Machine translation and computer-based translation aids:

systems and usage

John Hutchins

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

Libera Università degli Studi "S. Pio V", Rome
December 2002

Contents

- Processes of translation
- Types of linguistics rule-based systems
- Corpus-based systems
- Computer-based translation tools
- Translation workstations, translation memories
- Systems and uses by large organizations
  - post-editing and controlled languages
  - localization
  - lexical resources
- Systems for professional translators
- Systems for occasional (non-professional) use
- Web and online translation
- MT and other LT applications
- Evaluation and the future

Why use computers in translation?

- Too much translation for humans
- Technical materials too boring for humans
- Greater consistency required
- Need results more quickly
- Not everything needs to be top quality
- Reduce costs

- any one of these may justify machine translation or computer aids

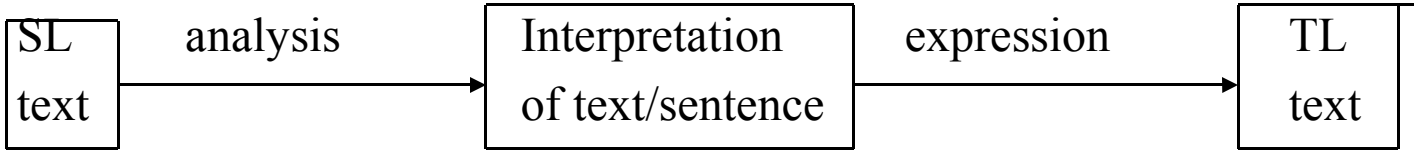
Basic distinctions

- 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

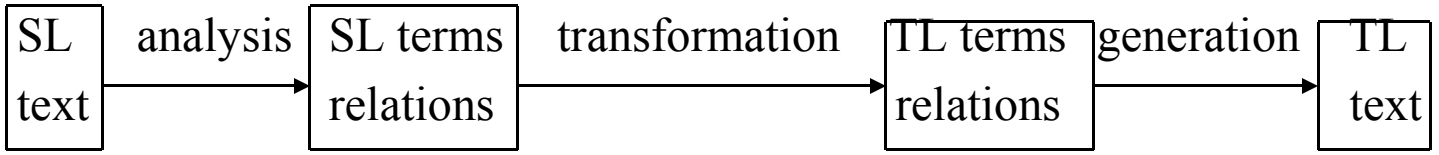
System architectures and strategies

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

Processes of translation (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)
- 

Processes of translation (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
- 

Processes of translation (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’
- often found in current low-priced commercial systems

Monolingual ambiguity

- morphological ambiguity:
  - German **-en**: noun plural, dative plural, weak noun non-nominative, adjective masculine non-nominative, etc.
- compound nouns:
  - coincide -> coin+cide, cooperate -> cooper+ate
- category ambiguity:
  - *round*: the first round (noun), to round up cattle (verb), the round table (adjective), go on a voyage round the Mediterranean (preposition), it measure three feet round (adverb), etc.
- homographs and polysemes:
  - *branch*: 'of a tree', 'of a bank'; *crane* (a bird or lifting machine)
  - *ball*: The ball rolled down the hill, The ball lasted until midnight

Bilingual lexical ambiguity

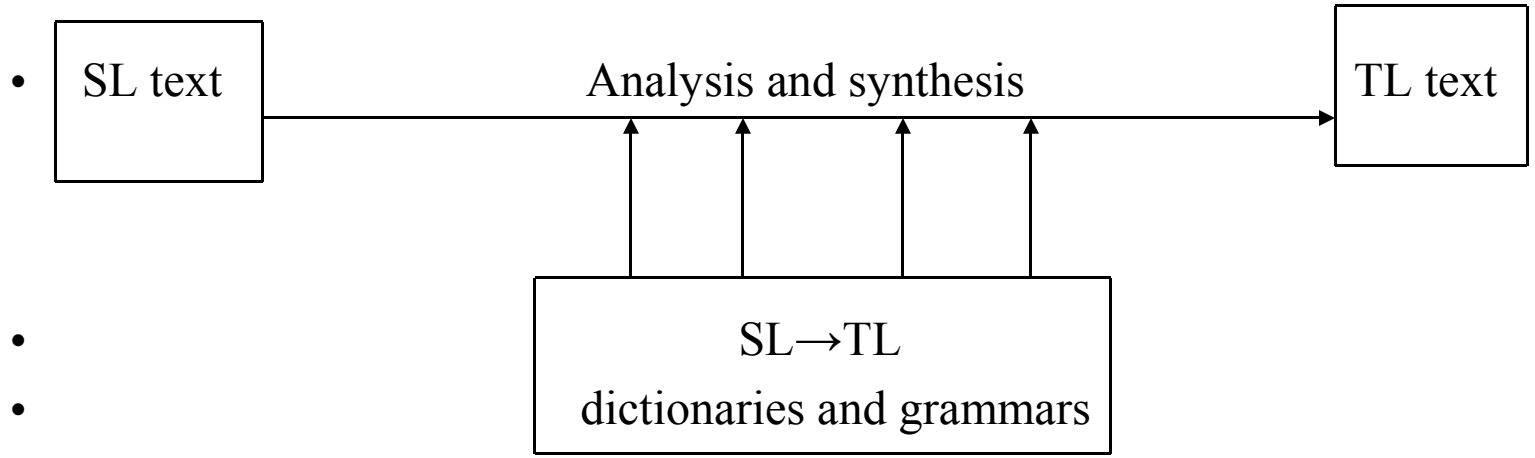
- English *wall*: German *Mauer* (outside) or *Wand* (inside)
- English *river*: French *fleuve* (major) or *rivière* (general term)
- English *leg*: French *jambe* (human), *patte* (animal, insect), *pied* (table), *étape* (journey)
- English *blue*: Russian *goluboi* (pale blue) or *sinii* (dark blue)
- French *louer*: English *hire* or *rent*
- German *leihen*: English *borrow* or *lend*
- English *wear*: Japanese *haoru* (coat/jacket), *haku* (shoes/trousers), *kaburu* (hat), *hameru* (ring/gloves), *shimeru* (belt/tie/scarf), *tsukeru* (brooch/clip), *kakeru* (glasses/necklace)

- resolvable by:
  - rules (indicating allowable or usual categories or types of subjects, objects, verbs, etc.)
  - collocations (specifying particular adjacent words)
  - frequencies (most probable adjacent or dependent words)

Structural ambiguity

- Flying planes can be dangerous
- The man saw the girl with a telescope
- John mentioned the book I sent to Mary
- I told everyone concerned about the strike
  - everyone concerned/involved/relevant, or: everyone disturbed/worried
- He noticed her shaking hands
  - either which were shaking from cold, or which were shaking other hands
- They complained to the guide that they could not hear
  - *that* as relative pronoun (‘whom they could not hear’) or as complementizer (‘that they could not hear him’)
- The mathematics students sat their examinations
- The mathematics students study today is very complex
  - difficulty of identifying noun compound vs. relative clause
- Gas pump prices rose last time oil stocks fell
  - each word potentially noun or verb

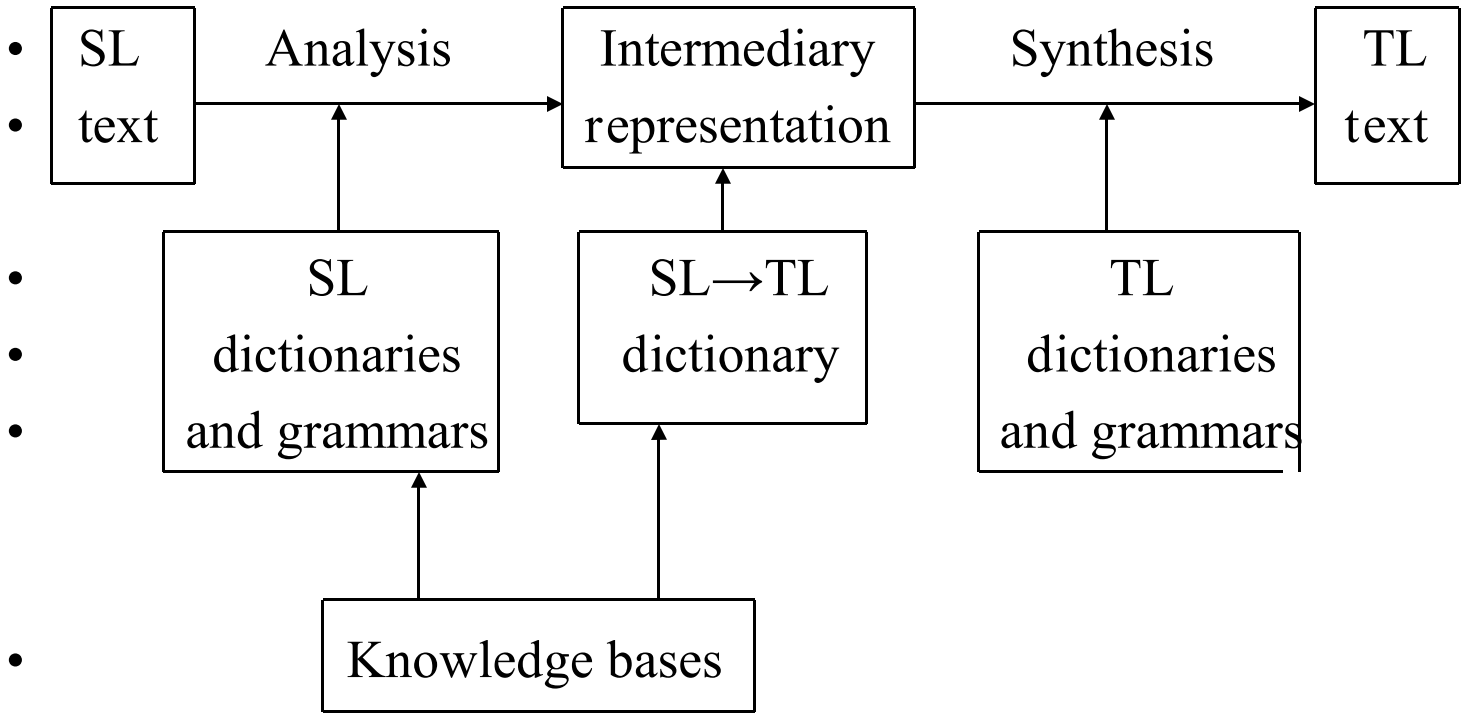
Direct translation



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
- Dictionary entries include TL rearrangement rules
- Use of ‘cover’ words
- 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”
- systems:
  - Univ.Washington, IBM (US)
  - Georgetown University (US)
  - Ramo-Wooldridge (US)
  - Institute for Precision Mechanics and Computer Technology (USSR)
  - National Physical Laboratory (UK)

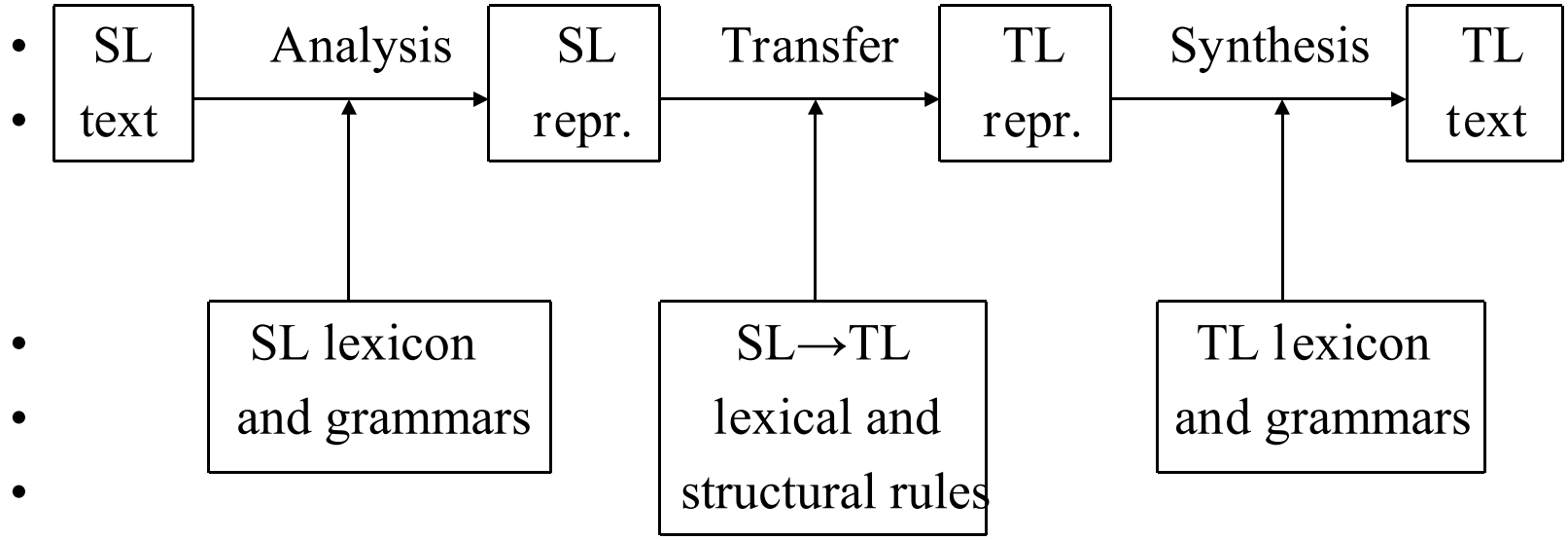
‘Interlingual’ system



Interlingua-based MT

- two independent stages: analysis, synthesis
- abstract language-neutral representation
- multistratal: morphology, syntax, semantics
- semantics-oriented (‘understanding’)
- domain-specific ‘knowledge bases’ (AI-oriented)
- projects:
  - Grenoble (CETA), Texas (METAL)
  - DLT, Rosetta, Pivot (NEC)
  - Carnegie-Mellon University (KBMT, KANT, CATALYST)
  - New Mexico State University (ULTRA, Pangloss)
  - Univ. Maryland (UNITRAN)

‘Transfer’ system



Transfer-based MT

- three stages: analysis, transfer, synthesis
- abstract semantico-syntactic interfaces/representations
- multiple level/strata: morphology, syntax, semantics
- syntax-oriented, tree-transduction
- batch processing, post-edited
- little/no discourse information (anaphora, etc.)
- projects/systems:
  - GETA-Ariane, Eurotra, LMT, Mu

Theories and formalisms

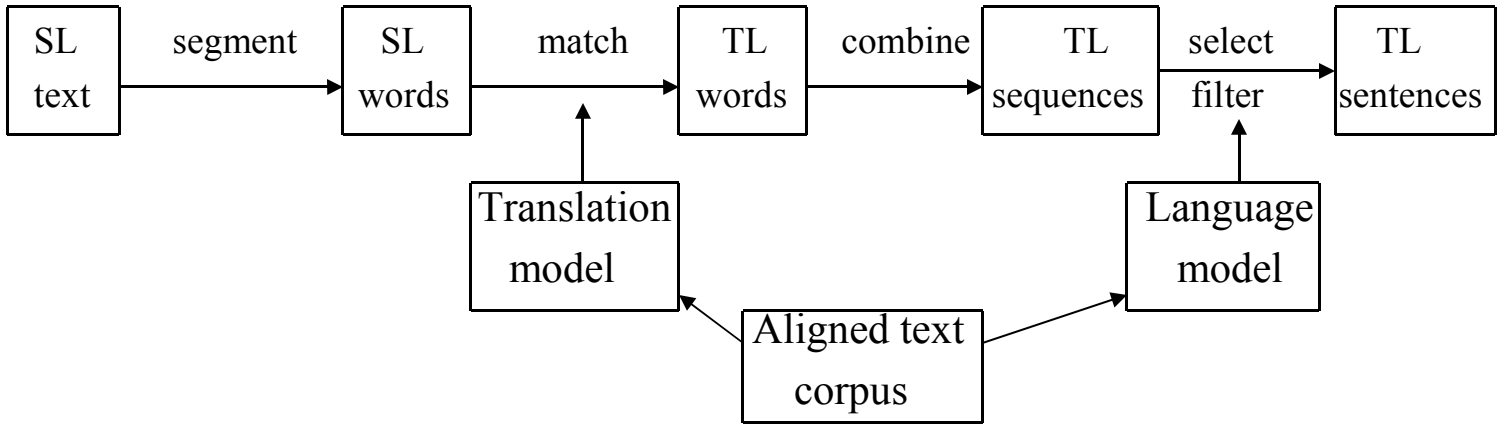
- Information theory (Shannon, Weaver, Yngve (MIT), Bar-Hillel, …)
- Transformational-generative grammar
- Dependency grammar
- Stratificational grammar (Lamb (UC Berkeley), Mel'chuk (MTM))
- Artificial intelligence (Wilks, Carnegie-Mellon)
- Lexical-functional grammar and Unification grammar
- Generalized phrase-structure grammar
- Definite clause grammar
- Principles and parameters, Government-binding theory (Univ.Maryland)
- Categorial grammar
- Montague grammar (Rosetta)
- Neural networks

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

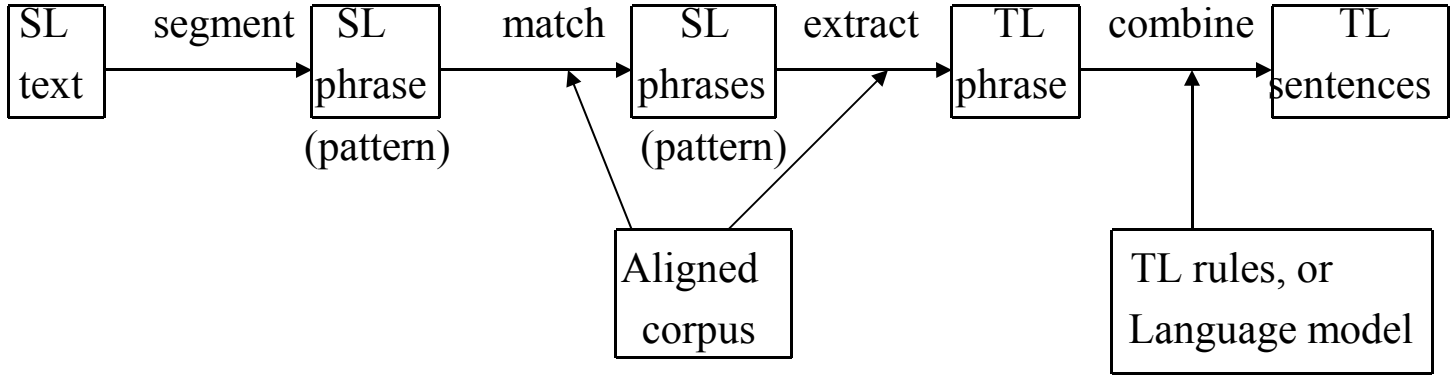
- Bilingual corpora: original and translation
- little or no linguistic 'knowledge', based on word co-occurrences in SL and TL texts (of a corpus), relative positions of words within sentences, length of sentences
- Sentences aligned statistically (according to sentence length and position)
- compute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

Example-based MT

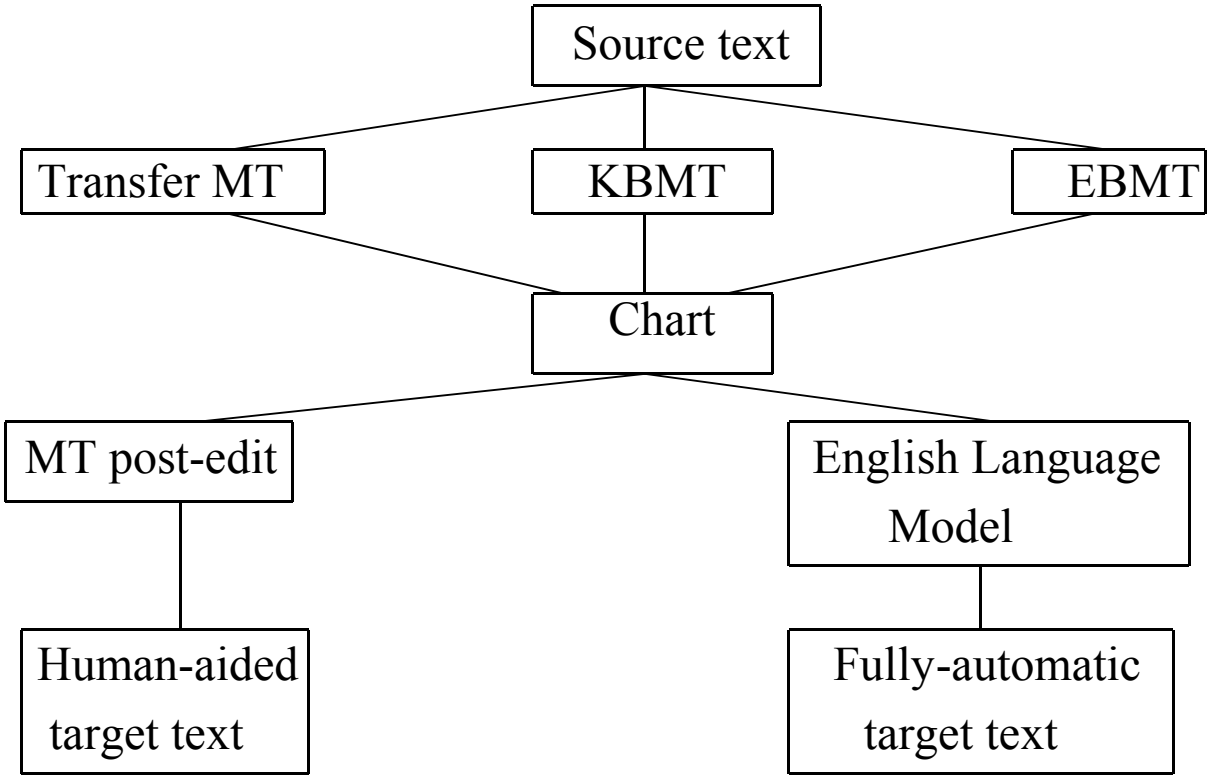
- Use of already translated sentences or phrases either from actual translations (corpus) or from data supplied by user or developer
- Sentences/phrases aligned in database (either by rule-based parser or statistically)
- matching algorithm (exact and close) of SL input and TL examples
- combination 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



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!

Computer-aided 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, 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

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 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)
  - searching for words in combination (e.g. ...take… + …a swipe at…)
  - re-combining phrases to produce sentences
- example-based MT research

Translation workstations
(often called Translation memory systems)

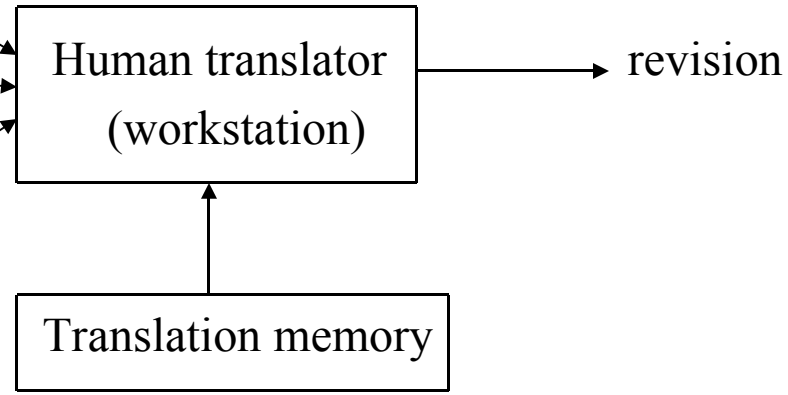
- Components and facilities controlled by users (translators)
- Terminology management
- Translation memory, and alignment
- Facilities for building dictionaries (e.g. from Internet)
- Augmented by MT systems
- Compatible with authoring systems (technical writers)
- Compatible with publishing systems

Workstations (TM systems) available

- Trados Translation Solution
- STAR Transit
- 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

Machine-aided human translation

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



The translation demand

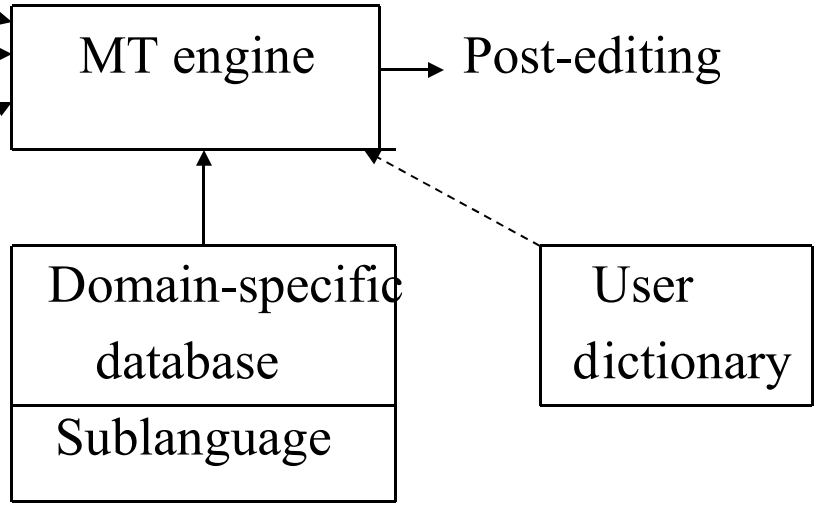
- Dissemination: production of ‘publishable quality’ texts
  - but, since raw output inadequate:
    - post-editing
    - control of input (pre-editing, controlled language)
    - domain restriction (reducing ambiguities)
- assimilation: for extracting essential information
  - use of raw output, with 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



- unlike MAHT, the human is not at the centre 'in control'

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.)

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]

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 corrections

- 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 corrections (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 (from MT dictionaries)
  - 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

Custom-built controlled-language systems

- Caterpillar Corp, [with Carnegie-Mellon Group; interlingua]
- LANTMARK [Xplanation b.v., Belgium; old METAL system]
- Smart Translator [Smart Corporation, New York]
  - clients: Citicorp, Chase, Ford, General Electric, Canadian Ministry of Employment
- WebTran [VTT Information Technology, Finland]
- Cap Volmac
- ESTeam Ltd. (Greece) [own statistics-based system]

In-house and special-purpose systems: examples

- Pan American Health Organization [medical, social, welfare]
- Japan Center for Science and Technology [abstracts]
- NHK [news broadcasts]
- IBM Japan
- CSK (Japan)
- PaTrans:[patents]
- GSI Erli
- Hook and Hatton [chemistry texts]
- Linguanet [Police, customs, air traffic control]
- ALTo [TV captions; English to Spanish; spoken language transcription
- DIPLOMAT [Military ‘field’ communication; Carnegie-Mellon]
- Phraselator [Military, government, tourism]

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

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**
- own association: Localization Industry Standards Association
- examples of 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)

EURAMIS: example of convergence

- 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

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), some examples for European languages:
  - ENGSPAN (PAHO): English→Spanish
  - ESI Professional (WordMagic): English↔Spanish
  - 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]
  - 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

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)

- First in 1980s: ALPS, Weidner, Microtac, Globalink, various Japanese systems
- Dictionaries (both as CD-Roms and downloadable from Internet)
- PC systems, examples for European languages
  - 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
  - LogoMedia Translate (LogoMedia): Chinese↔English, English↔French, English↔German, English↔Italian, English↔Japanese, English↔Korean, English↔Portuguese, English↔Russian, English↔Spanish
  - 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
  - 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 hand-held devices (Personal translation)

- Special devices (most little more than dictionaries)
  - 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 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

Free online 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

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) [English↔French]

MT in the marketplace

- retail availability
  - many only purchased direct from manufacturer
- 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 (many systems have failed)

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

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

Some future developments and expectations

- 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) - also online
- bilingual (multilingual) communication as much as translation

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, ViaVoice Translator, Vocal PeTra
- Online translation with voice output
  - Translation Wave
- Speech translation

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

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

How to judge MT

- 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

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’)

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* [directory of current commercial systems on EAMT website]
- 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