

完全自動の口頭日本語能力試験の開発

Developing a completely automated test of spoken Japanese

鈴木 正紀† Masanori Suzuki
 原田 康也‡ Yasunari Harada
 伴野 崇夫‡ Takao Tomono

1. まえがき

近年の日本の国際化に伴い、国内に在住する日本語を母語としない人々と、海外において日本企業や関連組織において就労する日本語を母語としない人々が急激に増加している。また、アニメや音楽などから日本語に関心を持ち、日本語の学習を始める海外の青少年も多数存在する。学習や就労を目的とする場合、一定の日本語運用能力が必要であるが、スピーキング能力測定にはこれまで面接試験等が必要であり、多くの人手とコストを必要とした。米国 Ordinate Corporation と早稲田大学情報教育研究所が共同開発中の日本語口頭能力自動測定試験は HMM をベースとした音声認識エンジンとコンピュータによる自動採点システムによって日本語学習者の口頭言語処理能力を完全自動で迅速かつ正確に測定することを目的としている。本稿では、本試験の開発過程と現状、自動テストシステムの仕組みについて報告する。

2. SJT 開発背景

国際化の流れが速まる中で、日本国内に在住する日本語を母語としない人の数が増えている。また、企業においても非日本語母語話者を採用するケースが増加している。また、アニメや音楽などから日本語に関心を持ち、日本語の学習を始める海外の青少年も多数存在する。文化庁[1]の調査によると、2003年現在、日本国内での日本語学習者は135,146人にのぼり、その数は過去最高となった。また国際交流基金[2]の調べによると、海外での日本語学習者の数は2003年現在2,356,745人となり、1979年に比べ約18.5倍増加し、この増加傾向を如実に反映している。また学習や就労には一定の日本語運用能力が必要となる。国際交流基金[3]が実施している2004年度の日本語能力試験では、300,000人以上が受験した。国内外において日本語学習者数は増加傾向にある。

日本語教育を始め外国語教育ではコミュニケーション力養成が授業の中心になって久しいが、従来では学習者のスピーキング力とリスニング力を測定する場合にはOPI (Oral Proficiency Interview)などに代表されるインタビューテストが一般的な方法であった。通常OPIでは、試験開始から終了まで受験者一人につき20-40分程度を要し、訓練を受けた試験官・評価官を必要とする。このため、一定期間に多数の学習者の口頭能力を測定したい場合には、時間・費用・評価者の確保に多大な負担がかかり、運用上の制約がある。また、一定の訓練を受け認定を受けた、採点者も定期的に再訓練を受けねばならぬ

いなど、採点要員の確保には多大なコストを要する。日本語学習者の増加に伴い、インタビューテスト以外に、客観的で迅速な日本語のスピーキング・リスニング能力測定試験への需要が高まっている。

アメリカ・カリフォルニア州にある Ordinate Corporation (www.ordinate.com) は、計算言語学・音声認識などの技術を活用し、外国語学習者が電話で受験することで口頭能力が自動測定される試験プラットフォーム、『Ordinate テストシステム』を開発した。現在このテストシステム上では、英語の口頭能力自動試験 SET-10 (Spoken English Test - 10 minutes) とスペイン語の口頭能力自動試験 SST (Spoken Spanish Test) が運営されている。

Ordinate Corporation と早稲田大学情報教育研究所では、この Ordinate テストシステム上で運用可能で、電話により受験でき、受験結果を音声認識システムを利用し自動採点ができる、完全自動な口頭日本語能力試験 SJT (Spoken Japanese Test) を共同開発中である。現在、項目開発が終了し、プロトタイプとなる SJT を開発した。現在このプロトタイプ型 SJT テストを用い、自動採点システム構築と妥当性検証に必要なデータ収集を行っている。本稿では、Ordinate のテストシステムについて説明し、テスト項目開発の詳細を含め、プロトタイプ型テスト作成まで過程と今後の予定について詳述する。

3. Ordinate テストシステム

SJT では Ordinate の他のテストと同様、受験者は個別化された試験用紙と固定電話 (携帯電話・ワイヤレス子機は不可) が受験の際に必要な。試験用紙の表面には試験の受験方法が説明されており、裏面が試験用紙となる。準備ができた段階で、試験用紙に印刷されている電話番号に電話をかけ、プッシュボタンを使い、印刷されている個別のテスト番号を入力するとテストが開始する。試験は約15分で終了する。現状の SJT はデータ収集の段階にあり、自動採点モデルが構築されていないため、テストは採点されない。SET-10 や SST などの通常の Ordinate のテストでは、試験終了後、通常数分以内で採点は終了し、Ordinate のウェブサイトからスコアレポートが入手可能となる。スコアでは、総合スコアのほか、文章構成、語彙、流暢さ、発音の4つのサブスコアも報告される。スコアは全て20点 (最低) から80点 (最高) の範囲で報告される。試験担当者 (教員・人事担当者) は、自分の担当する受験者のスコアを一覧として閲覧・入手でき、回答音声の一部を実際に聞くことができる。SJT でも最終的には SET-10 と同様なスコアレポートが入手可能になる予定である。

† Ordinate Corporation

‡ 早稲田大学

コンピュータを使用した Ordinate テストシステムは、大きく分けて、テスト配信システムと採点システムの2つの要素から構成されている。第一の要素、テスト配信システムは電話を通して受験者にテスト項目を音声で提示する。これらのテスト項目は予め対象言語の母語話者により録音される。後述の通り、開発中の SJT も日本語母語話者により録音されている。

受験者が試験用紙に印刷されている電話番号に電話をかけると、TDS (Test Delivery System) というテスト配信システムがテストを配信する。日本を含め一部の国々ではローカル TDS が設置されているため、受験者は国際電話をかける必要なく受験できる。日本では、フリーダイヤルの番号で日本全国から受験可能である。このローカル TDS システムが Ordinate 本社にあるテストシステムと交信し、テストの実施を行う。試験終了後、回答音声ファイルはデジタル化され、インターネット経由で本社テストシステムに転送される。これ以降の認識・分析・採点・蓄積などは本社システムが行う。

通常 Ordinate テストシステムに送信された受験者の回答は、第二の要素となる音声認識エンジンとコンピュータ化された自動採点システムにより自動採点される。

Ordinate の音声認識エンジンは、HMM (Hidden Markov Model) に基づいており、母語話者だけでなく、非母語話者 (対象言語学習者) からの音声データも用いて開発されていることが特徴である。こういった採点システムの構築するため、非母語話者の音声データは母語、年齢、語学力などを考慮し、バックグラウンドの異なる話者から収集される。SJT ではこの自動採点システムの開発を行うために現在データ収集を行っている。

収集された音声データのうちの一部は書き起こされ、音声認識エンジンを訓練するために利用され、非母語話者の発話も認識できるよう調整していく。既存の SET-10 では、これまで母語話者、非母語話者合わせて約 5,000 人からデータ収集を行い、そのうちの一部を書き起こし、音声認識エンジンの開発・訓練・調整に利用した。

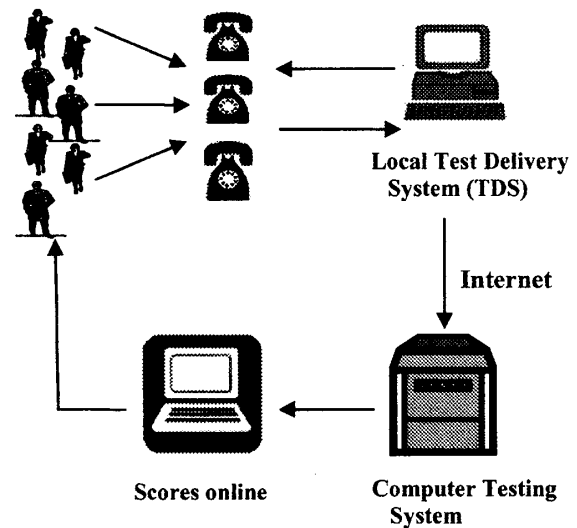
音声認識エンジンでの認識結果は自動採点システムに送られ、受験者の応答にどの単語が使用されたかとともに、応答のペース、なめらかさ、語句や文中で使われた単語の発音の正確さに基づきスコアが算出される。つまり、Ordinate の自動採点システムでは受験者が「何」を「どのように」に言ったか、という2つの側面が判定できる。採点結果はスコアレポートとなり、Ordinate 社の web ページから入手可能となる。前述の通り、通常、試験終了後数分以内にスコアレポートが入手可能となり、総合スコアのほか、文章構成、語彙、流暢さ、発音といった4つのサブスコアも報告される。SJT では次セクションで詳述する通り、日本語用自動採点システムを構築することが必要である。

図1はこの Ordinate テストシステムの流れを図式化したものである。

Ordinate の試験はコンピュータにより自動採点されるため、客観的であり、同時に高い信頼性も得られている。この共通のプラットフォームを利用した SET-10 と SST では、それぞれ 0.97[4] と 0.96[5] という高い信頼係数が得られることが Ordinate が行った研究から実証されている。

またインタビュー試験などの他のスピーキング試験とも高い相関関係が得られることも分かっている。詳しくは[4]と[5]を参照されたい。

図1. Ordinate テストシステム



4. 開発過程

Ordinate テストシステム上で日本語口頭能力試験 SJT が運用可能となるには、様々なステップが必要となるが、大きく3つに分けることができる。

- 1) テストデザイン・問題作成
- 2) データ収集
- 3) データ分析・妥当性の検証

以下ではこれらを詳述する。

4.1 テストデザイン・問題作成

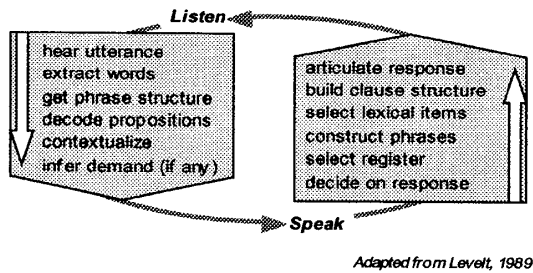
テスト構成概念

SJT では受験者の "Facility in spoken Japanese" を測定を目的としている。"Facility in spoken Japanese" とは「日常生活一般の話題について、日本語母語話者 (もしくは上級の日本語学習者) と同様なペースで日本語の会話を理解し、明瞭に受け答えすることができる能力」のことである。日常会話に効率良く参加するためには「聞いたことを実時間で理解し、応答する」という心理言語学的な基礎言語処理能力が必要となる。具体的には、話者は発生された単語・文を聞き取り、瞬時に意味を抽出し、そして聞かれていることに関連した内容で明瞭に応答する、ということである。Levelt[6]の発話モデルを改良し、図式化したものが図2である。

SJT ではすべての設問において「音声による指示または提示とそれに対する回答」という課題になっている。受験者はこうした指示または提示を聞き、理解し、答える必要がある。この「聞いて→話す」という課題に対処するためには、上記で示される心理言語学の視点からの、「実時間での言語処理能力」が必要になる。言い換えると、いかに文法構造の解析などに注意 (Attention) を使わずに、言

語処理能力を駆使できるか、つまり自動化(Automaticity)を測定することになる。

図2. 口頭言語処理モデル



Adapted from Levell, 1989

テストの構成

7つの問題がSJT用に作成され、データ収集の段階で使用されるプロトタイプ型SJTでは、作成した7つの全ての設問項目が使用されている。データ収集後のデータ分析の結果を基に最終的な設問項目数、各設問項目内に

表1 データ収集用試験問題

設問項目	問題数
パートA: 音読	8
パートB: 復唱	16
パートC: 反対語	8
パートD: 質問	16
パートE: 文の構築	10
パートF: 話の要約	2
パートG: 自由回答式質問	2
合計	62

おける問題数などを決定していく。表1に、データ収集用の試験用紙に含まれる問題をまとめた。パートAからパートGで構成され、音読、復唱、反対語、質問、文の構築、話の要約、そして自由回答式質問という設問項目で構成される。パートAの音読の問題だけ、実際の項目がテスト用紙に印刷される。全部で12文印刷され、各文には番号が振られ、受験者は電話からの音声で指示された番号の文を順次読み上げていく。各問題文では漢字が使用されているため、全ての漢字にはひらがなで振り仮名をつけ、漢字をあまり学習していない受験者でも文を読み上げられるように配慮してある。パートBからパートGまでは全ての問題が完全に音声のみで提示され、試験用紙には各設問項目の説明と例題だけが印刷されている。

問題作成

テスト構成概念にあるように、SJTでは受験者の日常生活一般の話題についての口頭能力の測定を目的としている。このため、各項目で使用される単語、文体、話題は日常会話で一般的に良く使用されるものを用いている。文体の面

では、日本語の会話体では67%の割合で主語を顯示しない[7]などの先行研究などにに基づき、主語を顯示しない項目も含んでいる。

また、日本語の日常会話では、母語話者であれば状況や話し手などにより敬語を使い分ける[7][8]。日本語の日常での言語会話を反映させるため、敬語を用いた項目も作成した。日本語での日常言語活動を反映するようなバランスの良い項目作成を心がけた。

項目作成後には、日本語言語学・日本語教育を専門とする日本在住の言語学者、米国で外国語としての日本語を教えている日本人日本語教師などに項目レビューを依頼した。基本としている言葉はいわゆる標準語である。使用されている単語、表現が日本語母語話者が日常生活で使用するものかどうか、日本語として適切で正しい表現かどうか、特定地域に固有の表現や日本文化に関する単語・表現がないかどうか、などの確認を依頼した。専門家からのコメントを基に、必要に応じ項目を修正し、約1,800から成る項目バンクが出来上がった。

項目の中には日本語母語話者同士での日常会話で頻繁に使用されるが、日本語の授業では必ずしも教えられていないと思われる単語もいくつか含まれる。「携帯」や「パソコン」などがその一例である。米国在住の日本語教師に項目レビューを依頼した際、これらの単語は米国では「携帯電話」、「コンピュータ」として教えているところが多いと予想されるため、使用しないほうが良いのではというコメントがあった。しかし、このSJTは運用能力試験(proficiency test)であり、到達度試験(achievement test)ではないという理由と、実生活で使用されている自然な日本語(Authenticity)の観点からこのような単語も含めることになった。

項目作成後は、日本語母語話者複数名により、録音された。実際の言語活動では標準語を話していても方言に影響され、アクセントがある標準語を話す人も多い。このような実際の言語状況を反映させるため、東京、千葉、大阪、広島、沖縄出身の人により項目は録音され、アクセントは許容範囲とした。男性、女性両方の声でも録音された。

データ収集

現在の開発状況は冒頭でも述べた通り、電話上で受験可能となっているが、自動採点はされていない。現在Ordnateと早稲田大学情報教育研究所では音声認識エンジンと自動採点システムの開発に必要なデータ、また妥当性検証のためのデータなど、一連のデータ収集を行っている。日本語非母語話者の発話をできるだけ正確に認識、採点できる音声認識エンジンと自動採点システムを開発するためには、母語話者だけではなく非母語話者から大量にデータを収集することが重要である。母語話者に関してはできるだけ東京だけでなく、様々な出身地の人たちからデータを集めることを心がけている。現在のところ、早稲田大学在籍中の学部生、大学院生、留学生などからデータ収集を行っている。また、日本のその他の大学に在籍する日本人学生や留学生にも現在協力を要請中である。母語話者からのデータは現時点で約150名程度に達している。

現時点では非母語話者のデータは日本在住の非母語話者から主に収集しているが、韓国、アメリカなど、日本国外にいる日本語学習者からも徐々にデータ収集を行っていく予定である。また、年齢、性別、母語、語学力などを考

慮に入れ、様々な非母語話者からデータを収集していくことも重要な点である。

Ordnate のテストシステムでは、受験者の解答は音響モデルや応答モデルなどに照らし合わせ統計的に処理され、一番確率の高いものが選ばれ、受験者が何をどのように答えたのかが判断される。現在 Ordnate の音声認識エンジンと自動採点システムは、英語とスペイン語にしか対応していないため、日本語用に開発しなければならない。収集されたデータのうちのかなりの量は日本語母語話者により書き起こしされ、回答音声データとともに音響モデル、発音辞書、予測反応ネットワークなどを構築し、非母語話者の発音にも適合した音声認識エンジンを開発・調整していく。

また、自動採点システムの開発のため、収集された回答データの一部は、評定者によりまず人為採点される。評定者は「流暢さ」と「発音」に関して採点基準が与えられ、その基準に基づいて各応答を採点していく。この採点結果を利用し、自動採点システムアルゴリズムを構築し、自動採点システムが人間の評定者と同程度の評価ができるように調整していく。

一部の非母語話者には SJT の他、口頭能力試験として一般的に広く認知されている ACTFL-OPI (OPI: Oral Proficiency Interview) といったインタビューテストを受験してもらうことで、併存的妥当性 (Concurrent Validity) 検証のためのデータとなる。

データ分析・妥当性検証

最後のステップでは、データ収集で得られたデータを基に様々な分析を行っていく。まず、項目応答理論 (IRT) を用いて、各項目の項目困難度を算出していく。また、最終的な項目バンクに残る項目は 80-90% 以上の母語話者が正解を答えられたもののみである。回答データを基に、この基準に照らし合わせ、不適切な問題を明らかにし、最終的な項目バンクから排除していく。また、総合スコアと4つのサブスコア (文章構成、語彙、流暢さ、発音) に関して個々に信頼係数を算出する。また、人間の評定者と自動採点システムとの採点結果を比較し、自動採点システムによるスコアが人間の評定者が与えるスコアとどのような相関関係があるかも検証していく。上述の通り、データ収集の段階で一部の非母語話者には ACTFL-OPI などのインタビューテストを同時に受験してもらうが、そのインタビュースコアと SJT の得点とを比較し、SJT とインタビューテストの間にもどのような関係があるが、つまり併存的妥当性の検証も行っていく予定である。

5. まとめ

本稿では、アメリカにある Ordnate Corporation と早稲田大学情報教育研究所で共同研究・開発・検証を進めている完全自動の日本語口頭能力試験 SJT の開発背景と開発過程を中心に述べてきた。自動認識・自動採点を可能とする技術の中核は音声処理システムであり、これを開発するためには日本語の母語話者からだけでなく、初級者から上級者まで様々なレベルの非母語学習者からも大量に音声データを集めることが必要である。現在そのためのデータ収集を行っている段階である。ここで収集されたデータを基にこのテスト自体の信頼性と妥当性を検証していくことが今後の課題である。

6. 謝辞

本研究開発費の一部は、文部科学省科学研究費萌芽研究 (課題番号16652040) の助成を受けている。

本稿で言及した SET-10 について、早稲田大学情報教育研究所では2000年度より継続的に実証実験を行っているが、1999年から2002年に実施されたKDD株式会社 (現KDDI株式会社)・株式会社KDDIコミュニケーションズ (現株式会社KCOM)・株式会社KDD研究所 (現株式会社KDDI研究所) と早稲田大学メディアネットワークセンターの共同研究生涯学習サポートシステムにおけるネットワーク利用環境技術に関する研究のサブテーマのひとつ「コンテンツ動的作成システムを利用したネットワーク上での自習環境の試作と学習効果の検証」の一環として実施されたものである。また、SJT の開発に関わる音声データの収集にあたって、MNC=KDDI 共同研究による支援を得ている。

文 献

- [1] 文化庁文化庁国語科, “平成15年度国内の日本語教育の概要”, 2004. http://www.bunka.go.jp/laramasi/15_kokunai_nihongokyouiku.html
- [2] 国際交流基金, “海外の日本語教育の現状—日本語教育機関調査・2003年—概要”, 凡人社, 2004.
- [3] 国際交流基金, 日本国際教育協会, “日本語能力試験結果の概要, 2004(平成16年度)”, 2005.
- [4] Ordnate Corporation, *SET-10 Test Description & Validation Summary*, Menlo Park, CA, 2004.
- [5] Ordnate Corporation, *Spoken Spanish Test: Description & Validation Summary*, Menlo Park, CA, 2004.
- [6] W. J. M. Levelt, *Speaking: from intention to articulation*, MIT Press, Cambridge, MA, 1990.
- [7] 水谷信子, “続日英比較話しことばの文法”, 東京, くろしお出版, 1999.
- [8] S. K. Maynard, *Japanese communication-language and thought in context*, University of Hawaii, Honolulu, 1997.