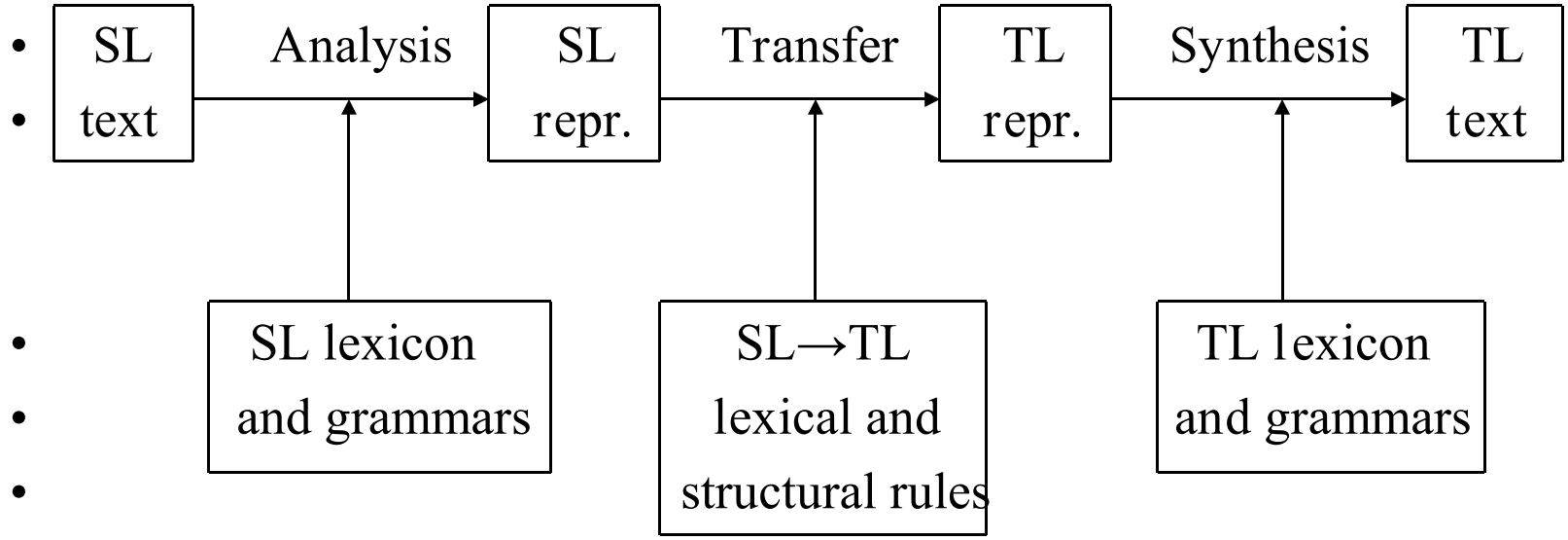
Interlingua-based MT

- two independent stages: analysis, synthesis
- abstract language-neutral representation
- multistratal: morphology, syntax, semantics
- semantics-oriented ('understanding'), hence domain-specific 'knowledge bases' (AI-oriented)
- problems:
  - nature of interlingua: natural, artificial, logical?; language-neutral or language-universal? (latter not feasible, most are 'neutral' for a few specific languages)
  - few 'pure' interlinguas: most only syntax, retain bilingual SL-TL lexicon
  - complexity of representations, complexity of fully disambiguating analysis
- projects:
  - [Trojanskij, 1933], Milan (Ceccato), Cambridge (CLRU), Grenoble (CETA), Texas (pre-METAL), [Mel'chuk (MMT)], Utrecht (DLT), Eindhoven (Rosetta), NEC (Pivot), Carnegie-Mellon University (KBMT, KANT, CATALYST), New Mexico State University (ULTRA, Pangloss), Univ. Maryland (UNITRAN), United Nations University (UNL)

Lexical entry in CMU system (*find*)

- (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 at home”)
  - (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))))))

‘Transfer’ system

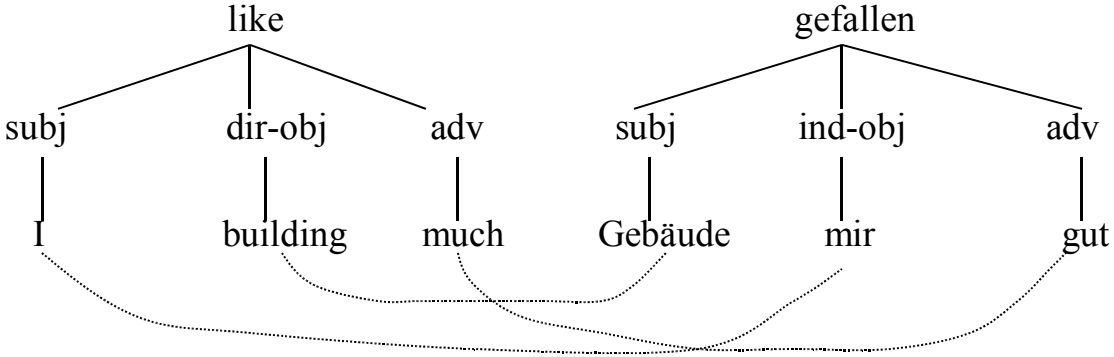


Transfer-based MT

- three stages: analysis, transfer, synthesis
- abstract semantico-syntactic interfaces/representations but basically syntax-oriented
- multiple level/strata: morphology, syntax, semantics
- problems:
  - failure at one analysis stage may mean no output
  - separation of morphological and syntactic analysis may not be relevant (in fact, many did not), similarly, distinction between syntax and semantics may not be helpful
  - distinction between interlingua-based and transfer-based often unclear (many combined features of both, e.g. Eurotra)
  - little/no discourse information (anaphora, etc.)
  - complexity of tree transduction rules
- projects/systems:
  - Georgetown University, MIT -- METAL (Texas), GETA-Ariane (Grenoble), SUSY (Saarbrücken) -- Eurotra, LMT (IBM), Mu (Japan), and many Japanese systems (JICST, Fujitsu, Toshiba…) -- many current PC systems

Tree transduction

- I like the new building very much ↔ Das neue Gebäude gefällt mir gut



- I like coffee ↔ ich trinke gern Kaffee
- He has just broken his leg ↔ il vient de se casser la jambe

Theories and formalisms

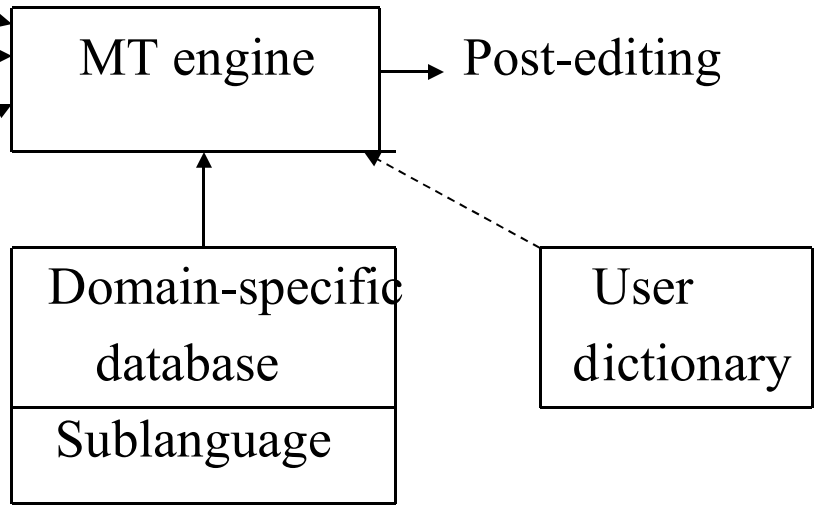
- Information theory (Shannon, Weaver, Yngve (MIT), Bar-Hillel, …)
- Transformational-generative grammar
- Dependency grammar
- Stratificational grammar (Lamb (UC Berkeley), Mel'chuk (MTM))
- Artificial intelligence (Wilks, Carnegie-Mellon)
- Lexical-functional grammar and Unification grammar
- Generalized phrase-structure grammar
- Definite clause grammar
- Principles and parameters, Government-binding theory (Univ.Maryland)
- Categorial grammar
- Montague grammar (Rosetta)
- Neural networks
- finite-state models
- adopted by fashions; single method failures; but none entirely superseded

Unification grammar:
example (LFG)

- SL f-structure *John likes Mary*
- [PRED like
- SUBJ [PRED John]
- OBJ [PRED Mary]]
- like, V:
- (↑ PRED) = like < SUBJ, OBJ >
- (↑ PRED FR) = plaire < SUBJ, OBJ >
- (τ↑AOBJ OBJ) = τ (SUBJ)
- (τ↑ SUBJ) = τ (OBJ)
- TL f-structure *Marie plaît à Jean*
- [PRED plaire
- SUBJ [PRED Marie]
- AOBJ [OBJ [PRED Jean]]]

Human-assisted MT

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



From 1967 to 1979

- 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
- 1979: Pan American Health Organization system (SPANAM)
- 1979: Eurotra project begins

MT research in 1970s and 1980s

- Rule-based systems:
  - involving long-term efforts compiling grammar rules (interlocking) and creating dictionaries
- Interlingua systems
  - DLT, Rosetta, Carnegie Mellon
- Transfer-based systems
  - GETA (Ariane), SUSY, Eurotra, Mu (Kyoto)
- Knowledge-based systems
  - Carnegie Mellon, New Mexico, Pangloss
- Speech translation
  - ATR, C-STAR, Verbmobil
- **Computer-based tools**

Changes since late 1980s

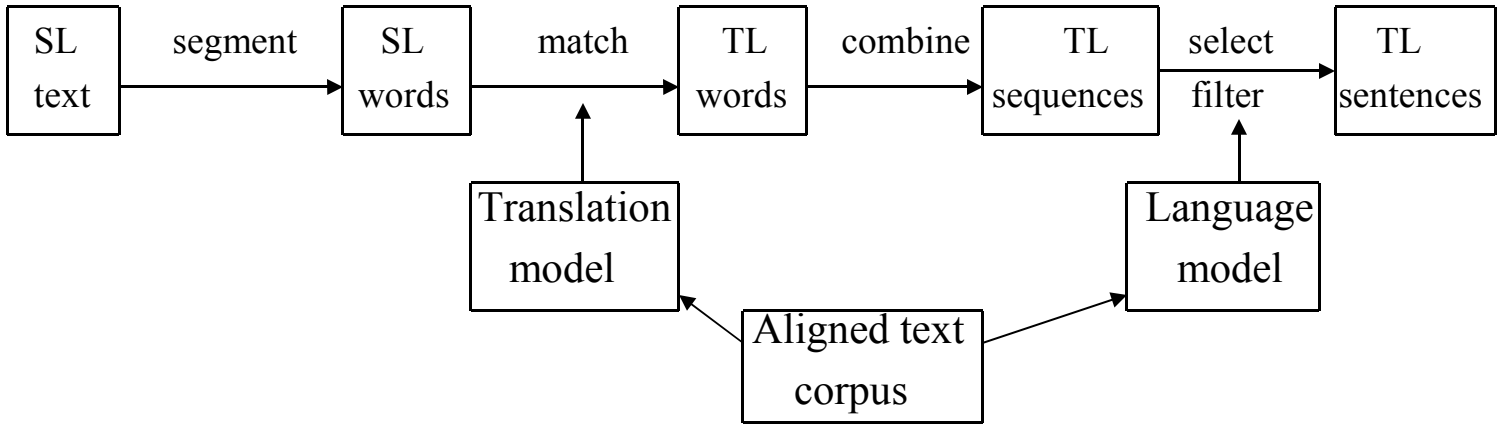
- Increasing use of MT by large enterprises
- Translation memory and translation workstations
- Localization
- Growth in PC systems
- The impact of the Internet
- Online translation
- MT and other language activities
- Research on corpus-based MT methods

Corpus-based systems

- Not rule-based: grammar rules (analysis, transfer, synthesis), multiple strata, 'deep' semantic analysis; complex dictionary entries
- based on bilingual text resources, e.g.
  - have a direct effect on… — ont une influence directe sur…
  - 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...
- Extraction of phrases for re-combination [Example-based MT]
- Statistical translation model (word-word frequencies), target language model (word co-occurrences) [Statistics-based MT]
- Text alignment methods enabled use of bilingual text corpora [Translation Memory]

Statistics-based MT

- Based on observations that translations observe statistical regularities
  - TL words are chosen as those most likely to correspond with the SL words in specific context
  - TL words are combined in ways most appropriate for the TL in a specific context/domain and style/register etc.



Statistics-based MT

- Bilingual corpora: original and translation
- little or no linguistic 'knowledge', based on word co-occurrences in SL and TL texts (of a corpus), relative positions of words within sentences, length of sentences
- Sentences aligned statistically (according to sentence length and position)
- compute probability that a TL string is the translation of a SL string ('translation model'), based on:
  - frequency of SL/TL co-occurrence in aligned texts of corpus
  - position of SL words in SL string, and TL words in TL string
- compute probability that a TL string is a valid TL sentence (based on a 'language model' of allowable bigrams and trigrams)
- search for TL string that maximizes these probabilities
- first example: IBM Candide (1988) on Canadian Hansard (English and French)

Statistics-based MT: problems

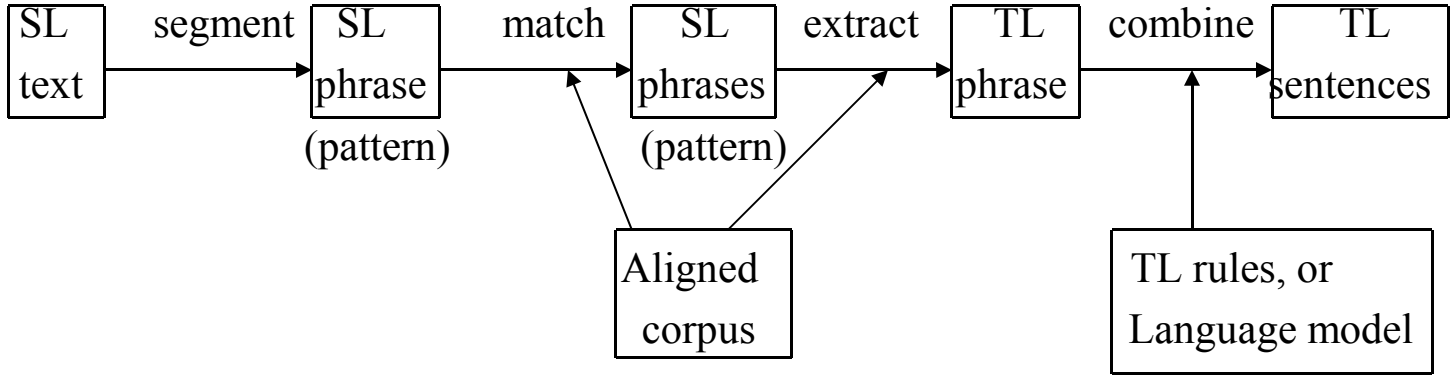
- still insufficient corpora
  - but Internet may solve this
- corpus must be aligned and analyzed before translation of (similar) text in same domain
  - unless large corpus for domain available
- word frequencies not sufficient: Candide intended to add morphological information, and some grammatical categories
  - some of this information may be statistically derived from large corpora
- most research aims to test how far purely statistical methods can go
  - laudable as research project, but not for developing working systems
  - in my view, some research needed on practicality of SMT for 'real' systems

Example-based MT

- Use of already translated sentences or phrases either from actual translations (corpus) or from data supplied by user or developer
- Sentences/phrases aligned in database (either by rule-based parser or statistically)
- matching algorithm (exact and close) of SL input and TL examples
- combination algorithm (for generating a TL sentence from extracted examples)
  - not trivial
- example (translating *no* from Japanese):
  - hoteru no juusho — address of the hotel
  - eigo no panfuretto — pamphlet in English
  - asu no tenkou — tomorrow's weather

Example-based MT

- 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



Example-based MT: some problems and issues (1)

- bilingual aligned corpora
  - size: adding examples may improve performance or may degrade performance
  - repetition of same or similar examples may reinforce selection or may be unnecessary clutter
  - suitability of examples: automatically compiled or manually compiled
  - need: phrases/clauses aligned (not sentences), length is open issue
  - stored: as word strings or as annotated trees(e.g. dependency or case grammar trees)
- analysis of corpus at run-time or in advance
- use of grammatical categories (patterns)
  - templates (e.g. <1st name><family name> flew to <city> on <date>)
  - X [pron] eats Y [noun/NP] ↔ X [pron] ga Y [noun/NP] o taberu
  - X o onegai shimasu → may I speak to the X (if X=jimukyoku ‘office’, … etc.); or: please give me the X (if X=bangō ‘number’, … etc.)

Example-based MT: some problems (2)

- matching by characters:
  - This is shown as A in the diagram ↔ This is shown as B in the diagram
  - The large paper tray holds up to 400 sheets <≠> The small paper tray holds up to 300 sheets
    - (because system does not know that *large* and *small* are similar/substitutable)
- matching by words via thesaurus (close in meaning)
  - English eat → Japanese taberu or okasu::
  - A man eats vegetables↔ Hito wa yasai o taberu
  - Acid eats metal↔ San wa kinzoku o okasu
- recombination (joining example phrases):
  - AIDS control programme for (…) ↔ programma contra el SIDA para (…)
  - (…) Ethiopia ↔ (…) Etiopia; (…) Spain ↔ (…) España; etc.
  - AIDS control programme for Ethiopia → programma contra el SIDA para Etiopia
- but problem of 'boundary friction'
  - that old man has died ↔ ce vieil homme est mort
  - that old woman has died ↔ (**not simple substitution**: ce viel femme est mort), **but**: cette vieille femme est morte

Example-based MT: some problems and issues (3)

- Examples in database:
  - (1e) The obstinate man refused all help
  - (1g) Der hartnäckige Mann hat alle Hilfe verweigert
  - (2e) Help was rejected by the stubborn man
  - (2g) Hilfe wurde von dem starrköpfigen Mann abgewiesen
- sentence to be translated:
  - (3e) Help was rejected by the obstinate man
- fail to relate verweigern and abweisen, and hartnäckig and starrköpfig
- and boundary friction if example for 'obstinate man' inserted into (2g):
  - (3g) * Hilfe wurde von der hartnäckige Mann abgewiesen.

- **In general, morphological variation handled more easily by rule-based systems than by corpus-based systems**
- **In general, EBMT research seeks practical solutions for working systems (whether fully example-based or not)**

Translation databases: lexical differences

- Translation of German adjective **stark**:

• Das ist ein **starker** Mann	This is a **strong** man
• Es war sein **stärkstes** Theaterstück	It has been his **best** play
• Wir hoffen auf eine **starke** Beteiligung	We hope a **large** number of people will take part
• Eine 100 Mann **starke** Truppe	A 100 **strong** unit
• Der **starke** Regen überraschte uns	We were surprised by the **heavy** rain
• Maria hat **starkes** Interesse gezeigt	Mary has shown **strong** interest
• Paul hat **starkes** Fieber	Paul has **high** temperature
• Das Auto war **stark** beschädigt	The car was **badly** damaged
• Das Stück fand einen **starken** Widerhall	The piece had a **considerable** response
• Das Essen was **stark** gewürzt	The meal was **strongly** seasoned
• Hans ist ein **starker** Raucher	John is a **heavy** smoker
• Er hatte daran **starken** Zweifel	He had **grave** doubts about it

Bilingual lexical differences

- bilingual lexical ambiguity (more than one equivalent, whether ambiguous in SL or not):
  - river: fleuve/rivière
  - Taube: dove/pigeon
  - Schraube: screw/bolt/propellor
  - corner: coin or angle; Ecke or Winkel
  - light: léger, clair, facile, allumer, lumière, lampe, feu
  - look: regarder, chercher, sembler
- lexical gaps
  - dacha, cottage, marmelade, vodka, etc.
  - snub: infliger un affront; verächtlich behandeln, or: derb zurückweisen
  - het Turks kennen: to know Turkish
  - kenner van het Turks: *knower of Turkish, someone who knows Turkish

- **Solved (?) by contextual rules (RBMT), or examples (EBMT), or frequencies and 'language models' (SMT)**

Structural ambiguity

- (1) Peter mentioned the book I sent to Mary
  - Peter mentioned the book which I sent to Mary
  - Peter mentioned to Mary the book which I sent [to Peter/David]
- (2a) We will meet the man you told us about yesterday
  - … the man you told us about yesterday
- (2b) We will meet the man you told us about tomorrow
  - we will meet tomorrow the man...
- (3a) pregnant women and children
  - des femmes et des enfants enceintes
- (4a) Smog and pollution control are important factors
- (4b) Smog and pollution control is under consideration
- (4c) The authorities encouraged smog and pollution control
- (5a) old men and women may usually mean ‘old men and old women’
- (5b) [but perhaps not in] Tickets were refunded for children, old men and women,
- **Problems (1), (2), (3), and (5a) may be ‘solved’ by SMT’language model’ and by EBMT databases. But problems (4c) and (5b) require ‘knowledge’ (i.e. rule-based KBMT)**

Bilingual structural differences

- (1) Young people like this music
  - Cette musique plaît aux jeunes gens
- (2) The boy likes to play tennis
  - Der Junge spielt gern Tennis
- (3) He happened to arrive in time
  - Er ist zufällig zur rechten Zeit angekommen
- (4) Le moment arrivé je serais prêt
  - When the time comes, I shall be ready

- **Difficult to specify rules (RBMT) to cover all circumstances and contexts; example-based (EBMT) and statistics-based (SMT) approaches yet to prove any better. Possibly examples like no.4 are inherently unsolvable**

Anaphora

- Die Europäische Gemeinschaft und ihre Mitglieder
  - The European Community and its members
- The monkey ate the banana because it was hungry
  - Der Affe ass die Banane weil er Hunger hat
- The monkey ate the banana because it was ripe
  - Der Affe ass die Banane weil sie reif war
- The monkey ate the banana because it was lunch-time
  - Der Affe ass die Banane weil es Mittagessen war
- Particular problem when translating from Japanese when it is good style to omit the subjects of verbs and to avoid repetition.

- **Sentence-orientation of all systems makes most anaphora problematic (unresolvable); possibly only a discourse-oriented 'language model' is the only chance**

Non-linguistic problems of 'reality'

- The soldiers shot at the women and some of them fell
- The soldiers shot at the women and some of them missed
  - must know what 'them' refers to e.g. if translating into French (ils or elles)

- **No solutions with linguistic rule-based approaches**
- **No solutions with corpus-based approaches**
- **Perhaps only solution using Artificial Intelligence approaches (Knowledge-based machine translation, e.g. Carnegie-Mellon University)**

- However, perhaps this problem is exaggerated: no need to understand what AIDS and HIV are in order to translate:
  - The AIDS epidemic is sweeping rapidly through Southern Africa. It is estimated that more than half the population is now HIV positive.

Problems of stylistic difference

- The possibility of rectification of the fault by the insertion of a valve was discussed by the engineers
- The engineers discussed whether it was possible to rectify the fault by inserting a valve

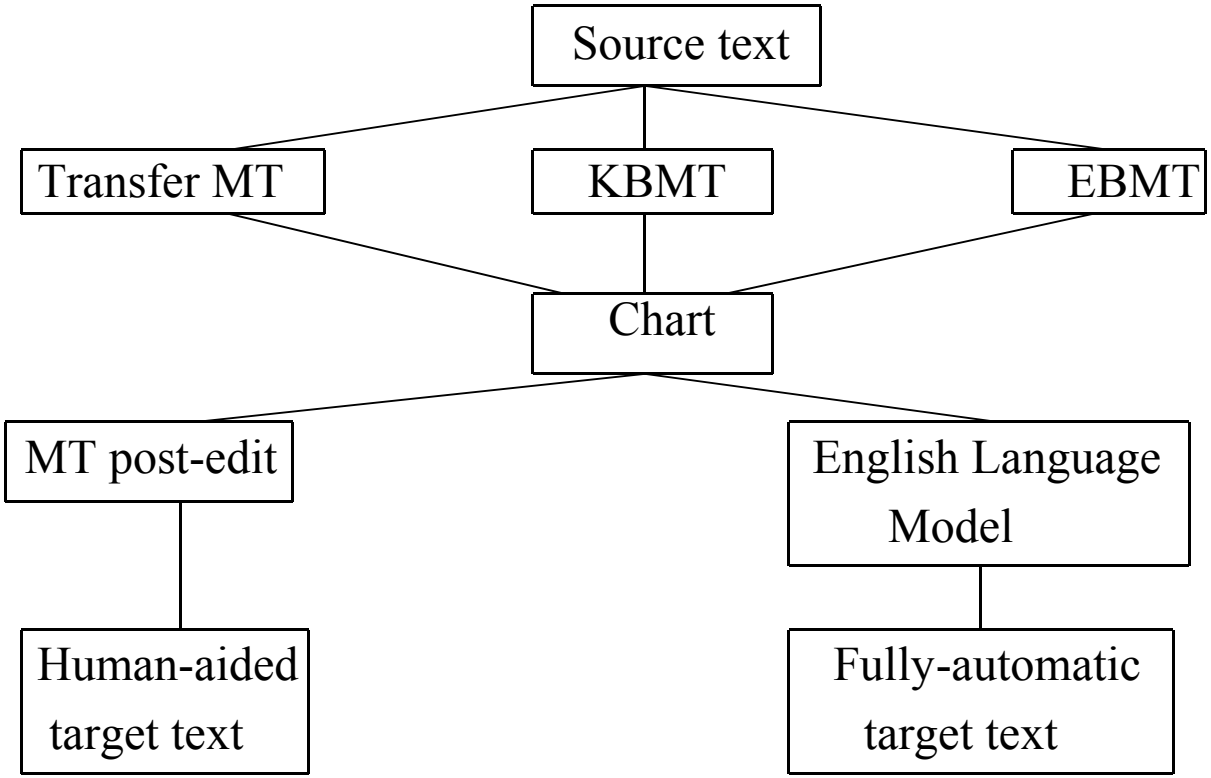
- [English] Advances in technology created new opportunities
- [Japanese] Because technology has advanced, opportunities have been created
- [or Japanese] Technology has advanced. There are new opportunities.

- **All methods of MT tend to retain SL structural features; however, theoretically SMT 'language model' approach should be more TL-oriented.**

Hybrid systems

- clearly, none of the current MT ‘models’ are capable of solving all problems
- hence search for hybrid architectures
- in theory, it would seem that (on average):
  - RBMT better for SL analysis
  - EBMT better for transfer
  - SMT best for TL generation
- Problem is that different approaches not easily compatible:
  - there are however research prototypes combining:
    - EBMT with statistical methods
    - EBMT using rules similar to those in RBMT systems
  - perhaps a version of EBMT will be the answer
- Currently ‘hybrid’ systems are parallel systems with a selection mechanism, as in:

Hybrid systems:
an example (Pangloss Mark III)



Speech translation

- Speech recognition, speech synthesis
- highly context dependent, use of ‘knowledge databases’
- discourse semantics, ‘ill-formed’ utterances
- ellipsis, use of stress, intonation, modality markers
- restricted domain (e.g. hotel booking by telephone)
- systems: ATR (Japan), JANUS (US, Germany), SLT (SRI, Cambridge), Verbmobil (Germany), DIPLOMAT (Carnegie-Mellon)
- examples:

– Der Montag, der passt mir gut	The Monday suits me fine
– Sie haben Zeit?	(do) you have time?
– Ich habe am Freitag bloss keine Zeit	But I don’t have time on Friday
– Ich habe bloss am Freitag Zeit	I have time only on Friday
– Ich muss nach Hannover	I have to go to Hanover
– klingt gut!	sounds OK!
– bin einverstanden!	agreed!

Speech translation: problems

- speech recognition, speech synthesis
- highly context dependent, use of ‘knowledge databases’
- discourse semantics, ‘ill-formed’ utterances
- ellipsis, use of stress, intonation, modality markers
- restricted domain (e.g. hotel booking by telephone)
- colloquial usage not yet investigated sufficiently (even in linguistics)

- half-way solutions (?) available with voice input/output

Voice input/output

- Word processing add-ons:
  - Dragon Naturally Speaking, IBM ViaVoice
- PC translation systems with voice input/output
  - Al-Wafi, CITAC, ESI, Korya Eiwa, Personal Translator PT, Reverso Voice, TranSphere, Vocal PeTra, ViaVoice Translator
- Online translation with voice output
  - Translation Wave

Computer-based tools (1980s)

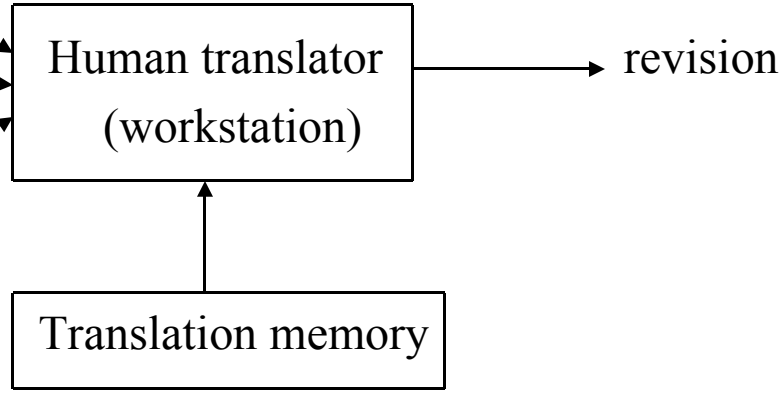
- Term banks: TEAM, LEXIS, TERMIUM, Eurodicautom
- Terminology management (Mercury/Termex)
- Text-related glossaries (Bundeswehr, ALPS)
- Translation databases ('translation memory')
  - first: Arthern (1978), Kay (1980), ALPS
- Melby's three levels (early 1980s)
  - word processor with integrated terminology aids, manual insertion of words
  - machine-readable input texts, concordance (to find occurrences of words in text), local term bank, automatic insertion of terms
  - integrated 'workstation' with MT system, and automatic 'quality' evaluation

Computer-aided translation and translation tools

- recognition that fully automatic translation not appropriate for professional translators
- PCs and multilingual word processing, desk top publishing
- Translator ‘in control’
- dictionaries (monolingual, bilingual): on-line access
- grammar aids, spelling checkers
- user glossary, terminology management, ‘authorised’ terms, standards, specialist glossaries
- input, output, transmission (OCR, pre-editing, controlled language)
- translation memory, alignment
- management support tools (project control, budgeting, workflow)
- previous antagonism of translators to MT diminished

Machine-aided human translation

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



Terminology management

- domain or customer specific; company or individual translator
- involvement: translators, terminologists, database managers
- extraction and selection (bilingual databases)
- content of entries for terms:
  - category/classification; definition; grammatical information; usage (country); standards; technical note; translation; context, example of use; source
- authorization
- updating and corrections
- sharing/transfer/exchange: MATER
- standards/conferences: InfoTerm
- examples: hundreds in Europe: TEAM, LEXIS, TERMIUM (early examples), Eurodicautom
- software: MultiTerm (Trados), MTX (Linguatech)

Translation memory

- based on sets of original texts and their ‘authorized’ translations
- particularly suitable for translation of revisions and for translating standardized documents
- most suitable for large (organizational) translation agencies/departments
- alignment of bilingual text corpora
- revised texts (i.e. updated documents) are checked against corpus for any changes; for unchanged source sentences, the ‘authorized’ translation is retained
- search of exact matches or ‘fuzzy’ matches
- extract target phrase for insertion and/or amendment (by human translator)
- still much post-editing, and there is need for programs to ‘meld’ or conflate extracted phrases (semi-automatically)
- problems of unnecessary examples (overload) and untypical or rare translations
- problems of fuzzy matching without linguistic information (e.g. morphological variants)

Translation workstations

(often called Translation memory systems): requirements

- Components and facilities controlled by users (translators)
- Terminology management
- Translation memory, and alignment
- Facilities for building dictionaries (e.g. from Internet)
- Augmented by MT systems
- Compatible with authoring systems (technical writers)
- Compatible with publishing systems

Workstations (TM systems) available

- Trados Translation Solution
- STAR Transit
- Déjà Vu (Atril)
- SDLX (SDL Corporation)
- Multilizer (Multilizer Inc.)
- LogiTerm (Terminotix)
- WordFast (Champollion)
- MultiTrans (MultiCorpora)
- MetaTexis (MetaTexis Software)
- WordFisher (K.Tibor)
- MemorySphere (AppTek)
- CATALYST (Alchemy)
- ForeignDesk (Lionbridge)
- Xerox XMS

EURAMIS

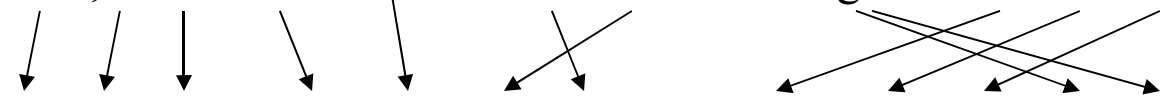
- European Commission's translation workstation network
- European Advanced Multilingual Information System
- Combination of tools for EC Translation Service, with single interface:
  - translation memory (Trados)
  - terminology extraction and management tool (MultiTerm)
  - Systran
  - Eurodicautom, other term bases
- documents transmitted over Commission internal network
  - from any EC administrator, etc.
  - accepted in Word, WordPerfect, Excel
  - automatic conversion to SGML
- Transmission by email
- post-editing by translators

Problems of alignment (1)

- bilingual corpora
  - suitability (i.e. appropriate domain, style, audience)
  - availability, e.g. for uncommon languages (lack of electronic resources)
- matching sentence lengths (for European languages, not for English/Japanese)
- matching words
  - cognates: first four letters and 'same' meaning (*mathematics* and *mathématique*)
    - - but failed for *government/gouvernement,* and *actual/actuel*
  - morphological patterns: *book/books, box/boxes, lady/ladies, wife/wives*, etc.
- using bilingual dictionaries (as seed for alignment: simple word pairs)

Problems of alignment (2)

- Work best for word-to-word alignment
  - well, I think if we can make it at eight on both days
  - ja, ich denke wenn wir das hinkriegen an beiden Tagen acht Uhr



- Difficulties when a SL word group (phrase) corresponds to TL word group
  - yes, then I would say , let us leave it at that.
  - Ja, dann würde ich sagen , verbleiben wir so.

- Problems with inadequate training corpus

Translation memories: weaknesses

- major gains (time saving, etc.) from retrieving already translated text
- sentence-based comparisons restrict potential use (no phrase matching)
- any TM likely to contain redundant, ambiguous versions
- any TM likely to contain conflicting translations (with little or no guidance)
- sentences are edited by translators outside TM environment and therefore not included in the database
- TM systems do not 'learn' decisions/choices made by users (e.g. which potential translations are preferred, which rejected)
- fuzzy matching often too complex, and translators opt not to use the facility
- combining extracted translation segments left entirely to user/translator
- developments needed:
  - **finding phrases (retrieval, fuzzy matching)**
  - **combining phrases; searching for words in combination**

Translation memory: searching

– Query: **take+…swipe+**

• The Conservatives not being satisfied • with the cuts the Liberals made to the • Established Programs Financing, have • **taken three successive swipes** at it.	Les conservateurs, insatisfaits des réductions apportés par les libéraux au Financement des programmes établis, s'y sont attaqué à trois reprises.
• Speaking more extemporary, yes, I did • **take a swipe** at the activities of the • President of the United States.	Dans mes propos un peu plus improvisés, je m'en suis effectivement pris aux activités du président des États-Unis.
• Every time we look around someone • else is **taking a swipe** at health care • in this country.	À tout moment, on porte une nouvelle atteinte aux soins de santé dans notre pays.

Translation memories: other problems

- Expensive to build in time and money
- Loss of context (beyond sentence), e.g. domain of document
  - may need to translate whole
- Fuzzy matching ineffective
  - occurrence of different (hidden) formatting tags
  - recombination of fuzzy matches is longer than translation from scratch
- Whole sentence repetition is rare
  - limited value for administrative documents, minutes of meetings, marketing texts, most reports, web sites
- **Repetition of phrases, clauses much more common. Therefore need 'example-based' approach**

The translation demand

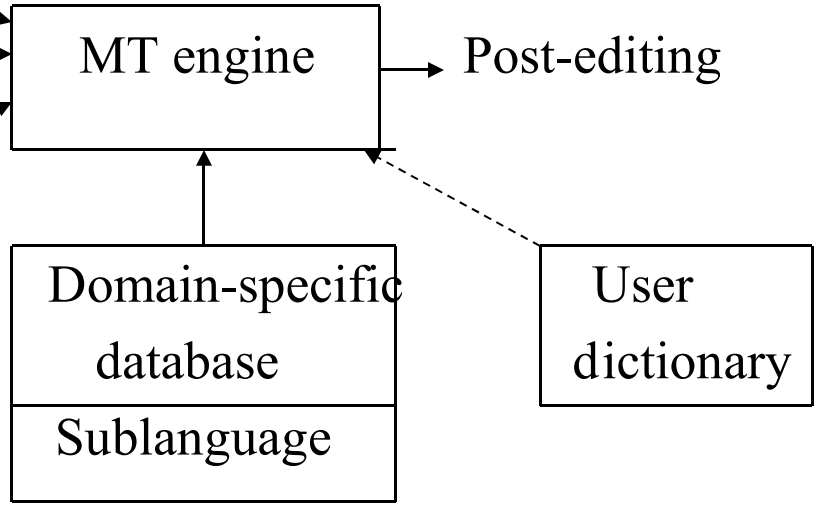
- Dissemination: production of ‘publishable quality’ texts
  - but, since raw output inadequate:
    - post-editing
    - control of input (pre-editing, controlled language)
    - domain restriction (reducing ambiguities)
- assimilation: for extracting essential information
  - use of raw output, with or without light editing
- interchange: for cross-language communication (correspondence, email, etc.)
  - if important: with post-editing; otherwise: without editing
- information access to databases and document collections
  - limited use before 1990

System types from the users' viewpoint

- The differences between system architectures and methods:
  - Direct translation
  - Interlingua-based translation
  - Transfer-based translation
  - Statistics-based translation
  - Example-based translation
  - 'Hybrid' systems
- are largely irrelevant.
- Users are normally only concerned with
  - compiling and/or augmenting dictionaries
  - storing texts for translation memory systems
- In theory any MT systems can be used for any of the functions (dissemination, assimilation, interchange, information access)

Human-assisted MT

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



Large-scale translation and MT

- accurate, good quality, publishable (dissemination)
- publicity, marketing, reports, operational manuals, localization
- technical documentation; large volumes
- repetitive, frequent updates; saving costs (and staffing?)
- multilingual output (e.g. English to French, German, Japanese, Portuguese, Spanish)
- available in-house terminological database; user (company) dictionaries
- backup resources (translated texts, personnel for dictionaries, etc.)
- human assistance for quality (controlled language input, post-editing)
- integrate with technical writing and publishing
- availability of in-house printing/publishing
- technical expertise (computers, printers, etc.)

Operational systems in 1980s: examples

- Systran
  - Ford, General Motors, Aerospatiale, Berlitz, US Air Force, National Air Intelligence Center, Foreign Broadcasting Information Service, Xerox, European Commission
- Logos
  - Ericsson, Lexi-Tech, Osram, Océ Technologies, SAP
- METAL
  - Boehringer Ingelheim, Philips, Union Bank of Switzerland, SAP

Systran at EC

- Uses and users:
  - administrators
    - browsing texts in unknown language, deciding whether to submit for human translation
    - fast rough translation of urgent texts, often with rapid post-editing; possible internal distribution
    - drafting texts in non-native languages
  - translators
    - as drafts (or basis) for polished translations
    - for post-editing of internal documents
  - interpreters
    - as basis for translation of complex oral reports

Systran at EC (contd.)

- languages:
  - English to French (1976), Italian (1978), German (1982), Dutch (1984), Spanish (1985), Portuguese (1985), Greek (1988)
  - French to English (1977), German (1982), Dutch (1984), Italian (1989), Spanish (1990)
  - German to French (1980), English (1988)
  - Spanish to English (1990), French (1991)
  - tested: French to Portuguese (1997), Greek to French (1993), more to come
- growth of demand: five times since mid 1990s, over 20% per annum
- and quality can be improved

Post-editing

- Why needed?
  - Misspelling in original not recognised, therefore not translated
  - missing punctuation
    - e.g. *The Commission vice president* translated as *Le président du vice de la Commission* (because no hyphen between *vice* and *president*)
  - complex syntax
- Always necessary?
  - More standardised, more jargon-full documents mean less correction
- Can it be avoided?
  - If rough version acceptable

Post-editing: types of errors

- What types of mistakes need correction?
  - prepositions:
    - …el desarrollo de programs de educación nutricional...
    - MT: …the development of programs of nutritional education
    - PE: …**in** nutritional education…
  - verb phrases:
    - …el procedimento para registrar los hogares…
    - MT: the procedure in order to register the households
    - PE: …the procedure for registering households

Post-editing: types of errors (contd.)

- inversions:
- …la inversión de la Argentina en las investigaciones de malaria
  - MT: …the investment of Argentina in the research of malaria
  - PE: Argentina's investment in malaria research
- reflexive verbs with inversions:
- Se estudiarán todos los pacientes diagnostocados como…
  - MT: There will be studied all the patients diagnosed as…
  - PE: Studies will be done on all patients diagnosed as…
- En 1972 se formuló el Plan Decenal de Salud para las Américas.
  - MT: In 1972 there was formulated the Ten-Year Health Plan for the Americas
  - PE: The year 1972 saw the formulation of the Ten-Year Health Plan for the Americas.

Translators and post-editors

- post-editing by translators:
  - not foreseen initially
  - skills acquired over time and practice in real working conditions
  - requires perseverance (initially post-editing takes longer than complete translation)
- advantages:
  - translators can maintain quality control
  - consistency of terminology
  - repetitive matter produced by MT, linguistic quality by HT
- disadvantages:
  - correction of ‘trivial’ mistakes; too often correcting same type of error
  - style too much SL oriented
  - translators as ‘slaves’ to machine
- need for special post-editing tools (not always provided)
- specially trained post-editors [still rare]

Adaptation of input

- MT-ese
  - writing with MT in mind (i.e. to avoid ambiguities)
- pre-editing
  - marking words for grammatical category
    - e.g. *convict* as noun or verb
  - indicating proper names
    - e.g. to ensure that *John White* is not translated as *Johann Weiss*
  - indicating compound nouns
    - e.g. to translate *light bulb* as *ampoule* and not *bulbe léger* or *oignon léger*
  - marking parenthetical phrases
    - e.g. *There are he says two options…* as *There are (he says) two options...*
  - dividing sentences into shorter clauses
  - in theory, need not know target language(s)

Adaptation of input (contd.)

- Pre-editing now not common
  - [except in some cheaper PC systems]
  - no change of style or grammar or vocabulary; no feedback to authors
- sublanguages
  - the success of Météo has led to search for other sublanguages
    - e.g. avalanche warnings -- (research project in Switzerland)
- adjusting systems to restricted domains
    - primarily via dictionary entries: single equivalents for SL terms
  - but without imposing constraints on original texts
- controlled language input
  - in practice, the more favoured approach

Controlled language

- Controlled authoring of the source text in standard manner, suitable for unambiguous translation
- Typical rules:
  - use only approved terminology, e.g. *windscreen* rather than *windshield*
  - use only approved sense: *follow* only as ‘come after, not ‘obey’
  - avoid ambiguous words: *replace*, either (a) remove and put back, or (b) remove and put something else in place; not *appear* but: come into view, be possible, show, think
  - only one ‘topic’ per sentence, e.g. one instruction, command
  - do not omit articles
  - do not use pronouns instead of nouns if possible
  - do not use phrasal verbs, such as *pour out*
  - do not omit implied nouns
  - use short sentences, e.g. maximum 20 words
  - avoid co-ordination of phrases and clauses

Controlled languages (contd.)

- Typical rules (contd.):
  - do not omit articles; use relative pronouns (*which, in order that*); avoid post-nominal gerundive forms (*wires connecting…* → *wires that connect…*)
  - avoid anaphora; prefer nouns instead of pronouns if possible
  - do not omit implied nouns; avoid ellipsis
  - avoid phrasal verbs, such as *pour out*
  - use short sentences, e.g. maximum 20 words
  - avoid co-ordination of phrases and clauses

  - advantage of controlled language is improvement of original SL text
  - sometimes translation no longer necessary
  - later revision can be faster

Controlled languages: examples

- Example sentences:
  - ***not***: After agitation, allow the solution to stand for one hour
  - ***but***: If you shake the solution, do not use it for one hour.
  - ***not***: It is very important that you keep all of the engine parts clean and free of corrosion.
  - ***but***: Keep all of the engine parts clean. Do not let corrosion occur.
- Old idea -- 'Model English' (Stuart Dodd, 1952):
  - she did be loved; I will send he to she
- Controlled languages:
  - AECMA
  - MCE (Xerox), using Systran
  - PACE (Perkins Engines), using Weidner system

Controlled language systems

- Caterpillar (CTE)
  - 800 pages per day; 12 languages regularly, 10 others some material; 6 months delivery cycle
  - Caterpillar Fundamental English (1000 words) - abandoned
  - since 1994: collaborated with Carnegie Group to create Caterpillar Technical English (70,000 terms and phrases; acceptable syntactic constructions)
  - for interlingua- (knowledge-) based KANTOO system
  - terms not unambiguous, uses SGML codes (during authoring) to help disambiguate
  - post-editing still needed

Custom-built controlled-language systems

- LANTMARK [Xplanation b.v., Belgium]
  - Dutch↔French, English↔French, English↔German, English→Spanish, German→French, German→Spanish
- Smart Translator [Smart Corporation, New York]
  - English↔French (European or Canadian), English↔German, English→Greek, English↔Haitian Creole, English→Chinese (Mandarin), English↔Portuguese (Brazilian), English↔Spanish (Castilian or Latin American)
  - clients: Citicorp, Chase, Ford, General Electric, Canadian Ministry of Employment
- WebTran [VTT Information Technology, Finland]
  - European languages
- Cap Volmac
- ESTeam Ltd. (Greece)
  - Danish, Dutch, English, French, German, Greek, Icelandic, Italian, Norwegian, Spanish, Swedish.

In-house systems: examples

- Pan American Health Organization [medical, social, welfare]
  - Spanish→English (SPANAM), English→Spanish (ENGSPAN), Portuguese→English (PORTENG)
- Japan Center for Science and Technology [abstracts]
  - English→Japanese
- NHK [news broadcasts]
  - Japanese→English
- IBM Japan
- CSK (Japan)
- PaTrans:[patents]
  - English→Danish
- GSI Erli
- Hook and Hatton [chemistry texts]
  - Dutch→English

Other special-purpose systems

- Police, customs, air traffic control (Linguanet)
  - Danish, Dutch, English, French, German, Italian, Portuguese, Spanish
- TV captions (ALTo: English to Spanish)
  - spoken language transcription
  - sentence segmentation, word identification, name recognition
  - robustness of grammar and lexicon
- Military ‘field’ communication (CMU: DIPLOMAT)
  - Croatian, Spanish, Haitian Creole, Korean
- Military, government, tourism (Phraselator)
  - Albanian, Arabic, Bengali, Cambodian, Chinese, Farsi, French, German, Haitian Creole, Hindi, Indonesian, Japanese, Korean, Pashtu, Polish, Portuguese, Russian, Serbo-Croatian, Singhalese, Spanish, Swahili, Tagalog, Thai, Turkish, Urdu

Controlled language and special-purpose systems: requirements and issues

- system developed by external agency (e.g. Smart, LANT) or in-house?
- special dictionaries (domain, company): existing, or to develop?
- terminology databases
- new or adapted from existing controlled languages
  - despite previous models, SAP developed own language (SKATE)
- grammar and style analysis (usual grammar checkers inadequate)
- lexicon
  - internal (company) and external (standard terminology)
- grammar
  - recommendations or obligations

Lexical acquisition

- dictionary building
  - hand-crafted (pre-1990) was expensive in time and effort
  - required information: morphological variants, grammatical categories, syntactic contexts, lexical co-occurrences, semantic conditions/constraints, translation options
  - generally more detailed than terminology information for human translation (and includes **all** words)
- major problem for all current (commercial and custom-built) systems
- providers: vendor vs. customer
  - basic dictionary, special dictionaries, user dictionary (customer-specific)
- corpus-based methods do not require detailed dictionaries (future prospect)

Lexical resources: requirements

- Resources for creating dictionaries
  - size (what is adequate? definition of domain)
  - use of lexical resources (printed dictionaries, Internet dictionaries)
  - extraction from electronic texts (monolingual/bilingual, internal, Internet, Web pages): word alignment
  - validating, checking
  - conversion into required formats for particular MT system
  - updating procedures
- access to resources:
  - EDR, ELRA/ELDA, LDC

Software (enterprises)

- Requirements: client-server (intranet) systems, customizable
- facilities: large basic dictionary, technical dictionaries, user dictionaries
- platforms: Windows NT, Unix, Sun Solaris; or browser (client) access to server
- languages:
- English, French, German, Italian, Portuguese, Spanish
  - Amikai, [Comprendium], LogoMedia Enterprise Solutions, m^2T (globalwords), PeTra Enterprise, Reverso Intranet, SDL Enterprise Translator, Systran Enterprise, WebSphere Translation Server (IBM)
- English, Japanese, Korean, Chinese
  - Amikai, ATLAS (Fujitsu), EWTranslate, Systran Enterprise, TranSphere (AppTek), WebSphere (IBM)
- other languages
  - TranSmart [Finnish], TranSphere [Arabic]

Localization

- Internationalisation, globalisation (e.g. software and Web pages)
  - estimated market (end 2006) is $3.5 billion and $3 billion resp. (ABI, 2001)
- Cultural and linguistic adaptation (not just translation)
  - currency, measurements, power supplies
- Screen commands and help files; users' guides; warranties; publicity, marketing; packaging; workshop manuals
- Large scale, multiple language output, fast results (days, not weeks)
- Repetitive (translation memory)
- Graphics, formatting, layout, etc. (to be preserved)
- **companies use both translation tools (workstations, translation memories) and MT systems**
- has its own associations: Localization Industry Standards Association; GALA
- Software companies (many in Ireland):
  - ALPNET; Berlitz; Compaq; Corel; Eastman-Kodak; IBM; Lotus; Microsoft; Oracle; SAP; Symantec

Management implications

- Terminology database: acquisition, consistency, management
- Translation memory: inclusion/exclusion policy, quality, access
- Text alignment: quality control
- Documentation flow (from author to publication): project management
- Technical authoring: interaction with translation systems
- Publishing, formatting: graphics, layout
- Personnel training: project manager, translators, reviewers
- Technical assistance: language engineer, computer technician (software development)
- Recruitment, supervision, etc. of translators and post-editors
- Administrative support (incl. legal aspects)
- Customer contact (quotes, orders, servicing, technical support)
- Management control systems
  - e.g. LTC Organiser, PASSOLO

Convergence of HAMT and MAHT

- increasingly, systems straddle different categories
- workstations (TM systems) include MT components (e.g. Trados, Atril)
- MT systems include TM components (e.g. globalwords)
- localization systems embracing, or as components of, either TM or MT systems
- common facilities:
  - terminology management; integration with authoring and publishing systems; project management; quality control; Internet access and downloading; Lexical acquisition; Web translation
- common aim: production of quality translations for **dissemination;** utilization of translator skills
- at present: both approaches in parallel rather than integrated
- in research: EBMT investigates merging of rule-based and database methods
- future: full integration (no distinctions)

MT for translators (office systems): requirements

- integration with other IT equipment
- cost-saving
- easy post-editing
- translation workstations still too expensive for individual translators
- functions of systems for large organizations but for stand-alone (PC) systems
  - i.e. include terminology management and use of translation database
- vendors either downsize client-server systems or upgrade cheaper PC systems
- other users?:
  - companies not able to afford (or without facilities for) client-server systems
  - smaller translation agencies
  - occasional translators (perhaps)

Software (Professional translation)

- Systems, designed specifically (for translators to produce 'publishable quality' translation):
  - CITAC Translator: Chinese→English
  - ENGSPAN (PAHO): English→Spanish
  - ESI Professional (WordMagic): English↔Spanish
  - HICATS (Hitachi): English↔Japanese
  - Honyaku Office (Toshiba): English↔Japanese
  - Hypertrans (D'Agostini): English↔French, English↔German, English↔Italian, English↔Spanish, French↔German, French↔Italian, French↔Spanish, German↔Italian, German↔Spanish, Italian↔Russian, Italian↔Spanish, Portuguese↔Spanish -- [patents]
  - LogoVista X Pro (LEC): English↔Japanese

Software for professionals (contd.)

- Personal Translator PT Office Plus (Linguatec): English↔German
- PeTra Expert (Synthema): English↔Italian
- ProMT Translation Office (ProMT): English↔Russian, French↔Russian, German↔Russian, Italian↔Russian
- Reverso Expert (Softissimo): English↔French, English↔German, English↔Spanish, French↔German
- SPANAM (PAHO): Spanish→English
- Systran Professional Premium/Standard: Chinese→English, English↔French, English↔German, English↔Italian, English↔Japanese, English↔Korean, English↔Portuguese, English↔Spanish, Russian→English
- Transcend (SDL International): English↔French, English↔German, English→Italian, English↔Portuguese, English↔Spanish
- TranSphere (AppTek): English→Arabic, English→Chinese, English→French, English→Japanese, English→Korean, English→Persian, English→Turkish

MT for translators (office systems): issues

- translation database -- ownership, copyright
- terminology management -- acquisition
- integration with other IT equipment
- translation workstations still too expensive for individual translators
- insufficient functionality in downsizing systems for large organizations onto stand-alone (PC) systems
- suitable project management tools (currently most for large agencies and companies)

Has MT improved?

- In what respect?
  - translation quality: general-purpose vs. domain-specific
  - usability (ease of use)
  - adaptability (integration with other software)
- Since when?
  - quality perhaps not in last ten years, but since 1980 it has
- Why not?
  - inherent problems of language
  - inherent problems of ‘cultural’ differences

USAF-IBM system

- Begin one should from that that in United States appeared new translation immortal novel L.N.Tolstogo “War and World/peace”. Truth, not all novel. But only several fragments of it, even so few/little, that they occupy all one typewritten page. But nonetheless this achievement. Nevertheless culture not stands/costs on place…
  - Mark I, August 1960 [a report in *Izvestiya* on the USAF-IBM system itself]

- Biological experiments, conducted on different space aircraft/vehicles, astrophysical space research and flights of Soviet and American astronauts with/from sufficient convincingness showed that short-term orbital flights lower than radiation belts of earth in the absence of heightened solar activity in radiation ratio are safe.
  - Mark II, 1966

Systran at EC example (English to French)

- [English original]
  - Since no request concerning changed circumstances with regard to injury to the Community industry was submitted, the review was limited to the question of dumping.
- [French 1987]
  - Puisqu'aucune demande concernant les circonstances changées en ce qui concerne la blessure à l'industrie communautaire n'a été soumise, l'étude était limitée à la question de déverser.
- [French 1997]
  - Comme aucune demande concernant un changement de circonstances en ce qui concerne le préjudice causé à l'industrie communautaire n'a été présentée, le réexamen était limité à l'aspect du dumping.

Systran at EC example (French to English)

- [French original]
  - Leur objet n'était pas de formuler des recommandations politiques, mais de servir de base analytique à la réflexion politique.
- [English 1987]
  - Their object was not to formulate of the political recommendations, but to be used as a basis analytical for the political reflexion.
- [English 1997]
  - Their object was not to make political recommendations, but to serve as an analytical base to political reflection.

Systran at EC example (English to Spanish)

- [English original]
  - No formal list of supporting arguments was compiled but a number of points were common to the papers and discussions, including the following: ...
- [Spanish 1987]
  - Ninguna lista formal de mantener las discusiónes fue compilada pero varios puntos eran comunes a los papeles y a las discusiónes, con inclusión del siguiente: ...
- [Spanish 1997]
  - No se compiló ninguna lista formal de argumentos favorables sino que varios puntos eran comunes a los documentos y a las discusiónes, incluida la siguiente: ...

Evaluation of systems

- Who needs to know?
  - potential purchasers, potential users (translators), service managers, system developers, researchers
- Quality control
  - fidelity, accuracy (of terminology), comprehensibility, intelligibility, readability, appropriate style
- Usability
  - adaptability (e.g. to new domains), extendibility (e.g. to other languages and operating systems), compatibility (software and hardware), error levels (e.g. post-editing effort)
- Task suitability
  - dissemination/assimilation: publishing, gisting, extraction, triage, detection, filtering
- Resources evaluation
  - suitability and quality of dictionaries, terminology resources, translation memories (databases)
- Methods
  - Black box vs. glass box; test suites (set of 'standard' texts); interviews

General considerations for evaluation

- MT is not *translation* as usually understood, it is merely a computer-based tool
  - for translators
  - for cross-language communication
  - for access to information resources
- Perfectionism is not necessary or essential
  - publishable quality will always require human editing/revision
  - assimilation/interchange can always tolerate imperfect communication
- MT should be used only as required to save costs/effort in appropriate circumstances
- Judgement should be based
  - ***not*** on whether system produces 'real' translations
  - and particularly not whether it produces 'good' translations
  - ***but***: whether the output can be *used*
  - and: whether its use will save time or money

MT for assimilation

- publication quality not necessary
- fast/immediate
- readable (intelligible), for information use
  - intelligence services (e.g. NAIC)
  - occasional translation (home use)
- as draft for translation
- aid for writing in foreign language
  - as used by EC administrators
- emails, Web pages
- systems can be any of those primarily designed for dissemination:
  - e.g. as Systran (at EC) and earlier systems
  - e.g. any PC system

Software (Personal translation)

- Dictionaries (both as CD-Roms and downloadable from Internet)
- PC systems, e.g.
  - Al-Wafi (ATA Software): Arabic↔English
  - CITAC Fastran: Chinese→English
  - Crossroad (NEC): English↔Japanese
  - Easy Translator (Transparent Language): English↔French, English↔German, English→Italian, English→Portuguese, English↔Spanish, Japanese→English
  - ESI Standard (WordMagic): English↔Spanish
  - Instant Spanish (Bilingual Software): English→Spanish
  - Korya Eiwa (LogoVista): English↔Japanese
  - LogoMedia Translate (LogoMedia): Chinese↔English, English↔French, English↔German, English↔Italian, English↔Japanese, English↔Korean, English↔Portuguese, English↔Russian, English↔Spanish
  - LogoVista Personal (LEC): English↔Japanese

Software for personal translation (contd.)

- NeuroTran (Translation Experts): Bosnian↔English, Croatian↔English, English↔French, English↔German, English↔Hungarian, English↔Polish, English↔Serbian, English↔Spanish
- PC Translator 2002: Czech↔English, Czech↔German, English↔Slovak, German↔Slovak
- Personal Translator PT Home (Linguatec): English↔German
- PeTra Word (Synthema): English↔Italian
- Pocket Transer (Nova): English↔Japanese
- PROMT Express (ProMT): English↔Russian
- Reverso Perso (Softissimo): English↔French, English↔Spanish
- Systran Personal (Systran): English↔French, English↔German, English↔Greek, English↔Italian, English↔Portuguese, English↔Spanish

MT and the Internet (personal translation of webpages and emails)

- CITAC: Chinese→English
- LogoMedia Passport (LogoMedia): Chinese↔English, English↔French, English↔German, English↔Italian, English↔Japanese, English↔Korean, English↔Portuguese, English↔Russian, English↔Spanish
- LogoVista Internet Plus: (LEC): English to Japanese
- Reverso Perso (Softissimo): English↔French, English↔Spanish
- Systranet (Systran): English↔French, English↔German, English↔Italian, English↔Portuguese, English↔Spanish
- Translingo (Fujitsu): English↔Japanese
- Transpad (AILogic): English↔Japanese
- WebTransSmart: Finnish↔English
- T-Mail [email only]: Chinese↔English, English↔French, English↔German, English↔Italian, English↔Japanese, English↔Korean, English↔Portuguese, English↔Spanish, French↔German, Russian→English

Free MT services

- [first systems: Minitel (1980s), Compuserve (from 1994), Babelfish on AltaVista]
- for English, French, German, Italian, Portuguese, Spanish: Babelfish, Free Translation, Gist-in-Time, InterTran, iTranslator Online, Lycos [=Systran], T1-testdrive, PT-Online; Sancho [Spanish], Systranet, T-Mail, T-Sail, Worldlingo
- for English, Russian, Polish, Ukrainian: PARS; PROMT-Online; Poltran; Rustran
- for English, Chinese, Japanese, Korean: Arcnet, Babelfish, T-Mail, T-Sail, Worldlingo
- for other languages: Ajeeb [Arabic], Amaro's Lab [Papiamentu], Arcnet, Parsit [Thai], Postchi [Persian], Tarjim [Arabic]
  - for email, chat: Gist-in-Time, IMTranslator, Word2word Chat, Yakushite
- MT portals: Foreignword, Translatum, Word2word

- **why free**? Vendors want to encourage sales of systems with more facilities, larger dictionaries, customised, etc. --- but others?
- **how long will they be free**? When specialised (domain) systems come online; the improvements from subject orientation will be paid for by users.

Charged online translation

- English, French, German, Italian, Portuguese, Spanish:
  - Automatic PlusTranslation (SDL), Bestiland, Compuserve, Hypertrans, LogoMedia
- English, Chinese, Japanese, Korean:
  - Bestiland, EWTransLite, JICST, LogoMedia
- Other languages:
  - CyberTrans [African languages], WebTranSmart [Finnish]
- Enhanced services (i.e. with human post-editing):
  - PlusTranslation (SDL), TranslationWave, XLT (Socatra)

MT and hand-held devices (Personal translation)

- Special devices
  - Partner (Ectaco): English↔French, English↔German, English↔Italian, English↔Portuguese, English↔Spanish
  - Gold Partner (Ectaco): English↔Russian and English↔Ukrainian
  - Universal Translator (Ectaco): English→French, English→German, English→Spanish
  - dictionaries only: Language Teacher (Ectaco) and Quicktionary (Seiko), and others...
- Text messages (mobile/cellnet phones)
  - MobileTran
  - Petra-SMS
  - PT-SMS

Online and PC translation: why so bad?

- old models (word for word, simple transformer architecture)
  - often single equivalents, no morphological analysis or target adjustment
- dictionaries too small, insufficient information, and difficult (or impossible) to update
- weak syntactic analysis/transfer
- poor disambiguation (little semantic information)
- general-purpose (not domain restricted)
- not designed for language/style of emails
- web page translations: graphics not translated, distorted, ignored; format lost
- need special functions, if used as aid for writing in foreign language
- language coverage uneven; many languages of Africa and Asia are lacking
- translation from English often poorer than into English

- **conclusion: of use/value only if source language unknown or known only poorly,and if essence and not full information is adequate**
- **the less the user knows of the source language, the more useful becomes fully automatic translation**

PC translation: example

- [German original]
- **Sprachtechnologie ist Basistechnologie … Geistiges Kapital hinter verschlossenen Türen**
- Dokumente gehören zu jedem Geschäftsprozeß und zu jedem Produkt. Täglich werden 5 Milliarden Dokumente in Europa erzeugt. Man schätzt die Kosten dafür auf 20 Milliarden ECU pro Jahr, das sind rund 8% der Unternehmensumsätze. Unternehmen sind so effizient wie ihre Informations- und Dokumentenflüsse: Anfragen, Angebote, Bestellungen, Mahnschreiben, Rechnungen sind in Umlauf. Papierberge stapeln sich auf Schreibtischen. Die Eingangspost geht durch viele Hände, bevor sie beim Sachbearbeiter landet. Unproduktive Transport- und Liegezeiten fallen an und es entstehen Kosten, die vermeidbar wären!
-

PC translation: example

- [translation by Personal Translator PT (Linguatec)]
- **Language technology is base technology ...Intellectual capital of doors locked backly**
- Documents are part of every business process and of every product. 5 billion documents are produced in Europe daily. One values the costs for it at 20 billion ECU per annum, these are about 8% of the enterprise sales.
- Enterprises are as efficient as their information and document rivers: Enquiries, offers, orders, reminders, invoices are in circulation. Paper mountains stack themselves on desks. The incoming mail goes by many hands before it lands with the clerk. Costs which would be avoidable attack unproductive transportation and lay days and arise it!

Online translation: example

- [translation by InterTran]
- **Sprachtechnologie am Basistechnologie … intellectual capital behind cagier doors**
- documents belong to everybody Geschäftsprozeß and to everybody product . daily become 5 milliards documents in Europe engenders . one cherishes the cost for it on 20 milliards ECU pro year, the are round 8% the Unternehmensumsätze . undertaking are so effizient how her information - and Dokumentenflüsse: inquiries, offers, orders, Mahnschreiben, calculations are in circulation . Papierberge batches himself on desks . the Eingangspost ambulates by a lot of hands, before she by specialist alights . unproductive transport - and Liegezeiten traps at and it arise cost, the avoidable wären!

Online translation: example

- [translation by Babelfish]
- **Language technology is fundamental technology... Mental capital behind locked doors**
- Documents belong to each business process and to each product. Daily 5 billion documents in Europe are produced. One estimates the costs of it on 20 billion ECU per year, that is approximately 8% of the enterprise conversions. Enterprises are as efficient as their information and document rivers: Inquiries, supplies, orders, printing reminder, calculations are in circulation. Paper mountains stack themselves on desks. The input post office goes through many hands, before it lands with the operator. Unproductive feed and downtimes result and it develop for costs, which would be avoidable!

MT in the marketplace

- retail availability
  - many only purchased direct from manufacturer
- promotion by vendors
  - confusion of terms:
    - ‘translation systems’ no more than dictionaries
    - ‘computer aided translation’ either HAMT or MAHT
    - combination of MT and support tools
    - translation memories either independent or components
- expectations of users
  - steady quality improvement
  - more languages
- suitability of system to expected use
- bench marks, consumer reports/reviews

Risks of marketplace

- Failures of previous products, e.g.:
  - ALPS Transactive, Weidner and Bravice
  - Intergraph and Transparent Language
  - Globalink (Microtac)
  - Lernout & Hauspie
  - Logos Corporation
  - Winger
  - Oki Electric (Pensee systems)
  - Sail Labs
- low profits, slow quality improvement, few differences between rivals
  - not helped by free online services
- is current system categorisation viable?
  - Enterprise systems, i.e. Client-server (intranet)
  - Workstations (TM systems)
  - Professional systems
  - Home systems

MT for interchange: what's needed?

- correspondence, emails, etc.
- in principle, any systems can be used for written interchange
  - many PC systems have specific facilities for email translation
- in future there may be special-purpose systems for business correspondence (e.g. with interactive authoring in controlled language)
  - has been subject of research (e.g. UMIST)
- interchange in military ('field') situations
  - e.g. systems for translating standard phrases (Diplomat, Phraselator)
- interchange in tourist situations
  - so far only dictionaries of words and phrases (hand-held devices)
- interchange with deaf and hearing impaired
  - translation into sign languages [mainly research so far]
- interchange by telephone or in business oral communication
  - still at research stage (speech translation)
- interpreting ex tempore (unlikely ever to be even semi-automated) , but:
  - interpreters (at EC etc.) do use rough MT of technical speeches to aid them

MT and other LT applications

- document drafting
  - Japanese researchers, EC administrators, school essays
- information retrieval (CLIR): translation of search terms
- information filtering (intelligence):
  - for human analysis of foreign language texts
  - document detection (texts of interest); triage (ranking in order of interest)
  - deciding whether text worth translating (discard irrelevant ones)
- information extraction: retrieving specific items of information (domain-tuned, captured by key words/phrases)
  - e.g. specific events, named people or organizations
- summarization: producing summaries of foreign language texts
- multilingual generation from (structured) databases
- localization of interactive commands (computers, mobile phones)
- television subtitling
- language teaching: MT as aid for teaching translation

MT and information analysis and extraction

- tasks for information analysis/filtering tasks
  - should be fully automated, with no pre- or post-editing
  - tuned for specific domains
  - should accept OCR input
  - should tolerate (and ideally correct) misspellings, missing diacritics, wrong transliteration, grammar mistakes, scanning errors
  - deal with mix of languages in same document
  - identify and retain all formatted information
  - provide facilities for easy updating of lexicon
  - specialist lexica for different domains
- additional tasks for information extraction
  - domain (scenario) templates for SL; presentation of completed template in TL
- additional tasks for 'translingual speech retrieval' (browsing radio broadcasts, information routing, automatic alerts)
  - generalised speech recognition
  - word detection; indexing of key terms

MT: when it works and when it doesn't

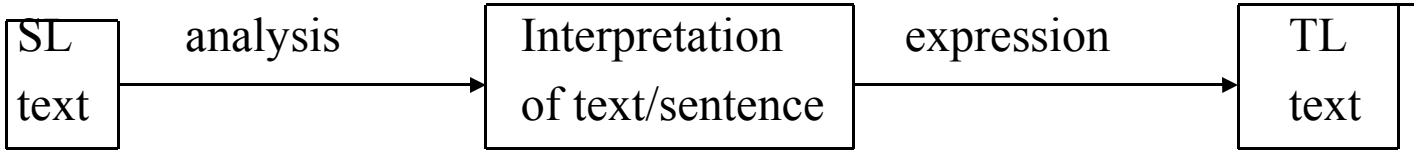
- cannot be both fully automatic (no pre- or post-editing) and general-purpose
- beyond its scope:
  - literature, philosophy, sociology, law
- large corporations, cost-effective if:
  - controlled input, standardised terminology, multilingual output, repetitive documentation, restricted domain
- occasional (information-only)
  - rough, not for publication; immediate (fast) production
- small-scale MT
  - 'formulaic' documents (business correspondence), restricted domain
  - interactive assistance

Why human (and machine) translation can fail

- Insufficient knowledge of (data covering) source language
- insufficient knowledge of (data covering) subject matter
- lack of knowledge of specialist vocabulary (access to specialist lexica)
- inadequate familiarity with cultural background (no background)
- inadequate knowledge of (data for) target language (in relevant domain)
- lack of translation experience (no ‘understanding’ or ‘learning’)

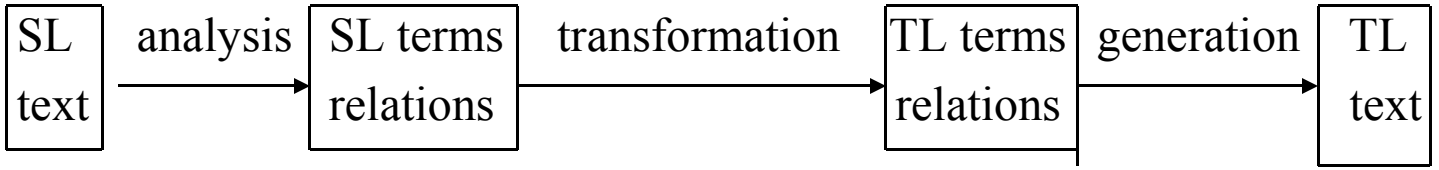
- may be analogies between processes of human translation and the various ‘models’ of machine translation
- but although the analogies may have didactic value, they remain hypotheses about ‘reality’. Language and translation are mysteries.

Analogies of HT and MT (1)

- Translation involves the understanding (interpretation) of a source text and its rendition in a target text
- Interpretation is a function of the meanings of parts of sentences (words, phrases) and the relationships between those parts (syntax)
- But words out of context have (often, usually) more than one meaning, and structures can have more than one interpretation
- Therefore, words and sentences have to be interpreted (analysed and disambiguated)
- 

- This is the ‘interlingua’ view or approach to MT

Analogies of HT and MT (2)

- Although complete interpretation (understanding) is desirable (ideal) for translation, sometimes difficult texts can be translated with minimal understanding
- If the translator can discover how particular technical terms are translated from the source language into the target language (capacitor → condensateur)
- If the translator knows how certain structures are rendered in the target
  - X likes to Y
- Translation in such cases involves the identification (analysis) of the relationships between lexical elements, the conversion of source words (compound words) into target words, and the generation of structurally equivalent sentences in the target language
- 

- This is the 'transfer' model of MT

Analogies of HT and MT (3)

- In some cases, syntactic transformation is not necessary, particularly between related languages
- in other cases, simple transformation of adjacent elements is sufficient
  - English Adj+N --> French N+Adj (with exceptions: grand, beau, vieux, etc.)
- these are only slightly more complex than simple word for word ‘translation’

- This is the ‘direct translation’ model of MT
- often used in the earliest days of MT research (to mid 1960s)
- still often found in current low-priced commercial systems

Machine translation and human translation 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?)
- HT for spoken language translation
- MT for integrating translation with other LT tasks

Human versus machine translation: Dissemination

•	HT	CAT(TM)	HAMT	MT
• Literary, legal	costly	no	no	no
• Technical, scientific	v.costly	yes	yes	no
• Weather reports	costly	?	yes	yes
• Localization	?	yes	yes	no
• Web localization	yes	yes	yes	no?
• Advertisements	?	yes	poss.	no
• Document drafting	no	yes	yes	no?

Human versus machine translation: Assimilation

–	HT	CAT	HAMT	MT
• Scientific, technical	rare	no	adeq	adeq
• Non-literary (occasional)	rare	no	poor	adeq
• Information monitoring		costly	no	adeq

Human versus machine translation: Interchange and information access

–	HT	CAT	HAMT	MT
• Business correspondence	yes	yes	yes	adeq?
• Personal correspondence	?	no	adeq	adeq
• Electronic mail	no	no	no	poor
• Web pages	yes!	no	no	adeq?
• Database searching	no	no	no	adeq
• Summarising (with translation)	rare	no	poss?	poss?
• TV captions	no?	no?	no?	adeq
• informal conversation	yes	no	no	no
• formal interpreting	yes	no	no	no
• telephone enquiries	rare	no	no	poss?

Some future directions and expectations

- merging of MT and TM for enterprise dissemination systems
- data-driven vs. (and) theory-driven -- hybrid systems
- Internet as resource
- rapid development of systems
  - particularly for assimilation/interchange
- improvements in quality
  - particularly PC commercial and online systems
- special-purpose systems (domain and function)
  - particularly on Internet (no longer free!)

- Reusability of resources (particularly dictionaries and translation memories)

Some future directions and expectations (contd.)

- Spoken language translation
- 'Minor' languages
  - languages of India, Africa, Asia
  - non-national ('official') languages (e.g. Welsh, Basque, Catalan)
  - languages of minorities (e.g. non-indigenous languages in Britain)
- Systems for monolinguals
  - from unknown source language
  - to unknown target language
- Further integration with other NLP systems
  - MT as option with summarization, information extraction, information retrieval, data retrieval, question-answering, Internet search tools
- bilingual (multilingual) communication as much as translation

MT as bilingual communication aid

- computer-produced draft translation (traditional post-edited MT)
- computer-based translation aids (dictionaries, terminology, translation memories, translator workstations)
- text assimilation aids (traditional use of ‘rough’ MT output)
- text production aids (multilingual generation, authoring aids)
- message dissemination aids (TV captions, public announcements, police messages)
- cross-language information access (information retrieval, information extraction, summarization)
- cross-language interchange (email, SMS, telephone, military ‘field’ communication, business negotiations, tourism, etc.)

Sources of information

- EAMT website (www.eamt.org) with links to other IAMT sites, etc.
- LISA website (www.lisa.org)
- Conferences: MT Summit, EAMT workshops, LISA Forums
- Journals:
  - *Language International*
  - *Machine Translation*
  - *Multilingual Computing and Technology*
  - *MT News International*
- Directory of current commercial systems: *Compendium of translation software* [on EAMT website]
- Books:
  - Hutchins, W.John and Somers, Harold L.: *An introduction to machine translation* (London: Academic Press, 1992)
  - Sprung, Robert C. (ed.): *Translating into success*. (Amsterdam: John Benjamins, 2000)
  - Esselink, Bert: *A practical guide to localization*. Rev.ed. (Amsterdam: John Benjamins, 2000)
- my website:
  - http://ourworld.compuserve.com/homepages/WJHutchins