Linguistic, computational and operational architectures of MT systems

Christian BOITET
GETALP-LIG-UJF, Grenoble, France
Christian.Boitet@imag.fr
Introductive words

• Thanks to
  NII for this kind invitation
  Pr Kitamoto for welcoming me in his group this April
  ... & for allowing PhD interns from our group to work on DSR
  all of you for coming

• Past
  scientific cooperation on MT with Japan since 1965 (Vauquois)
  Pr Nagao came 1 year, Pr Tsujii 9 months, Pr Okada 2 months...
  1 sabbatical, post-docs, interns, cooperation stays from GETA(LP)
    at KDD, ATR, NII, NiCT
  Dr Tomokiyo (linguist) invited researcher at GETA(LP) since 1996

• Present
  visiting professor at NII for 1 month
  advising 3 UJF PhD students staying in Japan (2 at NII, 1 at Kyodai)

• Future: purposes of this visit are to
  find convergence points in our research
  discuss possibilities of research cooperation (JSPS-CNRS Sakura?)
  same, concerning student exchanges (internships, post-docs)
Outline of the lectures

- Lecture 1: Linguistic, computational and operational architectures of MT systems
  Apr. 14, 2pm - 4pm

- Lecture 2: Linguistic architectures of MT systems
  Apr. 16, 2pm - 4pm

- Lecture 3: Computational architectures of MT systems
  Apr. 21, 2pm – 4pm

- Lecture 4: Engineering of MT and CAT systems
  (CAT = MT + TA, TA = translation aids)
  Apr. 23, 2pm – 4pm

- Lecture 5: Evaluation of MT and CAT systems
  for various operational architectures
  Apr. 28, 2pm – 4pm

- Lecture 6: Corpora for hybrid MT/CAT systems
  Apr. 30, 2pm – 4pm
Linguistic, computational and operational architectures of MT systems (Lecture 1)

• Fundamental facts about HT & MT
  ⇒ Nature of MT
  • scientific technology, not a science
  • variety of goals & associated evaluation methods
  • misconceptions about MT

• Independence of linguistic and computational architectures of MT systems
  • Linguistic: objects (representations)
  • Computational: processes (RBMT - SMT - EBMT - hybrid...)
  • Types & size of resources w.r.t. MT architectures

• Operational architectures

• Current evolution, recap & first conclusions
I. Fundamental facts about MT

• Varieties, difficulties, evaluation in HT (Human Translation)
  Translation is multiple!
  Translation is difficult!
  Evaluation of HT is also multiple and difficult!

• The place of MT in a translation process
  More or less automation: MT#1 >[ MT#2, HT >...] typology
  • Human Translation workflow
  • Automated Translation workflow
  What can be automated outside of the "MT module"
  What can be done by humans to help the MT module

• The CxAxQ theorem
  and ways of measuring Coverage, Automaticity, & Quality

• Dream & reality: misconceptions & achievements
Human Translation is multiple!

- Translating & interpreting are TRADES
  
speak \((P1, (L1, L2))\) \(\rightarrow\) able\((P1, \text{translate } (L1, L2))\)

- Variety of competences & tasks

```
<table>
<thead>
<tr>
<th>Human translator</th>
</tr>
</thead>
<tbody>
<tr>
<td>professional</td>
</tr>
<tr>
<td>translator or</td>
</tr>
<tr>
<td>interpreter</td>
</tr>
<tr>
<td>text</td>
</tr>
<tr>
<td>literature</td>
</tr>
<tr>
<td>news</td>
</tr>
<tr>
<td>technical manuals</td>
</tr>
<tr>
<td>program messages</td>
</tr>
<tr>
<td>speech</td>
</tr>
<tr>
<td>simultaneous,</td>
</tr>
<tr>
<td>liaison,</td>
</tr>
<tr>
<td>consecutive</td>
</tr>
<tr>
<td>interpreting</td>
</tr>
<tr>
<td>text</td>
</tr>
<tr>
<td>business letters</td>
</tr>
<tr>
<td>manuals (!)</td>
</tr>
<tr>
<td>speech</td>
</tr>
<tr>
<td>telephone</td>
</tr>
<tr>
<td>conversations</td>
</tr>
</tbody>
</table>
```
Human Translation is difficult!

- Sources of difficulties

  - language pairs
    - underspecification
    - L_S vs. L_T
    - ex: countability, determination
  - language in general
    - ambiguity,
    - fuzziness,
    - (probable)
    - unaxiomatizibility
  - different views of the world
  - work organization
    - division of document
    - term. consistency
    - tech. problems
    - (formats, tools)
  - language sets
    - multilingual dictionaries?
    - ==> databases
    - some double translations necessary
    - (Europe, India)
Evaluation of HT is multiple and difficult

- Grading by professors, no MCQ

```
  text
  /   \
 content form
  \
  task

  speech
  /   \
 content (fidelity) form
  \

  fidelity
  terminology
  language
  parallelism
  phraseology
  delay
  obligation of
  price

  themes (what was it about?)
  rhemes (what was said about it?)
  phemes (with which modalities?)
  fluency

  well-formedness
  consistency
  results
  means
```
What can be automated?

- The term MT is ambiguous
  - In French: MT#1 = TAO, MT#2 = TA

Diagram:
- MT#1 = Automated Translation
- MT#2 = Automatic Translation
- HT = Human Translation
- FAMT = Fully Automatic
- HAMT = Human Aided Translation
- DBMT = Dialogue-Based MT
- IMT = Interactive Machine Translation
- MAHT = Machine-Aided
- FHT = Fully Human
Professional Human Translation Workflow

- Only 50% time spent on "translation proper"
- Documents parts, translation units, segments
Automated Translation Workflow

- MT is only one module
- Human interaction possible, but translation proper is done by the system
Automation outside of MT module

• **Writing**
  - control
    - language appropriateness
      1. lexicon (unknowns -L_S, +L_S -L_T)
      2. grammar (anaphor, ellipsis, extrapolation, coordination, comb)
  - complexity
    1. simplification (split long sentences, use simple constructions)
    2. de-underspecification (fill ellipsis, 0-pronouns)

**pre-edition**
- mark segments, special chunks, groups (attachment)
- mark POS, Word sense  saw_V#eye

• **Preparation**
  - Complete/revise segmentation

• **Use of results**
  - Modify choices
    - rescoring, filters, tournament
  - During post-edition
    - adapt to user choices
    - integrate new equivalents & recompute translation
Human help to MT module

- **Typical pipeline process**
  - intermediate *representations* (data) and *phases* (processes)

  ![Diagram showing a typical pipeline process with steps and phases]

- **Steps = 1 or more phases** ($P_i$)
  - $\leftarrow$ Analysis $\rightarrow \leftarrow$ Transfer $\rightarrow \leftarrow$ Generation $\rightarrow \leftarrow$
  - $\leftarrow$ Analysis $\rightarrow \leftarrow$ Transfer+Generation $\rightarrow \leftarrow$
  - $\leftarrow$ Enconversion $\rightarrow \leftarrow$ Deconversion $\rightarrow \leftarrow$

  (2 or 3 lexical spaces)

- **Interaction possible**
  - during the phases
  - on multiple/factorizing IRs

  ITS-1, TAO, TransActive, ITS-2
  KBMT/KANT, JETS, Systran-5...
The CxAxQ MT theorem

- Coverage x Automaticity x Quality \(<\, 100\%

Quality here means linguistic quality (judged by translators)
That limitation is in the nature of the problem

\[ \Rightarrow \] FAHQTMT for all domains is impossible

- But 2 factors can be \( \approx 100\% \) if compromising on 3rd

  C x A \( \approx 100\% \) web translation
  A x Q \( \approx 100\% \) METEO, ALTFLASH
  C x Q \( \approx 100\% \) DBMT (JETS, CATALYST, LIDIA)
What are C, A, Q & how to measure them?

• **Automaticity** (MT module only)
  \[ A = 1 - \frac{T(\text{human\_interaction})}{T(\text{human\_first\_draft})} \]
  \[ A = 83.3\% \text{ if 10mn interaction, 1h 1st draft} \]

• **Quality (wrt HT)**
  \[ Q = (100 - 2 \times T(\text{post\_edition\_MT}))\% \]
  \[ Q = 40\% \text{ if 30 mn PE} \]
  \[ Q = 60\% \text{ if 20 mn PE (= Q(HT))} \]
  \[ Q = 90\% \text{ if 5 mn PE (cf. LucySoftware/METAL Spain) [ex Comprendium]} \]

• **Coverage**
  much more difficult to measure. What, how?
  • Words/compounds, Terms, Collocations
  • Grammatical constructions
  • For all:
    1. frequency & importance (FI)
    2. General Meanings (GM)
    3. Specific Meanings (SM)
Coverage computation

Estimate coefficients, then use normalized sums

<table>
<thead>
<tr>
<th></th>
<th>FI frequency + importance</th>
<th>GM general meanings ($\gamma$)</th>
<th>SM specific meanings ($\sigma$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$W_i$ words &amp; compounds</td>
<td>FI_W_i</td>
<td>GMW_i_1...</td>
<td>SMW_i_1...</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GMW_i_g_i</td>
<td>SMW_i_s_i</td>
</tr>
<tr>
<td>$T_i$ terms</td>
<td>FI_T_i</td>
<td>GMT_i_1...</td>
<td>SMT_i_1...</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GMT_i_g_i</td>
<td>SMT_i_s_i</td>
</tr>
<tr>
<td>$C_i$ collocations</td>
<td>FI_C_i</td>
<td>GMC_i_1...</td>
<td>SMC_i_1...</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GMC_i_g_i</td>
<td>SMC_i_s_i</td>
</tr>
<tr>
<td>$S_i$ standard constructions</td>
<td>FI_S_i</td>
<td>GMS_i_1...</td>
<td>SMS_i_1...</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GMS_i_g_i</td>
<td>SMS_i_s_i</td>
</tr>
<tr>
<td>$P_i$ particular constructions</td>
<td>FI_P_i</td>
<td>GMP_i_1...</td>
<td>SMP_i_1...</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GMP_i_g_i</td>
<td>SMP_i_s_i</td>
</tr>
</tbody>
</table>
Dream & reality: misconceptions & achievements

• The dream of FAHQMT *with large coverage* (from 1951!)
  Fully Automatic High Quality Machine Translation
• Proof of impossibility by Bar-Hillel (1960)
  ... impossibility of an *angelic* goal, of course
• Concrete operations, *useful* systems (since 1967!)
  1972, Systran, Ispra, Euratom: 20% ling. quality, 80% usefulness
• FAHQMT *on sublanguages* does exist
  CETA 1965-70, METEO 1977—, GETA-Ariane 78—, METAL,
  ENGSPAN/SPANAM (PAHOMTS), ALTFLASH... (many)
• You don't see them because they are *tailored!*
• Rough MT exists (web translators...)
  and SMT has introduced Machine Learning in MT
  enlarging the Coverage (and # languages), but not the Quality
Next...

• The various tasks of MT and their difficulty

• Independence of linguistic and computational architectures of MT systems
  Linguistic: objects (representations)
  Computational: processes (RBMT - SMT - EBMT - hybrid...)
  Types & size of resources w.r.t. MT architectures

• Example of a heterogeneous, IL-based system: UNL
  Introduction to UNL hypergraphs
  An experiment to evaluate UNL real potential
  Current developments & perspectives
Main goals of MT & associated measures

- 4 main "translational situations", in increasing order of difficulty of automation:
  1. production of high-quality (HQ) translation by bilinguals
     dissemination by bilinguals ==> human time needed
  2. understanding text or speech in an unknown language
     assimilation ==> adequacy / # buying acts (e-commerce)
  3. production of HQ translation from an unknown language
     HQ assimilation by monolinguals ==> human time needed
  4. production of HQ translation into unknown languages
     HQ dissemination by monolinguals ==> human time needed
Misconceptions about MT

- "Pivot MT" ⇒ "Rule-Based MT"
- "Transfer MT" + N languages ⇒ N(N-1) transfers
- "Rule-Based MT" ⇒ high cost
- "Statistical MT" ⇒ low cost
- "Statistical MT" ⇒ majority of operational systems
- "Statistical MT" ⇔ "Rule-Based MT"
- BLEU measures the quality of translations
- Linguistic quality of MT outputs has increased with SMT
- Adequacy should be measured by a positive number
- "MT with interlingual pivot (IL) cannot work and scale up":
  
  But,
  - ATLAS-II (Fujitsu) has been the best for J↔E since 20 years
  - It has more than 7.5M dictionary entries (v.14, Dec. 2008)
To sum up about MT, MT systems, variety

- MT is not a science, but a scientific technology
deep and hard problems come from the confrontation with reality
- MT without concrete goals is not MT
  —> automating various translation tasks for various users
  USSR in the 70's: MT without machines and without translations!
- There is not ONE unique goal / task for MT
  analogy with transportation
  translational situations          transport situations
  web surfing                       bicycle
  diffusion by bilinguals (general) car
  technical translation            train
  restricted translation (ALTFLASH) automatic train
  (difficulties are not parallel!)
- The SAME system can be judged very good/bad
  + if used when a user wants (a help) to understand a content
  - if "pushed" on the user by the originator of a content
3 types of MT architectures

linguistic
computational
operational
## Linguistic vs. computational architectures

<table>
<thead>
<tr>
<th>Linguistic architecture</th>
<th>Computational architecture</th>
</tr>
</thead>
<tbody>
<tr>
<td>objects</td>
<td>automatic processes</td>
</tr>
<tr>
<td>see Vauquois' triangle</td>
<td>human interaction, if any</td>
</tr>
</tbody>
</table>

**Intermediate representations**
- direct, semi-direct, transfer ($\geq 7$ variants)
- 2 lexical spaces
- IL ($\geq 2$ variants)
- 3 lexical spaces

**Programming paradigms**
- direct programming
- RBMT (rules, automata...)
- corpus-based
  - SMT, PSMT (unsupervised)
  - EBMT ($\geq 3$ variants)
  - $\pm$ supervised

**Hybrid**
Linguistic architectures in MT: Vauquois' triangle

- **Deep understanding level**
  - Ontological interlingua
- **Interlingual level**
  - Semantico-linguistic interlingua
- **Logico-semantic level**
  - SPA-structures (semantic & predicate-argument)
- **Mixing levels**
  - Multilevel description
- **Syntactico-functional level**
  - F-structures (functional)
- **Syntagmatic level**
  - C-structures (constituent)
- **Morpho-syntactic level**
  - Semi-direct translation
- **Graphemic level**
  - Direct translation
  - Tagged text
  - Text
Computational architectures

Intermediate Representation i -> Intermediate Representation i+1
Computational architectures

Intermediate Representation i → Intermediate Representation i+1

Computational type of the phase

- Expert
  - Procedural: Well-formedness grammar rules
  - Rule-based: Transitions of transduction automata
- Empirical
  - Statistical (probabilistic): Rewriting rules of transformational grammars
  - Example-based: Annotated parallel corpora trees, S-SSTC
  - Raw parallel corpora analogy-based MT: ABMT
Operational architectures for various translational situations
Operational architectures for various translational situations

- Users / tasks
  - help bilinguals produce good translations
  - help people understand an unknown or little known language
  - help people communicate (chat, spoken translation...)

Operational architectures for various translational situations

- **Users / tasks**
  help bilinguals produce good translations
  help people understand an unknown or little known language
  help people communicate (chat, spoken translation...)

- **Language pairs / volumes / kinds**
  1→1 (ALT/JE)  1←1
  1→N (MedSLT),  N←1
  1↔1 (Converser for Healthcare)
  1↔N (US Army, Phraselator)
  N↔N (debates, chat, multilingual peace forces)
Operational architectures for various translational situations

- **Users / tasks**
  - help bilinguals produce good translations
  - help people understand an unknown or little known language
  - help people communicate (chat, spoken translation...)

- **Language pairs / volumes / kinds**
  - $1\rightarrow 1$ (ALT/JE) $1\leftarrow 1$
  - $1\rightarrow N$ (MedSLT), $N\leftarrow 1$
  - $1\leftrightarrow 1$ (Converser for Healthcare)
  - $1\leftrightarrow N$ (US Army, Phraselator)
  - $N\leftrightarrow N$ (debates, chat, multilingual peace forces)

- **Possible involvement of humans**
  - authors (controlled language, rewriting, interactive disambiguation)
  - professional/occasional translators (post-editing)
  - readers (guessing from multiple factorized output)
Operational architectures for various translational situations

• Users / tasks
  help bilinguals produce good translations
  help people understand an unknown or little known language
  help people communicate (chat, spoken translation...)

• Language pairs / volumes / kinds
  1→1 (ALT/JE) 1←1
  1→N (MedSLT), N←1
  1↔1 (Converser for Healthcare)
  1↔N (US Army, Phraselator)
  N↔N (debates, chat, multilingual peace forces)

• Possible involvement of humans
  authors (controlled language, rewriting, interactive disambiguation)
  professional/occasional translators (post-editing)
  readers (guessing from multiple factorized output)

• Available resources
  data: huge parallel corpora necessary for SMT
  humans: computational linguists, lexicographers needed for "expert" MT
Linguistic and computational MT architectures are independent
MT architectures on which we work
Rule-based MT (symbolic) for sub-languages
Rule-based MT (symbolic) for sub-languages

All-domain MT via UNL
Rule-based MT (symbolic) for sub-languages

All-domain MT via UNL

Translation Memory based MAHT
Rule-based MT (symbolic) for sub-languages

All-domain MT via UNL

Translation Memory based MAHT

Statistico-structural MT
Rule-based MT (symbolic) for sub-languages

Translation Memory based MAHT

Statistico-structural MT

All-domain MT via UNL

Analogy-based MT
Rule-based MT (symbolic) for sub-languages

Translation Memory based MAHT

Statistico-structural MT

All-domain MT via UNL

Ontological interlingua

Semantico-linguistic interlingua

SPA-structures (semantic & predicate-argument)

Multilevel description

F-structures (functional)

C-structures

Tagged text

Text

Example-based MT (S-SSTC)

Analogy-based MT
## Direct translation systems

<table>
<thead>
<tr>
<th>Type</th>
<th>Steps</th>
<th>Method</th>
<th>Comments</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>RBMT</td>
<td>segmentation word for word translation</td>
<td>FST (rules + dict.) rules</td>
<td>OK for very near languages</td>
<td>ATLAS-I</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Japanese ↔ Korean Hindi ↔ Urdu</td>
<td>Fujitsu, 76-78</td>
</tr>
<tr>
<td>SMT</td>
<td>segmentation, reordering…</td>
<td>alignment + &quot;decoding&quot; statistical</td>
<td>SMT = first idea about MT from war cryptographers (W. Weaver 1949)</td>
<td>Many SMT</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>IBM… 1980-</td>
</tr>
<tr>
<td>EBMT</td>
<td>no preprocessing &quot;pure&quot; EBMT</td>
<td>analogy resolution + n-gram filtering analogical</td>
<td>Results ≈ those of SMT Nagao 1984 (similarity MT) Lepage 2000 (real analogy)</td>
<td>ALEPH ATR 2000-</td>
</tr>
</tbody>
</table>
# Semi-direct translation systems

<table>
<thead>
<tr>
<th>Type</th>
<th>Analysis</th>
<th>Transfer</th>
<th>Generation</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1G-MT 1950—</td>
<td>Program-based segmentation &amp;</td>
<td>dictionary consult. +</td>
<td>tables + string macros</td>
<td>GAT (Georgetown)</td>
</tr>
<tr>
<td></td>
<td>lemmatization procedural</td>
<td>reordering &quot;macros&quot;</td>
<td>procedural</td>
<td>Ispra, 1965—69</td>
</tr>
<tr>
<td>SMT 1990—</td>
<td>segmentation &amp;</td>
<td>alignment +</td>
<td>language model statistical</td>
<td>SPANAM-1, PAHO, 1975?—</td>
</tr>
<tr>
<td></td>
<td>lemmatization statistical</td>
<td>&quot;decoding&quot;</td>
<td>statistical</td>
<td>GLOBALINK ← Spanam-1, PAHO</td>
</tr>
<tr>
<td>Pidgin</td>
<td>segmentation snobol4</td>
<td>transduction +</td>
<td>morphological generation</td>
<td>Candide IBM, 1980—</td>
</tr>
<tr>
<td>translation</td>
<td>lemmatization rules</td>
<td>reordering</td>
<td>rules</td>
<td>Many SMT systems</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Q-systems rules on</td>
<td>formatting</td>
<td>NIST, IWSLT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>tree charts)</td>
<td>snobol4</td>
<td>Google (?)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Idea of B. Harris (TAUM, translatologist)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>rus → eng, fre Boitet 1972</td>
</tr>
</tbody>
</table>

**Morpho-syntactic level**  **Semi-direct translation**  **Tagged text**

**Graphemic level**

© Ch. Boitet — NII MT lectures 14/4/09
## Descending surface syntactic transfer systems: RBMT (+SMT / LanguageWeaver?)

<table>
<thead>
<tr>
<th>Type</th>
<th>Analysis</th>
<th>Transfer + syntactic generation</th>
<th>Morphological generation</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>RBMT</td>
<td>ATN</td>
<td>recursive descent</td>
<td>grammar+dict. rules</td>
<td>ENGSPAN, SPANAM-2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>AS-Transac</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Toshiba, 1982—</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reverso ProMT, 1986—</td>
</tr>
<tr>
<td>RBMT</td>
<td>ECFG (+decorations)</td>
<td>recursive descent</td>
<td>grammar+dict. rules</td>
<td>METAL Austin, 1982—</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Shalt-1 IBM-Jp, 1982—</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Kate KDD, 1983—</td>
</tr>
<tr>
<td>RBMT</td>
<td>Lemmatization +Slot-grammars</td>
<td>recursive descent</td>
<td>dictionary</td>
<td>LMT (IBM-US, 1983—)</td>
</tr>
<tr>
<td>1984—</td>
<td>prolog</td>
<td>logic programming</td>
<td>+tables+ prog. prolog</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>in prolog</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Syntagmatic level

### Morpho-syntactic level

### Graphemic level

### C-structures (constituent)

### Descending transfer

### Tagged text

### Text
# Descending deep syntactic transfer systems

<table>
<thead>
<tr>
<th>Type</th>
<th>Analysis</th>
<th>Transfer + synt. generation</th>
<th>Morphological generation</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>RBMT 1985—</td>
<td>Segm.+ lemmatization procedural</td>
<td>recursive descent procedural</td>
<td>grammar+dict. rules</td>
<td>JETS (IBM-Jp, 1985-90)</td>
</tr>
<tr>
<td></td>
<td>Dependency gramm. rules + constraint progr.</td>
<td></td>
<td>tables+prog. procedural</td>
<td></td>
</tr>
<tr>
<td>1.5G-MT 1990—</td>
<td>Lemmatization FST (+ dictionaries)</td>
<td>recursive descent procedural</td>
<td>grammar+dict. rules</td>
<td>Systran 1990—</td>
</tr>
<tr>
<td></td>
<td>Dependency graph procedural (C macros)</td>
<td></td>
<td>tables+prog. procedural</td>
<td></td>
</tr>
<tr>
<td></td>
<td>deterministic</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Syntactico-functional level**

**Morpho-syntactic level**

**Graphemic level**

**Ascending transfer**

**Tagged text**

**Text**

**F-structures (functional)**

often dependency structures
Horizontally surface syntactic transfer systems:
RBMT & Phrase-Based SMT

<table>
<thead>
<tr>
<th>Type</th>
<th>Analysis</th>
<th>Transfer</th>
<th>Generation</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>RBMT 1995—</td>
<td>lemmatization + Slot Grammars prolog</td>
<td>treelet dictionary prolog</td>
<td>recursive descent grammar + dict. rules</td>
<td>PT (from LMT) Linguatech, 1995—</td>
</tr>
<tr>
<td>EBMT 2000—</td>
<td>Initial data: bilingual corpus dictionary</td>
<td>Preparation: build S-SSTCs improve (hum)</td>
<td>Translation: A//T//G bottom-up</td>
<td>EBMT (Banturjah) UTMK, USM, 2000—</td>
</tr>
</tbody>
</table>

- **Syntagmatic level**: Syntactic transfer (surface) → C-structures (constituent)
- **Morpho-syntactic level**: Tagged text
- **Graphemic level**: Direct translation → Text

© Ch. Boitet — NII MT lectures
### Horizontal deep syntactic transfer systems

<table>
<thead>
<tr>
<th>Type</th>
<th>Analysis</th>
<th>Transfer</th>
<th>Generation</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>RBMT 1975—</td>
<td>grammar + dictionary dependency analysis rules</td>
<td>tree transformations rules</td>
<td>tree flattening grammar+ dictionary rules</td>
<td>ETAP-2, ETAP-3</td>
</tr>
<tr>
<td>RBMT 1995—</td>
<td>lemmatization + Slot Grammars ProLog</td>
<td>treelet dictionary ProLog</td>
<td>recursive descent grammar+ dictionary rules</td>
<td>PT (from LMT)</td>
</tr>
<tr>
<td>RBMT+ SMT 1999—</td>
<td>MSR (Microsoft ) analyzers rules (in G)</td>
<td>Learned from pairs (lf_s, lf_t) statistical</td>
<td>Microsoft generators rules(in G)</td>
<td>MTS-1 (on technical documentation)</td>
</tr>
</tbody>
</table>

**Graphemic level**  
**Morpho-syntactic level**  
**Syntagmatic level**  
**Syntactico-functional level**  

**Syntactic transfer (deep)**  

**F-structures (functional)**  
**C-structures (constituent)**  
**Tagged text**  
**Text**
## Horizontal multilevel transfer systems

<table>
<thead>
<tr>
<th>Type</th>
<th>Analysis</th>
<th>Transfer</th>
<th>Generation</th>
<th>Examples</th>
</tr>
</thead>
</table>
| RBMT 1990— | Lemmatization  
ECFG (govt & binding)  
gram. rules  
interactive disambiguation if not enough space | dictionary + tree transformations rules | tree transformations rules  
MG: dictionary + grammars rules | ITS (Geneva, 1990—)  
Perhaps PT-2 (rather than SF) |

### Logico-semantic level

- **N levels in 1 structure**
  - (abstract constituent tree)

### Syntactico-functional level

### Syntagmatic level

### Morpho-syntactic level

### Graphemic level

Multilevel transfer

Multilevel description
- F-structures (functional)
- C-structures (constituent)
- Tagged text

Text
Ascending multilevel transfer systems

<table>
<thead>
<tr>
<th>Type</th>
<th>Analysis</th>
<th>Transfer</th>
<th>Generation</th>
<th>Examples</th>
</tr>
</thead>
</table>
tree transformations rewriting rules | dictionary                | tree transformations rules | Ariane-G5-based  
ru-de→ru, en→my-th 80-87  
fr→en (BV/aero) 85-92  
fr→en-de-ru (LIDIA) 90-96 |
|            |                              |                           | MG: dict. + gram. rules  | HICATS Hitachi (1990—-)  
Jemah USM, NUS (1990—-)                            |

Logico-semantic level

N levels in 1 structure

Syntactico-functional level

Syntagmatic level

Morpho-syntactic level

Graphemic level

SPA-structures (semantic & predicate-argument)

Multilevel description

F-structures (functional)

C-structures (constituent)

Tagged text

Text
### Semantic transfer systems

<table>
<thead>
<tr>
<th>Type</th>
<th>Analysis</th>
<th>Transfer</th>
<th>Generation</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>RBMT 1982—</td>
<td>segmentation lemmatization direct programming tree transformations rules</td>
<td>dictionary + tree transformations rules</td>
<td>tree transformations rules MG: dict. + gram. rules</td>
<td>MU Kyodai, 82-87 MAJESTIC JICST, 87—</td>
</tr>
</tbody>
</table>

**Logico-semantic level**

**Syntactico-functional level**

**Syntagmatic level**

**Morpho-syntactic level**

**Graphemic level**

Semantic transfer

SPA-structures (semantic & predicate-argument)

F-structures (functional)

C-structures (constituent)

Tagged text

Text
# Conceptual transfer systems (IL with separate lexicon)

## Interlingual level

<table>
<thead>
<tr>
<th>Type</th>
<th>Enconversion</th>
<th>Conceptual transfer</th>
<th>Deconversion</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>RBMT 1980—</td>
<td>Lemmatization</td>
<td></td>
<td>graph-string transformations rules</td>
<td>ATLAS-II</td>
</tr>
<tr>
<td></td>
<td>direct or rules</td>
<td></td>
<td></td>
<td>Fujitsu, 1980—</td>
</tr>
<tr>
<td></td>
<td>string-graph</td>
<td>in principle none</td>
<td></td>
<td>PIVOT</td>
</tr>
<tr>
<td></td>
<td>transformations rules</td>
<td></td>
<td></td>
<td>Nec, 1983—</td>
</tr>
<tr>
<td>RBMT 1980—</td>
<td>DCG (?) rules</td>
<td></td>
<td></td>
<td>ULTRA NMSU, 89-95</td>
</tr>
<tr>
<td>RBMT 1997—</td>
<td>depending on partners</td>
<td>navigation in set of</td>
<td>depending on partners rules</td>
<td>UNL 1996—</td>
</tr>
<tr>
<td></td>
<td>rules (until now!)</td>
<td>UWs</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>UNL lexicon</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Semantico-linguistic interlingua

- **ATLAS-II**: Fujitsu, 1980—
- **PIVOT**: Nec, 1983—

## Graphemic level

- **ULTRA NMSU**: 89-95

## Text
## Knowledge-based systems: explicit understanding (IL linked with an ontology)

### Deep understanding level

<table>
<thead>
<tr>
<th>Type</th>
<th>Enconversion</th>
<th>Mapping into $\Omega$</th>
<th>Deconversion</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>KBMT</td>
<td>lemmatization &amp; EPSG+f-structures +pseudo-unification rules (using UP)</td>
<td>all but discourse elements dict.+rules + interactive disambiguation</td>
<td>planning deep-str rec. descent rules</td>
<td>KBMT-89</td>
</tr>
<tr>
<td>1980—</td>
<td></td>
<td></td>
<td></td>
<td>CMU, 1989—91</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>KANT/Catalyst</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CMU+Caterpillar, en→fr-sp-de-? 1992—</td>
</tr>
<tr>
<td>RBMT</td>
<td>dictionary + FST rules</td>
<td>IF is only a pragmatico-semantic representation no mapping to $\Omega$</td>
<td>dictionary + FST rules</td>
<td>CSTAR-II &amp; Nespole!</td>
</tr>
<tr>
<td>1997—</td>
<td></td>
<td></td>
<td></td>
<td>GETA 97-03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ETRI (Korea) 97-99</td>
</tr>
<tr>
<td>SMT</td>
<td>learned from (string,IF) KB statistical</td>
<td>no mapping to $\Omega$</td>
<td>learned from (IF, string) KB statistical</td>
<td>CSTAR-II &amp; Nespole!</td>
</tr>
<tr>
<td>2003—</td>
<td></td>
<td></td>
<td></td>
<td>IRst 98-03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Mastor-1 IBM 2003</td>
</tr>
</tbody>
</table>

### Ontological interlingua

- Ontological interlingua
- Semantico-linguistic interlingua
- SPA-structures (semantic & predicate-argument)
- F-structures (functional)
- C-structures (constituent)

### Graphemic level

- Graphemic level
- Text

© Ch. Boitet — NII MT lectures

14/4/09
# Size & cost of resources / MT architectures

<table>
<thead>
<tr>
<th>Type</th>
<th>Sentences</th>
<th>6.5 w/s BTEC, Meteo</th>
<th>25 w/s News</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMT</td>
<td>0.9—3 Mw</td>
<td>50—200 Mw</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3.6—12 K pages</td>
<td>200—800 K pages</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.15—0.5 M sentences</td>
<td>2—8 M sentences</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.4—8 m*y (already done!)</td>
<td>100—400 m*y (available?)</td>
<td></td>
</tr>
<tr>
<td>PSMT</td>
<td>N/A for short sent.</td>
<td>4—12.5 Mw</td>
<td></td>
</tr>
<tr>
<td>analogical EBMT</td>
<td>Supervised learning</td>
<td>15—50 K pages</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1h/page?</td>
<td>0.15—0.5 M sentences</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>10—40 m*y (to do!)</td>
<td></td>
</tr>
<tr>
<td>EBMT with trees</td>
<td>N/A for short sent.</td>
<td>4—12.5 Mw</td>
<td></td>
</tr>
<tr>
<td>MST</td>
<td>Supervised learning</td>
<td>15—50 K pages</td>
<td></td>
</tr>
<tr>
<td>Mastor-1</td>
<td>1h/page?</td>
<td>0.15—0.5 M sentences</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>10—40 m*y (to do!)</td>
<td></td>
</tr>
<tr>
<td>EBMT with trees and S-SSTCs</td>
<td>N/A for short sent.</td>
<td>4—12.5 Mw</td>
<td></td>
</tr>
<tr>
<td>Banturjah (USM)</td>
<td>Supervised learning</td>
<td>15—50 K pages</td>
<td></td>
</tr>
<tr>
<td></td>
<td>15 h/page!</td>
<td>0.6—1 K pages</td>
<td></td>
</tr>
<tr>
<td></td>
<td>dict. (50 K) available</td>
<td>0.006—0.01 M sentences</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>6—10 m*y (to do!)</td>
<td></td>
</tr>
<tr>
<td>RBMT</td>
<td>Dict. 3-10 K 0.6—2 m*y</td>
<td>Dict. 50-500 K 15—150 m*y</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total 1—3 m*y (to do!)</td>
<td>Total 40—175 m*y</td>
<td></td>
</tr>
</tbody>
</table>
Operational architectures

• Translational situations...
  
  *Graph of needed translations*, with weights (flows, importance)
  Possibility/necessity of human help to MT
  Possibility/necessity of pre-edition or post-edition
  Time constraints
  Available resources (no SMT if no corpora, no EMT if no experts)

• may influence linguistic & computational architectures
  Reusing existing steps to build new translation pairs
  Ex: transfer approach (if N languages, #transfers still linear)

\[
\begin{align*}
L_1 \rightarrow & L_0 \\
L_0 \rightarrow & L_0 \\
L_0 \rightarrow & L_2 \\
L_1 \rightarrow & L_1 \\
L_0 \rightarrow & L_0 \\
L_2 \rightarrow & L_2
\end{align*}
\]
Current evolution

• deeper linguistic architectures
  word-based → phrase-based → DeepSynt → LogSem → SemPrag

• hybrid computational architectures
  coupled with user involvement

• use more Machine Learning
  use it also with deep transfer or pivot architectures (through an IL)
  (more on that in lecture 3)
What kind of IL to choose if that is the choice?

<table>
<thead>
<tr>
<th>IL+ontology</th>
<th>restricted domain</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>high precision applications</td>
</tr>
<tr>
<td></td>
<td>cf. CLang (Mooney)</td>
</tr>
<tr>
<td></td>
<td>beware, Ø costlier than gram+dict!</td>
</tr>
<tr>
<td></td>
<td>machine learning possible</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pragmatico-semantic IL</th>
<th>task- and domain-related applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF of CSTAR/Nespole!</td>
<td>reservations in tourism</td>
</tr>
<tr>
<td></td>
<td>medical assistance</td>
</tr>
<tr>
<td></td>
<td>both MUST be restricted</td>
</tr>
<tr>
<td></td>
<td>works very well then</td>
</tr>
<tr>
<td></td>
<td>machine learning possible</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Semantico-linguistic IL</th>
<th>all domains/tasks: IL has to be grounded on a NL understandable by most developers anywhere amenable to machine learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATLAS-II, PIVOT</td>
<td>better: UNL</td>
</tr>
</tbody>
</table>
Recap & first conclusions

- The distinctions RBMT, EBMT, an-EBMT, (P)SMT... concern the computational architecture only (PROCESSES)
- The rawer the corpora, the larger they must be
  SMT/PSMT is for niches for the rich (languages, texts)
  - few parallel corpora of 200—800 K pages
  - to build them from scratch is 2 to 3 times more expensive than to build a classical large RBMT system
- IL-based MT can use any computational framework statistical, analogical, rule-based, hybrid
  all depends on available corporal / linguistic / human resources
- Many applications need an adequate IL
  all applications needing to manipulate content in a strongly multilingual setting