< 提言 >

統合的日本手話学習者コーパスの開発構想

小谷 克則^{*1} 吉見 毅彦^{*2} 小谷 志穂^{*1} 井佐原 均^{*3}

^{*1} 関西外国語大学 ^{*2} 龍谷大学

^{*3} 豊橋技術科学大学

kkotani@kansaiidai.ac.jp

1. はじめに

近年、情報通信技術や言語処理技術の発展に伴い、言語資源の一形態であるコーパスも様々な発展を遂げてきた。コーパスの発展の一例として、コーパスに収集される言語の多言語化や言語媒体の拡張が挙げられる。初期のコーパスでは、大規模なデータの収集や処理が容易であるという文字データの利点から、文学作品や新聞記事といった文字を媒体とする言語（書記言語）が主たる対象言語であった。その後、音声データ、映像データの収集・処理技術の進展に伴い、音声や映像を媒体とする言語（音声言語、手話言語）も扱われるようになった（知識情報処理技術専門委員会 2012）。

コーパスの発展は、言語媒体の拡張だけでなく、その利用形態の拡張からも伺える。初期のコーパスは、計量言語学などを中心に、ある言語における形態素や統語構造の分布調査などに利用されていた。近年では、基礎研究だけでなく、情報工学などを中心に、第二言語学習支援システム、機械翻訳システム、危機管理対策支援システムなどの開発研究にも利用されるようになった（安田他 2009, Kotani et al. 2011, Utiyama & Isahara 2003, Neubig 2011）。

手話言語を対象とするコーパス（手話言語コーパス）に限定しても、多言語化、利用形態の拡張といった発展がみられる。まず、手話言語コーパスの多言語化としては、アメリカ手話（Lu & Huenerfauth 2011）を始め、イギリス手話（Schembri 2008）、日本手話（Koizumi et al. 2002, 神田他 2006）、イタリア手話（Geraci et al. 2010）、アラビア手話（Almohimeed et al. 2010）などを対象言語とするコーパスの開発が挙げられる。また、手話言語コーパスの利用形態の拡張としては、言語学的分析の研究資料としての利用（神田他 2006）だけでなく、手話生成システム（Lu & Huenerfauth 2010）や機械翻訳システム（Bungeroth et al. 2008）の開発資料としても利用されるようになった。

手話言語コーパスの次なる課題の一つとして、手話言語話者の多様化への対応が挙げられる。書記言語コーパスや音声言語コーパスでは母語話者を対象とするコーパスだけでなく、第二言語としての学習者を対象とするコーパス（学習者コーパス）も開発されている（Granger et al. 2009, 和泉他 2004, 安田他 2009, Meretus 2010, Kotani et al. 2011）。既存の手話言語コーパスは主に母語話者を対象としてきたが、Mertzani（2011）も指摘するように、第二言語としての手話教育の環境整備の一環として、手話言語を対象とする学習者コーパス（手話言語学習者コーパス）の開発は重要な課題と位置づけられる。

したがって、手話言語コーパスの発展のため、本研究は日本手話を対象とする学習者コーパスの開発に取り組み、統合的日本手話学習者コーパスを提案する。統合的日本手話学習者コーパスとは、学習者の言語産出と言語理解の両方の言語領域における言語使用の処理結果と処理過程の両方をデータとして収集するコーパスである。まず、2節では既存の学習者コーパスを概観し、どのようなデータや情報が収集されてきたかを確認する。3節では2節での既存の学習者コーパスの概観に基づき、日本手話学習者コーパスとして収集すべきデータを提案する。このデータには学習者の言語使用に関するデータだけでなく、学習者の習熟度を示すデータも含まれる。そこで、学習者の習熟度を判定するための言語テストも提案する。4節では本稿のまとめを示す。

2. 学習者コーパスの概要

学習者コーパスとは、学習者の言語使用における様々な言語的特徴の分析に利用されるデータである。学習者コーパスのデータには二種類ある。主たるデータ（一次データ）は、学習者の言語使用を直接的に提示するデータである。一次データには、学習者が発話や作文などの言語使用を通じて生成した、音声データや文字データなどが含まれる。一次データの捕捉的データ（二次データ）として、言語使用の分析に利用される様々な情報データがある。二次データには、学習者の習熟度を示すデータや一次データの解析結果データ（品詞情報や統語情報など）などが含まれる。

本節では、既存の学習者コーパスにおいて収集されてきた一次データと二次データを概観する。一次データの概観においては、どのような言語使用が対象とされてきたかを確認する。二次データの概観においては、どのような指標を用いて学習者の習熟度が示されてきたかを確認する。まず、2.1節では音声言語学習者コーパスを概観し、次の2.2節では手話言語学習者コーパスを概観する。

2.1. 音声言語学習者コーパス

音声言語学習者コーパスでは多言語化が進み、英語を始め、日本語、フランス語、ドイツ語といった様々な言語を対象としてコーパスが開発されてきた。本稿では、

統合的日本手話学習者コーパスの開発構想

英語を対象とする学習者コーパス (Granger et al. 2009, 和泉他 2004, 安田他 2009, Kotani et al. 2011) を中心に概観するが、ドイツ語を対象言語とする学習者コーパス (Meretus 2010) も対象とする。

音声言語学習者コーパス (Granger et al. 2009, 和泉他 2004, 安田他 2009, Meretus 2010, Kotani et al. 2011) は、言語使用の領域 (言語領域) として発話や作文といった言語産出を一次データの対象とし、学習者が生成した音声データや文字データを収集してきた。この一次データに基づき、学習者の言語産出における言語的特徴を分析することが可能となる。分析対象となる言語的特徴には、発話量や語彙の使用頻度といった定量的な特徴と、音声・音韻、形態・統語、意味・語用に関する学習者特有の現象や習熟度に応じた現象といった定性的な特徴がある。音声・音韻に関する特徴は語や句の発音や音韻表象の形成における特徴であり、形態・統語に関する特徴は形態変化や統語構造の生成における特徴であり、意味・語用に関する特徴は文や場面の意味上の制約における特徴である。したがって、学習者コーパスの音声データや文字データを用いると、学習者特有の言語的特徴に基づき学習者の言語技能を言語産出の観点から分析可能となる。

Kotani et al. (2011) が指摘するように、学習者の言語技能は、言語産出だけでなく、言語理解の観点からも分析されるべきである。なぜなら、母語話者であれば言語技能は言語産出と言語理解の両方の言語領域において均衡的に発達すると想定できるが、学習者の場合は不均衡な発達が想定されるからである。例えば、読めても (言語理解における優れた言語技能) 書けない (言語産出における乏しい言語技能) という学習者の事例がある (Ellis 1994, Pawlak 2011)。そのため、Meretus (2010) と Kotani et al. (2011) は聴解や読解といった言語理解における言語技能にも着目し、学習者が聴解、あるいは読解した結果である理解度データを収集してきた。

Kotani et al. (2011) によると、母語話者コーパスは、特に言語領域が明示されないが、厳密には言語産出と言語理解の両方の言語領域を対象としてきた。日常会話やホームページなどの多くの母語話者が生成する音声データや文字データを収集するコーパスは言語産出を対象としている。一方、ニュース放送や新聞記事などの多くの母語話者にとって聴解や読解の対象である音声データや文字データを収集するコーパスは言語理解を対象としている。母語話者の場合は、一定の理解度が想定できるため明示的に理解度データを収集する必要はない。しかし、学習者の場合は、理解度は習熟度に応じて異なるため、明示的に理解度データを収集する必要がある。Meretus (2010) では、理解度データは学習者の作成した読解レポートを評価することで得られる。Kotani et al. (2011) では、理解度データは聴解と読解の理解度を問う設問の解答から正答率を算出することで得られる。

表1 音声言語学習者コーパスの言語領域

	言語産出		言語理解	
	発話	作文	聴解	読解
Granger et al. (2009)		*		
和泉他 (2004)	*			
安田他 (2009)		*		
Meretus (2010)		*		*
Kotani et al. (2011)	*	*	*	*

表1では音声言語学習者コーパスが対象としてきた言語領域を示している。Granger et al. (2009) はエッセイ課題の作文による言語産出を対象とした。和泉他 (2004) は面接課題の発話による言語産出を対象とした。安田 (2010) は翻訳課題の作文による言語産出を対象とした。Meretus (2010) は読解による言語理解、及び、読解レポート課題の作文による言語産出を対象とした。Kotani et al. (2011) は言語テスト課題の発話と作文による言語産出、及び、聴解と読解による言語理解を対象とした。Kotani et al. (2011) における発話の課題は、会話などの学習者が自由に発話を生成する課題ではなく、同じ英文を音読する課題であった。この課題は、音声・音韻に関する特徴を学習者間で直接比較することを目的としている。

音声言語学習者コーパス (Granger et al. 2009, 和泉他 2004, 安田他 2009, Meretus 2010, Kotani et al. 2011) は、発話、作文、聴解、読解における言語処理の結果として音声データ、文字データ、理解度データを収集してきた。これらの一次データに基づき、学習者の言語処理の精度を分析することが可能となる。言語処理の精度とは、処理結果の発音、文、理解の適確さである。

しかし、Kotani et al. (2011) が指摘するように、学習者の言語技能は、処理結果の観点からだけでなく、その処理過程の観点からも分析されるべきである。処理過程を示すデータとは言語処理に要した時間 (処理時間) や言語処理の負荷の程度 (難易度) であり、これらの処理過程を示すデータに基づき、発話、作文、聴解、読解の処理過程の流暢さ、つまり、効率を分析することが可能となる。母語話者であれば言語処理に関する技能が精度と効率の両方の観点において均衡的に発達すると想定できるが、学習者の場合は不均衡な発達が想定されるからである。例えば、適確な文を発しても (高精度) 迅速な対応ができない (低効率)、あるいは、適確な文を発するのは (高精度) 非常に難しく感じる (低効率) という学習者の事例がある (Schmidt 1992)。そのため、安田 (2010) と Kotani et al. (2011) は言語処理の過程にも着目し、処理結果の一次データだけでなく、処理過程の一次データも収集してきた。Kotani et al. (2011) における処理時間データは、一文当たりの処理に要した時間 (100 分の

統合的日本手話学習者コーパスの開発構想

1 秒単位で計測)であった。安田他 (2009) と Kotani et al. (2011) における難易度データは、学習者が一文の言語処理の負荷を五段階により評価したスコア (1-5) であった。

表2 音声言語学習者コーパスの処理段階とデータ要素

	処理結果			処理過程	
	音声	文字	理解度	処理時間	難易度
Granger et al. (2009)		*			
和泉他 (2004)		*		(*)	
安田他 (2009)		*			*
Meretus (2010)		*			
Kotani et al. (2011)	*	*	*	*	*

表2では音声言語学習者コーパスが対象としてきた処理段階とそのデータ形式を示している。Granger et al.(2009)は作文の処理結果を対象に文字データを収集した。和泉他 (2004) は発話の処理結果を対象に文字データを収集した。発話の処理結果は音声データであるが、音声データに対して文字化を施し、文字データへと変換を行った。また、この音声データは一般公開されていないが、音声データから発話に要した処理時間データの抽出が可能であるため、発話の処理過程を対象に処理時間データを収集したとみなせる。このみなしデータは表2において括弧付けで表示する。安田他 (2009) は作文の処理結果と処理過程を対象に文字データと難易度データを収集した。Meretus(2010)は読解と作文の処理結果を対象に文字データと理解度データを収集した。Kotani et al. (2011) は発話、作文、聴解、読解の処理結果と処理過程を対象に音声データ、文字データ、理解度データ、処理時間データ、難易度データを収集した。但し、処理時間データは学習者が処理時間を制御できる発話、作文、読解の言語領域においてのみ収集し、学習者が制御不能な聴解の言語領域においては収集しなかった。したがって、発話、作文、読解では処理時間データと難易度データを収集し、聴解では難易度データのみを収集した。

音声言語学習者コーパス (Granger et al. 2009, 和泉他 2004, 安田他 2009, Meretus 2010, Kotani et al. 2011) は学習者の習熟度を示す情報を二次データの対象とし、学習歴や言語テストの結果を収集してきた。言語テストにはその信頼性と妥当性が認められている言語テスト、例えば、英語学習者が対象の場合は TOEFL (Test of English as a Foreign Language) や TOEIC (Test of English for International Communication) が含まれる。これらの二次データに基づき、学習者の習熟度別に一次データを比較分析することが可能となる。さらに、学習者数が、初級、中級、上級の各習熟度において (ほぼ) 同じであれば、一次データを初級から上級へと習熟度が発達する過程を示

す偽時系列データ (pseudo-longitudinal data) としてみなせる (Granger 2004)。

2. 2. 手話言語学習者コーパス

Mertzani (2011) は、手話言語学習者コーパスの開発が遅々として進捗しない状況において、イギリス手話言語学習者コーパス (British Sign Language Learner Corpus) の開発に取り組んだ。筆者が知る限りにおいても、手話言語学習者コーパスに関連する研究がいくつかある程度である。

まず、手話言語学習者コーパス関連研究として、神田他 (2006) が挙げられる。神田他 (2006) は日本手話学習者対象の言語テストの設問を収集し、テストコーパスを開発した。さらに、テストコーパスの設問に対する日本手話学習者の解答も収集した。したがって、テストコーパスと解答データを合成することが可能であり、合成されたデータは学習者の言語使用を示すという学習者コーパスの基本的な機能を果たすため、本稿では学習者コーパスとみなす。

次に、母語話者の言語獲得過程を示すコーパスに関する研究 (Lillo-Martin & Pichler 2008) が挙げられる。Lillo-Martin & Pichler (2008) は言語獲得過程における統語的特徴の表出 (二語文の生成) を調査する目的で、アメリカ手話母語話者 (月齢 18 ヶ月から 42 ヶ月の 4 名) を対象に、24 ヶ月間、計 210 時間 (一人当たり 16 時間から 79 時間) に渡り日常生活を記録し、コーパスとして開発した。しかし、母語話者を対象とするコーパスであるため、本稿では議論を省く。

その他、Boyes-Braem (2001) や Leeson & Nolan (2008) を始め、手話言語コーパスの対象者として手話学習者が含まれるデータは存在する。しかし、学習者の言語使用に特化した研究ではなく、手話表現のデータ化を目的とすることから、本稿では議論を省く。

表 3 では手話言語学習者コーパスの一次データが対象としてきた言語領域、及び、処理段階とそのデータ形式を示している。神田他 (2006) は単語読み取り課題の言語理解を対象に、その処理結果を示す理解度データを収集した。Mertzani (2011) は表現課題の発話による言語産出を対象に、その処理結果を示す映像データを収集した。また、この映像データから発話に要した処理時間データの抽出が可能であるため、発話の処理過程を示す処理時間データが収集されたとみなせる。このみなしデータは表 3 において括弧付けで表示する。

表 3 手話言語学習者コーパスの対象言語領域と対象処理段階

	言語領域		処理結果		処理過程	
	言語産出	言語理解	映像	理解度	処理時間	難易度
神田他 (2006)		*		*		
Mertzani (2011)	*		*		(*)	

統合的日本手話学習者コーパスの開発構想

手話言語学習者コーパス（神田他 2006, Mertzani 2011）も、音声言語学習者コーパス（Granger et al. 2009, 和泉他 2004, 安田他 2009, Meretus 2010, Kotani et al. 2011）と同様、二次データの対象として学習者の習熟度を示す情報を収集してきた。神田他（2006）は学習者の習熟度を示す情報として手話技能検定の受験区分を用いた。対象の受験区分は6級から3級までの4段階であり、7級、2級、1級は、類型語の学習難易度調査には不适当であったため対象外とされた。Mertzani（2011）は学習者の習熟度を示す情報として学習歴を用いた。この学習者コーパスの目的は初級学習者の言語使用の分析であったため、対象の学習歴は1年未満であった。

3. 日本手話学習者コーパスの概要

本節では、統合的日本手話学習者コーパスの概要を述べる。3.1節ではコーパス開発の目的を示し、3.2節ではその目的を達成するために収集すべき一次データの形式を示す。3.3節では一次データを用いての学習者の言語使用の分析例を示す。3.4節では一次データ収集の問題点とその解決策を示す。3.5節では学習者の習熟度を示す情報として収集すべき二次データの形式を示す。

3.1. 課題と目的

統合的日本手話学習者コーパスは、第二言語としての日本手話言語の言語的分析、並びに教材開発の基礎資料としての利用が主たる目的である。

また、統合的日本手話学習者コーパスの開発において、本研究が取り組むべき課題は以下の三点である。

既存の手話言語学習者コーパス（神田他 2006, Mertzani 2011）は、言語産出、あるいは言語理解のいずれかの言語領域を対象としてきたが、両方の言語領域を統合的に対象とする手話言語学習者コーパスは、筆者の知る限り未開発である。これが本研究が取り組むべき第一の理論的課題である。

また、既存の手話言語学習者コーパス（神田他 2006, Mertzani 2011）は、言語使用の処理結果を対象に映像データや理解度データを収集してきたが、明示的に処理結果と処理過程を統合的に対象とする手話言語学習者コーパスは、筆者の知る限り未開発である。特に、処理時間データは既存の手話言語学習者コーパス（Mertzani 2011）から抽出することは可能であるが、難易度データは未着手といえる。これが本研究が取り組むべき第二の理論的課題である。

さらに、既存の手話言語学習者コーパス（神田他 2006, Mertzani 2011）では、学習者の習熟度を示す指標として学習歴とテストの受験区分が用いられてきたが、これらの指標は手話学習者の資格を示す指標として有効であるが、様々な習熟度の学習者を対象に同じテストを用いて学習者の習熟度を相対的に示す指標としては利用できない。音声言語学習者コーパス（安田他 2009, Kotani et al. 2011）の場合、相

対的に学習者の習熟度を示すテストの結果を習熟度の指標として用いてきた。これが本研究が取り組むべき第三の理論的課題である。

これらの理論的課題を解決すべく、手話言語学習者コーパスの一次データの補完を図る。具体的には、言語産出と言語理解の両方の言語領域を統合的に対象とし、処理結果と処理過程についても両方の処理段階を統合的に対象とする日本手話学習者コーパス、つまり、統合的日本手話学習者コーパスを開発する。統合的日本手話学習者コーパスでは、学習者の習熟度を示す指標として、信頼性と妥当性のより高い指標を開発し、二次データの改良を図る。

これらの理論的課題を解決することを通じて、日本手話研究と日本手話教育における実践的な課題の解決に貢献することも期待できる。日本手話研究における実践的な課題としては、第二言語学習者の言語使用における定量的、および定性的な言語的特徴を示すこと、また、これらの言語的特徴を習熟度に応じて示すことも挙げられる。日本手話教育における実践的な課題としては、教授法の検証支援やコンピュータ支援言語学習システムの開発が挙げられる。

教授法の検証支援の一例としては、学習者の習熟度に応じて定性的な言語的特徴の難易度の体系化が挙げられる。これまで言語的特徴の難易度は教員の経験的知識として集積されてきたため、日本手話教員にとって共有される情報ではなかった。そこで、統合的日本手話学習者コーパスに基づき言語的特徴の難易度を体系化し、共有化を図る。また、この難易度を経験的知識の豊富な教員が検証することで、その経験的知識が具現化されるといった効果も得られる。さらに、この難易度に基づき既存の教授法や教材を検証することが可能となる。

コンピュータ支援言語学習システムの開発の一例としては、学習者の習熟度や教材の難易度を自動推定するシステムの開発が挙げられる。このシステムは、統合的日本手話学習者コーパスのデータを利用して、学習者の習熟度、言語使用における言語的特徴、言語処理の結果と過程を、言語使用の三要素とし、要素間の関係を統計的に推定できる。例えば、これから学習者が読み上げようとする文や文章における言語的特徴とその学習者の習熟度を独立変数としてシステムに入力すると、言語処理の過程として発話に要する時間が従属変数として出力される。このシステムを用いると、システムが推定する発話時間を規定値として、学習者の発話時間がこの規定値から大きく乖離する文を抽出することが可能となる。このような文には学習者にとって未習得の項目などが含まれている可能性が高く、教員による指導が必要な文だといえる。また、独立変数として、学習者の習熟度と発話に要した時間（言語処理の過程）をシステムに入力すると、学習者の習熟度が出力される。つまり、学習者の習熟度を推定するシステムとして利用できる。

統合的日本手話学習者コーパスの開発構想

3. 2. 収集すべき一次データ

統合的日本手話学習者コーパスの一次データは、学習者の言語使用における処理結果データと処理過程データから構成される。言語使用の対象は単語、文、談話の三要素であり、学習者は与えられた課題に応じてこれらの要素を生成、あるいは理解する。言語産出における課題は、与えられた単語や文（刺激）を再現する課題（再現課題）、及び、与えられた指示に従って発話を生成する課題（生成課題）である。言語理解における課題は、提示された単語や文を認識する課題（認識課題）、及び、提示された文や談話を読解する課題（読解課題）である。

課題の特性に応じて、言語使用の三要素が異なる。例えば、再現課題と認識課題は理論的には単語、文、談話の三要素を対象とできるが、統合的日本手話学習者コーパスの場合、対象学習者に初級学習者が含まれるため、単語と文は対象とするが、談話は対象としない。また、読解課題も理論的には三要素を対象とできるが、単語の場合、認識課題との明確な相違点がないため割愛する。表4では統合的日本手話学習者コーパスの一次データの対象要素を課題ごとに示す。

表4 統合的日本手話学習者コーパスの一次データの対象要素

	言語産出		言語理解	
	再現課題	生成課題	認識課題	読解課題
単語	*	*	*	
文	*	*	*	*
談話		*		*

再現課題では、学習者は与えられた刺激（単語、文）を再現する。この再現発話の処理結果を映像データとして、処理過程を処理時間データと難易度データとして収集する。再現課題の刺激は、後述の生成課題と異なり、学習者間で語彙選択や文生成を統一し、さらに、刺激として映像を示すだけでは読み取り能力が関与するため、認識課題と異なり、静止画像上に矢印などの読み取りの支援となる情報を提示し（図1）、学習者の認識能力の影響を最小限に留める。また、刺激として日本語を提示する方法もあるが、語彙選択や文生成といった形態・統語の特徴や意味・語用的特徴の関与が生じるため、この手法は用いない。

生成課題では、学習者は日本語による指示に従って発話（単語、文、談話）を生成する。この生成発話の処理結果を映像データとして、また、処理過程を処理時間データと難易度データとして収集する。

認識課題では、学習者は提示された刺激（単語、文）を認識し、その意味を解答する（選択式）。この認識理解の処理結果を理解度データとして、処理過程を処理時

間データと難易度データとして収集する。但し、認識課題において学習者は処理速度を制御できないため、処理時間については直接測定しない。処理時間の代替データとしては、刺激の提示速度を変え（低速や高速）、学習者が理解可能と判断した速度をデータとして用いる。また、同一刺激を反復して提示し、学習者が理解可能と判断した回数も用いる。

読解課題では、学習者は提示された文や談話を読解し、その内容に関する設問に対して解答する（選択式）。この読解理解の処理結果を理解度データとし、処理過程を処理時間データと難易度データとして収集する。但し、認識課題と同様、処理速度が学習者により制御不能であるため、処理時間については直接測定しない。処理時間の代替データとしては、学習者が理解可能と判断した速度、あるいは回数をデータとして用いる。

図1 再現課題の付記情報

3. 3. 一次データに基づく分析例

統合的日本手話学習者コーパスの一次データに基づき、言語使用における様々な言語的特徴を分析することが可能である。例えば、発話量や頻度といった定量的な言語的特徴や、音声・音韻、形態・統語、意味・語用的特徴といった定性的な言語的特徴が分析対象である。言語使用には言語理解と言語産出が含まれるため、これらの言語的特徴はいずれの言語使用においても基本的に分析可能となる。したがって、以下に示す分析例は基本的にいずれの言語使用にも当てはまる。

再現課題の一次データを利用し、学習者が再現した単語と文の音声・音韻、形態・統語的特徴を精度と効率の観点から分析する。例えば、文の再現課題では統語構造

統合的日本手話学習者コーパスの開発構想

の複雑さ（図2）といった統語的特徴を精度と効率の観点から分析する。統語構造は図2が示すように、単語数（終端記号の数）が同じであっても構造の深さは異なり、より深い構造の方がより複雑であるといえる。この場合、精度は評価者の評価スコア（0.0-1.0）により示され、効率は処理速度（形態素数 / 処理時間（秒））、難易度（1-5）により示される。また、学習者の習熟度とも併せて分析し、精度や効率が習熟度に応じてどのように変化するかを分析する。例えば、統語構造が複雑になると文の再現精度が低下すると考えられ、また、その低下の程度は初級学習者の方が大きいと考えられる。図3では低精度化を示す予想結果を示す。学習者のうち初級Aから初級Eは、中級Aから中級E、上級Aから上級Eと比べ、4層構造での精度と5層構造での精度の乖離の程度が大きい。

図2 統語構造の複雑性

図3 統語構造の複雑性による精度の違い（仮定データ）

再現課題の刺激は後述の生成課題や認識課題の刺激とは異なり、学習者が生成した単語、文、談話を対象に音声・音韻的特徴に特化して精度と効率の観点から分析する。例えば、図4で示すように手指記号の位置や動きを捕捉、可視化（Cooper & Bowden 2009）することにより、学習者間での比較分析が可能となる。

図4 手指動作解析結果例

生成課題の一次データを利用し、学習者が生成した単語、文、談話を対象に定量的言語的特徴と音声・音韻、形態・統語、意味・語用的特徴を精度と効率の観点から分析する。例えば、談話の生成課題における発話量の習熟度別比較、発話量における誤用の割合の算出、発話速度（単語数 / 処理時間（分））の算出、音声・音韻的特徴の精度などを分析する。

認識課題の一次データを利用し、学習者が理解した単語と文を対象に形態素数や単語数といった定量的言語的特徴、及び、音声・音韻、形態・統語、意味・語用的特徴を精度と効率の観点から分析する。例えば、単語の認識課題における、手形、位置、動作、方向性といった形態的特徴における最小対立の理解度が分析対象として挙げられる。また、これらの形態的特徴の複雑性と理解度の関係や、非手指記号の理解度なども分析対象として挙げられる。

読解課題の一次データを利用し、学習者が理解した文と談話を対象に単語数や文数といった定量的言語的特徴、音声・音韻、形態・統語、意味・語用的特徴を精度と効率の観点から分析する。例えば、談話の読解課題において、談話の参照や場面の展開などによる意味的複雑性や談話構造の複雑性と理解度の関係などを分析する。また、これらの複雑性との関係は理解度だけでなく、処理速度や読解効率（処理速

度と理解度の積)との関係からも分析する。

3. 4. 一次データ収集の問題点とその解決策

コーパスは開発することよりも、広く利用されることにより大きな意義がある。すなわち、折角開発したコーパスであっても一般公開されず利用に供される機会が少ないと、その有用性が下がるといえる。

統合的日本手話学習者コーパスの一次データには、言語産出の映像データ、言語理解の理解度データ、言語使用の処理時間データ、及び、言語使用の難易度データが含まれる。これらのデータの中で、Crasborn (2008) が指摘するように、映像データには個人情報の問題が発生する。まず、データ収集に際し、映像データとして自らの顔が収録されることを対象学習者に拒まれるといったデータ収集時の問題がある。次に、仮にデータ収集時に対象学習者の了承を得られた場合は、コーパスデータとして研究利用やその結果を配信する際の個人情報保護といったデータ利用時の問題がある。特に、後者の問題から手話言語コーパスは一般公開が容易でないといえる。

このような個人情報の問題は、手話言語コーパスの発展において解決すべき課題の一つである。書記言語コーパスや音声言語コーパスの場合、個人情報の問題は伏せ字を始め様々な解決策により処理されているため、コーパスの流通が容易となる。したがって、言語資源の流通を担う機関、例えば、言語資源協会 (GSK、日本)、The Linguistic Data Consortium (LDC、北米)、European Language Resources Association (ELRA、欧州) などを通じて、様々なコーパスが流通している。しかし、絶対数の問題もあり、これらの機関を通じて流通している手話言語コーパスは非常に少数であり、ELRA から配布されている Czech Sign Language Corpus と SIGNUM Database の二種類だけである。一方、ELRA から配布されている書記言語コーパスと音声言語コーパスはそれぞれ 100 種類以上、手話言語以外の映像コーパスは 18 種類もある。今後、日本国内では個人情報保護法など法律的な観点も踏まえ、手話言語コーパスの流通経路の整備が期待される。

統合的日本手話学習者コーパスでは、映像データに関する問題を回避する手段として、学習者の言語使用として映像データにより収集される言語産出だけでなく、言語理解も対象として一次データを収集している。また、言語産出においても映像データ以外に処理時間データや難易度データを収集している。仮に映像データの収集や利用が困難な局面に陥った場合でも、映像データ以外のデータを収集し、利用できる。

3. 5. 収集すべき二次データの形式

統合的日本手話学習者コーパスの二次データには、習熟度別比較を行うために、学習者の習熟度を示す情報を含める。学習者の習熟度を示す情報としては、手話技能検定の受験区分と学習歴といった既存の手話言語学習者コーパス (神田他 2006,

Mertzani 2011) と異なり、手話言語テストの結果を指標として利用する。

言語テストは、評価基準により目標基準準拠テストと集団基準準拠テストに分類される (Brown 1996)。前者のテストは目標とする特定の知識や技術の達成を評価基準とする絶対評価であり、後者のテストは受験者の能力をある母集団における相対的な位置として示す。統合的日本手話学習者コーパスの場合、対象とする学習者の習熟度について初級から上級までの相対評価を行う必要があるため、集団基準準拠テストを用いる。

既存の日本手話を対象とする集団基準準拠テストには、手話技能検定、全国手話検定がある。しかし、これらのテストは各習熟度に応じて出題内容が異なり、合否判定によりテストの結果が示される。合否判定によるテストの場合、学習者の習熟度を判定するためには複数のテストを受験する必要がある。例えば、4 級のテストを合格した学習者の場合、3 級のテストを受験した結果に応じて 4 級程度かそれ以上かが判明する。不合格の学習者の場合も同様であり、4 級のテストを不合格の場合は 5 級のテストを受験する必要がある。したがって、統合的日本手話学習者コーパスの開発においては、学習者の習熟度の情報を収集する際、級判定によるテストではなく、スコア判定によるテストを用いる。

第二言語としての英語を対象とするスコア判定による集団基準準拠テストとして TOEFL や TOEIC がある。また、テストの対象者や対象領域が異なるが、手話言語テストにもスコア判定による集団基準準拠テストはある。日本手話文法理解テスト (武居 2010) は、日本手話母語話者の文法理解を対象とする。ドイツ手話テスト (ATG: The German Sign Language test) は、Haug (2011) によると、ドイツ手話の母語話者を対象とする。手話習熟度面接テスト (SLPI: The Sign Language Proficiency Interview) は、Caccamise & Samar (2009) によると、アメリカ手話話者を対象とする。これらのテストを参照にテストを作成する。

また、テストの結果を得点化する方式として、正答数に基づく得点 (NRS: Number-right score) と正答数と設問の難易度に基づく得点 (Logit score) がある。前者の得点は順序尺度であり、得点差や上限・下限付近の天井効果・床面効果の問題があることが知られている (大友 2009)。得点差の問題とは、50 点の学習者と 60 点の学習者の 10 点差と 70 点と 80 点の学習者の 10 点差の等価性が保証されないことである。天井効果・床面効果の問題とは、100 点以上の実力を有する学習者は 100 点に、0 点以下の実力を有する学習者は 0 点に集約されることである。一方、後者の得点は間隔尺度であるためこれらの問題は回避できる。さらに、テストの信頼性や妥当性も前者の得点方式よりも高い (大友 2009)。既存の英語を対象とする言語テストの場合、前者のテストとして英語検定、後者のテストとして TOEFL がある。したがって、統合的日本手話学習者コーパスの場合、間隔尺度のスコアに基づくテストを用いる。

4. おわりに

本稿は、手話言語コーパス発展の一環として、第二言語として日本手話を学ぶ学習者を対象とする統合的日本手話学習者コーパスを提案した。このコーパスには学習者の言語使用に関するデータだけでなく、学習者の習熟度を示すデータも含まれる。そこで、学習者の習熟度を判定する言語テストも提案した。このコーパスは、既存の手話言語学習者コーパスの理論的な課題を解決できるだけでなく、近年大学機関を始め様々な機関において積極的に行われている第二言語としての手話教育への貢献が期待できる。

このコーパスの一次データは、言語産出と言語理解の両方の言語領域を統合的に対象とするだけでなく、処理結果と処理過程の両方の処理段階も統合的に示す。この一次データが示す言語使用は、単語、文、談話レベルにおける課題への解答である。言語産出の課題は再現課題と生成課題であり、言語理解の課題は認識課題と読解課題である。

このコーパスは、二次データとして学習者の習熟度を示す情報、つまり、日本手話のテストの結果を収集する。このテストは、学習者の習熟度を相対的に示す必要があるため、集団基準準拠テストとする。さらに、テストの信頼性や妥当性の観点から、間隔尺度のスコアに基づくテストとする。

今後の課題は、本稿の設計項目に基づく実際のデータ収集が挙げられる。データ収集を行った後、データが学習者の習熟度を適切に反映しているかどうか、その信頼性と妥当性を検証する必要がある。また、データの拡張として、参照データとして母語話者のデータ収集も行うことが挙げられる。

謝辞

本研究の一部は科研費（22300299）の助成を受けたものである。また、本研究は電子情報技術産業協会知識情報処理技術専門委員会言語資源分科会の支援も受けたものである。

参考文献

- Almohimeed, A., M. Wald, and R. Damper (2010). An Arabic Sign Language Corpus for Instructional Language in School. In: Proceedings of 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies, the International Conference on Language Resources and Evaluation (LREC 2010), 81-82. Valletta, Malta.
- Boyes-Braem, P. (2001). A Multimedia Bilingual Database for the Lexicon of Swiss German Sign Language. *Sign Language & Linguistics* 4(1/2), 133-143.
- Brown, J. D. (1996). *Testing in Language Programs*. Upper Saddle River, NJ:

Prentice Hall.

- Bungeroth, J., D. Stein, P. Dreuw, H. Ney, S. Morrissey, A. Way, and L. V. Zijl (2008). The ATIS Sign Language Corpus. In: Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), 2943-2946. Marrakech, Morocco.
- Caccamise, F. C. and V. J. Samar (2009). Sign Language Proficiency Interview (SLPI): Prenegotiation Interrater Reliability and Rater Validity. *Contemporary Issues in Communication Science and Disorders* 36, 36-47.
- 知識情報処理技術専門委員会 (2012)「言語資源分科会」,『知識情報処理技術に関する調査研究報告書』,電子情報技術産業協会,印刷中.
- Cooper, H. and R. Bowden (2009). Sign Language Recognition: Working with Limited Corpora. In: Proceedings of the International Conference on Universal Access in Human-Computer Interaction. Addressing Diversity, Volume 1, 472- 481. Orlando, FL, USA.
- Crasborn, O. (2008). Open Access to Sign Language Corpora. In: Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora, the International Conference on Language Resources and Evaluation (LREC 2008), 33-38. Marrakech, Morocco.
- Ellis, R. (1994). *The Study of Second Language Acquisition*. Oxford, UK: Oxford University Press.
- Granger, S. (2004). Computer Learner Corpus Research: Current Status and Future Prospects. In: In U. Connor and T. Upton (eds.), *Applied Corpus Linguistics: A Multidimensional Perspective*, 123-145.
- Granger, S., E. Dagneaux, F. Meunier, and M. Paquot (2009). *International Corpus of Learner English, Version 2*. Louvain-la-Neuve, Belgium: Presses Universitaires de Louvain.
- Geraci, C., R. Bayley, C. Branchini, A. Cardinaletti, C. Cecchetto, C. Donati, S. Giudice, E. Mereghetti, F. Poletti, M. Santoro, S. Zucchi (2010). Building a Corpus for Italian Sign Language: Methodological Issues and Some Preliminary Results. In: Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies, the International Conference on Language Resources and Evaluation (LREC 2010), 98-102. Valletta, Malta.
- 和泉絵美・内元清貴・井佐原均 (編) (2004).『日本人 1200 人の英語スピーキングコーパス』. 東京: アルク .

- 神田和幸・木村勉・原大介 (2006). 「手話コーパスにおける類型語分類」, 『電子情報通信学会技術研究報告: WIT, 福祉情報工学』, 105(508), 79-84.
- Koizumi, A., H. Sagawa, and M. Takeuchi (2002). An Annotated Japanese Sign Language Corpus. In: Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002), Volume 3, 927-930. Las Palmas, Spain.
- Kotani, K., T. Yoshimi, H. Nanjo and H. Isahara (2011). Compiling Learner Corpus Data of Linguistic Output and Language Processing in Speaking, Listening, Writing, and Reading. In: Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011), 1418-1422. Chiang Mai, Thailand.
- Lillo-Martin, D. and D. C. Pichler (2008). Development of Sign Language Acquisition Corpora. In: Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora, the International Conference on Language Resources and Evaluation (LREC 2008), 129-133. Marrakech, Morocco.
- Leeson, L., Nolan, B. (2008): Digital Deployment of the Signs of Ireland Corpus in Elearning. In: Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora, the International Conference on Language Resources and Evaluation (LREC 2008), 112-122. Marrakech, Morocco.
- Lu, P., and M. Huenerfauth (2010). Collecting a Motion-Capture Corpus of American Sign Language for Data-Driven Generation Research. In: Proceedings of the 1st Workshop on Speech and Language Processing for Assistive Technologies (SLPAT), Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2010), 89-97. Los Angeles, CA, USA.
- Lu, P., M. Huenerfauth. (2011). Collecting an American Sign Language Corpus through the Participation of Native Signers. International Conference on Universal Access in Human-Computer Interaction (UAHCI), 81-90. Orlando, Florida, USA.
- Mertzani, M. (2010). British Sign Language Learner Corpus. [online] Available at: <http://www.bris.ac.uk/deaf/english/research/active/active_05.html> [Accessed 10 April 2012].
- Meurers, D., N., Ott, and R. Ziai (2010). Compiling a Task-based Corpus for the Analysis of Learner Language in Context. In: Proceedings of Linguistic Evidence 2010, 214-217. Tübingen, Germany.

- Neubig, G., Y. Matsubayashi, M. Hagiwara, and K. Murakami. (2011). Safety Information Mining: What Can NLP Do in a Disaster. In: Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011), 965-973. Chiang Mai, Thailand.
- 大友賢二 (2009).「項目応答理論」,『電子情報通信学会誌』92(12), 1008-1012.
- Pawlak, M. (2011). Students' Successes and Failures in Learning Foreign Language Pronunciation: Insights from Diary Data. In Arabski, J., and A. Wojtaszek The Acquisition of L2 Phonology, 165-182. Bristol, UK: Multilingual Matters.
- Schembri, A. (2008). The British Sign Language Corpus Project: Open Access Archives and the Observer's Paradox. In: Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora, the International Conference on Language Resources and Evaluation (LREC 2008), 1-5. Marrakech, Morocco.
- Schmidt, R. (1992). Psychological Mechanisms Underlying Second Language Fluency. Studies in Second Language Acquisition 14(4), 357-385.
- 武居渡 (2010).「日本手話文法理解テスト実用版の開発」,『第4回ことばと教育研究助成研究成果論文集』,博報児童教育振興会, 56-76.
- Utiyama, M. and H. Isahara (2003). Reliable Measures for Aligning Japanese-English News Articles and Sentences. In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL 2003), Volume 1, 72-79. Sapporo, Japan.
- 安田圭志・喜多村圭祐・山本誠一・柳田益造 (2009).「多重タグ付き英語学習者コーパスの開発と英語能力自動測定への応用」,『自然言語処理』16(4), 47-63.

－ 2012 年 6 月 1 日受理、2012 年 11 月 2 日掲載決定－

Development of Integrated Learner Corpus of Japanese Sign Language

**Katsunori Kotani¹, Takehiko Yoshimi², Shiho Kotani¹,
Hitoshi Isahara³**

¹Kansai Gaidai University, ²Ryukoku University,

³Toyohashi University of Technology

kkoatni@kansaigaidai.ac.jp

ABSTRACT

With the goal of advancing research on sign language corpora, the present paper proposes an integrated learner corpus of Japanese Sign Language that not only offers solutions to conceptual problems that have until now remained unsolved, but also contributes to Japanese Sign Language linguistics and education. The corpus has four main characteristics. First, the corpus covers linguistic abilities in the linguistic domains of generating and understanding. Second, the corpus enables examination of linguistic abilities with regard not only to the linguistic output generated by language learners, but also how linguistic output is generated. Examination of linguistic abilities with regard to these aspects reveals both the precision and process efficiency of linguistic output. Third, the corpus enables examination of basic linguistic abilities such as the manipulation of phonological-phonetic, morpho-syntactic, and semantic-discourse properties, because the designed language use situation prevented the use of advanced linguistic abilities including stylistic and rhetorical techniques, as well as non-linguistic abilities in specific academic or business domains. Finally, the corpus allows the development of linguistic abilities to be examined, as it can be regarded as quasi-longitudinal data. The corpus also provides data on the proficiency of language learners measured by a norm-referenced test that yields a scaled score. The proposed learner corpus is expected to serve as a useful language resource for research on Japanese Sign Language.