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 - 1Q94 Evaluation

- Evaluation materials
- Comprehension Evaluation
- Fluency Evaluation
- Adequacy Evaluation
- Data points tallied

Agenda - Part 3 - System Test Results

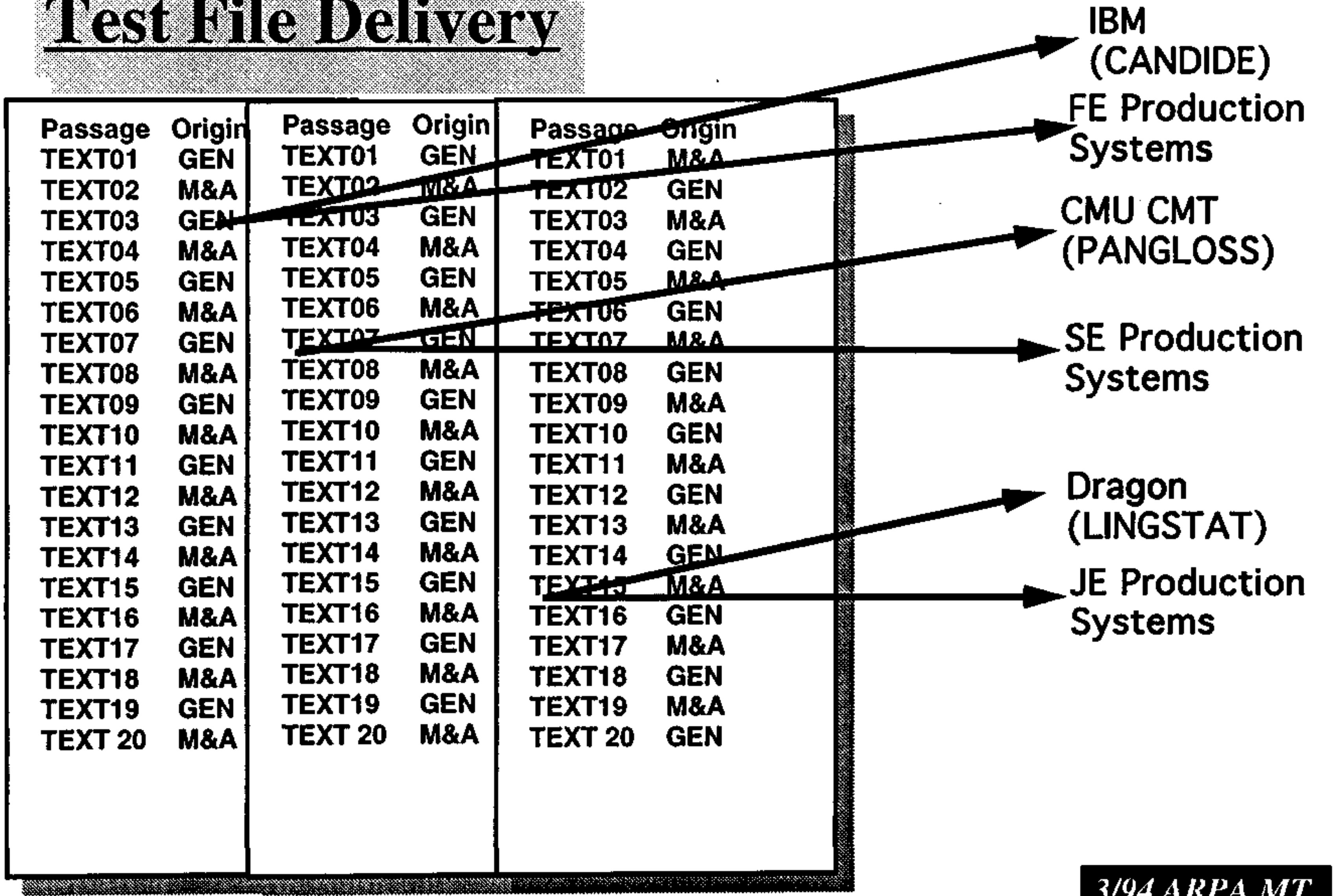
- Computation Methods
- Comprehension, Fluency, Adequacy
  - Comparison with 2Q93
  - FAMT systems by language
- Sensitivity

Part One - 1Q94 System Tests

Test Materials

- Source passages
  - 20 in each language, 300-500 words (800 Japanese Chars)
  - 10 Business mergers & acquisitions (M&A) texts
  - 10 General news texts

Test File Delivery



Passage	Origin	Passage	Origin	Passage	Origin
TEXT01	GEN	TEXT01	GEN	TEXT01	M&A
TEXT02	M&A	TEXT02	M&A	TEXT02	GEN
TEXT03	GEN	TEXT03	GEN	TEXT03	M&A
TEXT04	M&A	TEXT04	M&A	TEXT04	GEN
TEXT05	GEN	TEXT05	GEN	TEXT05	M&A
TEXT06	M&A	TEXT06	M&A	TEXT06	GEN
TEXT07	GEN	TEXT07	GEN	TEXT07	M&A
TEXT08	M&A	TEXT08	M&A	TEXT08	GEN
TEXT09	GEN	TEXT09	GEN	TEXT09	M&A
TEXT10	M&A	TEXT10	M&A	TEXT10	GEN
TEXT11	GEN	TEXT11	GEN	TEXT11	M&A
TEXT12	M&A	TEXT12	M&A	TEXT12	GEN
TEXT13	GEN	TEXT13	GEN	TEXT13	M&A
TEXT14	M&A	TEXT14	M&A	TEXT14	GEN
TEXT15	GEN	TEXT15	GEN	TEXT15	M&A
TEXT16	M&A	TEXT16	M&A	TEXT16	GEN
TEXT17	GEN	TEXT17	GEN	TEXT17	M&A
TEXT18	M&A	TEXT18	M&A	TEXT18	GEN
TEXT19	GEN	TEXT19	GEN	TEXT19	M&A
TEXT 20	M&A	TEXT 20	M&A	TEXT 20	GEN

IBM
(CANDIDE)

FE Production
Systems

CMU CMT
(PANGLOSS)

SE Production
Systems

Dragon
(LINGSTAT)

JE Production
Systems

3/94 ARPA MT
Workshop

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

FRENCH TO ENGLISH

Research System
- Candide FAMT
- Candide HAMT
- Level 2 Manual

Expert

Production Systems FAMT
- French Assistant
- METAL
- Power Translator
- Systran
- Winger
- XLT

SPANISH TO ENGLISH

Research System
- Pangloss FAMT
- Pangloss HAMT
- Level 2 Manual

Expert

Prodution Systems FAMT
- PC Translator
- Power Translator
- SPANAM
- Spanish Assistant
- Systran

JAPANESE TO ENGLISH

Research System
- Lingstat FAMT
- Lingstat HAMT
- Level 2 Manual

Expert

Production Systems FAMT
- PIVOT
- Systran

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
- 17 ~~16~~ Fluency passages + 1 practice passage = 17
- 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

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

Fluency & Adequacy

	Passage 1	Passage 2	Passage 3
Evaluator 1	PRACTICE	317 PAN	112 MAN
Evaluator 2	PRACTICE	210 MAN	105 CAN
Evaluator 3	PRACTICE	208 LIN	211 LIN
Evaluator 4	PRACTICE	104 WIN	114 GLO
Evaluator 5	PRACTICE	304 GLO	309 LPR
Evaluator 6	PRACTICE	118 SIE	108 SIE
Evaluator 7	PRACTICE	111 GLO	DUMMY GEN

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

Measure differences [illegible]
[illegible] (MT vs. HT)

- 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 [illegible]

- Derived question set from reference translations

- Six questions for each text

Comprehension Sample

Start Time ______

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

[illegible]
Howalong it takes to read aloud
[illegible] well-formedness is a subset
[illegible] criteria

- 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

Fluency Sample

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 épiscopalienne 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. ____

"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 we. ____

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

Adequacy Sample

right 2 mistakes

[Funeral Service for Michael Jordan's Father]	________	**Funeral service for the father of Michael Jordan**
[Family and close friends of American basketball star Michael Jordan gathered together on Sunday]	________	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.
[for a memorial service for Jordan's father, James.]	________	
[There was considerable security.]	________	The security was important and the press had been put secluded of the church Methodist épiscopalienne African (African Methodist Episcopal Church) placed nearly Teachey (Caroline of the north), where took place the service.
[and the press had been kept away from the African Methodist Episcopal Church near Teachey, North Carolina, where the service was held.]	________	
[Reporters had received a program of 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.
[which included a message from James Jordan's widow, Deloris,]	________	
[and from his 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 = #Correct/6
  - Fluency = (∑((Decision point - 1)/5-1))/#sent. in sys.
  - Adequacy = (∑((Decision point - 1)/5-1))/#frags. in sys.
- STD DEV calculated for system value over 20 passages
- STD DEV calculated over all system scores
- F-Ratio as ((VAR of scores / mean of VARS * SQRT(20))