Uses and Applications of Machine Translation

John Hutchins

Categories of systems

- Machine translation – for enterprises

- Machine translation – for professional translators

- Machine translation – for casual/home use

- Machine translation for bilingual communication

- Translation memory systems

- Translation workbenches – for professional translators

- Bilingual dictionaries

Basic types of use

- Dissemination (for publication)

    – Enterprise systems

    – Involving controlled input, company terminology, post-editing, etc.

- Assimilation

    – Acceptable lower quality

- Bilingual communication

    – Interchange, with feedback and elucidation

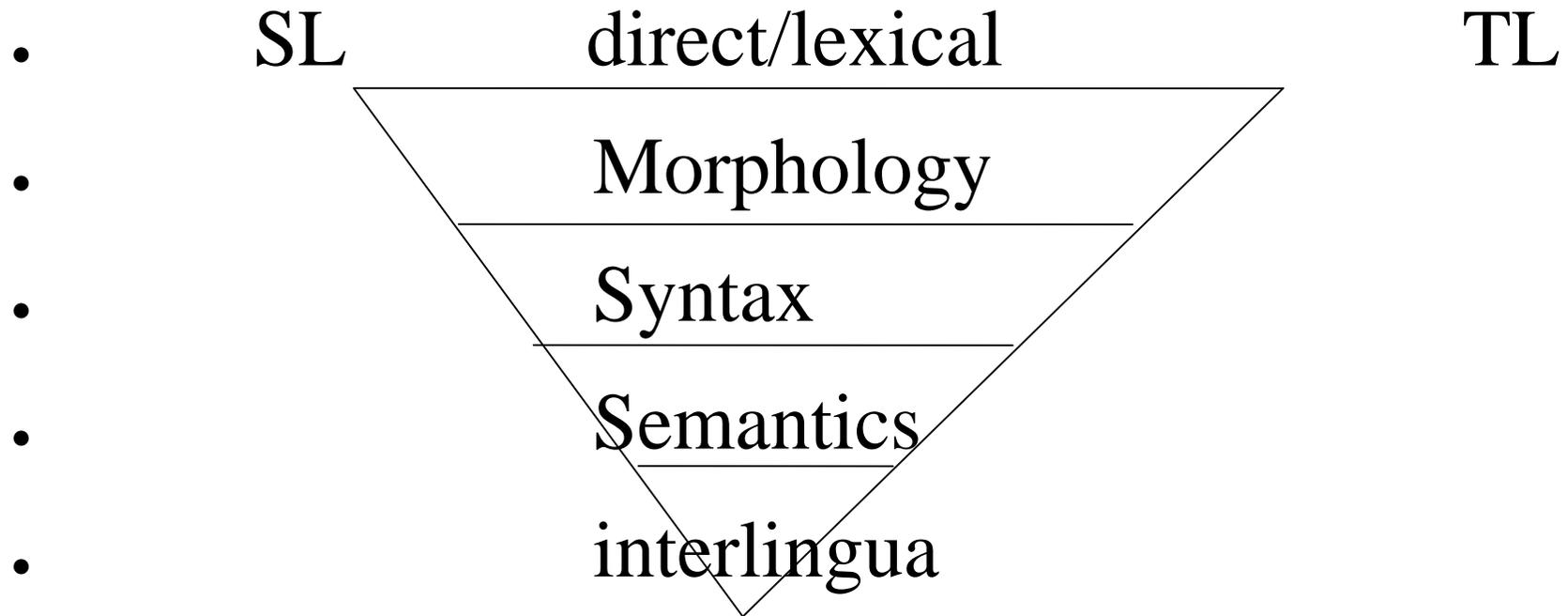- Translation aids

    – Drafts, dictionaries

General-purpose vs Subject-specific

- General purpose systems

  - General dictionaries with all translation options (or selection of most common only)

- Subject-specific systems

  - Fewer ambiguities within subjects (sublanguages)

  - Subject-specific system dictionaries

  - User dictionaries

  - Controlled language input

  - Terminology

Basic architectures

- Direct translation: Segment, substitute, rearrange

  - Early dictionary-based systems

  - Statistical machine translation

- Indirect translation (rule-based): multiple levels

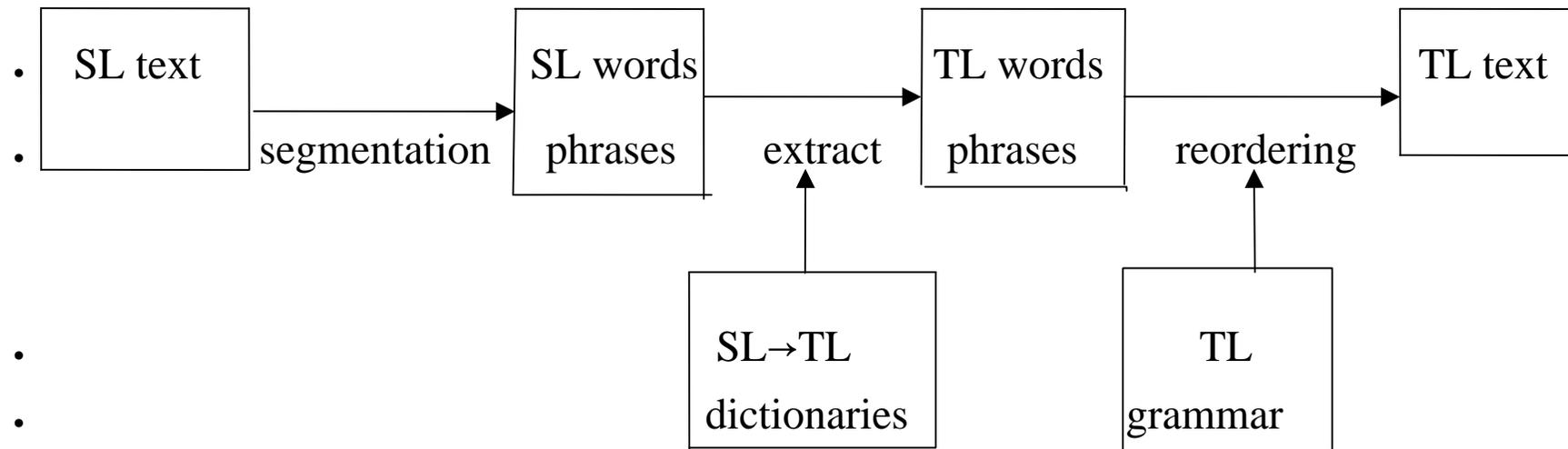 - Interlingual approach: two stages

  - Transfer: three stages

Main MT system types

- SL   direct/lexical   TL

-               Morphology

-               Syntax

-               Semantics

-               interlingua

Historical outline

- earliest systems (1950s and 1960s)

- ALPAC report (1966)

- 'quiet' decade (1966 to 1975)

- adoption of Systran by CEC (1976); Meteo

- coming of PC systems (1980s); coming of translation aids

- increasing use (since mid 1980s): companies, localisation, etc.

- dominance of 'transfer' framework (1980s)

- interlingua systems (late 1980s)

- translation memory (since late 1980s)

- corpus-based MT research (from late 1980s)

- online MT (since late 1990s)

'Direct' translation model

| SL text | → segmentation → | SL words phrases | → extract → | TL words phrases | → reordering → | TL text |

SL→TL dictionaries → extract
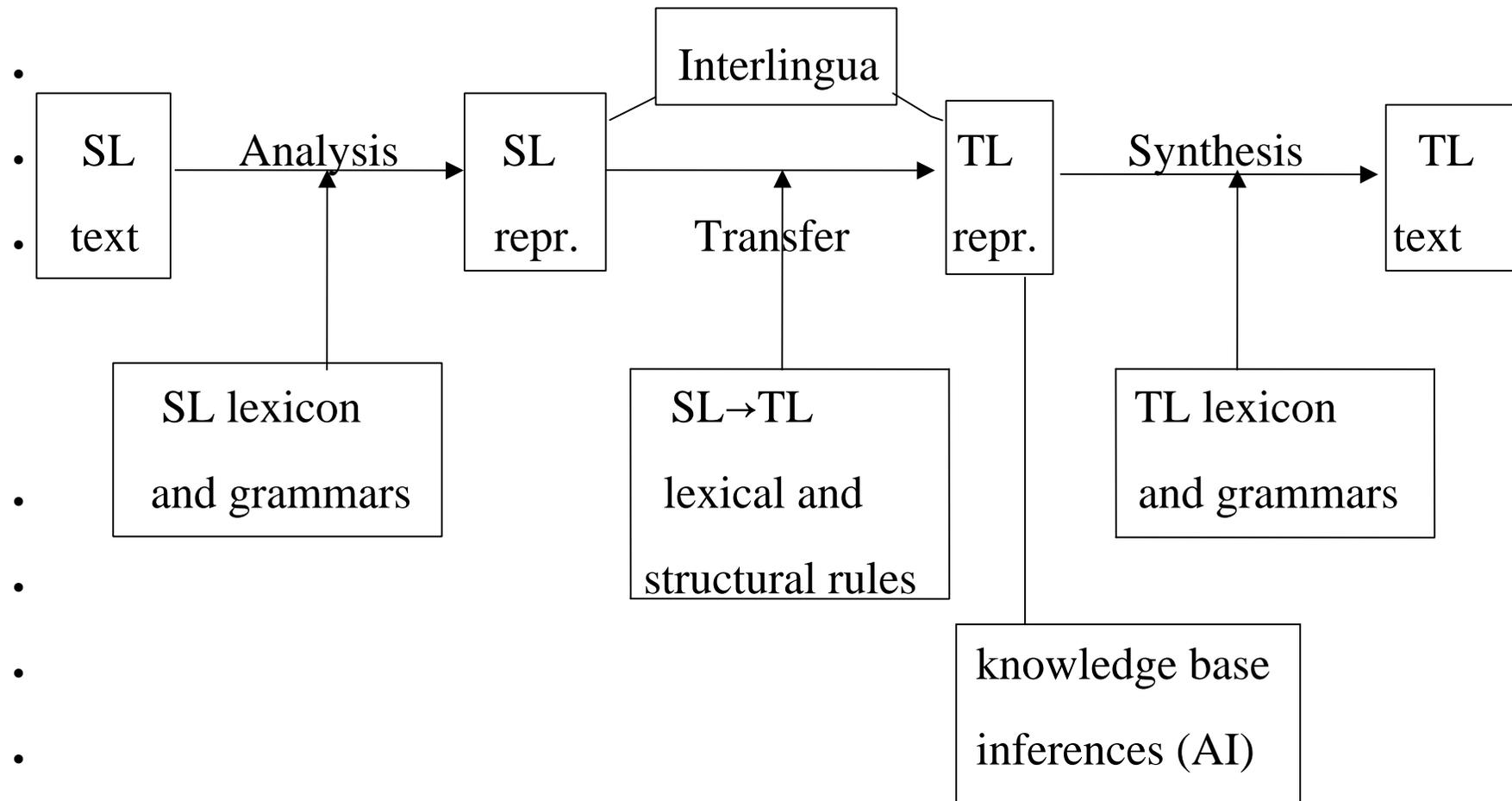
TL grammar → reordering

- model: segment, extract, combine/reorder

- original/early MT: word for word, some morphology, some reordering of TL

- later added some syntactic analysis (e.g. Georgetown systems)

Direct translation

- analysis of SL only as much as necessary for conversion into particular TL; dictionary lookup followed by TL word-for-word output, then TL rearrangement based on dictionary entries

- use of 'cover' words (most frequent not most appropriate)

- no analysis of SL syntax or semantics

- output too close to SL structure

- problems of direct translation systems:

    – too complex for modification and enhancement (not just computationally)

    – mixture of lexical rules and syntactic rules (no linguistic or translation 'theory')

- systems:

    – Univ.Washington, IBM (US), Georgetown University (US), Ramo-Wooldridge (US), Institute for Precision Mechanics and Computer Technology (USSR), National Physical Laboratory (UK)
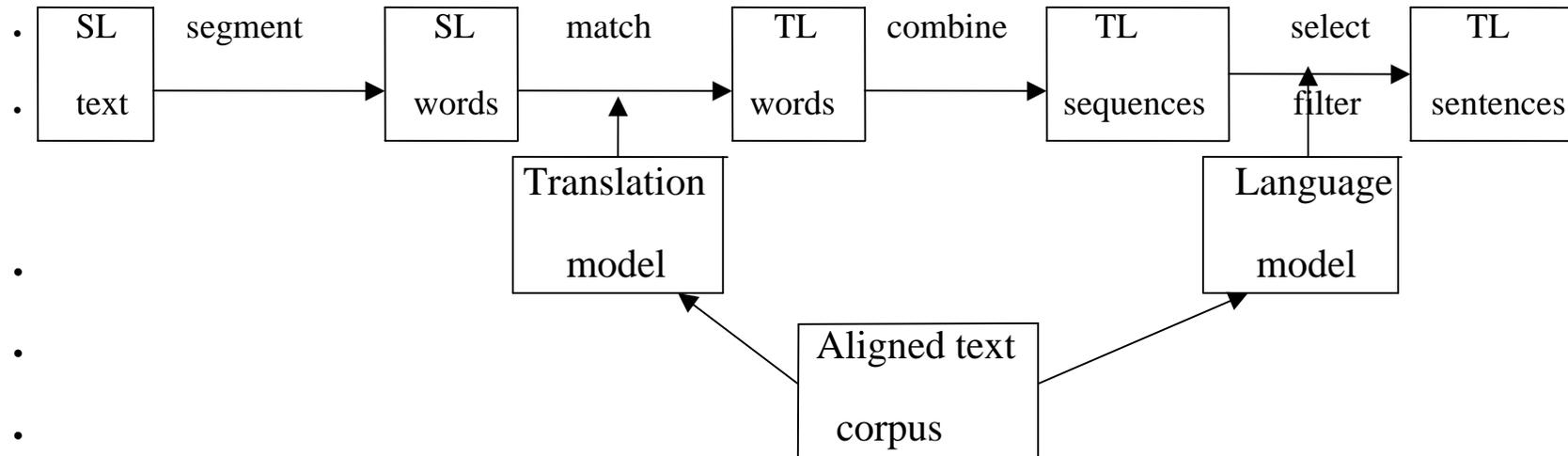
'Transfer' and interlingua systems

-

-

-

```
                              ┌─────────────┐
                              │ Interlingua │
                              └─────────────┘
                             ╱               ╲
┌──────┐              ┌──────┐              ┌──────┐              ┌──────┐
│ SL   │  Analysis    │ SL   │              │ TL   │  Synthesis   │ TL   │
│      │ ───────────▶ │      │ ──────────▶  │      │ ───────────▶ │      │
│ text │      ▲       │ repr.│      ▲       │ repr.│      ▲       │ text │
└──────┘      │       └──────┘   Transfer   └──────┘      │       └──────┘
              │                     │           │          │
      ┌───────────────┐    ┌───────────────┐    │   ┌───────────────┐
      │  SL lexicon   │    │    SL→TL       │    │   │  TL lexicon   │
      │               │    │                │    │   │               │
      │ and grammars  │    │  lexical and   │    │   │ and grammars  │
      └───────────────┘    │                │    │   └───────────────┘
                           │ structural rules│   │
                           └───────────────┘     │
                                        ┌───────────────┐
                                        │ knowledge base │
                                        │                │
                                        │ inferences (AI)│
                                        └───────────────┘
```

-

-

-

Transfer and interlingua systems

- Transfer systems
  - early example (1960s): 'syntactic transfer' at MIT
  - archetypal (1980s): Ariane, Eurotra, METAL, ETAP

- Interlingua systems
  - early examples (1960s): CLRU, Milan
  - Texas U (abandoned in favour of transfer approach)
  - later examples (1980s): Rosetta, DLT
  - AI/ knowledge based (late 1980s early 1990s): CMU, Pangloss
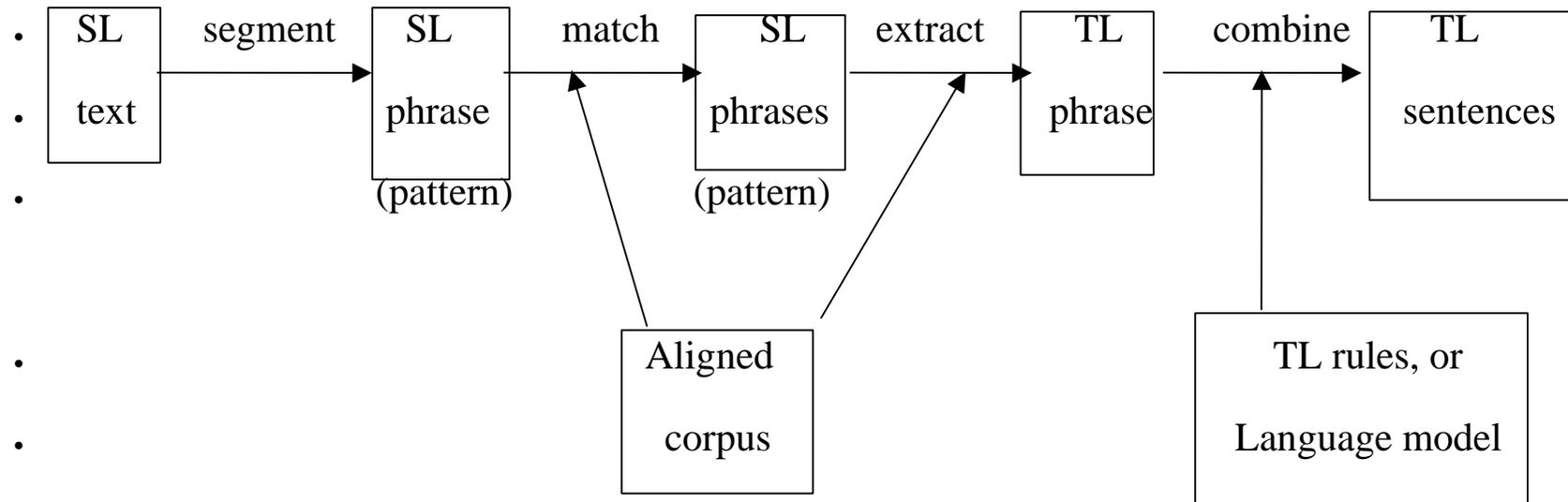 - recent (1990s to date): UNL

Statistics-based MT

- Research since early 1990s (first IBM Candide 1988)

- 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.

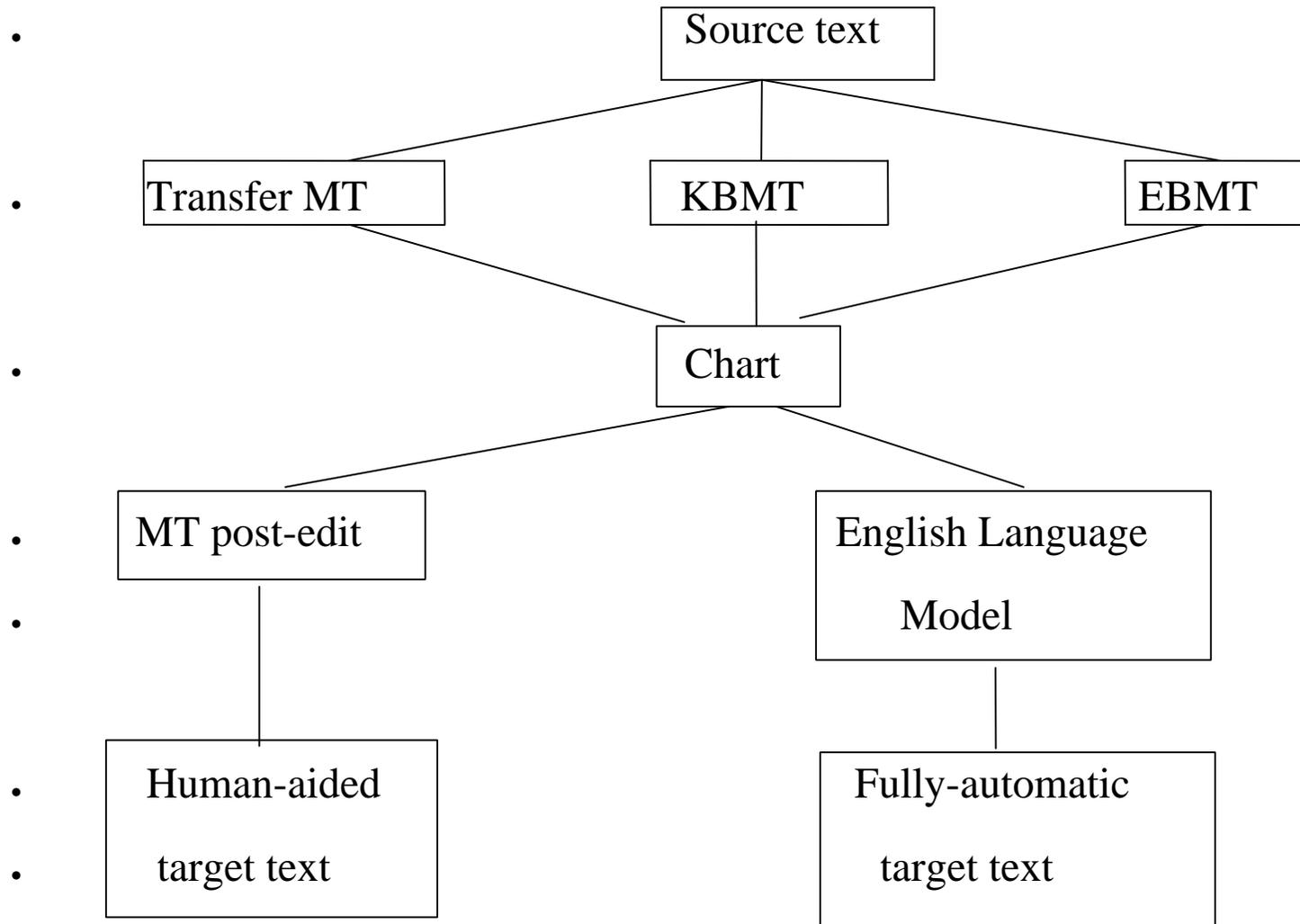| SL text | segment → | SL words | match | TL words | combine → | TL sequences | select filter | TL sentences |

```
SL        segment      SL        match      TL       combine      TL          select      TL
text      ──────→     words    ──────→     words    ──────→    sequences     filter    sentences
                          │                  ↑                       ↑
                          ↓                  │                       │
                     Translation                              Language
                       model                                    model
                          ↖                                    ↗
                            Aligned text
                               corpus
```

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
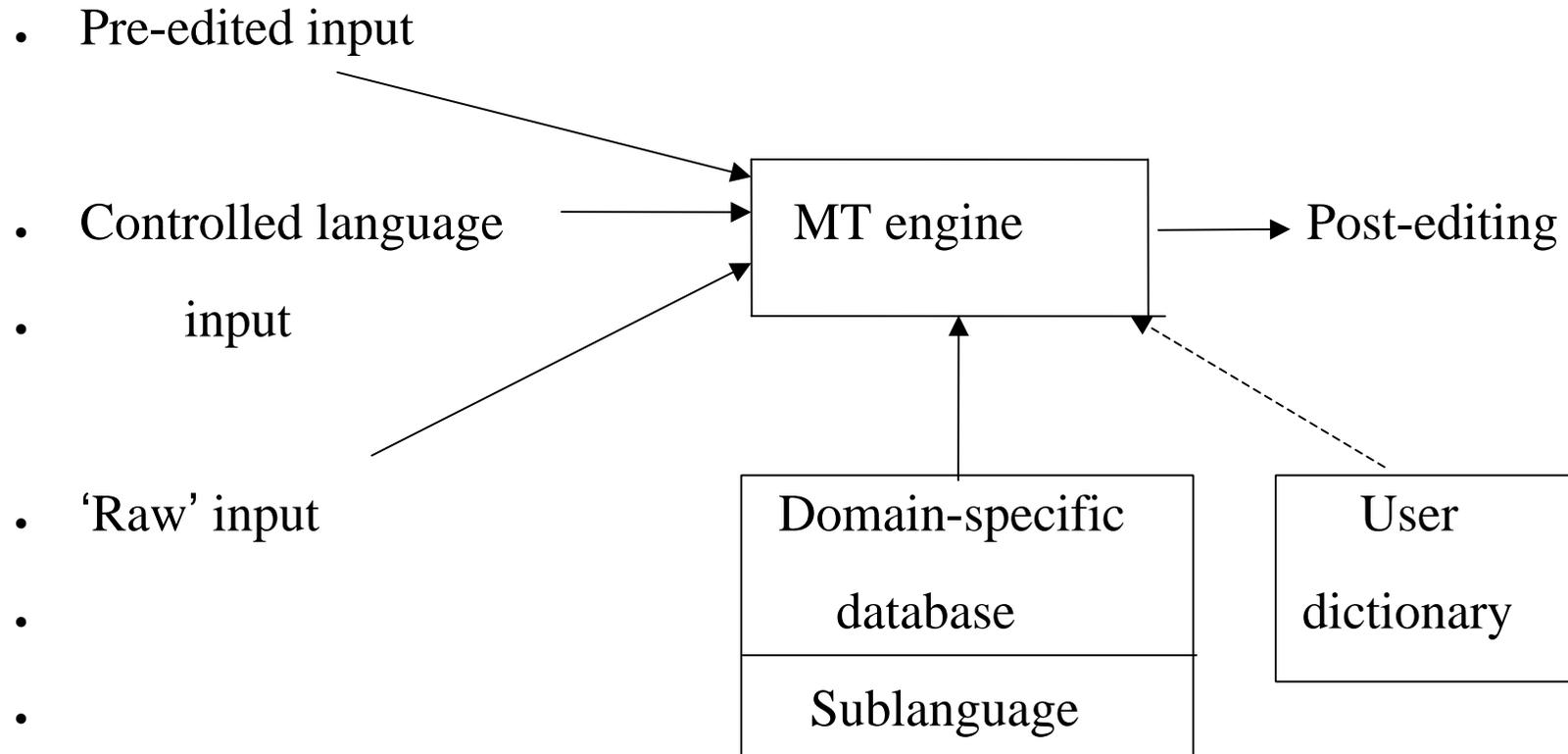
  - research since late 1980s

| SL text | segment → | SL phrase (pattern) | match → | SL phrases (pattern) | extract → | TL phrase | combine → | TL sentences |

Aligned corpus

TL rules, or Language model

Hybrid systems

- RBMT, EBMT, SMT
  - Each with strengths: RBMT analysis, EBMT transfer, SMT language model
- Multi-engine
  - Chart integration of outputs
- Statistical post-editing
  - RBMT with SMT language model

Hybrid systems:
an example

- Source text
- Transfer MT   KBMT   EBMT
- Chart
- MT post-edit   English Language Model
- Human-aided target text   Fully-automatic target text

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)

- (since late 1980s) 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!

Operational MT: Human-assisted MT

- Pre-edited input

- Controlled language

- input

- 'Raw' input

-

-

MT engine → Post-editing

Domain-specific database

Sublanguage

User dictionary

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)

Post-editing:
types of errors for correction

– 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

– prepositions:

- …el desarrollo de programs de educación nutricional...

- MT: …the development of programs of nutritional education

- PE: …**in** nutritional education…

– verb phrases:

- …el procedimento para registrar los hogares…

- MT: the procedure in order to register the households

- PE: …the procedure for registering households

Post-editing: types of errors (contd.)

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

Translators and post-editors

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

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

MT dictionaries

- For rule-based systems

    - Direct systems

        - Bilingual conversion, TL rearrangement

    - Indirect systems

        - Morphology, syntax, semantics, [knowledge]

- For statistical systems

    - Bilingual conversion (frquency data from corpora)

    - Monolingual 'language model'

Lexical acquisition: RBMT systems

- dictionary building

  - hand-crafted (pre-1990) was expensive in time and effort

  - required information: morphological variants, grammatical categories, syntactic contexts, lexical co-occurrences, semantic conditions/constraints, translation options

  - generally more detailed than terminology information for human translation (and includes **all** words)

- major problem for all current (commercial and custom-built) systems

- providers: vendor vs. customer

  - basic dictionary, special dictionaries, user dictionary (customer-specific)

Lexical resources

- Resources for creating dictionaries

    – size (what is adequate? definition of domain)

    – use of lexical resources (printed dictionaries, Internet dictionaries)

    – extraction from electronic texts (monolingual/bilingual, internal, Internet, Web pages): word alignment

    – validating, checking

    – conversion into required formats for particular MT system

    – updating procedures

- access to resources:

    – EDR, ELRA/ELDA, LDC

Dictionary information in linguistics-based MT

- Morphological variation

  – full forms vs patterns/paradigms

- Phrasal verbs (e.g. *turn up*, *turn out*, *turn in*, …)

- Collocations (compound nouns

- Idioms

- bilingual diversification (e.g. *wall* = Wand, Mauer, Wall)

- Syntactic categories (in detail, to aid syntactic analysis)

- Semantic features

  – subject-verb constraints (e.g. The car drinks petrol)

Lexical entry in CMU system (*find*)

- (find
    - (make-frame
    - +find-v1
    - (CAT (value v))
    - (STUFF
    - (DEFN "to discover by chance, to come across")
    - (EXAMPLES "found X in the bedroom", "found X sleeping upstairs", "found that X was sleeping at home")
    - (MORPH
        - (IRREG (*v+past* found) (*v+past-part* found))
    - (SYN-STRUC
        - *OR*   ((root $var0)
            - (subj (root $var1)(cat N))
            - (obj (root $var2)(cat N))
            - ((root $var0)
                - (subj (root $var1)(cat N))
                - (xcomp (root $var2)(cat N)(form pres-part)))
            - ((root $var0)
                - (subj (root $var1)(cat N))
                - (comp (root $var2)(cat V)(form fin)))))
    - (SEM
        - (LEX-MAP
            - (%involuntary-perceptual-event
                - (experiencer (value ^$var1))
                - (theme (value ^$var2))))))

SMT lexical data

- based on bilingual and monolingual corpora (e.g. websites)

- morphological information (automatically generated, manually edited)

- annotations:

  - syntactic information (from parser of database)

  - semantic features (from WordNet)

- collocations (= 'phrases')

- bilingual correspondences (statistical)

- bilingual dictionaries (traditional, online)

- language model (for TL)

User dictionaries

- Supplement basic (system) dictionaries

  - subject-specific, company terminology

  - system dictionaries are generally untouchable

- for RBMT

  - provide morphological, syntactic (and semantic) information -- via patterns/paradigms (offered by system)

- for SMT

  - limited to inclusion of subject-specific corpora

Computer-based tools: origins

- (since 1966) recognition that fully automatic translation not appropriate for professional translators

- Term banks (since 1970): TEAM, LEXIS, TERMIUM, Dicautom, Eurodicautom

- Terminology management (Mercury/Termex)

- Text-related glossaries (since 1970s: Bundeswehr, ALPS)

- Translation databases ('translation memory')

  – first: Arthern (1978), Kay (1980), ALPS

- Melby's three levels (early 1980s)

  – word processor with integrated terminology aids, manual insertion of words

  – machine-readable input texts, concordance (to find occurrences of words in text), local term bank, automatic insertion of terms

  – integrated 'workstation' with MT system, and automatic 'quality' evaluation

Computer-aided translation and translation tools: 1980s

- (since 1980s) PCs and multilingual word processing, desk top publishing

- Translator 'in control'

- dictionaries (monolingual, bilingual): on-line access

- grammar aids, spelling checkers

- user glossary, terminology management, 'authorised' terms, standards, specialist glossaries

- input, output, transmission (OCR, pre-editing, controlled language)

- translation memory, alignment

- management support tools (project control, budgeting, workflow)

- previous antagonism of translators to MT diminished

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

- software: MultiTerm (Trados), MTX (Linguatech)

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: usage

- 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

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, untypical and rare translations

- 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, e.g. without linguistic information such as mirphology, and translators opt not to use the facility

- combining extracted translation segments left entirely to user/translator

- need for programs to 'meld' or conflate extracted phrases (semi-automatically)

Translation memory: searching (TransSearch)

- 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.

Translation workstations

(incorporating 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

Localization

- Internationalisation, globalisation (e.g. software and Web pages)

  - estimated market (end 2006) $3.5 billion and $3 billion resp. (ABI, 2001)

- Cultural and linguistic adaptation (not just translation)

  - currency, measurements, power supplies

- Screen commands and help files; users' guides; warranties; publicity, marketing; packaging; workshop manuals

- Large scale, multiple language output, fast results (days, not weeks)

- Repetitive (translation memory)

- Graphics, formatting, layout, etc. (to be preserved)

- **companies use both translation tools (workstations, translation memories) and MT systems**

- has its own associations: Localization Industry Standards Association; GALA

- Software companies (many in Ireland):

  - ALPNET; Berlitz; Compaq; Corel; Eastman-Kodak; IBM; Lotus; Microsoft; Oracle; SAP; Symantec

The translation demand (again)

- 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

MT for companies and government organisations (dissemination)

- originally only use (e.g. pre-1966 systems, Systran at USAF)

- Usually general-purpose systems (Systran, SDL)

    - adapted for subject-specific terminology

- often with controlled language input (earliest: Xerox)

    - closely integrated with authoring software

- usually with post-editing

    - closely integrated with publishing software

MT for translators (office systems): requirements

- translation workstations still too expensive for individual translators

- PC offer easier integration with other IT equipment

- cost-saving

- easy post-editing (familiar word processors)

- functions of systems as for large organizations

  - i.e. include terminology management and use of translation database

- vendors either downsize client-server systems or upgrade cheaper (home) PC systems

- other users of such systems?:

  - 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
- any system type can be used (including those originally for mainframes and PCs)
    - online MT has all types of rule-based systems - and now also SMT

MT for interchange: examples

- 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 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 in the marketplace

- retail availability

  - most products only purchasable direct from manufacturer (online ordering)

- promotion by vendors can be misleading by 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

- need for 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

- low profits, slow quality improvement, few differences between rivals

    - not helped by free online services

- current system categories used by vendors - are they understood?

    - Enterprise systems, i.e. Client-server (intranet)

    - Workstations (TM systems)

    - Professional systems

    - Home systems

Free Online MT

- Limited lengths of text input (e.g. 100 words)

- No user dictionaries, but can be restricted to subject areas

- Free, vendors hope for sales of products

- 'Value-added' post-editing services (charged)

- Raised profile of MT, but at a cost...

Online MT

- For many users:
  - First use of MT
  - Unaware of PC products
  - Unaware of limitations
  - Tested with 'inappropriate' texts, back-translation
  - Howlers of 'first generation' MT
    - The spirit is willing, but the flesh is weak; Out of sight, out of mind
  - Often disappointed with results
- Quality improvements?
  - Desirable but not commercially attractive

Online and PC translation:
why so bad?

- old models (word for word, simple transformer architecture)

    - often single equivalents, no morphological analysis or target adjustment

- dictionaries too small, insufficient information, and difficult (or impossible) to update

- weak syntactic analysis/transfer

- poor disambiguation (little semantic information)

- general-purpose (not domain restricted)

- not designed for language/style of emails

- 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

Online MT usage

- No data on users: ages, background knowledge, types of texts, etc.

- Used by translators as rough drafts?

- Average length 20 words; 50% of submitted 'texts' just one or two words

- Very few webpages (unexpected!)

- Overall usage continues to grow exponentially

- The less the language knowledge of users , the more useful

Webpages

- Colloquial, culture-dependent language

- Texts in graphic format cannot be translated (very common in Japanese webpages)

- But website developers often recommend users to online MT services – do they know the dangers to their reputations?

- Website localisation for companies, etc. (IBM Websphere)

Electronic mail

- On PCs

  - initially mainly Japanese systems, now standard

- On intranets

  - basic facility of company ('enterprise') systems

- Commercial systems specifically for emails (e.g. Translution)

  - access online or via intranets

  - adapted to company terminology

Spoken Language

- PC systems with voice input/output, i.e. speech-text-text-speech

- Genuine speech translation

  - only research systems: ATR, CMU, JANUS, C-STAR, Verbmobil

- 'bottleneck' is speech recognition: limited range of speakers

- Domain restriction

  - telephone, hotel booking, business communication (ATR, Verbmobil)

  - military (DIPLOMAT, Phraselator)

  - medical (MedSLT)

  - tourism (ATR)

Rapid development of MT systems

- For languages poorly covered

- For languages of interest to 'intelligence' services

- Rule-based systems: not feasible because of:

  – Complex grammar, large dictionaries

  – Slow costly development

- Statistical MT

  – Based on large corpora

  – Little additional data required

  – open source systems and components

  – Commercialisation, e.g. Language Weaver

PC systems

- For assimilation – not concerned with quality

- Post-editing rare

- Drafts for dissemination

- despite low quality often trusted for dissemination without editing

  - often users ignorant of target language

- No data on usage! (used how often after purchase?)

- Sales continue to be high...

Hand-held devices

- 'Pocket translators' (special equipment)
  - Ectaco, Phraselator
  - Mechanised phrase books for military, tourism
  - Sold in large numbers (but how successful?)
- Mobile (PDA) devices
  - Text messaging
  - Only for common languages
  - Direct access to online MT services

MT for minorities

- May be widespread globally, but minor in particular country (e.g. Hindi in UK)

- Not commercially significant

- Poor resources (dictionaries, grammars)

- Not even word-processing (alphabets)

- Lack of bilingual corpora

  - even SMT rapid development not an option

- MT or other 'low-level' aids?

Embedding MT

- Information retrieval
  - Multilingual access to information
- Information extraction
  - data mining, text mining
- Intelligence
- Summarization
- Transliteration
- Question answering
- Authoring software

Subject-specific MT systems

Sublanguage systems (few successors of Meteo)

    e.g. police, drug enforcement, news

Commercial PC systems for medicine/patents (Japanese)

Availability of special glossaries (ranked for preference)

    e.g. medicine, law, Bible, business

Wide range available (but how many sold/used?)

Other applications

- Subtitles, broadcast transcripts, syndicated feeds

- chatrooms, social networking (Facebook, etc.)

    - Problems comparable to spoken language translation

- distance education, language teaching

- emergency services

- MT for the deaf

- Photocopier-MT; Scanner-MT ('pen' scanner)

- Camera-MT (menus, road signs?)

- MT for robots (spoken?)

- decipherment (back to MT origins!)

Summary: Current usages of MT

- Systems for dissemination (publication)
    - traditional use by corporations, agencies, localisation
    - rough drafts for authors
- Systems for assimilation
    - 'unedited' MT, intelligence/analysis, online MT
- Systems for interchange
    - electronic mail, correspondence, Web pages, tourism
- Language coverage
    - good (usable) for English, French, German, Spanish, Japanese, Chinese, Korean, Arabic
    - poor for: African, Indian, S.E.Asian, E.European, UK minorities

Future expectations: summary

- merging of MT and TM for enterprise dissemination systems
- Internet as major (chief) resource - not only SMT
- rapid development of systems (SMT)
- reuse of MT components (for closely related languages)
- improvements in quality (hybrid, multi-engine systems)
- minor (and minority) languages
    - i.e. languages not of major commercial or military interest
- special-purpose systems (domain and function) - also online
- spoken language MT, domain-specific only [not general-purpose]
- embedding of MT in other LT systems
- bilingual (multilingual) communication as much as translation

Evaluation: general considerations

- 'inherent' limitations of MT:

  - complexity (linguistic and computational)

  - constant creation of neologisms and new meanings for words and word sequences

  - cinstant creation of new idioms

  - cultural background

  - encyclopedic knowledge (Bar-Hillel and FAHQT)

- 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

Evaluation: criteria, methods

- 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)
- Usefulness
    - for particular tasks, for particular needs (assimilation, dissemination), for specific subjects
- Automatic evaluation
    - comparison of MT output and HT versions - emphasis on exact matches and close similarity of structures (statistical methods): tends to favour statistical MT systems
    - implies there are single 'best' translations

Machine translation (MT) and human translation (HT) in complementation

- HT for literature, and other 'culturally-sensitive' translation

- MT for technical, scientific, medical (etc.) texts which are culturally neutral

- HT (with translation aids) and human-aided MT for dissemination (publishable quality)

- MT for assimilation (rough 'gist')

- MT for real-time on-line translation (is this its 'real' niche?)

  - **the less the user knows of the source language, the more useful becomes fully automatic translation**

- HT for spoken language translation (except in highly restricted domains)