ARPA MT Initiative
1Q94 Test and Evaluation
17 March 1994

John White
Theresa O'Connell
Matthew Hopwood
PRC Inc.
ARPA MT Initiative Statement of Purpose

- Revolutionary advances in MT technology
- Promote development of core technology
- Advances throughout MT marketplace
- Using FAMT evaluation techniques
Agenda - Part One - 1Q94 System Tests

- Test materials
- Participants
- Test conduct
- Outputs
Agenda - Part Two - 1094 Evaluation

- Evaluation materials
- Comprehension Evaluation
- Fluency Evaluation
- Adequacy Evaluation
- Data points tallied
Agenda - Part 3 - System Test Results

- Computation Methods
- Comprehension, Fluency, Adequacy
  - Comparison with 2Q93
  - FAMT systems by language
- Sensitivity
Part One - 1Q94 System Tests
Test Materials

- Source passages
  - 20 in each language, 300-500 words (800 JapaneseChars)
  - 10 Business mergers & acquisitions (M&A) texts
  - 10 General news texts
### Test File Delivery

<table>
<thead>
<tr>
<th>Passage</th>
<th>Origin</th>
<th>Passage</th>
<th>Origin</th>
<th>Passage</th>
<th>Origin</th>
</tr>
</thead>
<tbody>
<tr>
<td>TEXT01</td>
<td>GEN</td>
<td>TEXT01</td>
<td>GEN</td>
<td>TEXT01</td>
<td>M&amp;A</td>
</tr>
<tr>
<td>TEXT02</td>
<td>M&amp;A</td>
<td>TEXT02</td>
<td>M&amp;A</td>
<td>TEXT02</td>
<td>GEN</td>
</tr>
<tr>
<td>TEXT03</td>
<td>GEN</td>
<td>TEXT03</td>
<td>GEN</td>
<td>TEXT03</td>
<td>M&amp;A</td>
</tr>
<tr>
<td>TEXT04</td>
<td>M&amp;A</td>
<td>TEXT04</td>
<td>M&amp;A</td>
<td>TEXT04</td>
<td>GEN</td>
</tr>
<tr>
<td>TEXT05</td>
<td>GEN</td>
<td>TEXT05</td>
<td>GEN</td>
<td>TEXT05</td>
<td>M&amp;A</td>
</tr>
<tr>
<td>TEXT06</td>
<td>M&amp;A</td>
<td>TEXT06</td>
<td>M&amp;A</td>
<td>TEXT06</td>
<td>GEN</td>
</tr>
<tr>
<td>TEXT07</td>
<td>GEN</td>
<td>TEXT07</td>
<td>GEN</td>
<td>TEXT07</td>
<td>M&amp;A</td>
</tr>
<tr>
<td>TEXT08</td>
<td>M&amp;A</td>
<td>TEXT08</td>
<td>M&amp;A</td>
<td>TEXT08</td>
<td>GEN</td>
</tr>
<tr>
<td>TEXT09</td>
<td>GEN</td>
<td>TEXT09</td>
<td>GEN</td>
<td>TEXT09</td>
<td>M&amp;A</td>
</tr>
<tr>
<td>TEXT10</td>
<td>M&amp;A</td>
<td>TEXT10</td>
<td>M&amp;A</td>
<td>TEXT10</td>
<td>GEN</td>
</tr>
<tr>
<td>TEXT11</td>
<td>GEN</td>
<td>TEXT11</td>
<td>GEN</td>
<td>TEXT11</td>
<td>M&amp;A</td>
</tr>
<tr>
<td>TEXT12</td>
<td>M&amp;A</td>
<td>TEXT12</td>
<td>M&amp;A</td>
<td>TEXT12</td>
<td>GEN</td>
</tr>
<tr>
<td>TEXT13</td>
<td>GEN</td>
<td>TEXT13</td>
<td>GEN</td>
<td>TEXT13</td>
<td>M&amp;A</td>
</tr>
<tr>
<td>TEXT14</td>
<td>M&amp;A</td>
<td>TEXT14</td>
<td>M&amp;A</td>
<td>TEXT14</td>
<td>GEN</td>
</tr>
<tr>
<td>TEXT15</td>
<td>GEN</td>
<td>TEXT15</td>
<td>GEN</td>
<td>TEXT15</td>
<td>M&amp;A</td>
</tr>
<tr>
<td>TEXT16</td>
<td>M&amp;A</td>
<td>TEXT16</td>
<td>M&amp;A</td>
<td>TEXT16</td>
<td>GEN</td>
</tr>
<tr>
<td>TEXT17</td>
<td>GEN</td>
<td>TEXT17</td>
<td>GEN</td>
<td>TEXT17</td>
<td>M&amp;A</td>
</tr>
<tr>
<td>TEXT18</td>
<td>M&amp;A</td>
<td>TEXT18</td>
<td>M&amp;A</td>
<td>TEXT18</td>
<td>GEN</td>
</tr>
<tr>
<td>TEXT19</td>
<td>GEN</td>
<td>TEXT19</td>
<td>GEN</td>
<td>TEXT19</td>
<td>M&amp;A</td>
</tr>
<tr>
<td>TEXT20</td>
<td>M&amp;A</td>
<td>TEXT20</td>
<td>M&amp;A</td>
<td>TEXT20</td>
<td>GEN</td>
</tr>
</tbody>
</table>
Research Systems

- CANDIDE (French-English)
  - IBM Thomas Watson Research Laboratory
- LINGSTAT (Japanese - English)
  - Dragon Systems
- PANGLOSS (Spanish-English)
  - Carnegie Mellon University Center for Machine Translation
  - New Mexico State University Computing Research Laboratory
  - University of Southern California Information Sciences Institute
Production Systems

- POWER TRANSLATOR (French-English, Spanish-English)
- PC TRANSLATOR (Spanish-English)
- FRENCH/SPANISH ASSISTANT (French-English, Spanish-English)
- PIVOT (Japanese-English)
- SPANAM (Spanish-English)
- METAL (French-English)
- XLT (French-English)
- SYSTRAN Translation Systems, Inc. (French-English, Japanese-English, Spanish-English)
- WINGER A/S (French-English)
Test Conduct

- Systems frozen upon start of test
- Restrictions on systems
- No human intervention
- Lexical development (Not)
Human-assisted Coordination at Research Sites

- "Level 2" translators
- Manual/HAMT of different subset
- CANDIDE and LINGSTAT - 2 translators
- PANGLOSS - 4 translators
- Issues: "Level 2" standards
Translations

- 20 Passages
- 3 Languages
- 9 Versions of French (excluding reference translations)
- 8 Versions of Spanish (excluding reference translations)
- 5 Versions of Japanese (excluding reference translations)
- = 440 Passage-Versions (500 including reference translations)
Part Two - 1Q94 Evaluation

- Evaluation materials
- Comprehension Evaluation
- Fluency Evaluation
- Adequacy Evaluation
- Data points tallied
Goals of ARPA MT Evaluation

- Optimize for individual strengths/purposes
  - Internally useful results
  - Not biased against approach/application
- Provide for fit of these methods/results into the MT discipline
  - Use external benchmark systems
  - Design/evolve tests for portability, objectivity, sensitivity, economy
# Inputs to 1Q94 Evaluation

<table>
<thead>
<tr>
<th>FRENCH TO ENGLISH</th>
<th>SPANISH TO ENGLISH</th>
<th>JAPANESE TO ENGLISH</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Research System</strong>&lt;br&gt;Candide FAMT&lt;br&gt;Candide HAMT&lt;br&gt;Level 2 Manual</td>
<td><strong>Research System</strong>&lt;br&gt;Pangloss FAMT&lt;br&gt;Pangloss HAMT&lt;br&gt;Level 2 Manual</td>
<td><strong>Research System</strong>&lt;br&gt;Lingstat FAMT&lt;br&gt;Lingstat HAMT&lt;br&gt;Level 2 Manual</td>
</tr>
<tr>
<td><strong>Expert Production Systems FAMT</strong>&lt;br&gt;French Assistant&lt;br&gt;METAL&lt;br&gt;Power Translator&lt;br&gt;Systran&lt;br&gt;Winger&lt;br&gt;XLT</td>
<td><strong>Expert Production Systems FAMT</strong>&lt;br&gt;PC Translator&lt;br&gt;Power Translator&lt;br&gt;SPANAM&lt;br&gt;Spanish Assistant&lt;br&gt;Systran</td>
<td><strong>Expert Production Systems FAMT</strong>&lt;br&gt;PIVOT&lt;br&gt;Systran</td>
</tr>
</tbody>
</table>
2Q93 Evaluation Materials

• 11 Evaluation Books

  - 22 Fluency passages

  - 22 Adequacy passages (same output version passages)

  - 9,10, or 11 Comprehension sets

  - Based on 8 passages/language

  - Same versions as already seen, plus reference versions
1Q94 Evaluation Materials

- 30 Evaluation books
- 18 Comprehension passages
  - Includes 2 expert versions
- 16 Fluency passages + 1 practice passage = 17
- 16 Adequacy passages + 1 practice passage = 17
Evaluation Matrix Design

- Random distribution in matrix
  - Dummy passages to fill matrix
- 1/2 General 1/2 M&A
- Every passage-version seen at least once
- Evaluator sees only one version of a passage
- Master matrix for Comprehension Evaluation
  - Includes reference versions
- Adequacy/Fluency matrix subset of Comprehension matrix
  - Excludes reference versions
### Evaluation Book Matrix

#### Comprehension

<table>
<thead>
<tr>
<th>Evaluator</th>
<th>Passage 1</th>
<th>Passage 2</th>
<th>Passage 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluator 1</td>
<td>316 EXP</td>
<td>317 PAN</td>
<td>112 MAN</td>
</tr>
<tr>
<td>Evaluator 2</td>
<td>118 EXP</td>
<td>210 MAN</td>
<td>105 CAN</td>
</tr>
<tr>
<td>Evaluator 3</td>
<td>207 EXP</td>
<td>208 LIN</td>
<td>211 LIN</td>
</tr>
<tr>
<td>Evaluator 4</td>
<td>310 EXP</td>
<td>104 WIN</td>
<td>114 GLO</td>
</tr>
<tr>
<td>Evaluator 5</td>
<td>306 EXP</td>
<td>304 GLO</td>
<td>309 LPR</td>
</tr>
<tr>
<td>Evaluator 6</td>
<td>107 EXP</td>
<td>118 SIE</td>
<td>108 SIE</td>
</tr>
<tr>
<td>Evaluator 7</td>
<td>205 EXP</td>
<td>111 GLO</td>
<td>DUMMY</td>
</tr>
</tbody>
</table>

#### Fluency & Adequacy

<table>
<thead>
<tr>
<th>Evaluator</th>
<th>Passage 1</th>
<th>Passage 2</th>
<th>Passage 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluator 1</td>
<td>PRACTICE</td>
<td>317 PAN</td>
<td>112 MAN</td>
</tr>
<tr>
<td>Evaluator 2</td>
<td>PRACTICE</td>
<td>210 MAN</td>
<td>105 CAN</td>
</tr>
<tr>
<td>Evaluator 3</td>
<td>PRACTICE</td>
<td>208 LIN</td>
<td>211 LIN</td>
</tr>
<tr>
<td>Evaluator 4</td>
<td>PRACTICE</td>
<td>104 WIN</td>
<td>114 GLO</td>
</tr>
<tr>
<td>Evaluator 5</td>
<td>PRACTICE</td>
<td>304 GLO</td>
<td>309 LPR</td>
</tr>
<tr>
<td>Evaluator 6</td>
<td>PRACTICE</td>
<td>118 SIE</td>
<td>108 SIE</td>
</tr>
<tr>
<td>Evaluator 7</td>
<td>PRACTICE</td>
<td>111 GLO</td>
<td>DUMMY GEN</td>
</tr>
</tbody>
</table>
Training Evaluators

- 30 Evaluators

- Spoken instructions
  - Group instructions for comprehension
  - Individual instructions for fluency and adequacy
  - Understanding to same level

- Written instructions
  - Examples for guidance only
  - Examples in degrading order
Evaluator Performance

- Consistency of judgement
  - Context effects
    - Bias minimized by random matrix
  - Practice passage
    - Promote consistent judgements across evaluation set
- Fatigue
  - Average 6 to 8 hours to complete evaluation
  - Two planned breaks plus a lunch break
  - Recommended additional break during each component
  - Unlimited personal breaks
  - Fluency component least fatiguing
Comprehension Evaluation

- Like SAT Comprehension Test

- Derived question set from reference translations

- Six questions for each text
Comprehension Sample

TEXT 101

1. A memorial service was held for
   a. James Jordan's father
   b. James Jordan's widow
   c. Michael Jordan
   d. Michael Jordan's father
   e. None of the above

2. The program included a message from
   a. family and close friends
   b. James Jordan
   c. the press
   d. the widow and children
   e. None of the above

3. The coffin was covered with
   a. an American flag
   b. flowers
   c. messages from his five children
   d. the Air Force insignia
   e. None of the above

4. The victim's body had been discovered
   a. close to the church
   b. in a river
   c. in his car
   d. on the route 95 exit
   e. None of the above
Fluency Evaluation

- Judge fluency of English "sentences" in context
- 1 to 5 scale
- Each version of each passage seen once
- No one sees more than one version of any passage
Funeral service for the father of Michael Jordan

The family and the near the star of the American basketball Michael Jordan develop are gathered Sunday for a funeral service to the memory of his/her/its father James.

The security was important and the press had been put secluded of the church Methodist episcopalian African (African Methodist Episcopal Church) placed nearly Teachey (Caroline of the north), where took place the service.

The journalists had receipt a program of the ceremony, that comprehended a message of the widow of James Jordan, Deloris, and of the his/her/their five children Michael, James Ronald, Deloria, Larry and Roslyn.

"all those that have been touched by the heat and the depth of this special man can comprehend the depth of the loss felt by the family", said the message.

"Papa is no longer with us.

But the lessons that they we has learnt will remain forever and they we will give the force of going of the prior to", added the text.

The coffin was re-covered of the American flag, in souvenir of the services returned by James Jordan in the army of the air.
Adequacy Evaluation

• Judge degree to which meaning present in expert translation is present in version

• By "fragment"

• 1 to 5 scale

• Compare fragments of reference translation against output

• De-emphasizes fluency of the translation

• Use same distribution matrix as in Fluency
Adequacy Fragments

- Derived by externally motivated trade-off
  - High-level syntactic constituent
  - Between 5 and 20 words (or so)
  - Fragment contains information sufficient to find match in translation
Funeral service for the father of Michael Jordan

The family and the near the star of the American basket-ball Michael Jordan develop are gathered Sunday for a funeral service to the memory of his/her/its father James.

The security was important and the press had been put secluded of the church Methodist épiscopaienne African (African Methodist Episcopal Church) placed nearly Teachey (Caroline of the north), where took place the service.

The journalists had receipt a program of the ceremony, that comprehended a message of the widow of James Jordan, Deloris, and of the his/her/their five children Michael, James Ronald, Deloris, Larry and Roslyn.
Data Points Tallied in Evaluation

- Dummy passages not tallied
- 6,744 in Fluency (Avg. 225 judgments/evaluator tallied)
- 13,316 in Adequacy (Avg. 444 judgments/evaluator tallied)
- 3,000 in Comprehension (Avg. 100 judgments/evaluator tallied)
- 23,060 Data points total (Avg. 769 judgments/evaluator tallied)
Part 3 - System Test Results

- Computation Methods

- Comprehension, Fluency, Adequacy
  - Comparison with 2Q93
  - FAMT systems by language

- Comparison of sensitivity
Time Computation

- Time axis = Translation time / AVG (manual 93 time + manual 94 time)

- Mean of operator time, manual time for each system

- Presumes indirect comparability
Comprehension, Fluency and Adequacy Computation

- Values between 0 and 1
  - Comprehension = \( \frac{\text{#Correct}}{6} \)
  - Fluency = \( \frac{\sum(\text{(Decision point - 1)/5-1})}{\text{#sent. in sys.}} \)
  - Adequacy = \( \frac{\sum(\text{(Decision point - 1)/5-1})}{\text{#frags. in sys.}} \)

- STD DEV calculated for system value over 20 passages
- STD DEV calculated over all system scores
- F-Ratio as \( \left( \frac{\text{VAR of scores}}{\text{mean of VARS} \times \text{SQRT(20)}} \right) \)