Machine Translation

MAI-HLT

MERITXELL GONZÀLEZ

Content

Historical background

MT problems/challenges

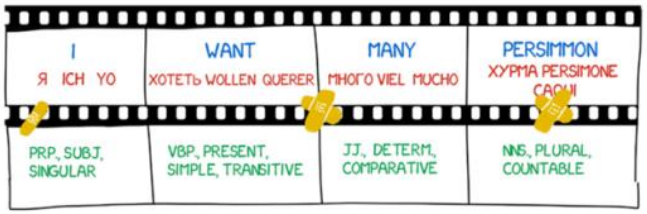
Classical MT

Statistical MT

MT Evaluation

1933: Artsrouni and Troyanskii

- Origins traceable to 17th century (Leibniz, Descartes)
- Pre-computer proposals (ENIAC, 1946)
- The first explicit proposals for ‘translating machines’. Patents presented by:
  - Georges Artsrouni (Georgian based in France)
  - Petr Petrovich Troyanskii (Russian)
- Both no more than mechanized bi- or multilingual dictionaries
- Although Troyanskii included codes (Esperanto based)



Using cards (4 languages), typewriter, and film camera
Considered useless by USSR.

1948: Richens and Booth

word for word dictionary + automatic grammatical analysis

French input with segmentation:

Il n'est pas étonn*ant de constat*er que les hormone*s de croissance ag*issent sur certain*es espèce*s, alors qu'elles sont in*opér*antes sur d'autre*s, si l'on song*e à la grand*e spécificité de ces substance*s

English output:

v not is not/step astonish v of establish v that/which? v hormone m of growth act m on certain m species m, then that/which? v not operate m on of other m if v one dream/consider z to v great v specificity of those substance m.

The stars (*) indicate automatic segmentations

v is untranslated French word, m multiple/plural, z unspecific

1949: Weaver

- Warren Weaver (Rockefeller Foundation) writes memorandum on MT
  - general strategies and long-term objectives of MT
- At the time, collaborating with Claude Shannon on **information theory**
- Discussed ideas with Andrew Booth of Birkbeck College (1946-47)
- Four main suggestions:
  - Disambiguation by examining adjacent words
  - The logical basis of language
  - Use of cryptographic methods: decoding of source text
  - Universal language

When I look at an article in Russian, I say to myself: This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.

Warren Weaver

1952: first conference

- Bar-Hillel appointed at MIT in May 1951, surveyed MT
- Convened first MT conference at MIT
- Topics covered:
  - Pre-editing, post-editing (Reifler)
  - Controlled language (Dodd's Model English)
  - Domain restriction (Oswald's microglossaries)
  - Syntactic analysis (Bar-Hillel's categorial grammar)
  - Computer hardware, programming
  - Funding

1954: first public demonstration

- Leon Dostert determined to show 'technical feasibility' of MT
- Collaboration of Georgetown University and IBM
- Public demonstration in New York, 7th January 1954 of Russian-English system
- Linguistic foundations by Paul Garvin of Georgetown U.
  - 250 words, 6 grammar rules
- Programming by Peter Sheridan of IBM
- Reported widely, worldwide interest

BUT carefully selected examples, excluding ambiguity

- Beginning of government funding - in both US and Soviet Union

IBM 701 at New York headquarters

1955: beginnings of MT in Soviet Union

- 1953: death of Stalin, March 1953 opened access to science of the west: cybernetics, structural linguistics, and computers
- 1954: news of Georgetown-IBM demonstration
- 1955: first attempts using BESM (Bolshaya Elektronno-Schetnaya Mashina / Large Electronic Computing Machine)
- 1956: foundation of groups: Inst Precision Mechanics, Steklov Mathematical Institute, Inst Linguistics, Leningrad University

Cold War

English vs. Russian

- Spying
- The conquest of the Moon
- ……

Science and Technology research for military reasons.

Dartmouth Summer Research Project on Artificial Intelligence

- 1956, summer workshop
- Organised by John McCarthy
- The beginning of Artificial Intelligence

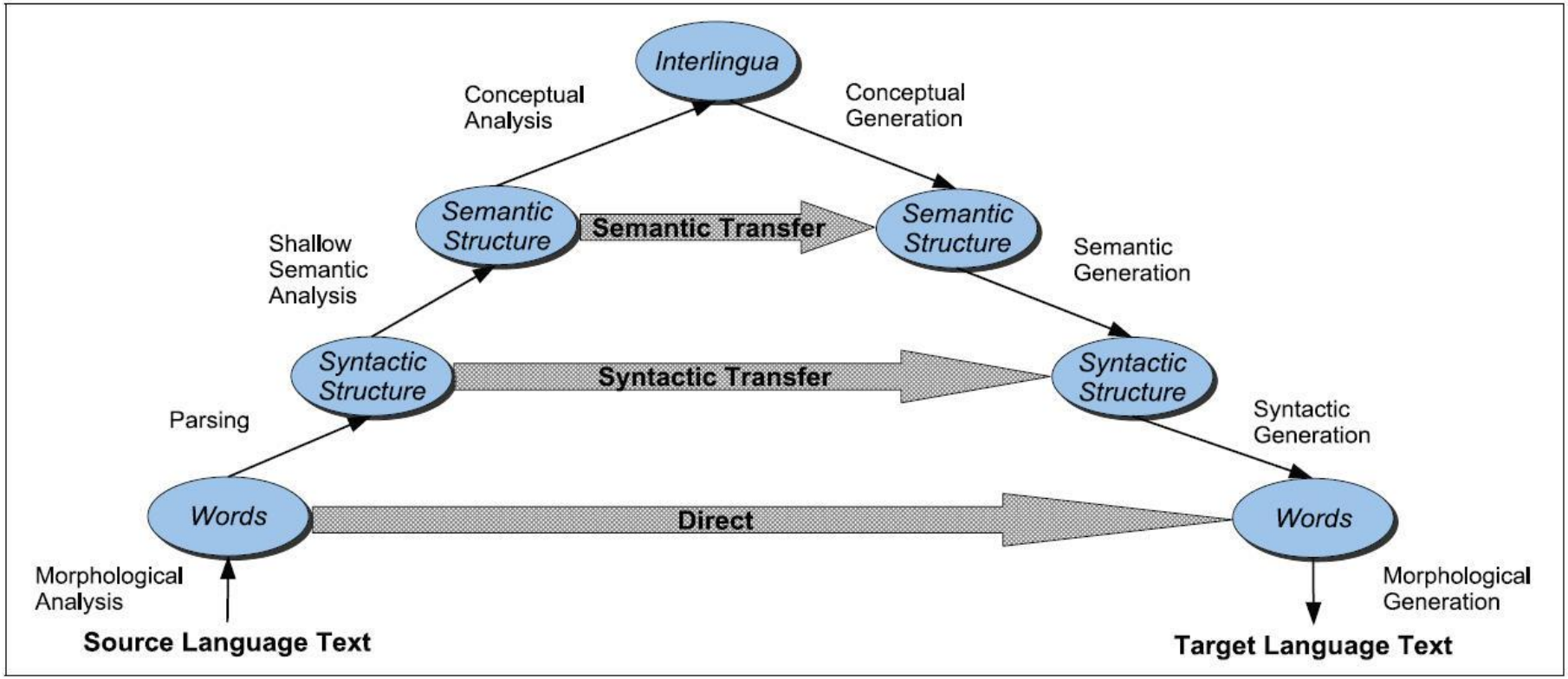
- One of the expectations for the next 10 years:

Fully machine translation

Main groups in 1960s

- University of Washington, IBM (**direct translation**) [Reifler, King]
- Harvard (**massive dictionary**, predictive syntax) [Oettinger]
- Massachusetts Institute of Technology (**syntactic transfer**) [Yngve]
- Georgetown (multiple levels of analysis) [Dostert, Zarechnak]
- Cambridge Language Research Unit (**interlingua**, lattices) [Masterman]
- Birkbeck College London [Booth]
- National Physical Laboratory, Teddington
- Milan University (**interlingua**) [Ceccato]
- Institute of Precision Mechanics and Computer Technology [Panov]
- Leningrad University (**interlingua**) [Andreev]
- Leningrad University (**statistical**) [Piotrowski]
- Institute of Linguistics, Moscow (**syntax**, **semantics**) [Kulagina, Mel'chuk]

Vauquois Triangle (1968)



Georgetown University

1954 founded by Leon Dostert

Largest MT group in USA (over 20 researchers), funded by CIA

Variety of methods for Russian-English system examined: code-matching, syntactic analysis (Paul Garvin), sentence-by-sentence (Antony Brown), general analysis (Michael Zarechnak)

GAT eventually adopted **multiple levels of analysis**: morphological, idiom identification, syntagmatic analysis (agreements, government), syntactic analysis (subject-predicates)

Implemented on SERNA (Peter Toma) **=> Systran !**

1961 demonstrated at Pentagon;

1963 installed at Euratom (Ispra, Italy);

1964 installed at Oak Ridge National Laboratory (USA)

Massachusetts Institute of Technology

1951: appointed Bar-Hillel, convened first MT conference

1953-1965:directed by **Victor Yngve**

Fundamental research, not “short-cut methods” (other researchers included Chomsky)

Linguistic analysis: German-English

Programming language (COMIT): first non-numerical, string-processor

Syntactic transfer (SL tree representation to TL trees)

Sentence production (**first generator system**)

Other languages: Finnish, Arabic

1964: reached “semantic barrier”

Yngve editor of “Mechanical Translation”, co-founder of Association of Computational Linguistics

Harvard University

1954 founded by Anthony Oettinger

Large-scale Russian-English dictionary, producing word-for-word translations and a research tool

1959 **Syntactic analysis**: based on a predictive analyzer developed at National Bureau of Standards by Ida Rhodes

1963-1966: Further developments at Harvard: pushdown store, multiple-path predictive analyzer (Susumu Kuno, Warren Plath)

Birkbeck College

1948: Booth's collaboration with Richard Richens on morphology

1953: first research by Booth on **dictionary**:

first test of procedure on APE(X)C built at Birkbeck

1955: fast dictionary lookup, 'binary division'

1955: funding by Nuffield Foundation for **French-English** system

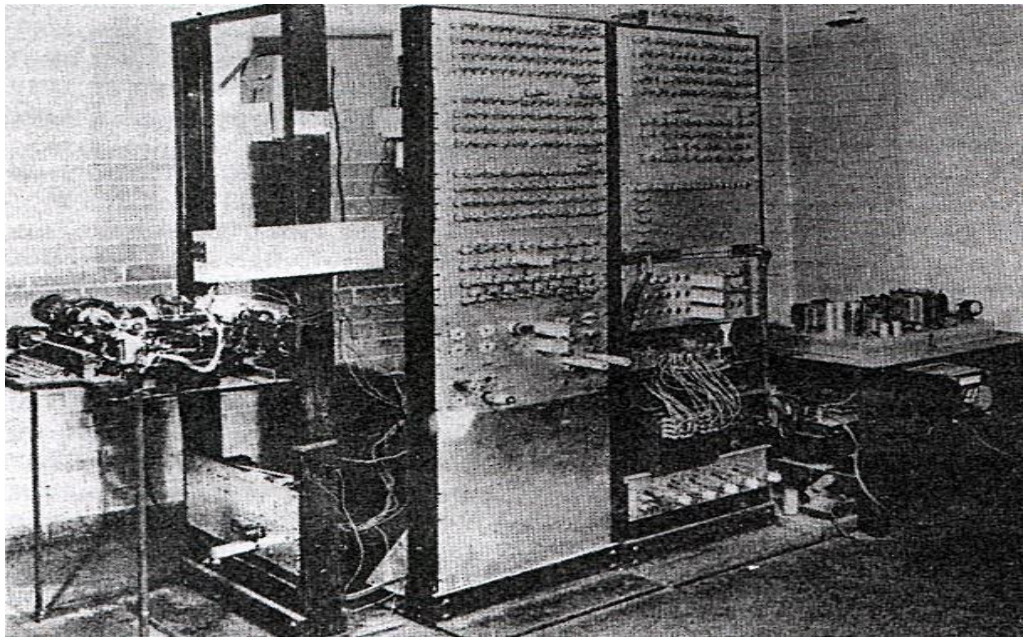
tested on APE(X)C 1958

1958: **syntactic research on German** (not programmed because of computer storage limitations)

1962: Departure for University of Saskatchewan

1965: Trial for Canadian MT system

Booth's APE(X)C about 1952



APE(X)C: All Purpose Electronic (agency identification) Computer

Cambridge Language Research Unit

UK. Founded 1956. Director Margaret Masterman

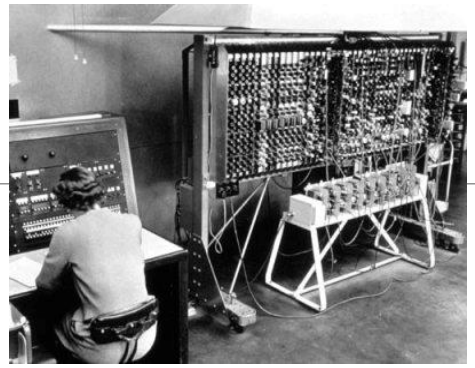
Interlingua: continuation of Richens ideas

- 51 elements (semantic classifiers):
  ASK BANG BE BEAST CAN CAUSE CHANGE COUNT DO …
- Influence on Artificial Intelligence (e.g. Schank)

Thesaurus approach: lexical items under 'heads', multiple meanings under different heads, e.g. 'plant' under place, vegetable, agriculture, trick, tool

Pidgin translation: interim results of thesaurus analysis

National Physical Laboratory



UK. Leader: John McDaniel

1959 began on ACE computer

Russian-English scientific texts

Use of Harvard Russian-English dictionary (18,000 words) – not operational until 1963

'international' words transliterated

syntactic analysis: noun groups, verb groups, however

results little better than word-for-word translations

evaluation in 1966: “slightly less than good”

1961 organized international conference

1966: The ALPAC report

- Set up by NSF for US sponsors of MT research

Concluded: No effective MT despite massive funding, and none in prospect

- Poor quality output
- Criticised at time for short sightedness
- Brought to end US funding for many years
- Affected funding elsewhere

ALPAC: Automatic Language Processing Advisory Committee (US 7 experts)

Consequences of ALPAC

- Identification of user needs:
  - Dissemination vs. assimilation
- Recognition that 'perfectionism' had neglected:
  - Operational factors and requirements
  - Expertise of translators
  - Machine aids for translators
- **Henceforth three strands of research**
  - Translation tools and aids
  - Operational tools: post-editing, controlled languages, domain-specific systems
  - New systems, new approaches, new methods

From 1967 to 1978

Continuation of research in US (Texas, Wayne State), Soviet Union, UK, Canada, France

1970: Systran (Peter Toma) installed at US Air Force

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

1978: Xerox Corporation uses Systran with controlled language (Caterpillar English)

1978: transfer-based MT

Beginning of research on :

- **ARIANE** system at Grenoble University (France) [Vauquois, Boitet] – Russian-French, English-French, German-French
- **Eurotra** system funded by European Commission
- Logos (USA) [Scott] - German-English
- Mu system, Kyoto University (Japan) [Nagao] - Japanese/English
- METAL, University of Texas (USA) [Lehmann] - German-English
  Meaning-Text Model (Moscow) [Mel'chuk]
- ETAP (Moscow) [Apres'yan]

ARIANE

Founded 1960 in Paris and Grenoble as CETA (Centre d'Etudes de la Traduction Automatique)

Interlingua model 1960-1970. Director: Bernard Vauquois

Failures: reduction to interlingua representations, destruction of useful Source Language (SL) information

Strict separation of linguistic data and programming

stages: morphological analysis, syntactic analysis, intermediate structure, lexical transfer, SL structure with TL units, structural transfer, syntactic generation, morphological generation

One of most influential MT system

Eurotra

Funded by European Commission 1979-1992

Intended to replace Systran (adopted by EC in 1976)

80 researchers in eight member states (UK, France, Germany, Belgium, Denmark, Netherlands, Italy, Spain)

Multilingual transfer design, intended to be operational as soon as possible

Required precise specification of analysis, transfer and synthesis programs; ambiguities dealt with by monolingual analysis programs; transfer not interlingual by 'Euro-versals'

Excellent and influential linguistic research; but neglected dictionary construction, industrial prototype not delivered

1981: MT for personal computers

Previously all MT systems for mainframe computers

ALPS (computer-assisted translation system)

Weidner Communications / Bravis (for Japanese)

Subsequently (in 1980s and 1990s):

- ESI, Instant Spanish, LogoMedia, Personal Translator, PeTra, PROMT, Systran

Many Japanese systems

- e.g. Crossroad, LogoVista

1982: AI and interlinguas

- Beginning of 'Fifth Generation' (AI) program in Japan; influence on US research
- Research on interlingua systems
  - At Philips (Rosetta) – implementing Montague grammar
  - At Utrecht (DLT) – modified Esperanto, bilingual knowledge bank
- Research on knowledge-based systems
  - At Colgate University, Carnegie-Mellon University, New Mexico State University (PANGLOSS)

1986: speech translation

- ATR in Japan, JANUS at Carnegie-Mellon, Verbmobil (at various German universities)
- speech recognition, speech synthesis
- discourse semantics, ‘ill-formed’ utterances
- ellipsis, use of stress, intonation, modality markers
- colloquial usage not yet investigated sufficiently (even in linguistics)
- restricted fields (telephone booking of hotels and conferences)
- Still continuing

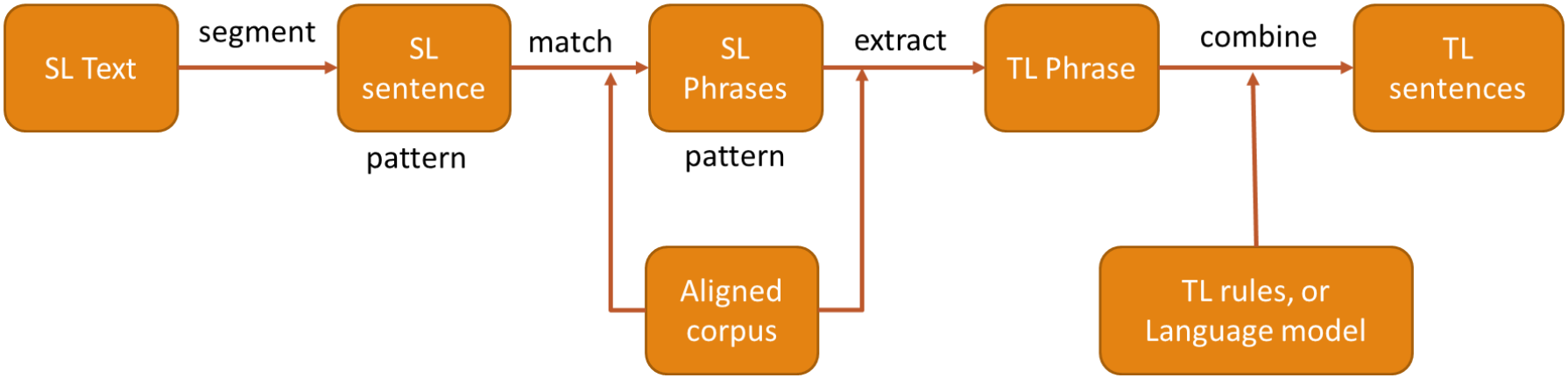
1988: corpus-based MT

- Availability of large bilingual corpora
- Beginning of Example-based MT research, 1988-89
  - First proposed in 1981 by Makoto Nagao
- First article on Statistical MT, 1988 (research at IBM, Candide system)
  - Revival of Warren Weaver's idea ('decoding' SL as TL)

Example-based MT model

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
- in essence, continuation of 'transfer' model, with statistical methods

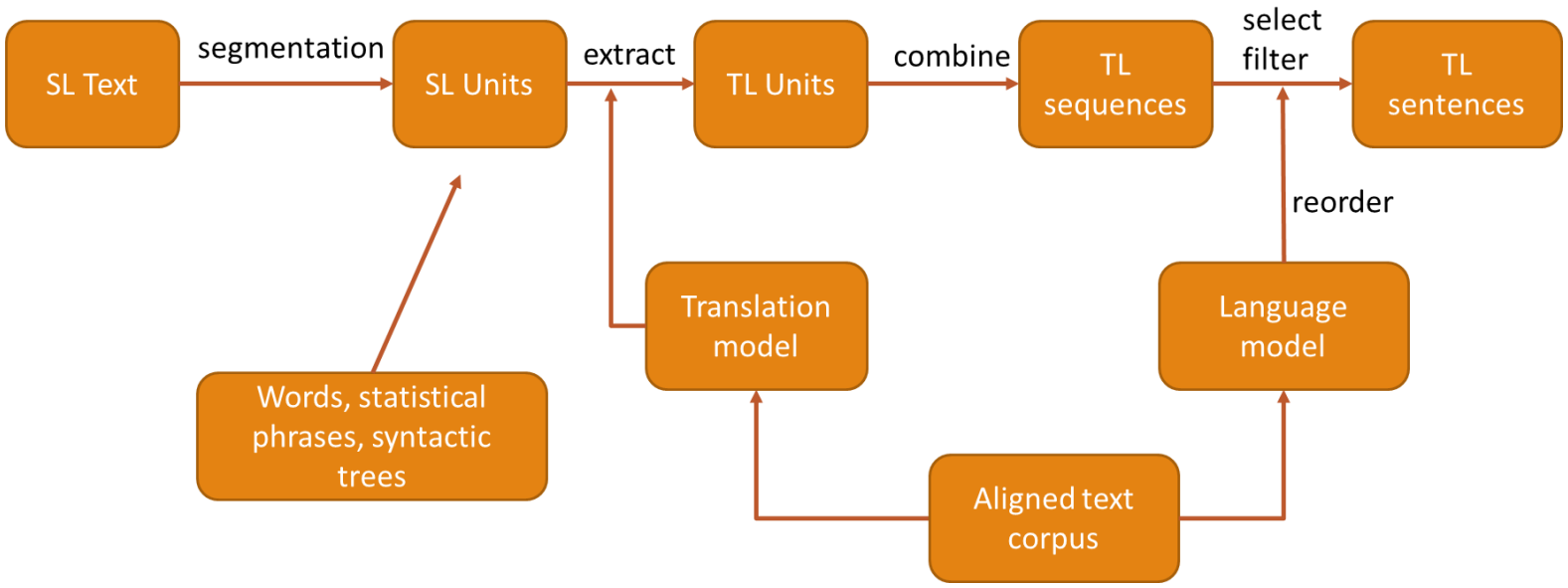


Statistics-based MT model

TL words/phrases are chosen as those most likely to correspond with the SL words/phrases in specific context (**probabilities, frequencies**)

TL words/phrases are combined in ways most appropriate for the TL in a specific context/domain and style/register etc. (**maximizing probabilities**)

- minimal use of linguistic information (morphology, syntax)



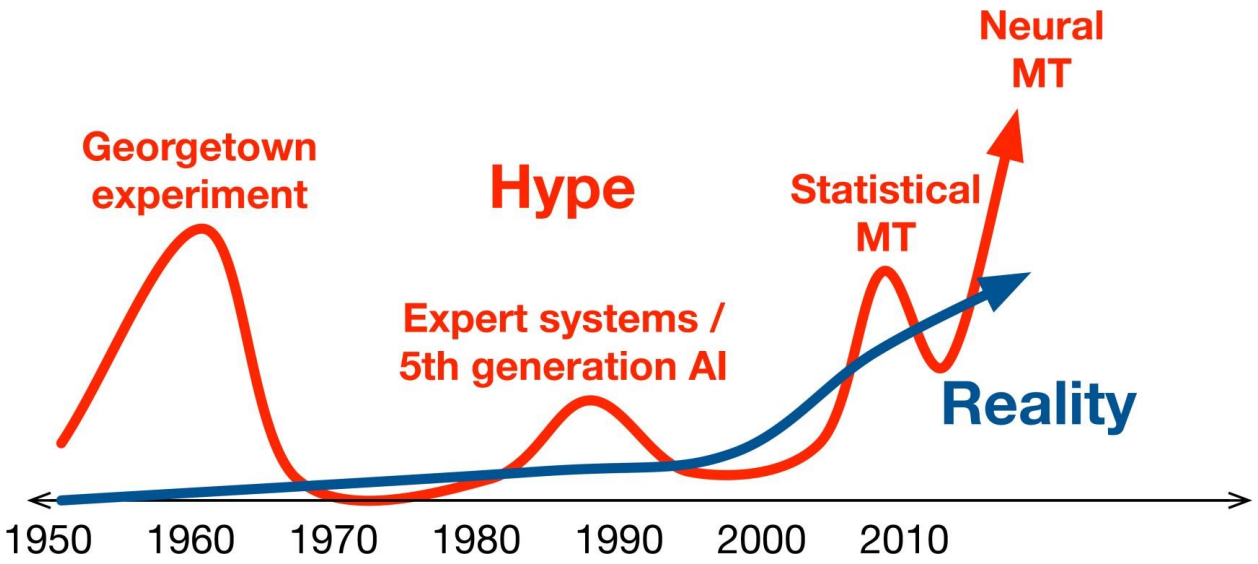
1993: translation memories

- Previous tools: dictionaries, termbanks, concordances
- in 1993 launch of first commercial system: **Trados**
  - later followed by Transit, Déjà Vu, ProMemoria, WordFast, ...
- using aligned bilingual corpora (of human translation), searchable by words and phrases
- Attractiveness for translators:
  - Components and facilities controlled by users
  - Terminology management
  - Facilities for building dictionaries (e.g. from Internet)
  - Compatible with authoring and publishing systems

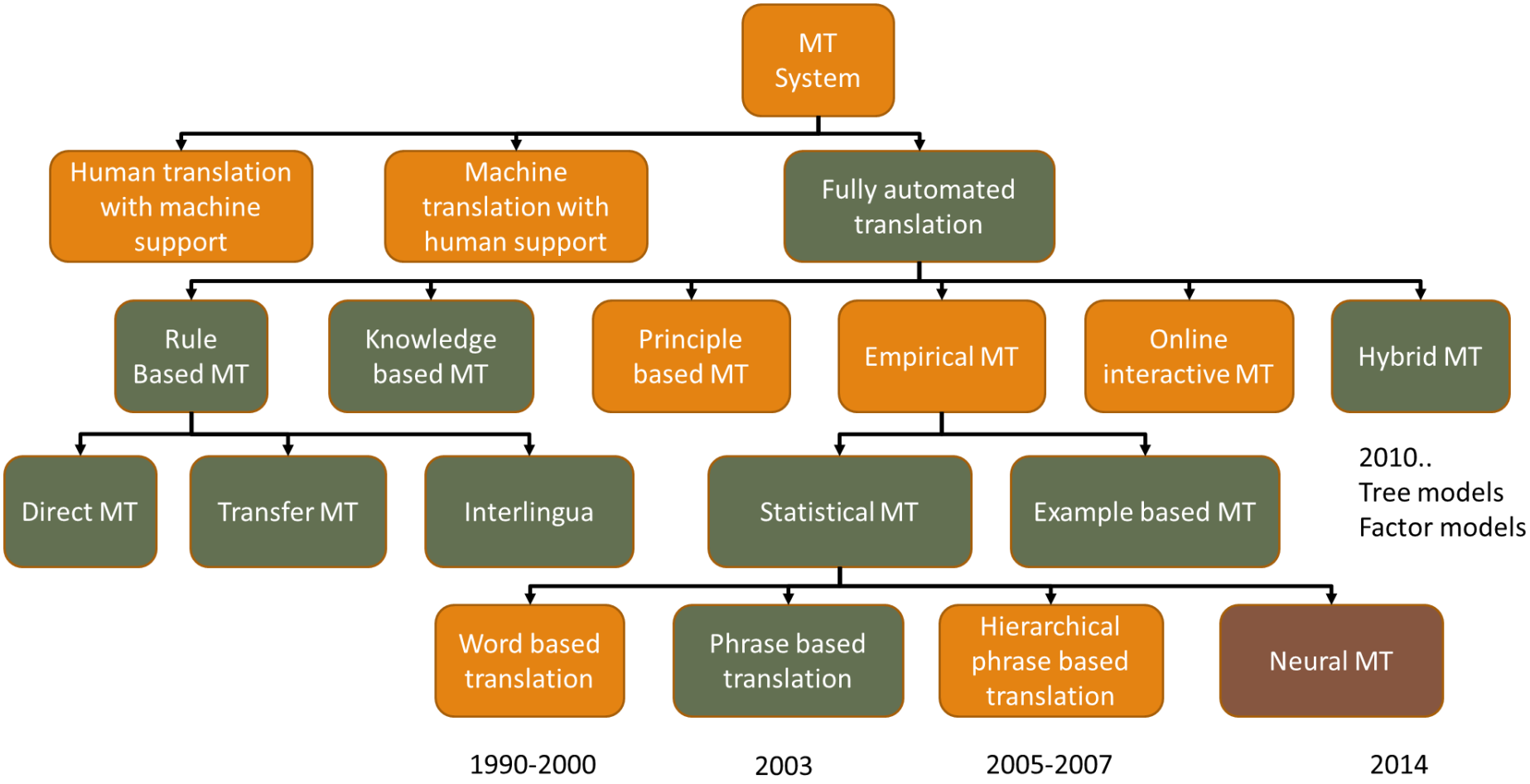
Since 2004: open source toolkits

- GIZA ++: tool for alignment in SMT
- MOSES: platform for building SMT systems
- Joshua: decoder for syntax-based (hierarchical) SMT
- Apertium: platform for building rule-based MT
- META-SHARE: data for EU projects
- LetsMT: cloud-based resource for supporting MT research
- Paracrawl/Bitextor: platform for creating parallel corpora by crawling multilingual websites

MT History: Hype vs. Reality



Approaches of MT



Content

Historical background

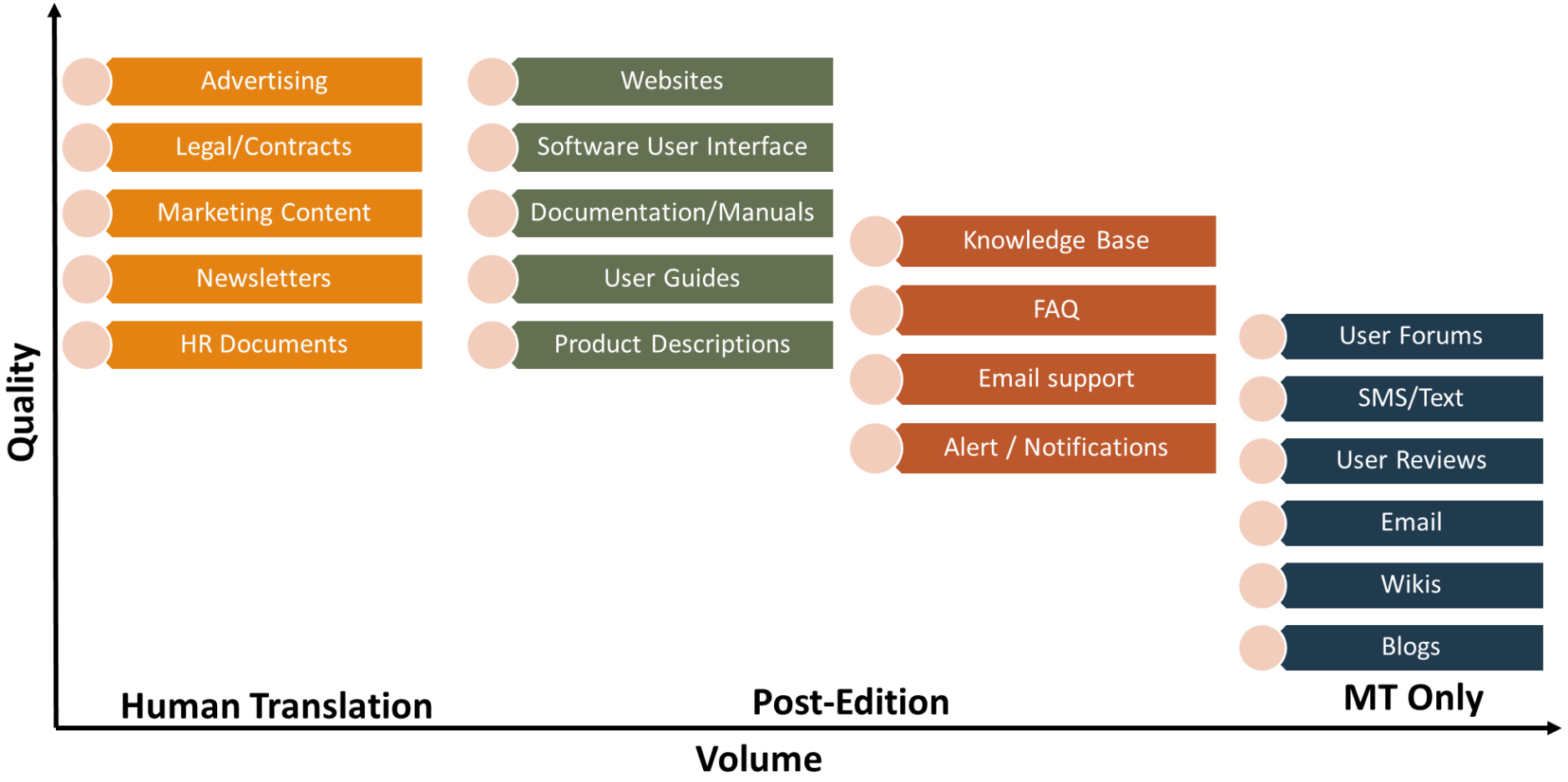
MT problems/challenges

Classical MT

Statistical MT

MT Evaluation

What to translate




information
ANA
In flight service
おむつ交換台
Baby diaper exchange


PLEASE REMOVE THE ROOM
ПРОСЬБА УБРАТЬ НОМЕР
БРЯНСК
PLEASE DO NOT WORRY
ПРОСЬБА НЕ БЕСПОКОИТЬ
БРЯНСК


EXTREME CAUTION
WATCH FOR ICE
EL RELOJ EXTREMO
DEL CUIDADO
PARA EL HIELO


Exit
Only
STARBUCKS
COFFEE
Éxito
Aquí


TAKE LUGGAGE
OF FOREIGNER
NO CHARGE
提行李處

What is MT good (enough) for?

- **Assimilation**: reader initiates translation, wants to know content
  - User is tolerant of inferior quality
  - Focus of majority of research
- **Communication**: participants in conversation don't speak same language
  - Users can ask questions when something is unclear
  - Chat room translations, hand-held devices
  - Often combined with speech recognition
- **Dissemination**: publisher wants to make content available in other languages
  - High quality required
  - Almost exclusively done by human translators

MT and 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
- MT for integrating translation with other LT tasks

Translation difficulties

- Typology
  - cross-linguistic similarities and differences
- Idiosyncratic differences
  - language or language pair
- Lexical divergences

Typology

- Morphology
  - Number of morphemes per word
    - isolating languages (ex. Vietnamese, one-one)
    - Polysynthetic languages (ex. Siberian Yupik, one-many)
  - Segmentable?
    - agglutinative languages (Turkish, German, clear boundaries)
    - fusion languages (ex. Russian, Italian, French, Spanish, .. one affix-many morphemes)

Typology

- Syntax
  - Order of Subject, Verb, and Object
    - SVO (English, Spanish)
    - SOV (Hindi, Japanese)
    - VSO (Irish, Arabic)
    - Similar word-order type -> other similarities (prepositions)

English: *He adores listening to music*

Japanese: *kare ha ongaku wo kiku no ga daisuki desu*

he music to listening adores

Typology

- Syntax
  - Argument structure and linking
    - head-marking (English, the man's house)
    - dependent-marking (Hungarian, the man house-his)
  - Marking the direction of motion
    - verb-framed (Spanish, e.g. subir, bajar, entrar)
    - satellite-framed (English, e.g. go up, go down, go inside)
  - Omission of some elements or explicit references

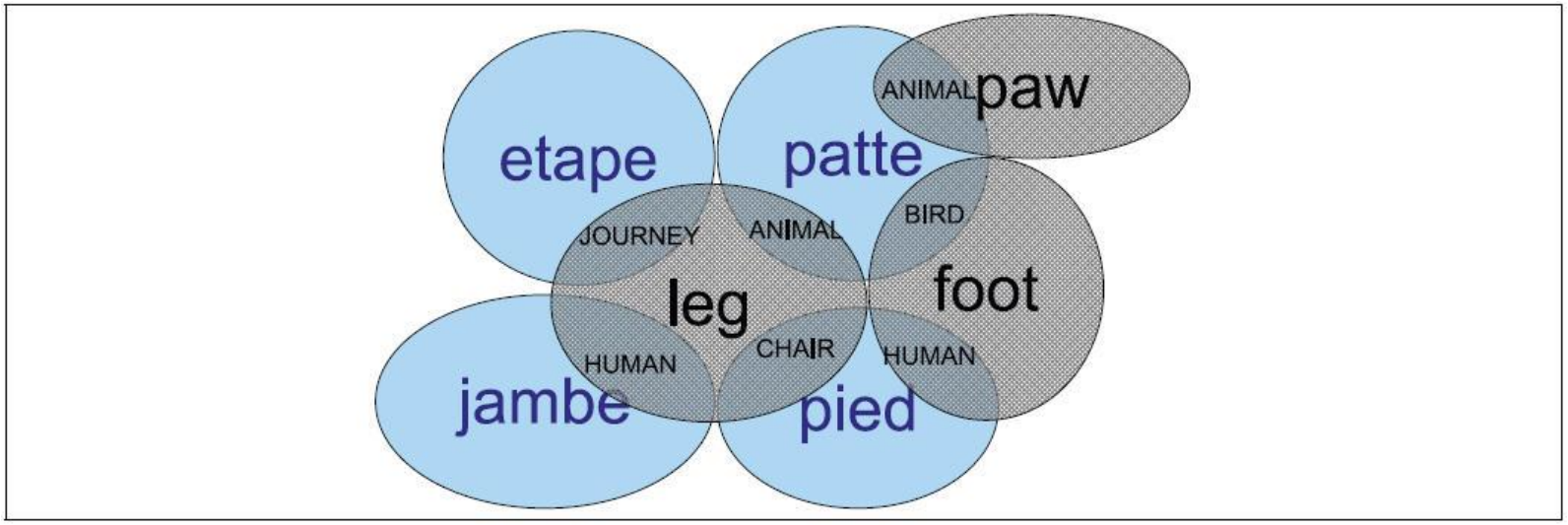
    I read a book // [Yo] Leo un libro
  - Adjectives and nouns
    - Spanish (casa azul)
    - French (maison bleue)
    - English (blue house)

Idiosyncratic differences

- Language specific constructions
- Dates
  - Different formats
  - Different calendars

Lexical divergences

- Polisemy, homonymy
  - English (bass) / Spanish (lubina or bajo)
  - English (know) / French (savoir or connaître)
  - Required disambiguation



English vs French, Hutchins and Somers (1992)

Lexical divergences

- Different granularity
  - English (wall) / German (Wand-inside- or Mauer-outside-)
- Different part of speech
  - ex. reflexive pronoun vs possessive determiner
- Differences on grammatical constraints
  - e.g. gender in adjectives, gender in pronouns, …