

Machine translation and computer- based translation aids

John Hutchins

(Email: WJHutchins@compuserve.com)

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

University of East Anglia, Norwich

November 2003

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

Mechanical translation from 1933 to 1956

- 1933: Troyanskii's patent proposal
- 1947: letter from Warren Weaver to Norbert Wiener, Weaver meets Andrew D. Booth
- 1949: memorandum from Warren Weaver; research by Reifler, Kaplan, Oswald, Fletcher, Huskey
- 1951: Bar-Hillel appointed to MIT
- 1952: first MT conference
- 1954: first MT systems demonstrated (IBM and Georgetown University); research begins in Soviet Union
- 1954: first MT journal established (Mechanical Translation)
- 1955: first book on MT published (ed. Booth and Locke)
- 1956: first international MT conference

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’ (FAHQ T) had neglected:
 - operational factors and requirements
 - expertise of translators
 - machine aids for translators
- henceforth three strands of MT:
 - translation tools
 - operational systems (post-editing, controlled languages, domain-specific systems)
 - research (new approaches, new methods)

System architectures and strategies

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

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

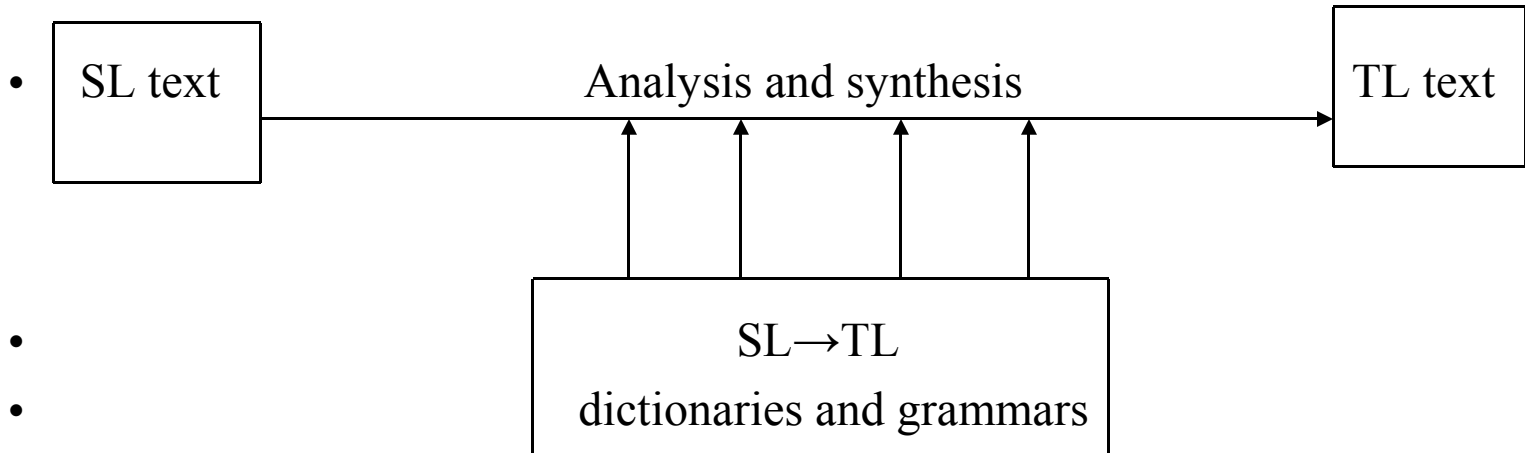
Bilingual lexical ambiguity

- 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

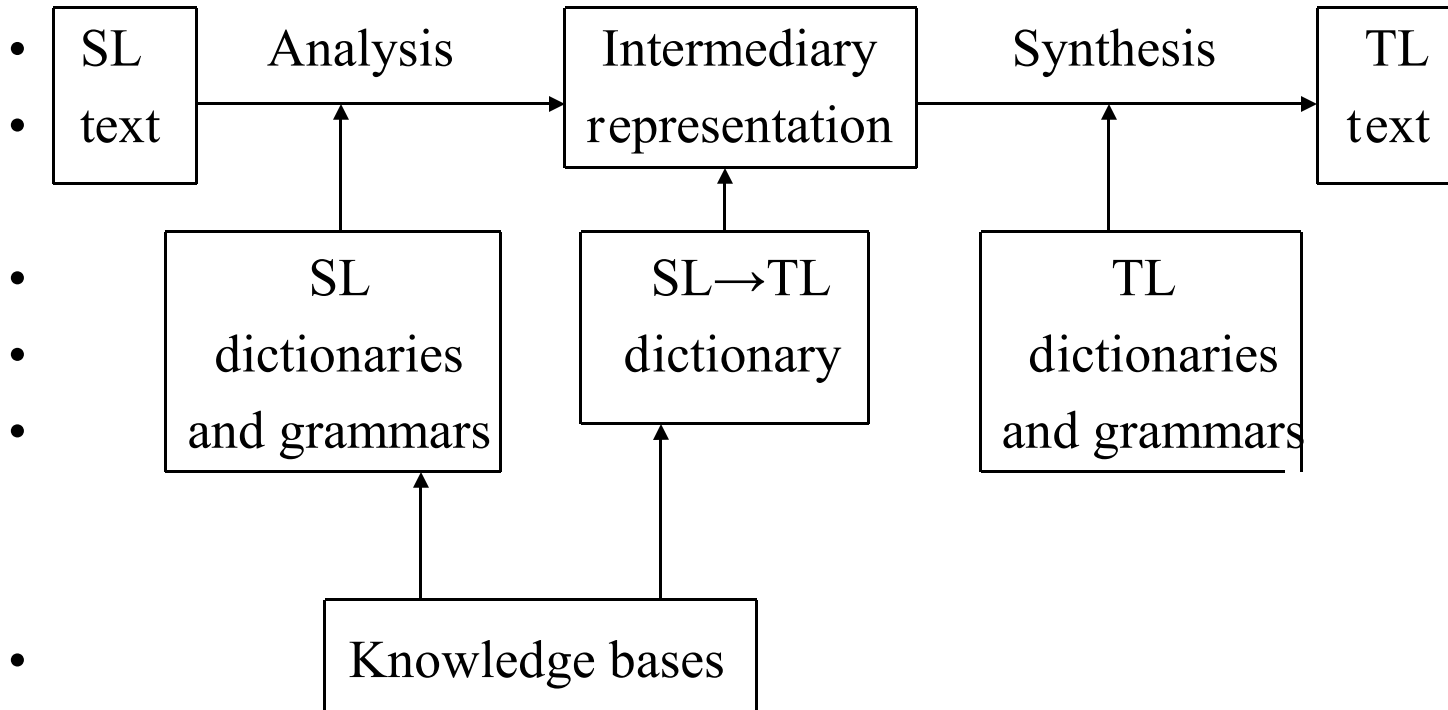
Direct translation



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



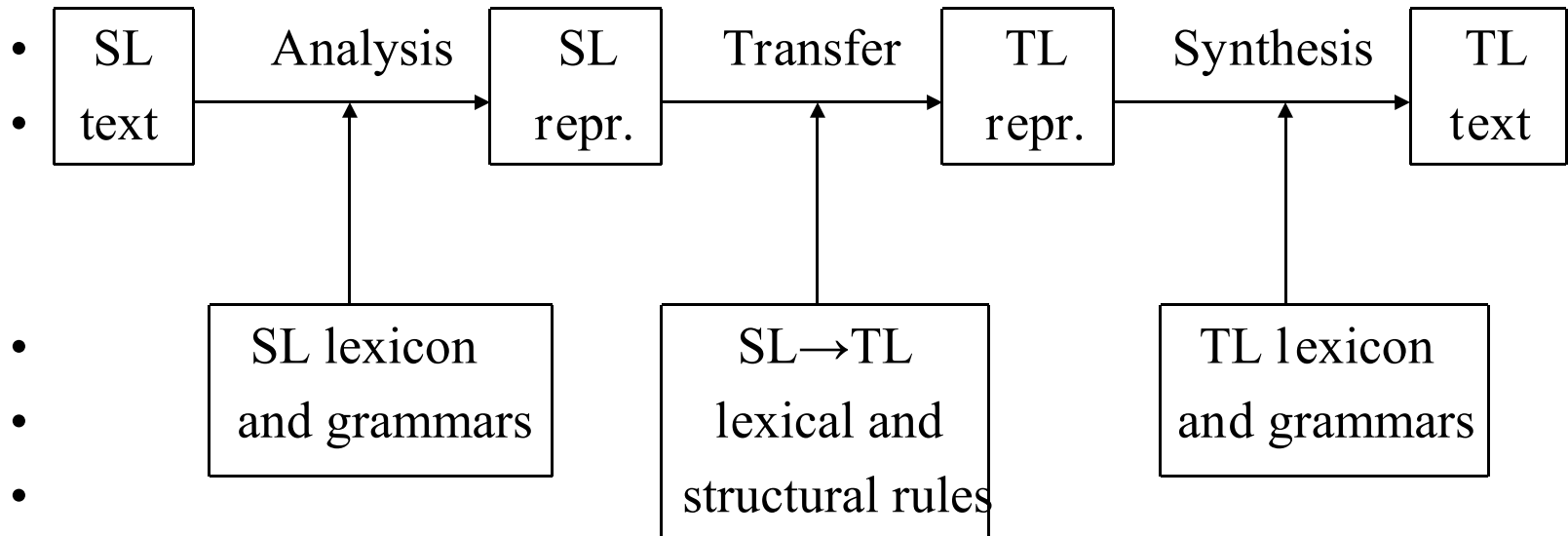
Interlingua-based MT

- 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)

Levels (strata) of analysis and synthesis

- Morphological analysis
 - identification of endings (e.g. -s for plurals, 3rd sing.; -ly for adverbs; French *-ment* for adverbs; German *-heit* for nouns, etc.)
- Syntactic analysis (surface)
 - adjective-noun modification, noun phrases, noun-verb modification, coordination, etc. (phrase structure)
- Syntactic analysis (deep)
 - relations of agent, object, indirect object (beneficiary), adverb to main verb, prepositional phrase to verb, etc. (case relations)
- Semantic analysis
 - acceptability of noun-type for verb-type (e.g. *drink* and animate noun)
- ‘Reality’ (domain) analysis
 - *tape* in IT context is ‘magnetic tape’ not ‘adhesive tape’
- Semantic synthesis
 - collocation of acceptable patterns
- Syntactic synthesis
 - construction of phrase structure and relationships
- Morphological synthesis
 - selection of correct word forms

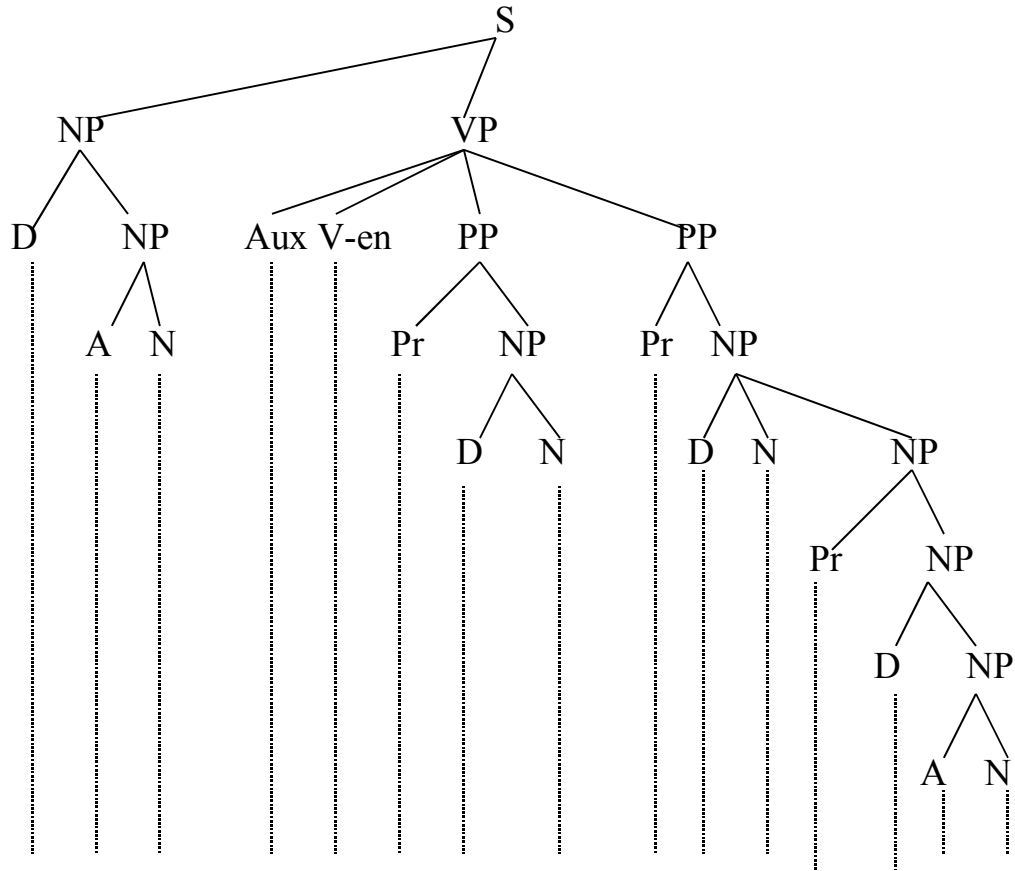
'Transfer' system



Transfer-based MT

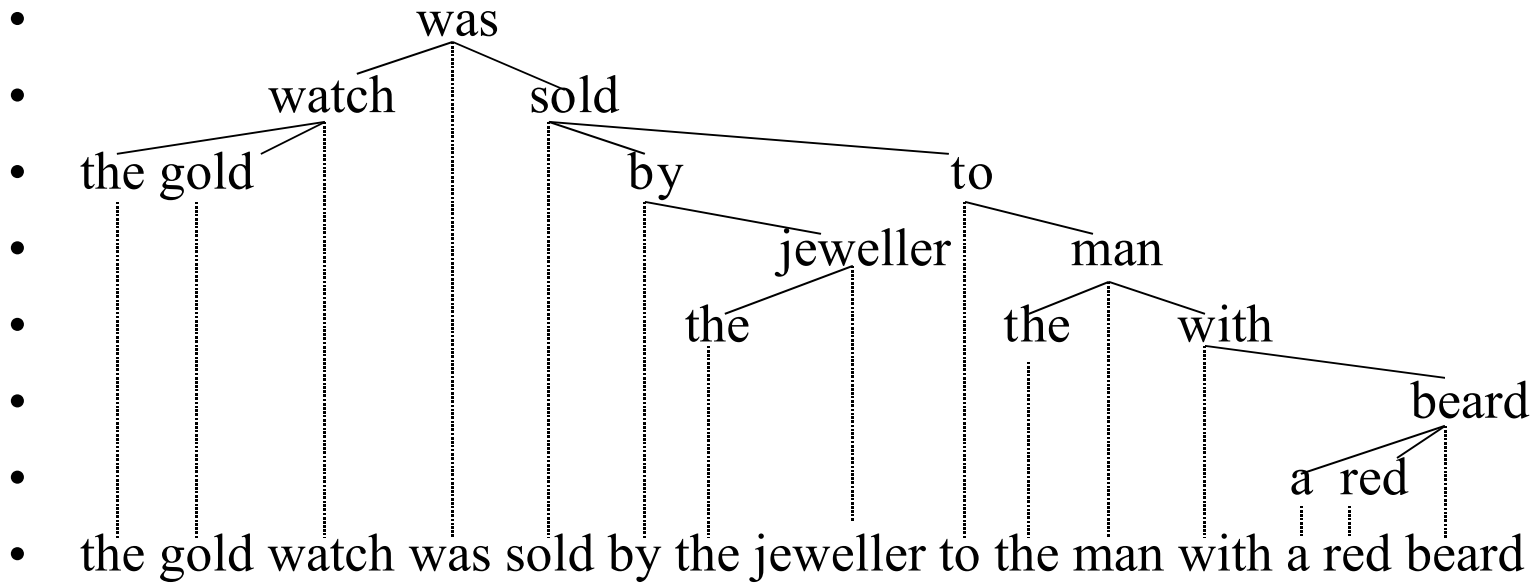
- three stages: analysis, transfer, synthesis
- abstract semantico-syntactic interfaces/representations
- multiple level/strata: morphology, syntax, semantics
- syntax-oriented, tree-transduction
- batch processing, post-edited
- little/no discourse information (anaphora, etc.)
- projects/systems:
 -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

Constituency ('phrase-structure') grammar

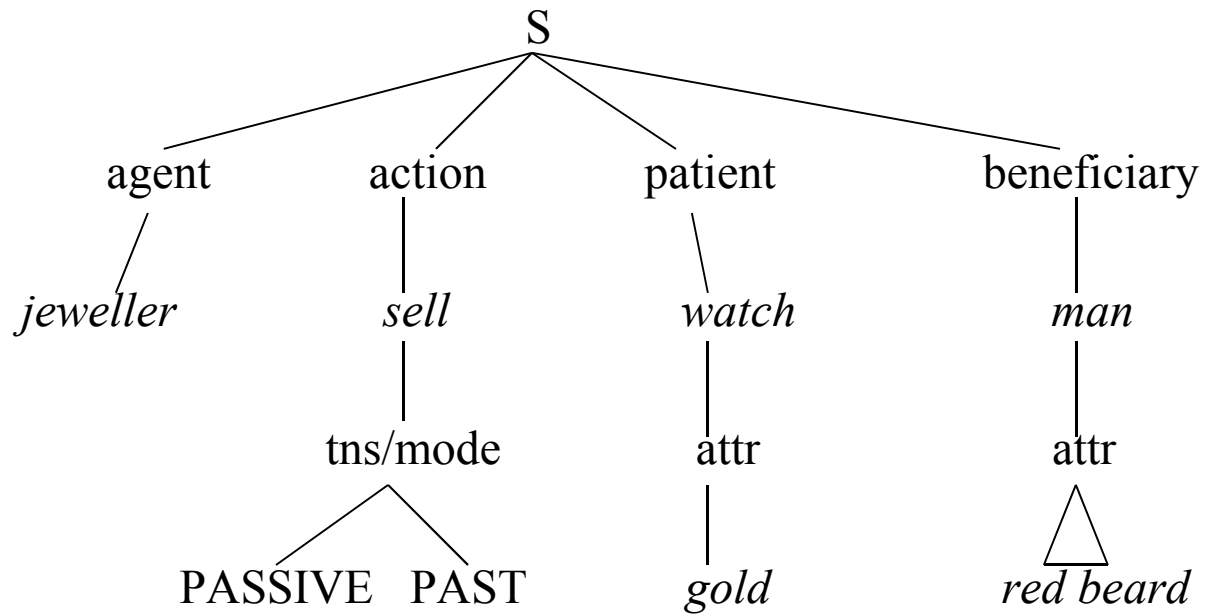


the gold watch was sold by the jeweller to the man with a red beard

Dependency grammar



Case grammar

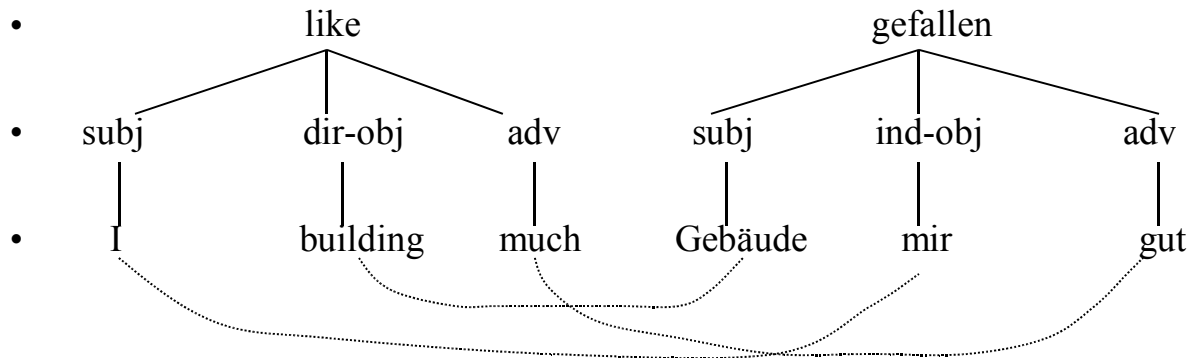


Bilingual structural differences

- (1) Young people like this music
 - Cette musique plaît aux jeunes gens
- (2) The boy likes to play tennis
 - Der Junge spielt gern Tennis
- (3) He happened to arrive in time
 - Er ist zufällig zur rechten Zeit angekommen
- (4) Le moment arrivé je serais prêt
 - When the time comes, I shall be ready
- **Need for complex rules of syntactic transformation, or rules/patterns for generating correct target language sentences**

Tree transduction

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



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

Anaphora

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

Non-linguistic problems of ‘reality’

- The soldiers shot at the women and some of them fell
- The soldiers shot at the women and some of them missed
 - must know what ‘them’ refers to e.g. if translating into French (*ils* or *elles*)
- **No solutions without non-linguistic context, i.e. probably outside competence of computational methods.**
- However, perhaps this aspect is exaggerated: no need to understand what AIDS and HIV are in order to translate:
 - The AIDS epidemic is sweeping rapidly through Southern Africa. It is estimated that more than half the population is now HIV positive.

Problems of stylistic difference

- The possibility of rectification of the fault by the insertion of a valve was discussed by the engineers
- The engineers discussed whether it was possible to rectify the fault by inserting a valve
- [English] Advances in technology created new opportunities
- [Japanese] Because technology has advanced, opportunities have been created
- [or Japanese] Technology has advanced. There are new opportunities.
- **All current methods of MT tend to retain SL structural features.**

From 1967 to 1979

- After ALPAC continuation of research in US (Texas, Wayne State), Soviet Union, UK, Canada, France
- **rule-based approaches: interlingua and transfer**
- 1970: Systran installed at USAF (Foreign Technology Division)
- 1970: TITUS installed (restricted language: textile industry abstracts)
- 1975: Météo ‘sublanguage’ English-French system (weather broadcasts)
- 1975: CULT Chinese-English (restricted language: mathematics)
- 1976: European Commission acquires Systran
- 1979: Pan American Health Organization system (SPANAM)
- 1979: Eurotra project begins

MT research in 1970s and 1980s

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

Changes since late 1980s

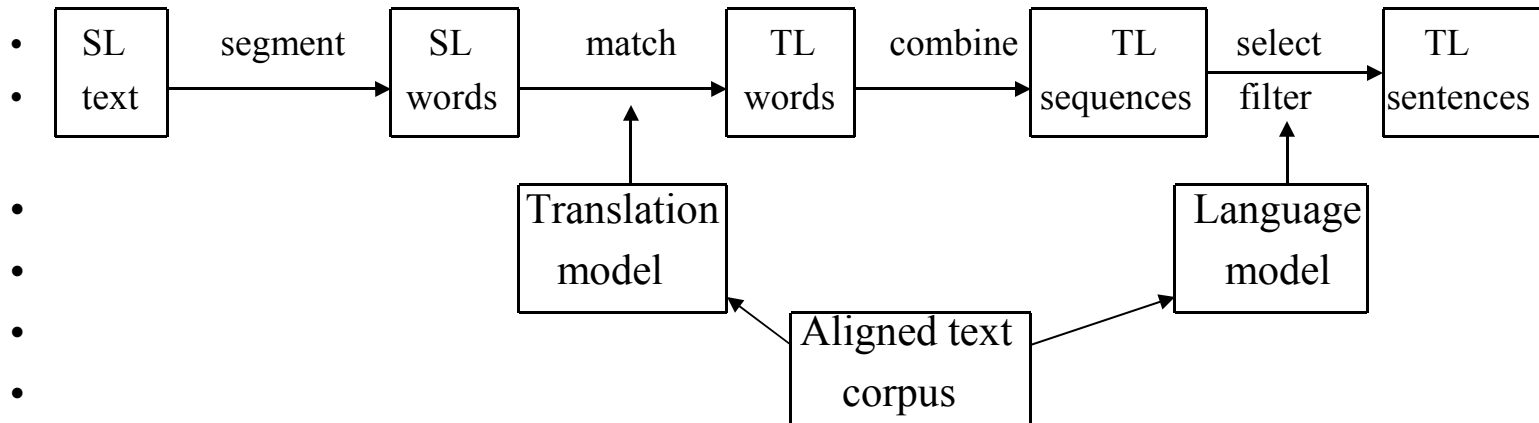
- Increasing use of MT by large enterprises
- Translation memory and translation workstations
- Localization
- Growth in PC systems
- The impact of the Internet
- Online translation
- MT and other language activities
- **Research on corpus-based MT methods**

Corpus-based systems

- Not rule-based: grammar rules (analysis, transfer, synthesis), multiple strata, ‘deep’ semantic analysis; complex dictionary entries
- based on bilingual text resources, e.g.
 - have a direct effect on... ont une influence directe sur...
 - have a direct effect on... intéressent directement
 - have a direct effect on... ont eu une répercussion directe sur...
 - has had a marked effect on... a largement influencé...
 - had a positive effect on... s’est avérée positive dans...
- Extraction of phrases for re-combination [Example-based MT]
- Statistical translation model (word-word frequencies), target language model (word co-occurrences) [Statistics-based MT]
- Text alignment methods enabled use of bilingual text corpora [Translation Memory]

Statistics-based MT

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

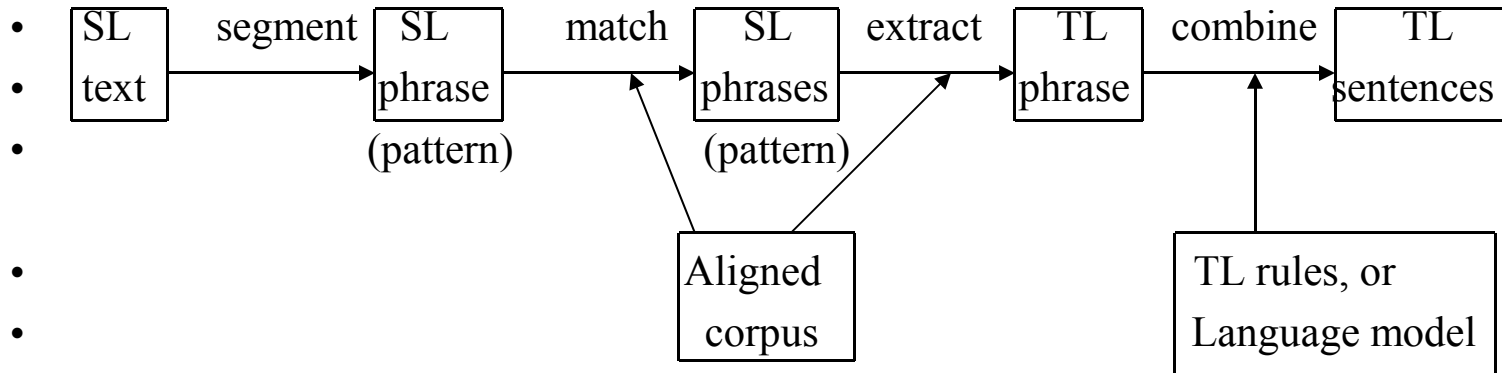


Statistics-based MT

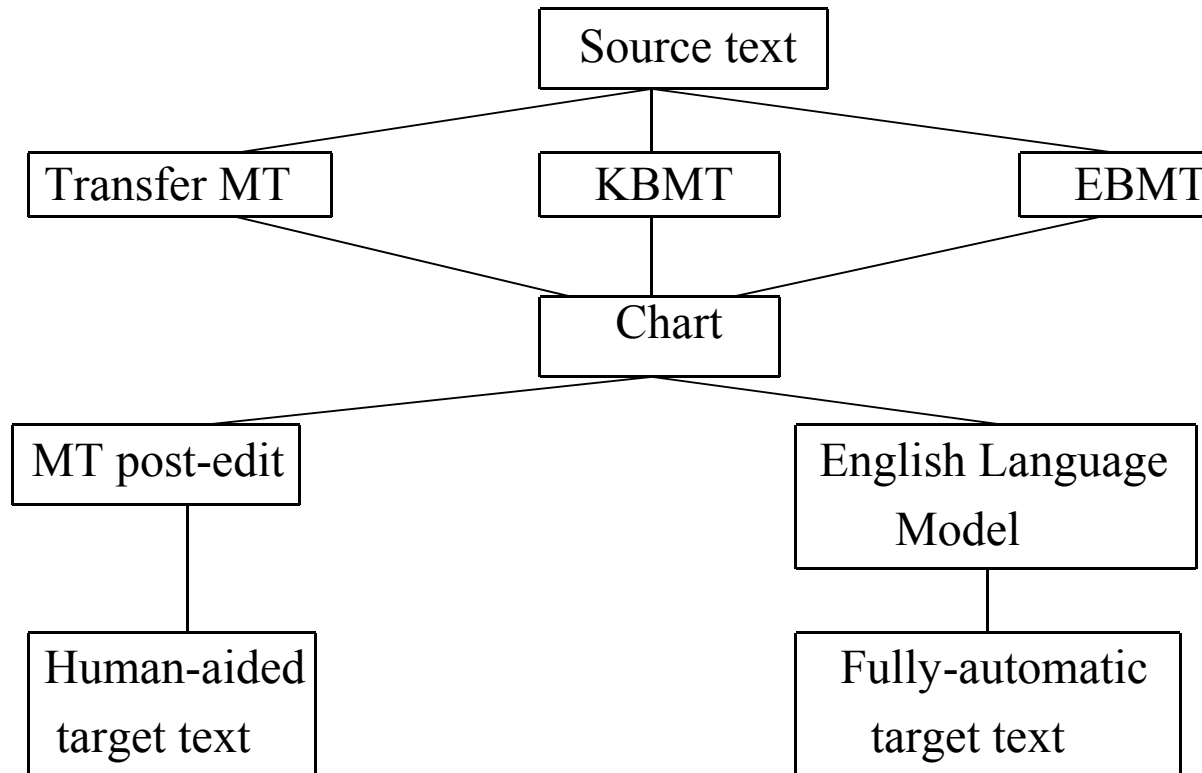
- Bilingual corpora: original and translation
- little or no linguistic ‘knowledge’, based on word co-occurrences in SL and TL texts (of a corpus), relative positions of words within sentences, length of sentences
- Sentences aligned statistically (according to sentence length and position)
- compute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

- 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)

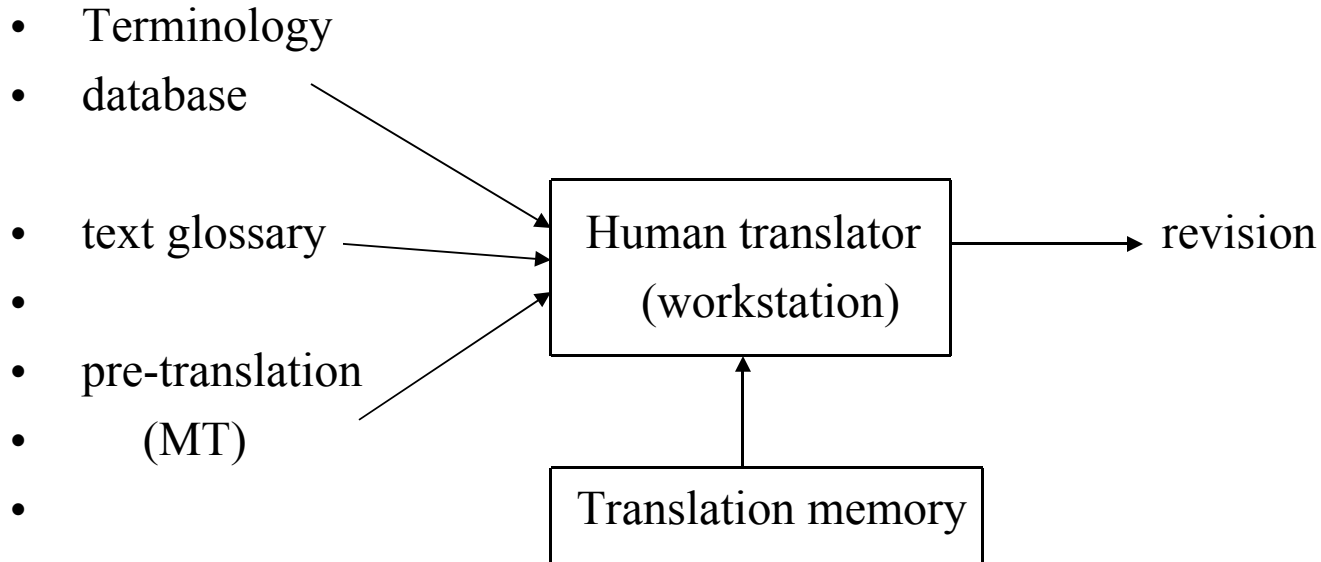


Speech translation: problems

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

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

Machine-aided human translation



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 (Linguattech)

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

- Although 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
- perhaps from example-based MT research

Translation workstations

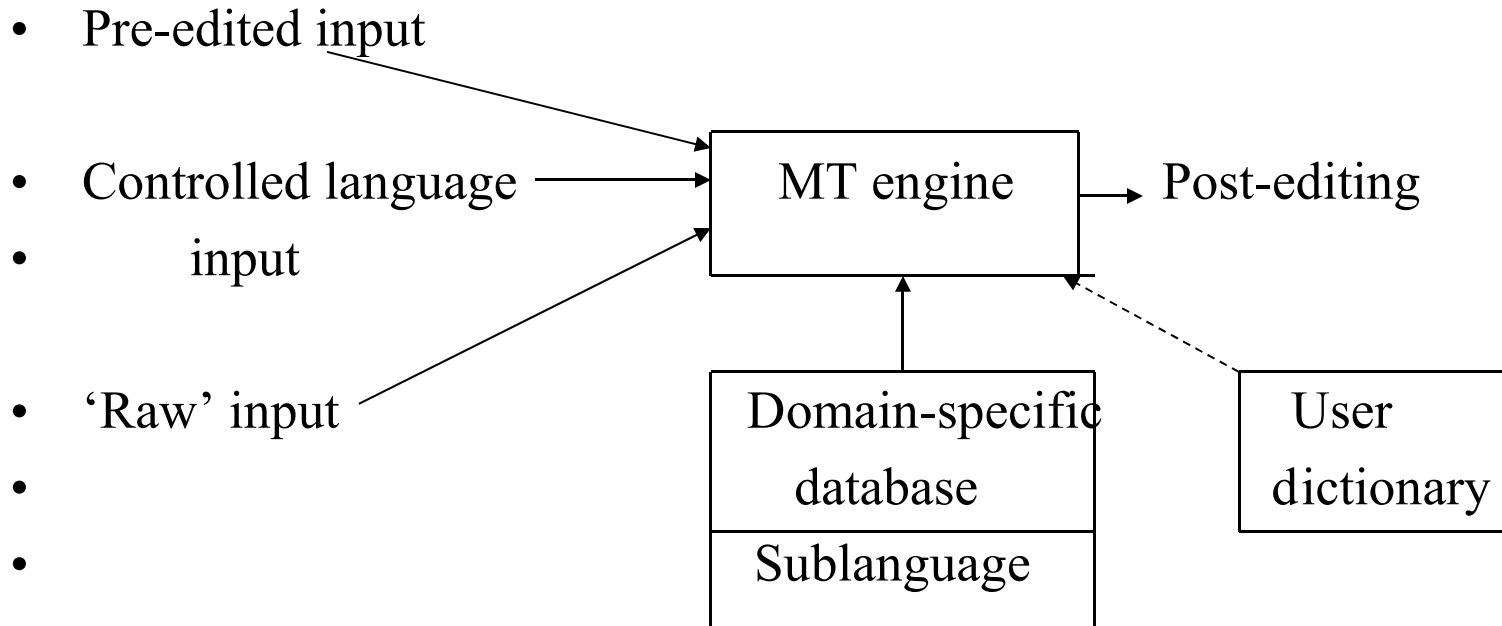
(often called Translation memory systems)

- 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

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

Human-assisted MT



- 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.)

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 procedimiento 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 diagnosticados 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.

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

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)

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

Convergence of HAMT and MAHT

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

MT for translators (office systems): requirements

- 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)

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

MT for personal translation

- Dictionaries (both as CD-Roms and downloadable from Internet)
- PC systems
 - first in 1980s (ALPS, Weidner, Globalink, Japanese systems)
- Hand-held devices (for tourism, text messages)
- online services (for emails, webpages)
 - free services (Minitel, Babelfish)
 - charged (with human postediting)

Online and PC translation: why so bad?

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

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

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

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 lexis)
- 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’)

Machine translation and human translation in complementation

- HT for literature, and other ‘culturally-sensitive’ translation
- MT for technical, scientific, medical (etc.) texts which are culturally neutral
- HT and human aid for dissemination (publishable quality)
- MT for assimilation (rough ‘gist’)
- MT for real-time on-line translation (is this its ‘real’ niche?)
- HT for spoken language translation
- MT for integrating translation with other LT tasks

MT as bilingual communication aid

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

Sources of information

- EAMT website (www.eamt.org) with links to other IAMT sites, etc.
- LISA website (www.lisa.org)
- Conferences: MT Summit, EAMT workshops, LISA Forums
- Journals: *Machine Translation*, *Multilingual Computing and Technology*, *MT News International*
- *Compendium of translation software* [directory of current commercial systems on EAMT website]
- Books:
 - Hutchins, W. John and Somers, Harold L.: *An introduction to machine translation* (London: Academic Press, 1992)
 - Sprung, Robert C. (ed.): *Translating into success*. (Amsterdam: John Benjamins, 2000)
 - Somers, Harold (ed.): *Computers and translation* (Amsterdam: John Benjamins, 2003)
- my website:
 - <http://ourworld.compuserve.com/homepages/WJHutchins>