Machine translation in the real world

John Hutchins

(Email: WJHutchins@compuserve.com)

[http://ourworld.compuserve.com/homepages/WJHutchins]

Universitat Internacional Menéndez Pelayo, Barcelona

18 July 2002

Outline

- Use of MT until late 1980s
- Systems and uses by large organizations
  - post-editing and controlled languages
- Translation workstations
  - terminology and translation memories
- systems for professional translators
- systems for occasional (non-professional) use
- the Web and online translation
- MT and other LT applications
- MT and human translation

Basic distinctions

- Wholly automatic systems
  - systems that (attempt to) translate texts and sentences as wholes
- Computer-based translation aids
  - systems that provide linguistic aids for translation:
    - dictionaries, grammars
    - previously translated texts

Types of translation demand

- dissemination
  - translation for publication, or translation of 'publishable' quality
  - usually by organisations and often involving professional translators
- assimilation
  - translation for internal 'monitoring', information 'filtering', occasional uses
- interchange
  - translation for communication between individuals, e.g. correspondence, email, telephone
- information access
  - translation for facilitating information retrieval/extraction, database searching, access to Web pages

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)

IBM-USAF Translator

- Earliest MT system put into operation
- Developed by IBM (director: Gilbert King) for the Foreign Technology Division of the US Air Force
- Mark I installed in June 1959
- used for translating Pravda, Izvestia and other Russian documents
- about 10,000 words daily
- found to be 'useful' as indication of potential items of interest (i.e. for assimilation)
- Mark II developed by 1964
- Demonstrated at New York World Fair, 1965

IBM-USAF examples

- Shakespeare Overspat/outdid…

- 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 [Izvestia, August 1960]

- Biological experiments, conducted on different space aircraft/vehicles, astrophysical space research and flights of Soviet and American astronauts with/from sufficient convincingness sowed 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 [ALPAC, 1966]

Georgetown University system

- Only other MT system of 'first generation' to become operative
- Public demonstration at IBM headquarters of joint GU-IBM model, January 1954
  - 'toy system' (250 words, 6 grammar rules, 49 sentences), but attracted wide interest
- Director: Leon Dostert; chief linguists: Paul Garvin, and Michael Zarechnak; chief programmers: Peter Toma, A.F.R.Brown, and John Moyne
- Funded by CIA
- Russian to English mainly (but also French)
- Demonstrated at Pentagon, January 1960
- Installed at Euratom (Ispra, Italy), 1963 [ran until 1976]
- Installed at US Atomic Energy Authority, Oakridge National Laboratory, 1964 [ran until 1980]
- Georgetown project ended 1963

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
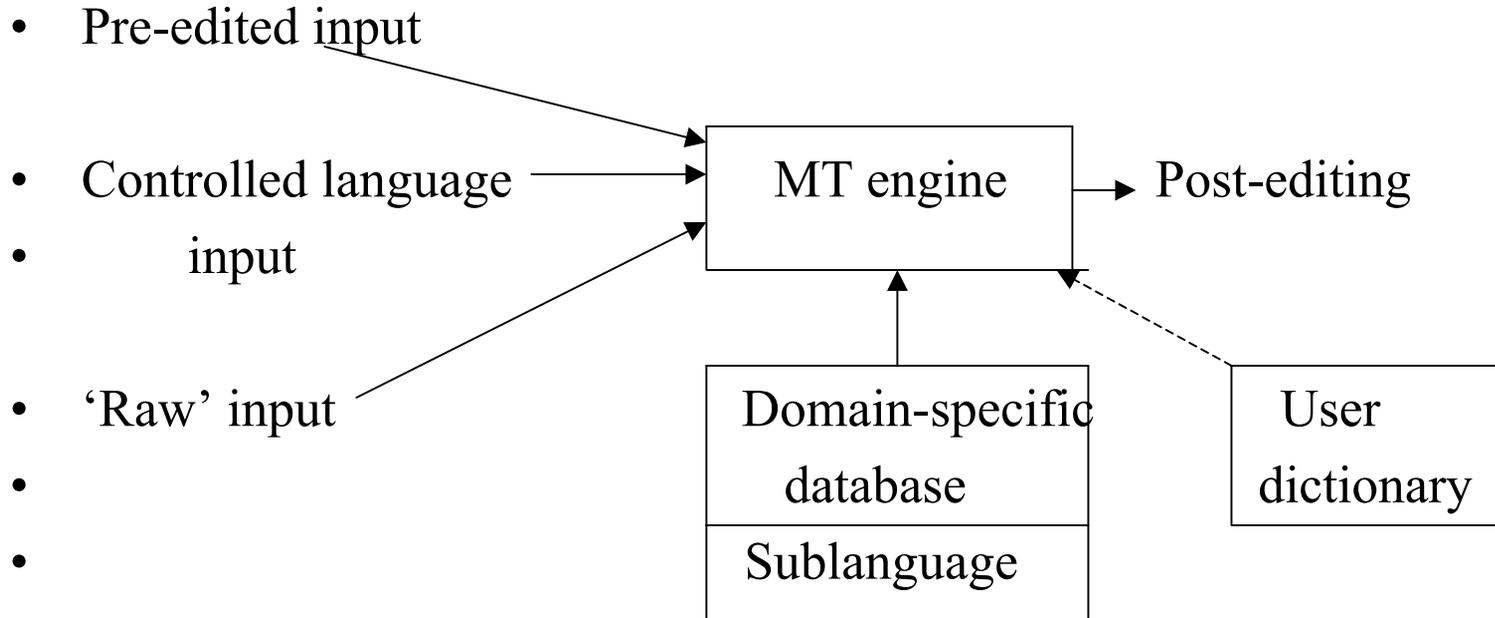
Consequences of ALPAC

- 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)
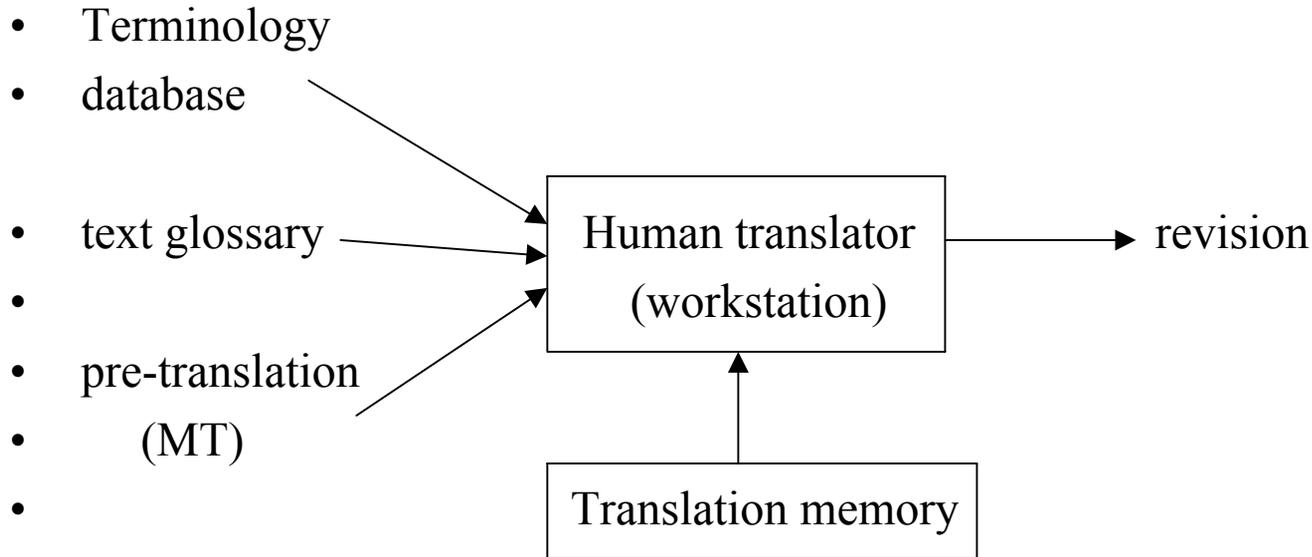
Meeting the translation demand

- dissemination, since raw output inadequate:
  - post-editing
  - control of input (pre-editing, controlled language)
  - domain restriction (reducing ambiguities)
- assimilation, use of raw output
  - with or without light editing
- interchange
  - if important: with post-editing
  - otherwise: without editing
- information access
  - limited use before 1990

Human-assisted MT

- Pre-edited input

- Controlled language
- input

- 'Raw' input

-

-

| MT engine | → Post-editing |

| Domain-specific database |
| Sublanguage |

| User dictionary |

Machine-aided human translation

- Terminology
- database

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

Human translator
(workstation)

→ revision

Translation memory

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

Systran example
(Russian to English)

- A helicopter, a flight vehicle heavier than air with vertical by takeoff and landing, lift in which is created one or by several (more frequent than two) rotors… A helicopter takes off upward vertically without a takeoff and it accomplishes vertical fitting without a path, motionlessly "will hang" above one place, allowing rotation around a vertical axis to any side, flight in any direction at speeds is produced from zero to the maximum...

Meteo 1976

- Developed at University of Montreal (TAUM project)
- weather reports (Canadian Meteorological Center)
    - repetitive, boring, difficult to retain staff
- sublanguage system
    - corpus of actual texts
- public broadcasting since May 1977
- revised version introduced 1984, and French-English system in 1989

Météo example

- LOWER ST JOHN VALLEY UPPER ST JOHN RIVER WIND WARNING ENDED BOTH REGIONS. SNOW AND BLOWING SNOW TONIGHT BECOMING INTERMITTENT NEAR DAWN FRIDAY. CLOUDY WITH PERIODS OF LIGHT SNOW FRIDAY. STRONG GUSTY NORTHEASTERLY WINDS TONIGHT BECOMING NORTHWESTERLY WINDS FRIDAY AFTERNOON.

- VALLEE DU BAS ST JEAN HAUT ST JEAN FIN DE L'AVIS DE VENT POUR LES DEUX REGIONS. CETTE NUIT NEIGE ET POUDRERIE DEVENANT PASSAGERES VENDREDI A L'AUBE. VENDREDI NUAGEUX AVEC FAIBLES CHUTES DE NEIGE PASSAGERES. CETTE NUIT VENTS FORTS DU NORD EST SOUFFLANT EN RAFFALES DEVENANT VENTS FORTS DU NORD OUEST VENDREDI APRES-MIDI.

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**

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)
- Melby's three levels
  - 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

First commercial PC (workstation) systems: 1980s

- ALPS
  - included 'reptitions processing' (to build translation database) and 'repstraction' (to find matching phrases
- Weidner
- Microtac
- Globalink
- Japanese systems:
  - Fujistu
  - Hitachi
  - NEC
  - Oki
  - Sharp

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]

MT for enterprises: requirements

- accurate, good quality, publishable (dissemination)
- large volumes
- saving costs (and staff?)
- restricted domain
- integration with other IT equipment
- types of document
  - external publications: publicity, marketing
  - internal reports
  - operational manuals
  - reference manuals
  - software localization

Large-scale translation and MT

- technical documentation (e.g. operating manuals)
- repetitive, frequent updates
- multilingual output (e.g. English to French, German, Japanese, Portuguese, Spanish)
- available in-house terminological database
- backup resources (translated texts, personnel for dictionaries, etc.)
- human assistance for quality (controlled language input, post-editing)
- integrate with technical writing
- 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

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
(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 discusiones, 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 discusiones, incluida la siguiente: ...

Systran at EC example
(English to German)

- [English original]
  - The objective of all three groups is to ensure the satisfactory implementation of the policy, strategy and measures of the fifth programme

- [German 1987]
  - Das Ziel der aller drei Gruppen soll die befriedigende Durchführung der Politik, Strategie und der Massnahmen des fünft Programms gewährleisten.

- [German 1997]
  - Das Ziel aller drei Gruppen besteht darin, die zufriedenstellende Durchführung der Politik, der Strategie und der Massnahmen des fünften Programms zu gewährleisten.

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
    - style too much SL oriented
    - translators as 'slaves' to machine
- 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.)

- 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: 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

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]

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)
  - but current corpus-based research seeking methods using minimal information
- providers: vendor vs. customer
  - basic dictionary, special dictionaries, user dictionary (customer-specific)

Lexical resources

- resources
  - 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)
  - validating, checking
  - conversion into required formats for particular MT system
  - updating procedures
- access to resources:
  - EDR, ELRA/ELDA, LDC

Computer-aided translation (MAHT): growth factors

- recognition that fully automatic translation not appropriate for professional translators
- terminology, standards
- translation databases: translation memories
- PCs and word processing
- Desk top publishing
- Translator 'in control'

Translation tools

- dictionaries (monolingual, bilingual): on-line access
- grammar aids, spelling checkers
- user glossary, terminology management, 'authorised' terms, specialist glossaries
- multilingual word processing
- 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

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 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

Translation database example

- 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: 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
- example-based MT research

Translation workstations
(often called Translation memory systems)

- 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

– [IBM TranslationManager/2 - no longer marketed]

– Déjà Vu (Atril)

– SDLX (SDL Corporation)

– Multilizer (Multilizer Inc.)

– Terminotix: LogiTerm

– Champollion: WordFast

– 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

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)
- Organisation:
  - Localization Industry Standards Association
- Software companies (many in Ireland):
  - ALPNET; Berlitz; Compaq; Corel; Eastman-Kodak; IBM; Lotus; Microsoft; Oracle; SAP; Symantec

Localization systems and support tools

- For project management, document control (formatting, etc.), personnel, and integrating workstations, translation memories, and terminology management
- support tools:
  - CATALYST (Alchemy), Convey Localization Suite, ForeignDesk (Lionbridge), GlobalSight, InstallShield, JCAT, LocalSphere, Lotus, PASSOLO, PowerGlot, RC-Wintrans, SDL Localization Suite, Uniscape GXT, WizTom
- management and quality assurance:
  - HelpQA, HtmlQA, LTC Organiser, SDLinsight, ToolProof, WebBudget
- web localization tools:
  - ArabSite, IBM WebSphere, InterTran Website Translation Server, SDL Webflow, SystranLinks, Worldlingo

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)

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

MT for translators (office systems): requirements

- translation database
- terminology management
- 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
- 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.)

- Pensee (Oki): English↔Japanese
- 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 (Systran): Chinese→English, English↔French (S), English↔German (S), English↔Italian (S), English↔Japanese, English↔Korean, English↔Portuguese (S), **English↔Spanish** (S), 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 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**

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!

- [Last sentence might be translated: Unproductive transport times and demurrage periods occur and costs arise that ought to be avoidable.]

MT and the Internet
(personal Web page translation)

- 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

MT and the Internet
(personal translation of emails)

- CITAC: Chinese→English

- LogoMedia: Chinese↔English, English↔French, English↔German, English↔Italian, English↔Japanese, English↔Korean, English↔Portuguese, English↔Russian, **English↔Spanish**

- Reverso Perso: English↔French, **English↔Spanish**

- T-Mail: Chinese↔English, English↔French, English↔German, English↔Italian, English↔Japanese, English↔Korean, English↔Portuguese, **English↔Spanish**, French↔German, Russian→English

- Translingo: English↔Japanese

Free MT services

- [first systems: Minitel (1980s), Compuserve (from 1994), Babelfish on AltaVista]

- 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

- English, Russian, Polish, Ukrainian: PARS; PROMT-Online; Poltran; Rustran

- English, Chinese, Japanese, Korean: Arcnet, Babelfish, T-Mail, T-Sail, Worldlingo

- 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

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!

Charged online translation

- English, French, German, Italian, Portuguese, Spanish:
  - Automatic PlusTranslation (SDL)
  - Bestiland
  - Compuserve
  - Hypertrans
  - LogoMedia
- English, Chinese, Japanese, Korean:
  - Bestiland
  - EWTransLite
  - FLM
  - JICST
  - LogoMedia

Charged online services (contd.)

- Other languages:
  - CyberTrans [African languages]
  - WebTranSmart [Finnish]
- Enhanced services (i.e. with human post-editing):
  - PlusTranslation (SDL)
  - TranslationWave
  - XLT (Socatra) [English↔French]

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

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

- 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

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
- Online translation with voice output
  - Translation Wave
- Speech translation

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!

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
- language teaching: aid for teaching translation
- translator training

MT embedded in other tasks

- tasks for information analysis/filtering tasks
  - should be fully automated, so 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

Types of translation demand (review)

- dissemination
  - external publications
  - internal reports
  - operational manuals
  - localization
- assimilation
  - internal 'monitoring', information 'filtering'
- interchange
  - correspondence, email
  - telephone
- information access
  - information retrieval/extraction, database searching, Web pages

MT: when it works and when it doesn't

- Beyond the scope
  - fully-automatic general-purpose
  - 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')

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	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?

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.)

New directions and challenges

- Spoken language translation
  - general-purpose?
- '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
- Improvement expectations
  - particularly PC commercial and Internet systems
- Reusability of resources (particularly dictionaries and translation memories)
- Integration
  - MT as option in word processing packages, on Web pages
  - MT as option with summarization, information extraction, information retrieval, data retrieval, question-answering, Internet search tools

The 'quality' of MT

- An early expectation (confirmed):
  - "the resulting literary style would be atrocious and fuller of 'howlers' and false values than the worst that any human translator produces."
  - "[computers] might be made to turn out a rough draft which a competent editor versed in the subject matter, though unacquainted with the foreign language, could then pull into shape."
  - (J.E.Holmström, 1951)
- Early evaluations:
  - FTD system (IBM-USAF Translator)
  - Georgetown system
  - ALPAC
  - Logos English-Vietnamese system
  - Systran (by EC)

Evaluation

- 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

Conclusions

- 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

The future

- merging of MT and TM for enterprise dissemination systems
- data-driven vs. theory-driven
- Internet as resource
- rapid development of systems
  - particularly for assimilation/interchange
- improvements in quality
- minor (and minority) languages
  - i.e. not of major commercial or military interest
- special-purpose systems (domain and function)
- bilingual (multilingual) communication as much as translation

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*
  - *Multilingual Computing and Technology*
  - *MT News International*

- *Compendium of translation software*

- Books:
  - 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