Development and Validation of an Automatic Spoken Spanish Test

Jared Bernstein, Isabella Barbier, Elizabeth Rosenfeld and John De Jong

Ordinate Corporation
www.ordinate.com

Abstract

A computer-based Spoken Spanish Test (SST) was developed to provide an accurate and convenient assessment of speaking and listening for learners of Spanish. The SST is intended to measure a candidate's receptive familiarity with a variety of forms of Spanish and the candidate's facility in speaking Spanish. The SST test takes between 15 and 20 minutes to complete, all the items are presented in spoken Spanish, and the test is administered and scored completely automatically. The SST has seven sections, a reading section, a series of elicited imitations, a sentence construction section, two vocabulary sections, a section with opinion questions, and a section with story retelling. The SST items were assembled into tests presented to 579 adult non-native Spanish learners (mostly university students) and to 435 native Spanish speakers. The paper describes validation of the automatic scoring system with reference to concurrent administrations of ACTFL, ILR, and SPT Oral Proficiency Interviews conducted by certified interviewers/raters.

1. Introduction

We describe the construction of a 15-minute-long spoken Spanish test, SST, which is delivered over the telephone by computer and automatically scored using speech recognition technology [1]. The SST test measures *facility in spoken Spanish* – that is, the ability to understand spoken Spanish on everyday topics and to respond appropriately at a native-like conversational pace in intelligible Spanish. Another way to express the construct *facility in spoken Spanish* is "ease and immediacy in understanding and producing appropriate conversational Spanish." This definition relates to what occurs during the course of a spoken conversation. While keeping up with the conversational pace, a person has to track what is being said, extract meaning as speech continues, and then, on occasion, formulate and produce a relevant and intelligible response. These component processes of listening and speaking are schematized in Figure 1, adapted from [2].

In the administration of an SST test, the Ordinate testing system presents a series of discrete prompts to the test taker at a native conversational pace. The prompts are drawn at random from an item pool. The prompts were recorded by a variety of different native speakers from several countries, producing a range of native accents and speaking styles. These integrated "listen-then-speak" items require real-time receptive and productive processing of spoken language forms, and the items are designed to be relatively independent of social nuance and high-cognitive functions. The SST test measures the

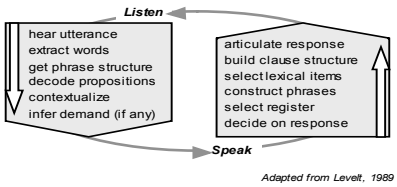

Figure 1: Conversational Processing Components in Listening and Speaking

test taker's control of core language processing components, such as lexical access and syntactic encoding. For example, in normal everyday conversation, speakers go from building a clause structure to phonetic encoding in about 40 ms [3]. Similarly, the other stages shown in Figure 1 have to be performed within the small period of time available to a speaker involved in everyday communication. The typical window in turn taking is about 500-1000 ms.

In this process, *automaticity* is required in order for the speaker/listener to be able to pay attention to what needs to be said/understood rather than to how the message is to be structured/analyzed. Automaticity entails the ability to access and retrieve lexical items, to build phrases and clause structures, and to articulate these without conscious attention to the linguistic code [4].

A test such as the SST probes the psycholinguistic elements of spoken language performance rather than the social and rhetorical elements of communication. Because during the test this probing is performed in real time, the SST measures the degree of automaticity in language performance. A person has to understand and produce language at some level of accuracy and fluency to participate in a spoken interchange. Since performance standards can be established for accuracy and fluency based on representative samples of language users, the SST checks the level of accuracy and fluency, and at the same time directly measures the rate and level of language process control.

To summarize, the SST measures basic encoding and decoding of oral language as performed in integrated tasks in real time. This performance predicts a more general spoken language facility, which is essential in successful oral communication. The reason for the predictive relation between spoken language facility and oral communication skills is basic. The language

structures that are shared among the members of a speech community are used to encode and decode various threads of meaning that are communicated in spoken turns. These threads of meaning that are encoded and decoded include declarative information, as well as social information and discourse markers. World knowledge and knowledge of social relations and behavior are also used in understanding the spoken turns and in formulating the content of spoken turns. However, these social-cognitive elements of communication are not represented in this model and not directly measured in the SST.

2. SST Content Design and Material

The SST test measures both listening and speaking skills, emphasizing the test taker's facility (ease, fluency, immediacy) in responding aloud to common, everyday spoken Spanish. All SST items are designed so that both native speakers and proficient non-native speakers find them very simple to understand and to respond to appropriately. The items cover a broad range of skill levels and skill profiles. Verification of these test characteristics will be reported in the SST validation report when the SST is available as a product. The vocabulary used in the test items and responses is restricted to the most frequent words in a corpus of Spanish text. In general, the language structures used reflect those that are common in everyday Spanish.

Each SST item is independent of the other items in the test and presents unpredictable spoken material in Spanish. Context-independent material is used in the test items for three reasons. First, context-independent items exercise and measure the most basic meanings of words, phrases, and clauses on which context-dependent meanings are based. Second, when usage is relatively context-independent, task performance depends less on factors such as world knowledge and cognitive style and more on the test taker's facility with the language itself. Thus, the test performance relates most closely to language abilities and not to other test taker characteristics that may be outside the target construct, i.e. *facility in spoken Spanish*. Third, context-independent tasks maximize response density; that is, within a given test administration time, the test taker has more time to demonstrate performance in speaking the language. Less time is spent in developing a background cognitive schema for the task.

Both the test items presented and the expected responses are constrained to contain common Spanish vocabulary and constructions that can be consistently understood and/or produced by at least 80% of a reference sample of educated native speakers of Spanish. In addition, these item types maximize reliability by providing multiple, fully independent measures. They elicit responses that can be analyzed automatically to produce measures that underlie facility with spoken Spanish, including phonological fluency, sentence comprehension, vocabulary, and pronunciation of rhythmic and segmental units.

Test Structure. The SST test consists of 60 items that are presented in seven separate sections. Each of the seven sections presents the test taker with a different task type, as shown in Table 1.

Table 1. Number of items presented.

Task	Presented
A. Readings	6 (of 8 printed)
B. Repeats	16
C. Opposites	8
D. Short Questions	16
E. Sentence Builds	8
F. Open Questions	3
G. Story Retellings	3
Total	**60**

In Part A, test takers are instructed to read 6 sentences from among a set of 8 numbered sentences printed on the test paper. In Parts B through G, the item materials are presented by voice only, with no direct support from the test paper. The first-item response in each part of the test is not scored, and the responses to the three open questions and the story retelling in Parts F and G are not scored automatically. Thus, 49 independent responses are scored in the SST.

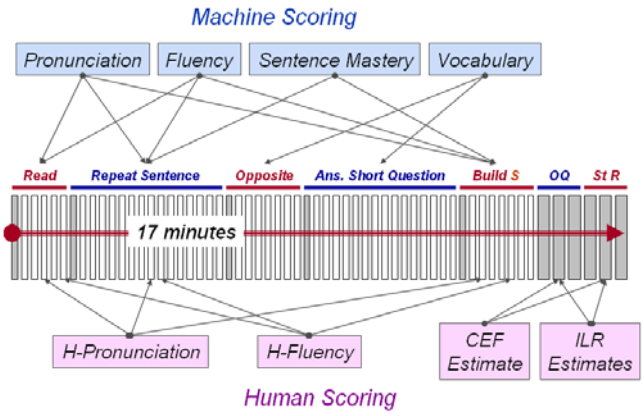

Figure 2. Test structure and test scoring

In Part A, the 8 printed sentences are grouped into two related sequential groups of four in order to provide some context and limit the reasonable readings. However, the system has the test taker read the sentences in random order. The items in Parts B and D are presented in a stratified random order so that the item difficulty generally increases over the sequence of items presented.

3. Results: Performance Data

Native Speaker Data Collection. Native samples were collected from adult (18 years old or older) native speakers of Spanish. Native speakers were roughly defined as individuals who spent the first twenty years of their lives in a Spanish-speaking country, were educated in Spanish through college level, and currently reside in a Spanish-speaking country. Samples were gender balanced, when possible. The native-speaker sample comprised 435 candidates: 140 from Argentina, 38 from Colombia, 222 from Mexico, 20 from Puerto Rico, and 15 from other Latin American countries.

Non-Native Data Collection. Ordinate contacted a number of Spanish departments at universities in the United States asking them to have students take the SST and, if possible, also an official ACTFL-certified Spanish OPI. Students/universities were remunerated for their participation and for the ACTFL test fee. In addition, test takers were recruited from other institutions. Apart from SST, subsets of each group took a form of an oral interview test. Three oral interviews were used:

1. The American Council on the Teaching of Foreign Languages Oral Proficiency Interview (ACTFL/OPI) administered by Language Testing International
2. The Interagency Language Roundtable Oral Proficiency Interview (ILR/OPI) administered by certified raters from the Defense Language Institute (DLI)
3. The telephone-administered version of the ILR/SPT, the Spoken Proficiency Test (SPT), administered by US Government-certified raters

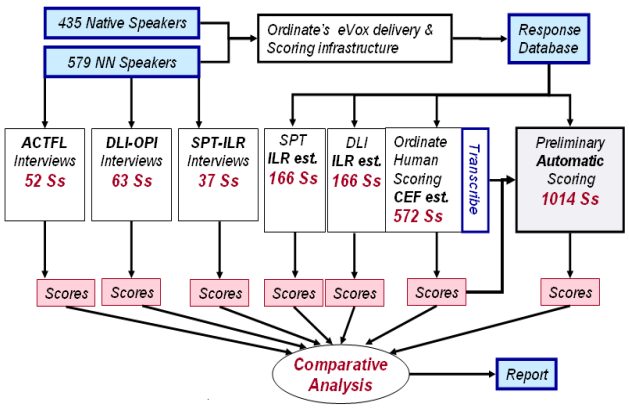

Figure 3. Data Paths.

Human Scoring. Human grading of recorded test-taker responses serves two main purposes: (1) to accumulate resources for automatic scoring, and (2) to accumulate evidence for validity. The development of automatic scoring for the pronunciation and fluency scores depends upon the availability of a set of human pronunciation and fluency ratings for a set of relevant calls in accordance with a set of rating criteria that are applied consistently across raters. Automatic pronunciation and fluency scores are 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. These measures are then scaled and combined so that they optimally predict the human judgments.

Once an algorithm for automatic scoring has been developed and implemented, the validity of the automatic scoring can be evaluated by comparing human ratings for test takers with the scores assigned automatically by the Ordinate system to these same test takers. Human ratings for this purpose were collected on two overall functional scales, an estimate of Oral Proficiency Interview ratings (E-OPI) and the Common European Framework level descriptors (CEF). The ratings on the functional scales were based on test takers' responses to the open-ended tasks that were not scored automatically and are therefore completely independent from the SST scores.

Data Analysis and Results. One would expect native speakers to obtain high scores on SST. On the other hand, for SST to operate as a measurement instrument, speakers of other languages who are learning Spanish would need to be distributed over a wide range of scores. Also, we wanted to know if SST scores correlate significantly with other, communicative, measures of spoken Spanish, particularly human ratings on the communicative ILR scale.

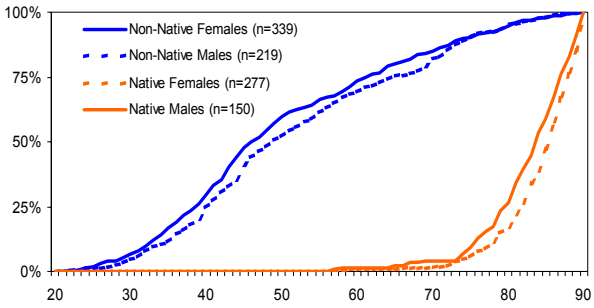

Figure 4. Cumulative density functions per gender for non-native speakers and native speakers.

SST scores by language background. In order to evaluate low discrimination of SST for native speakers of Spanish, Figure 4 presents cumulative density functions for two speaker groups: native and non-native. Figure 4 shows that the native speakers (male or female) have almost identical distributions that clearly distinguish them from the non-native sample (female or male). This suggests that the SST has high discriminatory power among learners of Spanish as a second or foreign language, whereas native speakers obtain near maximum scores irrespective of region of origin or gender. Furthermore, many comparisons are possible in this data set between the SST machine scores and certified human interview scores and score estimates produced by expert or certified raters. We include just three such scatter plots. Figure 5 displays

the relation between SST scores and ACTFL certified interview scores for a set of 52 Spanish learners, then the relation of SST scores to U.S. Government interview tests (SPT OPI) for 37 subjects, and finally the relation of CEF proficiency estimates to SST scores for a sample of 572 subjects (native speakers are plotted with "**x**").

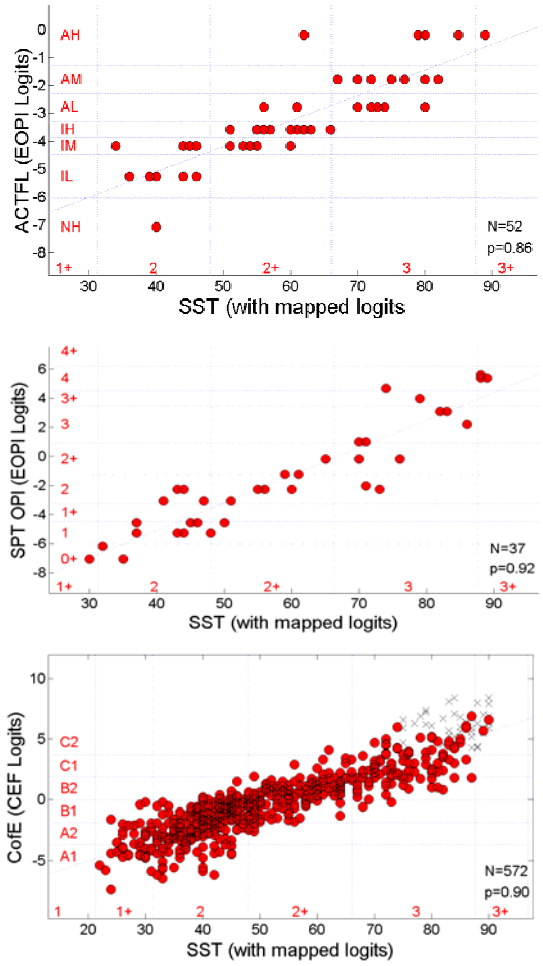

Figure 5. Comparison of machine and human scores.

In all three comparisons shown in Figure 5, the correlation between the machine scores and human ratings is extremely high, with no outliers.

4. Conclusion

Generally the SST scores will accurately predict 80% or more of the variance in US government oral language proficiency ratings. Given the fact that the reliability for ILR scales is reported in the literature to be around 0.90 [5], this is as high as the predictive power between two consecutive independent OPIs. The SST also predicts the ACTFL interview scores with about the same precision as two independent ACTFL OPIs would predict each other [6].

The SST instrument offers a method for measuring facility in spoken Spanish. The SST procedure elicits sufficient spoken language behavior on which to base a reliable and accurate human judgment of practical speaking and listening skills. Furthermore, automatic scoring of responses from an SST administration can produce reliable and useful information about these spoken language skills. Preliminary validation data suggest that the automatic scoring has the following four properties: (1) Native speakers consistently obtain high scores on SST; (2) Learners of Spanish are distributed over a wide score range; (3) SST scores correlate nicely with other measures of spoken Spanish; and (4) SST scores can predict ACTFL and ILR scale ratings.

The data presented here may be consistent with several conclusions. First, the data might support the assertion that functional models of language use are the best foundation for designing spoken language tests and for interpreting their results, and that the SST test procedures and scoring are just a convenient proxy for the communicative procedures and scoring. Conversely, the communicative interview tests could be seen as convenient, low-tech proxies for a crisp test of language processing skills. Second, these experimental results indicate that the two theoretical approaches to testing do not produce different patterns of proficiency scores for populations of second language speakers. The distributions of paired test scores over several populations and different methods are closely aligned and contain no outliers. Because of the stronger empirical grounding of the psycholinguistic testing, we can conclude that this pattern of data obviates the need to posit a 'communicative' basis for language test design.

References

[1] Bernstein, J., De Jong, J.H.A.L., Pisoni, D., & Townshend, B. (2000). Two Experiments on Automatic Scoring of Spoken Language Proficiency. In: P. Delcloque (Ed.), *Proceedings of InSTIL2000: Integrating Speech Technology in Learning* (pp. 57-61). University of Abertay Dundee, Scotland, August, 2000.

[2] Levelt, W. (1989). *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.

[3] Van Turennout, M., Hagoort, P., & Brown, C. M. (1998). Brain Activity During Speaking: From Syntax to Phonology in 40 Milliseconds. *Science, 280*, 572-574.

[4] Cutler, A. (2003). Lexical access. In L. Nadel (Ed.), *Encyclopedia of Cognitive Science. Vol. 2, Epilepsy - Mental imagery, philosophical issues about* (pp. 858-864). London: Nature Publishing Group.

[5] Stansfield, C.W., & Kenyon, D.M. (1992). The development and validation of a simulated oral proficiency interview. *Modern Language Journal,* 76, 129-141.

[6] Sieloff-Magnan, S. (1987). Rater reliability of the ACTFL Oral Proficiency Interview, *Canadian Modern Language Review,* 43, 525-537.