Developing an Automated Test of Spoken Japanese

Yasunari Harada
Institute for Digital Enhancement of Cognitive Development, Waseda University
1-6-1 Nishi-Waseda, Shinjuku-ku,
Tokyo, Japan 169-8050
harada@waseda.jp

Masanori Suzuki
Ordinate Corporation
1040 Noel Dr. Suite 102,
Menlo Park, CA 94025, U.S.A.
msuzuki@ordinate.com

Jared Bernstein
Ordinate Corporation
1040 Noel Dr. Suite 102,
Menlo Park, CA 94025, U.S.A.
jared@ordinate.com

Abstract

In order to assess spoken skills of learners of Japanese effectively and more efficiently the Institute for DECODE (Institute for Digital Enhancement of Cognitive Development) at Waseda University is collaborating with Ordinate Corporation to develop and validate an automated test of spoken Japanese, SJT (Spoken Japanese Test). The SJT is intended to measure a test-taker's *facility in spoken Japanese*, that is listening and speaking skills in daily conversation, in a quick, accurate and reliable manner. In this paper, we discuss the purposes for developing the SJT, the mechanism of a fully automated test, and the test development processes, including item development and implementation.

1 Introduction

According to the Japan's Agency for Cultural Affairs, in 2002, the number of learners studying Japanese as a second language in Japan was 126,350. This is twice as many students as 10 years ago. Similarly, the Japan Foundation (2003a) reported that a little over 2 million people learned Japanese outside Japan in 2003, which is 18.5 times more than in 1979. The Japan Foundation also administers the Japanese Language Proficiency Test. Approximately 227,000 people took the test in 2001, which is about 4 times more than the number of test-takers in 1996 (The Japan Foundation, 2003b).

Currently, one of the major focuses of language instruction is to enhance learners' ability to communicate, that is, to enhance their oral communication skills. Therefore language assessment should emphasize the competent use of language in spoken communication. Oral Proficiency Interviews (OPIs) are often viewed as assessments well-aligned with this goal. However, administering OPIs is time-consuming and often expensive because each interview may take 20-40 minutes and must be administered and scored by human raters. With the rapid increase of students learning Japanese, there is a growing need for a quick but reliable and accurate assessment instrument in the field of teaching Japanese. However, at present, no such test exists.

Ordinate Corporation, a language testing company in California, develops fully automated tests that measure the speaking and listening skills of non-native speakers. The Ordinate testing system is currently delivering tests that measure the spoken language skills of non-native speakers of English and non-native speakers of Spanish. A series of studies has proven that the both tests are highly reliable (The reliability of SET-10 (the Spoken English Test), is 0.97 and the reliability of SST (the Spoken Spanish test) is 0.96). Building on Ordinate's existing testing system, Ordinate Corporation in the U.S. and the

Institute for DECODE at Waseda University in Japan are collaborating to develop a fully automated test of spoken Japanese, the Spoken Japanese Test (SJT).

In this paper, we first describe Ordinate's testing system, in general, including the test development processes including test construct, then we describe the structure of the SJT test, and item development, data collections, and validation. Note that we refer to Ordinate's existing tests such as SET-10 and SST (existing English and Spanish tests) to provide more concrete descriptions, as necessary.

2 Ordinate's Testing System

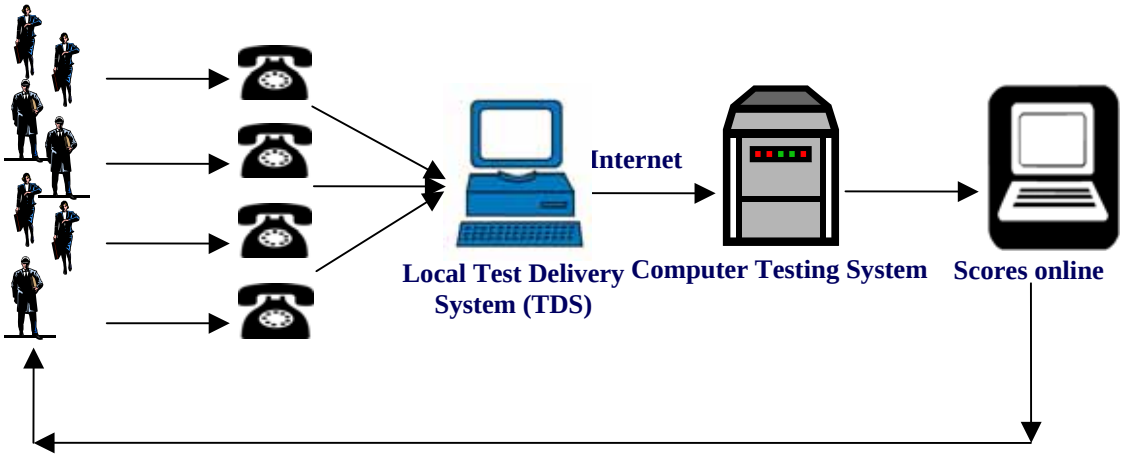
2.1 Ordinate's Test Administration

The SJT test will follow the same test administration procedures as the other Ordinate tests. Ordinate's tests are administered over a telephone by Ordinate's testing system. Prior to taking a test, a test-taker receives a test paper. One side of it has general test instructions and the other side has a unique Test Identification Number, a telephone number, the verbatim spoken instructions, and examples of tasks and items. When a test-taker is ready to take the test, the test-taker calls a telephone number on the test paper. Then, to begin the test, the test-taker is asked to enter the Test Identification Number printed on the test paper using the telephone keypad. Ordinate's tests take approximately 10-15 minutes. . The system presents a test-taker with a series of spoken prompts in the target language (e.g. Japanese), and the test-taker responds by speaking. For example, SET-10 takes approximately 10 minutes to complete. SJT will also be 10-15 minutes to complete.

A score report becomes available on Ordinate's website usually within a few minutes after a test has been completed. For example, both the SET-10 and the SST score report consist of one Overall score and four subscores: Sentence Mastery, Vocabulary, Fluency, and Pronunciation. In other words, the SET-10 measures two aspects of the spoken skills: *what* the test-taker said and *how* the test-taker said it. Sentence Mastery and Vocabulary are the *what* aspects of the scores and Fluency and Pronunciation are the *how* aspects of the scores. The scores are reported in the range of 20-80 and each aspect counts for 50% of the Overall score. A SJT score report will be similar to the SET-10 and the SST score report.

These Ordinate's general test administration procedures are schematized in Figure 1. These procedures will be applied to SJT as well.

Figure 1.



2.2 Key Components of Ordinate's Testing System

The Ordinate testing system is comprised of three key components. Test administration is performed by the first key component, the *Test Delivery System*. The test delivery is done over the telephone and via the Internet. As described above, each test-taker calls into the Ordinate testing system, listens to spoken prompts and answers them appropriately over the telephone. The test-taker's responses are stored in Ordinate's database system. In some countries such as Japan, Korea, China and some European countries, a local TDS (Test Delivery System) is set up and test-takers in those countries take tests using a local toll-free number. Test-takers' responses first go to the TDS and then are sent via the Internet to the Ordinate testing system for scoring.

The second key component is *the automated scoring system using automated speech recognition (ASR) and other computer algorithms*. Ordinate uses an HMM-based ASR. For example, in SET-10, one of the characteristics of Ordinate's speech recognizer is that it is trained to recognize speech not only from native speakers of English but also from non-native speakers. Ordinate's speech recognizer has been trained with a diverse sample of non-native speakers and optimized for various types of non-native speech patterns. Each incoming response is recognized automatically and the words, the pauses, the syllables, the phones, and even some subphonemic events are identified automatically and extracted from the recorded signal for measurement.

These recognition results are fed into the computer scoring system. The computer scoring system examines the two aspects of the speech: *what* the speaker said and *how* the speaker said it. The content of the response is scored according to whether the test taker used expected words in the correct sequence. The manner-of-speaking aspect is calculated by measuring the latency of the response, the rate of speaking, the position and length of pauses, the stress and segmental forms of the words, and the pronunciation of the segments in the words within their lexical and phrasal context. These measures are scaled according to native and non-native distributions and then combined so that they optimally predict human judgments. Because a machine, rather than a person, performs the scoring, scores can be objective and relatively free of biases and other artifacts of the human scoring process such as rater fatigue.

The third key component is *test equation using IRT (Item Response Theory)*. For example, the SET-10 test items are presented in a stratified random order so that the item difficulty generally increases over the sequence of items presented. The item difficulty for each item was calculated using IRT after the data collections conducted for SET-10. These items are assembled into tests from a larger item pool, so the likelihood of one particular test-taker seeing the same items over different test administrations is low. Each assembled test covers about the same range of item difficulty as measured by IRT.

These three key components will also be the key components in SJT. However, these components are currently available only for the English test and the Spanish test. These will need to be developed for the SJT following the steps described in the subsequent sections of this paper.

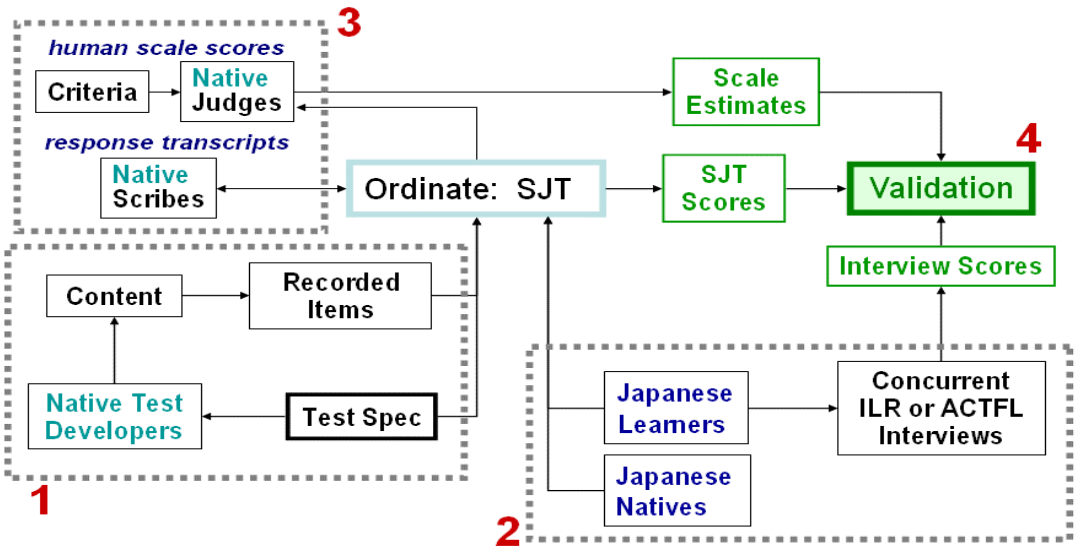
3. Test Development Process and Status

Developing a new test for automatic scoring on the Ordinate system requires the following:

(1) content development
(2) normative data collection
(3) data preparation, and
(4) validation analysis

These four processes are schematized in Figure 2.

Figure 2.



3.1. Content Development

3.1.1 Test Construct

The SJT is intended to measure *Facility in the spoken Japanese* – the ability to understand spoken Japanese on everyday topics and to respond intelligibly at a native-like conversational pace. There are many basic elements required to participate in a spoken conversation: a person has to track what is being said, extract meaning as speech continues, and then formulate and produce a relevant and intelligible response. These component processes of listening and speaking are schematized in Figure 3, adapted from Levelt (1989). All the test items are presented orally. Test-takers need to understand them and answer them intelligibly. Each of these *listen-then-speak* items requires real-time receptive and productive spoken language forms. The SJT is designed to measure the test taker's control of these core language processing components in real time. In other words, the SJT is designed to probe the psycholinguistic elements of spoken language performance rather than the social, rhetorical and cognitive elements of communication.

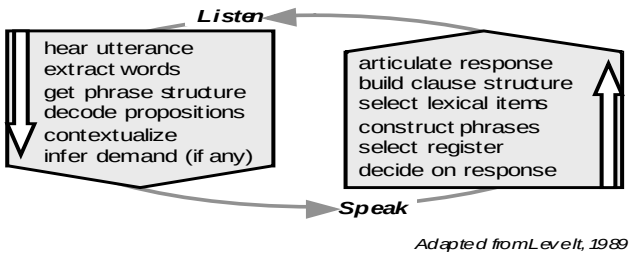


Figure 3: Conversational Processing Components in Listening and Speaking

3.1.2 SJT Test Structure

Seven tasks have been developed for the SJT following a similar format to that of the SET-10 and SST. The seven tasks are Reading, Repeat Sentences, Opposites, Short Answer Questions, Sentence Builds, Open Questions, and Story-Retellings. After collecting and analyzing the data, the final set of tasks and the final number of items to be presented in each of the tasks will be determined. Table 1 shows the seven tasks developed for the SJT.

Table 1. Tasks to be presented in data collection
Part A: Reading
Part B: Repeat Sentences
Part C: Opposites
Part D: Short Answer Questions
Part E: Sentence Builds
Part F: Open Questions
Part G: Story-Retellings

In Part A, test-takers are asked to read sentences at random from among the printed sentences on the test paper. In Part B, test-takers repeat sentences verbatim as they hear them. In Part C, test-takers are presented with a Japanese word (orally) and are asked to respond with a word that represents an opposite meaning. In Part D, test-takers are presented with a series of questions and they answer each question with a single word or a short phrase. Part E requires test-takers to make a reasonable sentence out of three short phrases that they hear. In Part F, test-takers hear a spoken prompt in Japanese asking for an opinion, and they provide an answer with an explanation in Japanese. In Part G, test-takers listen to a very short narrative and then are asked to re-tell what happened in their own words.

The Open Questions and Story-Retellings will not be part of the automatic scoring. The responses to these two parts allow score uses or test administrators to listen to test-taker's spontaneous speech.

3.1.3 Item Development

SJT test items were developed by native Japanese item developers. As described above, the SJT is intended to measure the ability to understand spoken Japanese on everyday topics and to respond intelligibility at a native-like conversational pace. In general, the vocabulary and sentence structure used in the SJT reflect common everyday Japanese. The vocabulary is selected with reference to the CALLHOME corpus of spontaneous spoken dialogue available from the Linguistics Data Consortium (LDC) at the University of Pennsylvania.

Some of the characteristics of the Japanese spoken language were taken into consideration in terms of sentence structure. For example, some SJT test items do not have a subject because, as Mizutani (2001) claims, 67% of the spoken Japanese utterances do not have a subject when a conversation takes place between two participants.

In addition, Sakata, Shinya, and Moriya (2003) assert that the difference in the use of sentence-final particles by male or by female is a structural characteristic of the Japanese language. Ide and Yoshida (2002) arranged sentence-final particles into male-favored particles (e.g. zo, ze) and female-favored particles (e.g. noyo, wa) depending on frequency of their use by men and women. Some of the sentence-final particles were found to be used almost equally by men and women. In SJT, the

sentence-final particles that are mostly used either only by male or only by female are excluded from test items. Only the gender neutral particles (e.g. yone, ne) were selected.

In daily conversation, speakers of Japanese use different types of honorifics depending on a number of factors such as age differences, relationships between the speaker and the listener, social status, settings, and so on (Maynard, 1997; Ide & Yoshida, 2002). Because the SJT test is intended to measure the daily conversational ability in spoken Japanese, the test includes items that use honorifics in order to reflect the language used in daily conversation. According to Marynard (1997), the plain polite form (-desu, -masu) is the most common honorific form observed in daily conversation. In addition, the -desu and -masu forms are taught by many of the textbooks for beginning learners of Japanese such as ***Genki*** published by *The Japan Times* and ***Japanese for Busy People I*** published by *the Association of Japanese-Language Teaching*. Therefore, SJT test items employ the -desu and -masu forms more than any other forms of honorifics. However, the other forms of honorifics, "respectful" form (Sonkeigo) and "humble" form (Kenjogo) and also informal sentence-finals (e.g. -da), were also used to reflect reality of language use. In summary, care was taken to create a balanced set of items that are common forms in daily conversation during item development.

The items developed for the SJT test have been reviewed by linguists in Japan and by two teachers of Japanese as a second language in the U.S., including an American teacher of Japanese as a second language. These reviewers were asked to examine the items to see if they conform to conversational Japanese of educated native speakers and use natural expressions. Reviewers were also asked to identify any test items that use expressions that are specific to only certain areas of Japan. Test items were modified based on the reviewers' comments as necessary. So far, about 1,800 items have been drafted and reviewed for development.

3.2 Normative Data Collection

The next step in the development process is to collect normative data from both native speakers of Japanese and non-native speakers of Japanese. As described as the second key component, Ordinate's testing system uses ASR and computer algorithms for automatic scoring. Ordinate's testing system computes the best hypothesis of what the test taker said, given the response model and acoustic models used for recognition. The ASR system uses statistical methods to formulate the best hypothesis. In order to properly develop such an ASR system for SJT, a sufficient amount of data has to be collected both from natives and non-natives. The data collections are expected to start by December 2004 and be finished in the spring of 2005.

Native data will be collected in Japan. The goal is to collect data from different regions of Japan such as the Kansai area, Tohoku area, etc. The collected data will allow us to see if the items can be answered correctly by natives regardless of their geographical differences. In addition, the data collection will provide us with different dialect and pronunciation samples. This is important because "Japanese consists of a number of different dialects and they have their own phonetic peculiarities." (Haraguchi, 2002, p.1)

The data from non-native speakers will be collected both from non-native speakers learning Japanese in Japan and from non-native speakers learning Japanese outside of Japan such as in China, Korea, the United States, and so on. Non-native data will include various first language backgrounds and different proficiency levels. In addition, the non-native sample will include both genders and various age groups.

The speech samples collected from native speakers as well as non-native speakers of Japanese will be used to develop, train, and optimize the speech recognizer specifically for the SJT test. To do this, the collected data will be transcribed by human transcribers who are native speakers of Japanese. The transcriptions and speech data will also enable Ordinate to develop acoustic models, response models,

pronunciation dictionaries, and expected-response networks that underlie the automatic response recognition and scoring processes.

The data collection is important in ensuring that the items are intuitive to native and advanced non-native speakers and that they can be answered correctly. For an item to be retained in the final item pool, it has to be understood and answered correctly by at least 80% - 90% of a reference sample of educated native speakers of Japanese. In addition, correct answers for some tasks such as Opposites and Short Answer Questions are pre-defined by the item writers and item reviewers. This data collection will allow us to ensure that the pre-defined correct answers are indeed the most common answers provided by virtually all of native speakers as well as by high proficiency non-natives.

3. 3 Data Preparation

In addition to collecting normative data from native and non-native speakers, for concurrent validation purposes, a subset of non-natives will be asked to take other well-accepted oral proficiency interview tests such as ACTFL-OPI (American Council of on the Teaching of Foreign Languages-Oral Proficiency Interview). These human-rated scores will be compared with machine-generated scores from SJT in the next step (Step 4: Validation Analysis).

A set of human raters will be trained to produce consistent ratings for fluency and pronunciation. They will be asked to assign scores to each of the responses they hear in terms of fluency and pronunciation. The raters will be provided scoring rubrics developed by Ordinate. These criterion-referenced scores will be used to train the automated scoring system which will optimally predict the human ratings.

3.4 Validation Analysis

After norming data have been collected, a series of validation analyses will be conducted. One of the analyses to be conducted will be test reliability. Reliability will be calculated for the Overall score as well as for the subscores.

Another analysis planned will be a comparison of the performance of native versus non-native speakers. We expect that native speakers obtain high scores on the SJT test, while non-native speakers of Japanese will be distributed over a wide range of scores. The test results are expected to show effective separation between native and non-native test-takers. If this expected distinction between the two known population holds, it will support the SJT test's validity.

Some human raters will listen to spontaneous responses to open questions and story-retellings and assign scores to these responses using a set of scoring criteria such as CEF (Common European Framework). These estimated scores of test-takers' responses will be compared with their machine-generated scores to see how highly correlated they are.

Finally, we will conduct concurrent validity analyses. As described in Step 3, we will have a subset of non-native speakers of Japanese take well-established speaking tests such as the ACTFL-OPI. The purpose of doing this analysis is to understand the relation of SJT scores to the scores obtained from other well-documented human-mediated measures of oral proficiency.

4 Conclusion

A core component of Ordinate's automated testing system is automated speech recognition and computer scoring. To develop the system specifically designed for the Spoken Japanese Test, data from a large sample of native and non-native speakers have to be collected. In addition, a

series of data preparation and validation studies will follow after the data collections to ensure that the test is reliable and valid. The SJT test is still in development and data collection is planned to be completed in the spring of 2005. If SJT shows high reliability and strong concurrent validity, the development will have been a success, and the process will have produced a reliable, accurate, and quick assessment instrument that can assess the core speaking and listening skills of non-native speakers of Japanese. Our hope is that SJT will make a significant contribution to the field of teaching and learning Japanese.

Acknowledgements

The research reported here is partly supported by Grant-in-Aid for Exploratory Research #16652040, provided by the Japanese Ministry of Education, Culture, Sports, Science and Technology, and by funding from Ordinate Corporation.

References

Agency for Cultural Affairs (2002, November). Heisei 14nendo kokunaino nihongokyoikuno gaiyo [Overview of the situations of Japanese teaching in Japan in 2002]. Retrieved on July 28, 2004. http://www.bunka.go.jp/1aramasi/14_kokunai_nihongokyouiku.html

Sakata, Y., Shinya, E., and Moriya, M., (2003). *Nihongo unyo bunpo* [Practical usage of the Japanese grammer]. Bojinsha, Tokyo.

Haraguchi, S. (2002). Accent. In Tsujimura, N. (Ed), *The handbook of Japanese linguistics* (pp.1-30). Malden, MA: Blackwell Publishers, Inc.

Ide, S., and Yoshida, M. (2002). Sociolinguistics: honorifics and gender differences. In Tsujimura, N. (Ed), *The handbook of Japanese linguistics* (pp.444-480). Malden, MA: Blackwell Publishers, Inc.

Japan Foundation (2003a). 2003 *Kaigai nihongo kyoiku kikan chosakekka gaiyo* [2003 summary of survey on Japanese Educational Institutions Overseas]. Retrieved on July 28, 2004 from http://www.jpf.go.jp/j/japan_j/news/0407/07-01.html

Japan Foundation (2003b). Heisei 13 nendo nihongonoryokushiken bunsekihyokani kansuru hokokusyo [The report of the analyses of the results from 2001 Japanese language proficiency test]. Tokyo: Authors

Maynard, Senko K. (1997). "*Japanese communication-language and thought in context*." Honolulu: University of Hawaii Press.

Mizutani, N. (1999). *Zoku nichiei hikaku hanashikotoba no bunpo* [Continued Japanese-English comparison of grammar spoken language]. Kuroshio, Tokyo.

Ordinate Corporation. (2004). *SET-10 test description & validation summary*. Menlo Park, CA: Author.