日英及び日中音声翻訳システムのコミュニケーション支援能力の評価
ー課題遂行型対話実験による試みー

○水島昌英，竹澤寿幸，清水徹，菊井玄一郎（ATR 音声言語コミュニケーション研究所）

1. はじめに

ATR では，実世界で利用可能な多言語の音声翻訳(S2ST)システムの実現を目指して研究を進めている[1]。著者らは，同システムの評価等を目的とした日本語話者と英語話者による課題遂行型の模擬対話実験を実施してきた[2]。これまでに，S2ST システムを介した対話音声は，会話例文を読み上げた音声に近い発話スタイルになることなどが分かっている[3,4,5]。

本研究の第一の目的は，我々の S2ST システムが“異言語話者間のコミュニケーションにどの程度使えるのか”，という問いに対する一つの目安を示すことである。それに加え，新たに開発した日本語ー中国語の S2ST システムの現時点での到達点を，従来から開発してきた日英システムと比較しながら示す。

2. S2ST システム

S2ST システムでは，話者の音声を音声認識部においてテキストに変換し，話者にその結果を表示すると共に機械翻訳部に渡し，相手言語テキストに変換する。そして，その翻訳テキストを相手側に表示すると同時に，音声合成部で合成された音声を出力する。対話システムでは，この逆側のシステムも必要となる。

音声認識には ATRASR[6]を使用した。機械翻訳は用例翻訳 HPATR2[7] と統計翻訳 SAT[8]に最良翻訳選択器 SELECTOR[9] を組み合わせたシステムを使用した。日本語及び中国語の音声合成は XIMERA[10]，英語には，AT&T Labs’ Natural Voices™を使用した。音声認識の言語モデル及び機械翻訳のモデルは，旅行会話例文集と海外旅行に関する模擬対話実験データから構成される日英約 100 万文，日中約 50 万文の大規模対訳コーパスから作成された[11]。

3. 対話実験の設計

3.1 基本方針

本研究では，S2ST システムのコミュニケーション支援能力を，話者が互いに持っている情報をいかに的確にそして速やかに相手に伝達できたかで測ることにする。そのためには，S2ST システムを介して，“十分達成が見込める課題設定”にする必要がある。そのことを踏まえて，以下のような方針で課題作成を試みた。

- 場面はコーパスの中に多く含まれているものから選ぶ。具体的には“**買い物**”，“**ホテル予約**”，“**ホテル及びレストランでの簡単なトラブル対応**”とした。
- 固有名詞の使用は必要最小限にとどめ，課題達成に必要なキーワードは全てコーパスの中にあるものとする。
- 伝えるべき情報の数で課題の難易度を，伝えられた情報の数で達成度を，そして，それに費やした発話数で効率を評価する。

3.2 課題設定の例

3.1 節の方針に従った課題設定を，具体例をあげて説明する。例えば買い物の場面において，客役に対して，“エスサイズのオレンジジュースを二杯注文すること。氷無しにするようにお願いすること。一杯の値段を確認すること。”のような課題を与える。これは“1：オレンジジュース，2：エスサイズ，3：2杯，4：氷無し”と，4つの情報を店員役の相手に伝えることになる。店員に正しく伝われば，あらかじめ店員にのみ渡してあるメニューからその値段を客役に伝えて，合計で5つの情報伝達が要求された課題となる。

3.3 実験方法

対話実験は，2節で述べた S2ST システムを使用して ATR 内の実験室で実施した。なお，音声収音にはヘッドセットマイクを使用した。

[被験者]一日に二人一組で，日英，日中対話を六日間ずつ実施した。英語，中国語話者は日本に比較的長期に滞在中の人たちで，日本語が堪能の人も多かった。また日本人も英語が分かる人は多い。過去の実験[12]で，相手言語が分かってしまうと，それが対話に影響することが明らかであったため，相手が発話する時に，ヘッドホンからマスキングノイズを流し，互いに相手話者の音声が直接聞こえないようにした。

[教示・学習効果]被験者には，実験の前に音声翻訳されやすい話し方として，“明りょうな声で”，“短く簡潔に”話すように教示した。さらに，あらかじめ指示された情報を過不足なく相手に伝え，その目的から外れる発話はなるべくしないように指示した。

被験者は一組あたり 20 課題以上繰り返し実験するが，本論文の結果は，10 課題前後実施して，システムに十分に慣れた後のものである。

[課題]課題は，3.1 節で示した三つの場面について，難易度の高いもの，低いものを織り交ぜ，日英，日中で通貨単位を除いて同一のものを実施した。話者は互いが店役と旅行客役を交代で担当する。各課題の制限時間は8分とした。

[翻訳の中止]誤認識されたテキストをそのまま翻訳すれば，誤訳する可能性は高い。そこで，認識結果を発話者自身が確認した後に翻訳する方式も試みた。誤認識があれば，話者は翻訳を中止して，発話しなおすことが可能となる。以後，この方法を手動モード，認識結果の良し悪しに関わらず翻訳する方法を強制モードと呼ぶ。

4. 実験結果

4.1 発話特性，認識率及び正訳率

表1に対話実験で収集した各言語の発話の基本特性，音声認識率，そして翻訳主観評価に基づく正訳率を示す。ここで正訳率とは，発話の元の内容をほぼ過不足なく相手に伝えることが出来ると思われる発話の頻度である。文法の間違いなどは修復可能であると評価者に判断されれば許容する。

手動モードは，話者によって翻訳が中止された発話を除いた結果である。つまり日本語から英語(JtoE)及び中国語(JtoC)では，2割強の発話が捨てられて，単語認識率が 95%を超え，約8割の発話が正しく翻訳された。英語から日本語(EtoJ)では，3割弱の発話が無駄になるが，正訳率は約 15%向上して 76%になる。中国語から日本語(CtoJ)では半分近くも捨てなければならないが，その結果，約5割だった正訳率が 64%となる。

4.2 課題達成率と発話効率

表1より，現状システムの音声翻訳の性能は，JtoE と JtoC が同程度，次に EtoJ，最も低いのが CtoJ ということが分かった。この差は実際のコミュニケーションに影響を与えた。表2は全体としてどの程度の課題達成（情報の伝達）ができたかをまとめたものである。

ここで**課題達成率（TAR）**，**全発話効率（UE_A）**及び**有効発話効率（UE_V）**は，対話システムの評価法[13,14]などを参考に，以下のように定義した。

$$TAR = \frac{TI_T + TI_C}{AI_T + AI_C}$$

$$UE_A = \frac{AU_T + AU_C}{TI_T + TI_C}, \quad UE_V = \frac{VU_T + VU_C}{TI_T + TI_C}$$

ここで，AI は実施した課題の中で相手に伝える必要があった情報の総数，TI は実際に伝えられた情報の総数で，添え字は旅行者役 T と店員 C を意味する（以下同じ）。AU は発話総数，VU は翻訳が実行された発話（有効発話と呼ぶ）の総数である。いずれも対話の冒頭や終了時にあるような定型的な挨拶発話や，明らかな発話の失敗は除いた。

課題達成率 TAR は互いに伝えるべき情報の中でどれだけの情報が制限時間内に伝達出来たかを表す。発話効率 UE は，1つあたりの情報を伝えるのに必要な互いの発話総数の平均を表す。そして，手動モードにおいて，有効発話だけを考慮するのが有効発話効率 UE_V，翻訳が中止された発話を含めた全発話を考慮するのが全発話効率 UE_A である。

日英の手動モードにおいて，課題全体の UE_A が 3.0，UE_V が 2.0 ということは，話者が二人で3発話する間に，1発話の翻訳が中止されて，一つの情報が伝わることを意味する。日中では，一つの情報が伝わる間に4発話弱が費やされ，1.5 発話の翻訳が中止される。日英と日中では，中国語の中止される発話数の多さが，特に手動モードで伝達が必要な情報数が多い場合の課題達成率に大きな影響を与えている。

4.3 正訳発話情報量と発話効率の関係

対話実験を観察していると，話者の組によって，課題を比較的スムーズにこなしていく組と，そうでない組が見られた。その違いは音声翻訳の成否に大きく依存するが，一般に発話に含まれる情報量が多い（発話長が長い）ほど誤認識や誤訳は生じやすい。一方，一つの発話により多くの情報を含めたほうが情報伝達の効率は良くなる。そこで，“正しく伝わったと推定される一発話あたりの情報量”を指標化することを考え，**正訳発話情報量（CUI）**を次式のように定義した。

$$CUI_A = \frac{\sum UI_{Tcorrect} + UI_{Ccorrect}}{AU_T + AU_C}$$

$$CUI_V = \frac{\sum UI_{Tcorrect} + UI_{Ccorrect}}{VU_T + VU_C}$$

AU, VU 及び添え字 T, C は 4.2 節の定義と同じである。$UI\{T,C\}correct$ は正訳と評価された発話の情報量である。真の情報量は不明であるので，言語モデル確率で推定した。その総和を全発話数で割った CUI_A を全発話の正訳発話情報量，有効発話数で割った CUI_V を有効発話の正訳発話情報

表1 対話実験音声の発話特性と認識及び翻訳評価結果

	強制モード				手動モード			
言語方向	JtoE	JtoC	EtoJ	CtoJ	JtoE	JtoC	EtoJ	CtoJ
平均発話長	7.3	7.1	6.3	5.4	6.8	7.0	5.9	5.4
パープレキシティ	25	20	39	91	25	24	31	63
単語正解精度[%]	89	93	78	74	96	97	88	84
発話正解率 [%]	66	75	51	44	82	87	64	55
正訳率 [%]	69	78	60	51	81	80	76	64
翻訳実行率 [%]	-	-	-	-	77	78	71	55

表2 課題の難易度別の課題達成率と発話効率

課題難易度（伝達必要情報数）			<4	4-5	6-8	>8	全体
日英	実施課題数	強制 | 手動	8 | 7	11 | 10	7 | 11	15 | 12	41 | 40
	課題達成率（TAR）[%]	強制 | 手動	87 | 100	94 | 89	98 | 81	79 | 84	86 | 85
	発話効率（UE）	強制 UE_A	3.2	2.3	2.7	2.5	2.5
		手動 UE_A | UE_V	2.6 | 1.8	3.3 | 2.4	3.5 | 2.4	2.6 | 1.7	3.0 | 2.0
日中	実施課題数	強制 | 手動	9 | 7	11 | 12	7 | 11	11 | 16	38 | 46
	課題達成率（TAR）[%]	強制 | 手動	80 | 94	88 | 96	82 | 78	78 | 65	81 | 75
	発話効率（UE）	強制 UE_A	3.8	2.7	3.6	2.9	3.1
		手動 UE_A | UE_V	4.3 | 2.0	3.0 | 2.3	3.6 | 2.3	4.2 | 2.4	3.8 | 2.3

量と呼ぶ。図1(a), (b)は手動モード時の各話者の組の正訳発話情報量と発話効率の関係を示したものである。CUI_Aに全発話効率 UE_Aを，CUI_Vに有効発話効率 UE_vを対応させ，CUI_A-UE_Aを●（日英），▲（日中）で，同様に CUI_V-UE_Vを○，△でプロットし，対応する話者の組のそれらを線で結んである。

縦軸を反転しているため，全体的に右上がりの（反比例）相関が見られ，正訳発話情報量の指標としての妥当性が示されたといえる。線の長さは翻訳の中止頻度に対応する。全体的に日英に比べて，日中のほうが長く，日中の翻訳中止頻度の多さが結果的に全発話効率を低下させていることがこの図からも分かる。

5. 考察

5.1 日英と日中の差

各エンジンの性能差の原因などについては本論文では議論しないが，結果として，中国語の音声認識率が英語と比較して低いことが日中対話の課題達成率や発話効率が日英よりも低くなった最大の要因と考えられる。手動モードにおける話者による翻訳中止率の高さ（英29%，中45%）に加えて，翻訳が実行された発話の正解率も低く（英64%，中55%），それが正訳率の差（英76%，中64%）を招いている（表1参照）。

5.2 強制モードと手動モード

表2を見ると，強制モードと手動モードの全体の課題達成率は日英ではほぼ同じだが，日中では強制のほうが良かった。難易度が低い（伝達が必要な情報数が少ない）場合には，手動モードの“確実性”が勝り，達成率が高いものの，難易度が高くなると，無駄になる発話数が多い手動モードでは，時間切れになる可能性が高まり，達成率が逆転したためと考えられる。

図2は，課題が未達成の要因の詳細を調べたものである。“～未伝達”は，相手に情報が伝わらなかった頻度，“～誤伝達”は誤った情報が伝わった頻度である。“一次～”は，その情報を伝達するために相手からの情報を必要としない場合，“二次～”は，相手からの情報（要望）を正しく受けられなかったために，伝達できなかったり（二次未伝達）誤って伝達したり（二次誤伝達）した頻度である。特に日中の場合が顕著に，強制モード時に誤伝達が多く，手動モード時は未伝達が多いことが分かる。偽の情報が伝わることの弊害を重要視すれば，手動モードを元に，その全発話効率を上げる（翻訳中止率を下げる）方向で性能改善や，インタフェース等の検討を進めるべきだと思われる。

5.3. 正訳発話情報量と発話効率

図1(a), (b)のCUI_V-UE_V（白色プロット）に着目すると，有効発話の正訳発話情報量（CUI_V）は，日英で25あたり，日中の良い組で30を超える程度で，対応する発話効率（UE_V）は，2発話を超える程度までであることがわかる。つまり我々の現状における音声翻訳システムの（有効）発話効率は，たかだか“2発話で1情報”程度と

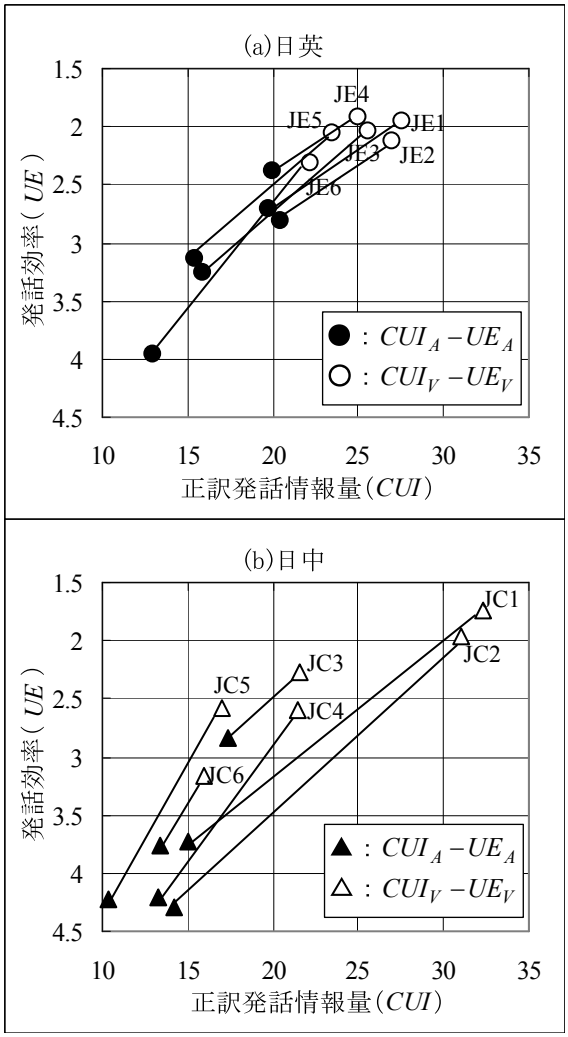

図1　正訳発話情報量と発話効率の関係

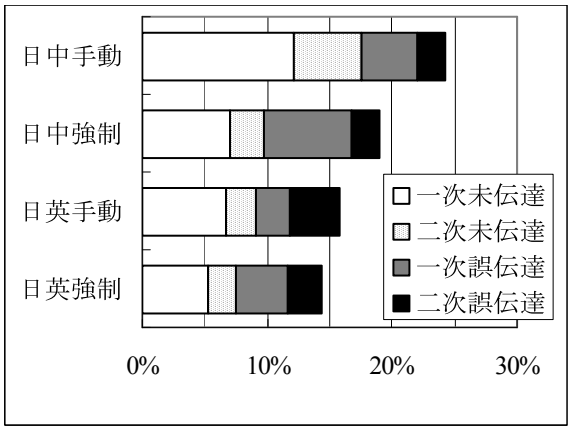

図2　課題未達成要因

いえる。発話種別の頻度を調べてみると，概ね約半分の発話が情報を提示する発話であることが分かった。つまり，一つの情報提示発話で一つの情報伝達が成功しているということになる。発話情報量とその中に含まれる情報の数から，情報一つあたりを伝えるのに必要な情報量を計算してみると，情報の“粒度”でばらつきはある

が，平均すると約 30 となった。日英全体の CUI_V の平均値は約 25 で，それよりも低いが，これは，発話情報量が情報提示発話よりも低い情報要求，了解発話などを含むためで，情報提示発話のみの CUI_V を計算してみると，29 となり，情報一つあたりに費やされる発話情報量にほぼ一致した。つまり，日英対話においては，情報提示発話の CUI_V が約 30 となったことが，“1情報提示発話で1情報”という発話効率を決め，そして発話全体に占める情報提示発話の頻度が約半分だったことで，対話全体の有効発話効率が 2.0 になったと説明できる。また CUI_V が 30 であることは，現状の日英システム性能の一端を示しており，例えば情報量が 40 の発話の正訳率の平均が，日英全体で8割弱であることに起因する。日中では，翻訳実行発話の中国語の認識率が低く，それにより CUI_V が低くなるために，全体の発話効率も低下した。

5.4 話者の話し方の課題遂行への影響

手動モードにおける全発話を考慮した発話効率（UE_A）の差は，誤認識による翻訳中止頻度の大小が大きく影響してくる。つまり，誤認識発話の頻度を減らすように，常に“短く簡潔な” 発話を心がけた組の効率が結果的に高くなっている。中国語の音声認識率が低い日中対話においてそれが顕著に見られる。表3にその例を示す。CUI_V が 30 を超える二組（JC1,2）の発話長は平均よりも長く，発話に含まれる情報量が多いために，正訳率の割に CUI_V が高くなり，結果 UE_v も良い。しかし，誤認識による翻訳中止率も高くなるため，UE_A は大幅に悪くなっている。逆に，唯一日中において UE_A が3発話以下の JC3 は，お互いの発話長がかなり短い。一発話の情報量が減ることで，正訳率の高さの割に，UE_v はそれほど良くない。しかし，翻訳を中止する発話を減らし，少しずつ着実に対話を進めていることで，結果的に課題達成率は高くなった。

表3　日中3組の比較

日|中（*UE*,*CUI*除く）	JC1	JC2	JC3
発話長	9.0 | 6.4	10.1 | 5.6	4.9 | 4.8
単語正解精度[%]	96 | 73	98 | 84	98 | 91
正訳率[%]	90 | 51	76 | 64	95 | 75
翻訳実行率[%]	55 | 40	100 | 29	95 | 68
CUI_V | CUI_A	32 | 15	31 | 14	22 | 17
UE_V | UE_A	1.8 | 3.8	1.9 | 4.2	2.3 | 2.8
TAR[%]	72	78	94

6. まとめと今後

ATR で開発中の日英及び日中音声翻訳システムを使った課題遂行型対話実験を実施し，同システムのコミュニケーション支援能力の評価を試みた。日英と比較すると，日中は中国語の音声認識性能が低く，それが全体の課題達成率を低下させたが，誤認識の影響を除外して考えれば，日英，日中ともに“2発話で1情報”という発話効率であることがわかった。また，一発話を短くし，情報を分割することで，誤認識を減らした話者が効率よく確実に課題を遂行できたことも分かった。

今後は今回の知見などを元に，“適切な発話”へ話者を誘導するための自動教示手法を検討するとともに，より自然なコミュニケーションのための音声翻訳性能の向上も必要であると考えている。

謝辞

本研究は情報通信研究機構の研究委託「大規模コーパスベース音声対話翻訳技術の研究開発」により実施したものである。

参考文献

[1] Yamamoto, S. “Toward Speech Communications beyond Language Barrier - Research of Spoken Language Translation Technologies at ATR -,” Proc. of ICSLP2000, Vol. IV, pp.406-411, 2000.

[2] Takezawa, T. and Kikui, G., “Collecting Machine -Translation -Aided Bilingual Dialogues for Corpus-Based Speech Translation,” Proc. of EUROSPEECH2003, pp. 2757-2760, 2003.

[3] 水島，竹澤，菊井，“翻訳システムを介した対話音声の発話スタイルについて-自然発話，朗読発話との関係-”，第 3 回話し言葉の科学と工学ワークショップ講演予稿集，pp.135-142, 2004.

[4] 水島，竹澤，菊井，“音声翻訳システムを介した対話の評価－誤認識及び誤訳が対話に及ぼす影響－”，言語処理学会第 11 回年次大会，229-332，2005.

[5] 水島，竹澤，菊井，“実環境における音声翻訳システムを介した対話実験－実験室環境との発話スタイルの比較－”，日本音響学会秋季研究発表会，1-P-17，2005.

[6] 伊藤，葦苅，實廣，中村，“音声認識統合環境 ATRASR の概要と評価報告”，日本音響学会秋季研究発表会，1-P-30，2004.

[7] K. Imamura, H. Okuma and E. Sumita, “Practical approach to syntax-based statistical machine translation,” Proc. MT Summit X, pp. 267-274, 2005.

[8] T. Watanabe and E. Sumita, “Example-based decoding for statistical machine translation,” Proc. MT Summit IX, 2002.

[9] Y. Akiba, T. Watanabe, and E. Sumita, “Using language and translation models to select the best among outputs from multiple MT systems,” Proc. COLING, pp. 8-14, 2002.

[10] H. Kawai, T. Toda, J. Ni, M. Tsuzaki, and K. Tokuda, “XIMERA: A new TTS from ATR based on corpus-based technologies,” Proc. 5th ISCA Speech Synthesis Workshop, 2004.

[11] G. Kikui, E. Sumita, T. Takezawa, and S. Yamamoto, “Creating Corpora for Speech-to-Speech Translation,” Proc. EUROSPEECH, pp. 381-384, 2003.

[12] 水島，竹澤，菊井，“翻訳システムを介した音声対話における相手話者音声と翻訳テキスト表示の影響について”，SLP-52,pp.99-106, 2004.

[13] J. Glass, J. Polifroni, S. Seneff, and V. Zue, “Data collection and performance evaluation of dialogue system: The MIT Experience”, Proc ICSLP2000, Vol. IV, pp.1-4, 2000.

[14] 橋田ほか，“DiaLeague-自然言語処理システムの総合評価-”，人工知能学会誌，Vol.12, No.3, pp.390-134, 1997.