2

Workable Models of Standard Performance in English and Spanish

JENNIFER BALOGH & JARED BERNSTEIN

1 Introduction

How well does Selma speak Spanish? Does your cousin know English? Practical questions abound, and any number of language tests might be applied to provide answers. However if you put yourself in the position of a test designer, and think through what it should mean to "speak Spanish" or "know English", choices need to be made sooner or later about what exactly a language *is*. Framing the issue another way, which actions from Selma would count as evidence that she does indeed speak Spanish? Suppose she speaks something that sounds a lot like Portuguese, but Spanish speakers can understand her and she seems to understand them as well. Facing variation in speaking and listening skills, the question arises as to how to define a *spoken standard* to be used to assess a specific spoken performance.

The issue can become important when considering high-stakes situations where a person seeks admission to a graduate program, applies for compensation, or is either granted or denied permission to enter a country based in part on one spoken language assessment.

Devising a workable test that decides whether or not a person speaks a language is proposing a feasible operational definition of the spoken language – its extent, its slope, and its limits. This operational definition includes structural elements, processes, and sampling procedures. It works more or less like a scientific hypothesis that links disparate phenomena (one performance sample and a domain of inferred performances), that requires

Diversity in Language: Perspectives and Implications
Yoshiko Matsumoto, David Y. Oshima, Orrin R. Robinson, and Peter Sells (eds.).

good-faith empirical grounding, and that needs to withstand potential empirical challenges.

The purpose of this paper is to describe a method for developing a workable model of a language to which spoken performances of speakers can be compared. First, we describe the method with concrete examples from models of American English and Latin American Spanish. Then, we evaluate each model with regard to its efficacy within an automatically scored spoken language test. We also report the results of an experiment investigating the adequacy of using the models with speakers of different regional dialects. The model building process produces data that may be of substantive or methodological interest when considering how a standard of a language should be defined. Specifically, we elucidate what should and should not be considered acceptable responses to a given prompt, based on distributions of response characteristics from appropriate samples of utterances in response to spoken prompts. Implications of the method are discussed in terms of its practical application to language testing and its contribution in defining the essential traits of a particular language.

2 Method for Developing a Workable Model of a Language

The models discussed here were built for the purpose of assessing *facility in a spoken language,* or the ability to understand a spoken language on everyday topics and to respond appropriately and intelligibly at a native-like conversational pace. In this regard, the models, along with scoring algorithms, are used in the context of spoken language tests that are administered over the telephone by computer. The models form part of Ordinate Corporation's Spoken English Test (SET-10, previously called "PhonePass") and the Spoken Spanish Test (SST) will be referenced as examples. These tests generate automatic scores of spoken language performance.

2.1 Background on Spoken Language Proficiency Testing

Before describing the models in detail, however, we address the topic of language proficiency testing, as such. There has been a broadening of constructs in applied linguistics. In the period from 1950 to 1990, a focus on narrow structural units like clauses and morphemes was replaced by emphasis on broader performance categories like speaking and reading, which were superseded, in turn, by even broader constructs under the rubric of communication. If we compare a work such as *Language Testing* by Robert Lado (1961) with Lyle Bachman's *Fundamental Considerations in Lan-*

guage Testing (1990), we can see the differences very plainly. In Lado's book, various narrow aspects of language like pronunciation and grammatical patterns are introduced and explained in some detail, and then techniques for testing them are described and evaluated. In Bachman's book, most of the traditional analytic categories of linguistics including vocabulary, syntax, morphology, and phonology are dismissed in a single paragraph (1990, p. 86) with a reference to a secondary source. There is no reference to Jesperson, nor to Sapir, nor to Bloomfield, nor to Jakobson, nor to Pike, nor to Chomsky. The current dominant view in the field of language testing, fairly represented in Bachman's much-cited book, is that language use encompasses multiple communicative competencies: grammatical competencies such as vocabulary and syntax; textual competencies such as cohesion; illocutionary competencies such as the ability to make a request; and sociolinguistic competencies such as sensitivity to register and naturalness (Bachman, 1990).

This chapter presents models that underlie tests that diverge from the current trend of communicative testing. We can distinguish communicative tests from language performance tests as there is a distinction in empirical traditions between sociolinguistics and psycholinguistics.

2.1.1 Spoken Language in Social Context

A social-communication view of language competence emphasizes that the spoken form of a language is used in social settings (according to a set of culture-specific norms) to accomplish explicit or implicit tasks of many sorts. In the current applied linguistics discourse, this insight was adapted from Hymes (1972) by Canale and Swain (1980) and has been further modified and refined in the language testing context by Bachman (1990), among others. Hymes had set out to extend the domain of linguistic analysis beyond the limits observed in the tradition of linguistics established by Bloomfield and Chomsky. In the view of Hymes, language as it has traditionally been studied should more properly be understood as a subset of a more general field – the study of communication. Corollaries of a communication-centric view that are relevant to testing may include:

1. Language and communication are not separable; therefore a test of language skill should be a test of the language as used in communication.
2. Structures above the sentence level are important for effective communication; therefore a language test should include long turns and multi-turn exchanges.

3. Language, as used, reflects and communicates social structures; therefore a test should let the candidate display a range of registers and illocutionary actions.

2.1.2 Spoken Language in Individual Performance

A psycholinguistic view of language competence emphasizes the development of first and second language skills and the real-time processes that underlie the performance of these skills. Historically, psycholinguistic research has focused on the phenomena of language performance, and used these phenomena to elucidate more general cognitive processes like memory, association, and pattern classification (Eysenck & Keane, 1995). Since the 1960s, many production and perception studies have also focused on the confirmation (or disconfirmation) of the "psychological reality" of structures and processes that have been hypothesized in linguistic research. This research has uncovered and quantified many robust phenomena of skilled (native-like) language performance, as well as a fairly clear structure of elementary processes that can be identified in appropriate experimental contexts. Eysenck & Keane review studies that show that complex cognitive skills that are used in language become automatic in skilled performance, and therefore do not absorb any of the attentional capacity of the speaker-hearers.

A psycholinguistic view of language competence emphasizes the development of first and second language skills and the real-time processes that underlie the performance of these skills. Corollaries of a psycholinguistic view that may be relevant to testing could include:

1. Attention to core language production limits attention to content; therefore the measurement of fluent, automatic control of core language will predict the complexity of content that can be produced or understood in real time.
2. Context-free language processing develops after context-bound processing; therefore context-free tasks should be at least part of a language proficiency battery.
3. Structures at or below the sentence level dominate what is well-understood in language; therefore these structures offer a firmer basis for scoring and diagnosis.

Even though the language develops in the context of communication, to serve the purposes of communication, it develops in individuals and finds expression in many activities (including games and songs) where communi-

cation, as such, is not primary goal. Skilled language performance has been experimentally analyzed outside of natural language use; therefore language processing skills can be measured outside a natural setting. Also, these studies are conducted exactly because high proficiency L1 and L2 speakers reach a stage of automaticity with the language in listening and speaking (and often in reading) that simply cannot be attended to by introspection or protocol methods.

There may be many reasons to favor of psycholinguistic tasks in language testing, but one important reason relates to sampling and precision: in a given set of linguistic performances from a test-taker, which aspects of the performance yield the most stable samples on which to base a measure?

2.1.3 Limits in Sampling and Linguistic Description

There are at least two sampling-related aspects of language and language use that support the measurement of smaller units in language testing. That is, the measurement of words and phrases and sentences, has advantages over measurement of turns, narratives or discourses.

First, the smaller the unit, the more of them one can observe in a practical amount of time and the more the sample will be stable and representative. As more independent measures are gathered, the resulting measures also become more and more reliable. Furthermore, the descriptive specificity of the expected, or correct, performance is much better understood for the smaller units. Most of what is known, and agreed on, in linguistic descriptions relates to units at the sentence level and below. For example, the core phrase structures of English are well described and generally agreed on, whereas the order, or even the limits, of acceptable dialogue turns or narrative structure are not well understood and are certainly not yet agreed upon among experts. One conclusion from these considerations could be to favor measures of performance on sentence-level units, because there are more such units available in a performance sample, and because the relative merit of alternative forms is better understood. In testing, it is good practice to be explicit about the nature of a *good* response and about the basis of deciding what a good response is.

If one is needed, John Carroll (1961) is a venerable source of psycholinguistic perspective applied to second language acquisition that supports the kind of workable model of a language described here. Carroll draws a distinction between a knowledge aspect of language performance and a control aspect. The knowledge aspect of language relates to the content of what is being said: the words that are used and mastery of the language's structure. The control aspect relates to the manner of speaking, or the way in which the spoken message is conveyed with regard to pronunciation and

phonological fluency. Phonological fluency, in this sense, refers to the rhythm, phrasing and timing evident in speaking and/or reading the language. In order to create a representation of both the knowledge and the control aspects of the language, many specific models must be created.

The models we propose comprise six components. Four of these relate to the knowledge aspect of the language and two relate to the control aspect. The components required for the knowledge aspect are the following: acoustic models, a lexicon, a set of spoken probes, and response models for these probes. For the control aspect of the language, the method calls for models of pronunciation and fluency. Descriptions of each of these representations, along with its development methodology, are given below.

2.2 Acoustic Models

The representation of the language at its most atomic level is a set of acoustic models. Each acoustic model is a representation of how a sound or phone in the language is produced in the context of other phones. The approach to developing the acoustic models is taken from automatic speech recognition. Thousands of utterances are recorded over the telephone. The speech is aligned with lexically guided phonetic transcriptions so that the sounds can be identified and grouped together with other sounds in the same category. For example, the /k/ sound in *cat* would be grouped with the /k/ sound in *catalog, decal,* and *decapitate*. The sections of the utterances associated with a specific sound are then transformed into feature vectors. A feature vector is a list of numbers representing measurable characteristics of the speech such as the amount of energy produced at various frequencies when making the sound. A statistical model (Hidden Markov Model) for each sound is created. The current method makes use of the structures implemented in HTK (the HMM Took Kit; see Young, Kershaw, Odell, Ollason, Valtchev, and Woodland, 2000). The acoustic model represents the variety of ways the sound is pronounced by a sample of speakers. Because the models can be trained on samples of native and non-native speakers with many different accents and speaking styles, the resulting acoustic models will accept (on approximately equal footing) spoken responses from males or females from Boston or Chicago, with or without features of an African American vernacular.

Once the acoustic model is built, it is used to determine how likely the sounds are in a specific utterance from a speaker, given the best word match. To do this, the system segments the speech signal into small frames and then compares each frame with the word-borne acoustic models to pinpoint which sound the speaker was most likely producing at that point in time.

The search through the models is constrained by what the system is expecting the person to say.

2.3 Lexicon

The next component for the workable model is a lexicon, or a representation of the words that are commonly known to any native speaker of the language. To develop the lexicon, a spoken corpus of the language is identified. A corpus of speech is preferable to a corpus based on text, since the model represents spoken as opposed to written language. For American English, the Switchboard Corpus was selected for this purpose. For Spanish, the corpus was Spanish Call Home, both of which are available from the Linguistic Data Consortium (LDC) at the University of Pennsylvania. From the corpus, the 8000 most frequent lemmas are identified and organized into a base lexicon. The lexicon approximates the most commonly used vocabulary of the spoken language and serves as a reference throughout the model development process. Of course, a typical adult native speaker of a language has knowledge of many more words, but the motivation for selecting the high-frequency lexical entries is to model the core of the language for the practical application of assessing speaking performances of learners. There are lexical entities beyond the word level such as idioms, frozen phrases, formulas, and collocations. These were neither purposefully written into items nor were they explicitly excluded. The models were built to represent language that occurs frequently in colloquial speech, therefore, an approach associated with isolated word frequency allows for inclusion of some of these frequently occurring conjoint lexical items as well.

2.4 Spoken Probes (the test items)

Next, a set of spoken probes are developed. First, native speakers of the language with the appropriate training write a set of items, often adapting naturally occurring utterances from everyday experience. The items are context-independent samples of the language that can be used to elicit spoken responses. For example, an item might be a question that is to be answered, or a sentence that is to be repeated. The vocabulary and syntactic structures in the items resemble colloquial language as spoken by educated members of the community. For the American English model, items were written by American item developers, and for the Spanish model, the items were drafted by an Argentine. In general, the syntactic structures used in the items reflect those that are common in everyday speech and are designed to be independent of social nuance and high-cognitive functions – again, reflecting the core of the language. The item writer might hear an utterance like “Hey, give me a break, nobody uses tachistoscopes any

more." Then, to simplify a bit and stay within the core-language lexicon, the writer might draft an item like "Give me a break, nobody uses oil lamps anymore."

The written form of each item is then reviewed by a separate set of experts who are also native speakers, but who live in different regions or countries. The purpose of the review is to ensure that the items conform to current colloquial usages in all the disparate regions where the language is indigenous. For American English, the items were reviewed by two linguists from each of the US, UK, and Australia. For Latin American Spanish, the reviewers were native Spanish speaking linguists from Chile, Colombia, Ecuador, Mexico, Puerto Rico, Spain and Venezuela. Any item that is too closely associated with a specific geographic area or cultural viewpoint is modified or discarded, producing a set of materials that should be colloquial, yet from the intersection of the various regional forms of the language. Also, all items are checked against the lexicon. Vocabulary items not present in the lexicon are either changed to other entries that are listed in the lexicon or the lexical item is kept and added to a supplementary vocabulary list, based on expert judgment. The changes proposed by the different reviewers are then reconciled and the original items are edited accordingly.

The reviewed items are then recorded by a diverse sample of speakers representing both genders and a variety of native accents and speaking styles. These recordings are the spoken probes.

2.5 Response Models

Once the spoken probe recordings are reviewed for quality, they are integrated into a central database and are presented to several hundred native speakers.

The speaker sample for the American English acoustic and response models was drawn from college graduates and was representative of the ethnic and geographic distribution of the US population, with an oversampling of African Americans. African American were intentionally oversampled for two reasons: a) they are underrepresented in the population of college graduates, and b) African American English is a salient American dialect that is not geographically situated. The sample was also balanced for gender. For the Latin American Spanish model, the native speakers were from Argentina, Colombia, Mexico, Puerto Rico and 13 other Latin American countries. The sample was roughly gender balanced across the geographic locations. Note that Iberian Spanish speakers did not contribute to the response models for Spanish (however, see Section 4, below).

Any item that is not understood or responded to correctly by at least 80 percent of the native speakers is excluded from the model. The reference sample responses are transcribed and vetted for quality by judges working independently. The models derived from the reference sample of educated native responses are used to estimate the appropriateness and quality of the responses from non-native speakers.

Responses from non-native speakers are also collected in order that the response models represent how non-native speakers behave when responding to the spoken probes. The responses of the native speakers to a spoken probe can be considered 'standard' for that probe. The non-native speakers, in contrast might respond to the same probe with a reasonable, but 'non-native' answer. For example, if the task is to provide an opposite word to a spoken probe /rait/, most native speakers say *wrong*, although some say *left*. Unlike a native sample, a high proficiency non-native sample of responses will include *read*. In this case, test design and pure model building diverge. The response *read* is definitely less native-like than *wrong*, but for test scoring, *read* must be counted as correct.

The response models allow the utterances from both native and non-native speakers to be recognized automatically. For this automatic recognition to take place, a recognition lexicon must be created. The recognition lexicon is a list of words that are expected in the responses, along with a pronunciation for each entry. The pronunciation indicates which acoustic models to string together to estimate the likelihood that the speaker said a specific word.

Since the models of the language are used to compare non-native utterances with standards of native speech, models that represent the extent to which a non-native speaker finds an item difficult to respond to are also developed. For example, a non-native speaker may be asked to repeat a sentence verbatim. To repeat a sentence longer than about seven syllables, the speaker has to recognize the words as produced in a continuous stream of speech (Miller & Isard, 1963). Highly proficient speakers of the language can generally repeat sentences that contain many more than seven syllables because these speakers are very familiar with the words of a language, as well as its phrase structures and other common syntactic forms. If a person habitually processes five-word phrases as a unit (e.g. "the really small white box"), then that person can usually repeat utterances of 15 or 20 words in length. Generally, the ability to repeat material is constrained by the size of the linguistic unit that a person can process in an automatic or nearly automatic fashion. As the sentences increase in length and complexity, the task becomes increasingly difficult for speakers who are not deeply familiar with the structure of the language. Likewise, if the non-

native speaker is presented with three short phrases and is asked to rearrange them to create a grammatical sentence, the length and complexity of the sentence that can be built is constrained by the size of the linguistic unit (e.g., one word versus a three-word phrase) that the person can manipulate automatically without conscious effort. By modeling the difficulty of the items, the system can automatically assess the speaker's level of proficiency at least with regard to the knowledge aspect of the spoken language.

2.6 Pronunciation Models

When developing a model that is the 'standard' of a language with which to compare non-native performances, the question arises as to what constitutes native versus non-native pronunciation. Although pronunciation patterns are encoded in the acoustic models for the language, the purpose of creating pronunciation models is to capture subjective perceptions of a speaker's pronunciation from the viewpoint of a native. A non-native speaker's accent may deviate from a native speaker's in many ways, but it is only those characteristics that are noticeable to a native speaker and that potentially interfere with intelligibility that should be considered in a model of pronunciation that is used for test scoring.

To create pronunciation models, several expert human raters listen independently to thousands of native and non-native utterances. The utterance samples are longer than a word or short phrase so that the human rater can hear the speaker's production of segments and words in the flow of continuous speech. The human ratings of pronunciation are used to train non-linear models that are optimized to predict the human judgments. The system extracts features of the speaker's speech including the stress and segmental forms of the words within their lexical and phrasal context. Some of these features are compared to native renditions of these segmental forms and the differences are quantified. These parameters are then used by the non-linear models to predict how the human rater would have scored the pronunciation ability of the speaker.

2.7 Fluency Models

As with pronunciation models, human raters listen to thousands of utterances from hundreds of non-native speakers and judge the fluency of their utterances. The human ratings are then used to train non-linear models that predict how the human raters would score the fluency of a given speaker. The information extracted from the system includes measures of the latency of the response, the rate of speaking, and the position and length of pauses in the utterance. Measurable aspects of the native speaker performances are

used to normalize some of the parameters by which non-native spoken performances are measured.

3 Evaluation of Workable Models of Spoken Language

Once the models for a language have been created and normalized on samples of both native and non-native speakers, the models are evaluated for their efficacy in scoring performances.

The model's quality is determined by three criteria: high reliability, ability to show effective separation between samples of native and non-native speakers, and strong correlations with other established measures of oral language proficiency.

To quantify these values, the models are evaluated in the context of spoken language tests developed by Ordinate Corporation: the Spoken English Test (SET-10) for American English and the Spoken Spanish Test (SST) for Latin American Spanish. Within the test, the models are used along with scoring algorithms to generate machine scores of spoken language performance. The test scores are reported on a scale from 20 to 80 for both tests, although the two scales are not calibrated to one another.

Data collected by Ordinate from both native and non-native speakers were used to evaluate the models. Native speakers were defined as literate adults (currently in university, or university graduates) who live in a prescribed set of regions and countries, and represent a range of age groups. Non-native speakers represented a broad range of proficiency levels and first languages.

3.1 Reliability

For each test, both native and non-native speakers took the test. For evaluation of test reliability for English, SET-10 scores from 50 non-native test takers were analyzed. The split-half reliability was 0.97. For the SST, the split-half reliability was 0.96 and was based on 150 test takers. The reliability of the scores of non-native performance for both tests is high. These results suggest that the models are providing consistent information regarding the performances of non-native speakers and how well these performances match the standard.

3.2 Separation of Native and Non-Native Performances

Next, performances of native and non-native samples are compared to determine whether or not the models distinguish the two groups. For the SET-10, fewer than 5 percent of the native sample scored below 68 (n=775);

whereas learners of English as a second or foreign language were distributed over a wide range of SET-10 scores. Only 5 percent of the non-natives scored above 68 (n=603). The scores show effective separation between native and non-native speakers.

A similar analysis was done for the SST. Figure 1 presents cumulative distribution functions which show the percentage of test takers in each group who received a given score or lower. Note that the range of scores displayed in Figure 1 is from 10 to 90, whereas the SST scores are reported on a scale from 20 to 80. Scores outside the 20 to 80 ranges are deemed to have saturated the intended measurement range of the test and are reported as 20 or 80.

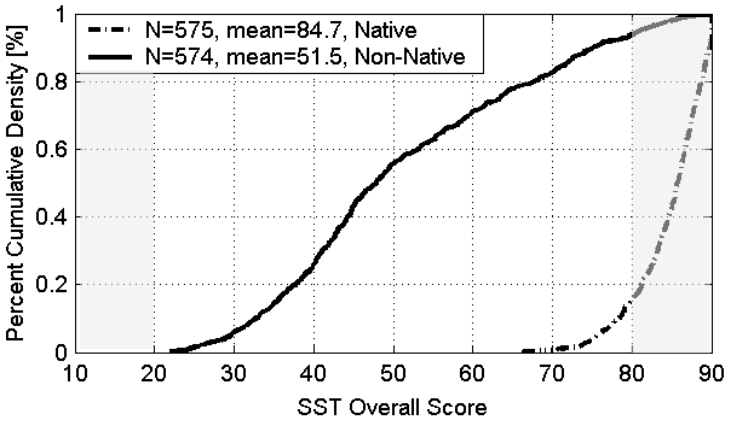

Figure 1. Cumulative distribution functions of SST scores for native and non-native speakers.

The distribution of the native speakers clearly distinguishes the natives from the non-native sample. Fewer than 5 percent of the native speakers received a score below 75, while only 10 percent of the non-native speakers received a score above 75. The results from this analysis suggest that the SST has high discriminatory power among learners of Spanish as a second of foreign language, whereas native speakers obtain near-maximum scores.

The results indicate that the models represent a standard of the spoken language and can distinguish the degree to which a non-native speaker complies or does not comply with this standard, both with regard to knowledge of the language and means of expressing it through pronunciation and fluency.

3.3 Correlation with Human Ratings

The third metric for evaluating the models is the correlation of the automatically generated test scores with other measures of spoken language performance from trained human raters.

First, each test taker's proficiency level is estimated by human experts. The raters listen to responses to open questions and story retellings and then these ratings are correlated with test scores.

For the SET-10, proficiency estimates were collected for 268 non-native speakers and 33 native speakers on an oral interaction scale based on the Common European Framework (Council of Europe 2001). Responses to open questions were assigned randomly to six raters who together produced 7,266 indepedent ratings in an overlapping design. Figure 2 shows the relationship between the SET-10 scores and the CEF levels. The correlation is 0.88. The graph also shows how both instruments (the SET-10 and CEF) clearly separate the native and non-native groups.

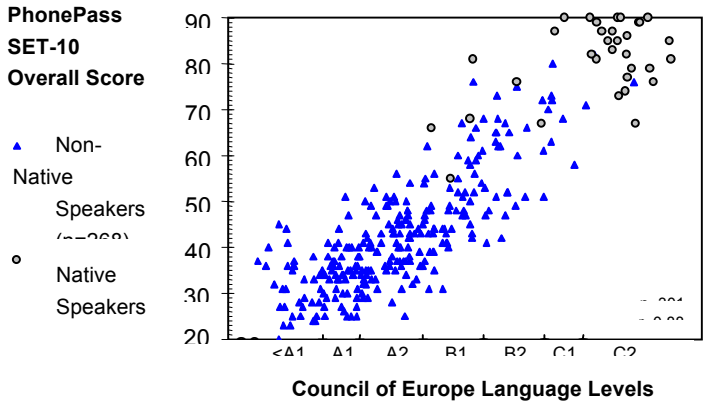

Figure 2. Correlation of SET-10 Overall scores and CEF estimates.

For the Spanish models, the estimates of spoken language proficiency were also based on the CEF level descriptors. For the CEF estimate scores, utter-

ances from 572 speakers were rated. Three native speakers of Spanish were selected to listen to 30-second recorded responses to open questions and story retellings. All three raters had degrees from universities in South America. Two were certified Spanish translator/interpreters. On average, the three raters together provided 11 independent scores for each speaker resulting in a total of 6125 ratings. Figure 3 is a scatter plot of the CEF estimate scores and SST scores.

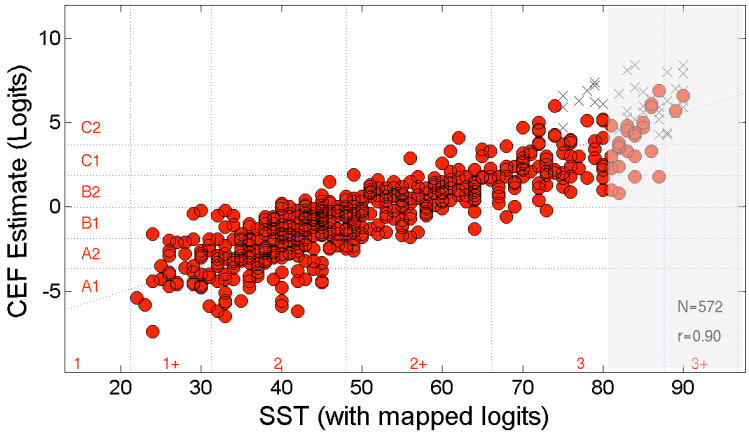

Figure 3. CEF estimates on spontaneous speech samples as a function of SST ratings based on the Spanish models. N=572, r=0.90.

For data analysis, the computer program FACETS (Linacre, 2003) was used to estimate rater severity, subject ability, and item difficulty (Linacre, Wright, and Lunz, 1990) based on a one-parameter Rasch model. The model expresses scores in a mathematical unit called a Logit. The bounda-

ries of the different CEF scale levels were mapped onto a continuous Logit scale.

The high correlation with the independent estimates of spoken language performance on the CEF level descriptors suggests that the models provide a reasonably accurate representation of judgments from native listeners on how closely natives and non-natives align with a data-driven standard model of the language.

In addition, results from the SST were correlated with two different human-conducted and human-rated Oral Proficiency Interviews (OPIs): official interviews by the American Council of the Teaching of Foreign Language (ACTFL), and telephone interviews with government-certified raters in accordance with the Spoken Proficiency Test (SPT) procedure, with scores reported on the Interagency Language Roundtable (ILR) scale. The standard ACTFL interviews were administered with at least two official ACTFL ratings per interview. For the ACTFL interviews, 52 scores were submitted, one for each of the 52 participants. For the SPT interviews, each rater independently provided ILR-based proficiency level ratings for each of the 37 candidates, for a total of 74 ratings. For both studies, test takers participated in the interview within a day of the SST administration. Figure 4 is a scatter plot of the ACTFL OPI scores as a function of SST ratings for 52 Spanish learners.

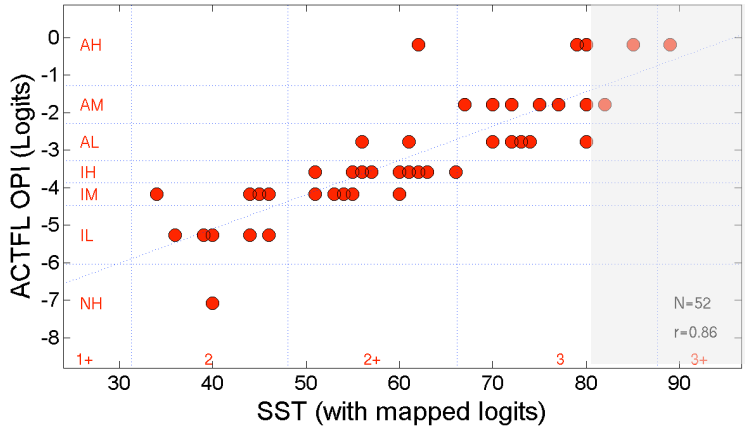

Figure 4. ACTFL OPI scores as a function of SST ratings.

The correlation for these two tests is 0.86, indicating a strong relation between the machine-generated scores and the human-rated interviews.

The other comparison of the machine scores with an oral proficiency interview was between the SST test and the SPT Interview on the ILR scale.

Figure 5 shows the scatter plot of ratings for 37 non-native Spanish speakers for these two measures.

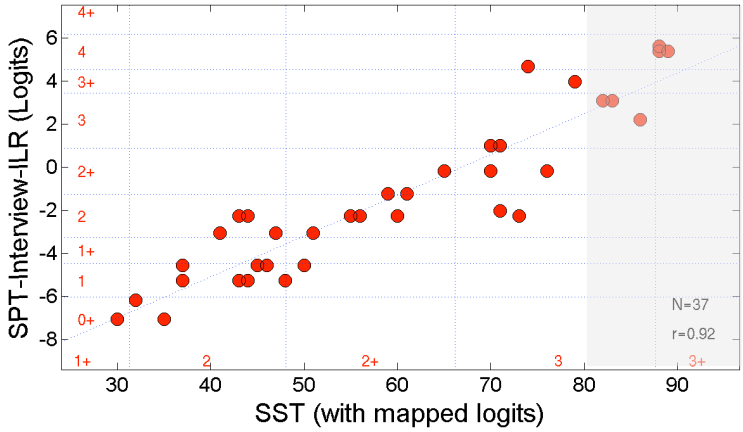

Figure 5. SPT Interview scores as a function of SST ratings.

The correlation between the SST scores and the ratings from the SPT Interview is 0.92. Like the ACTFL-SST data set, the SPT-SST data set contains no outliers. Together, the close alignment between the SST scores and the scores from these two certified tests of oral proficiency indicate that the models underlying the SST align with native judgments of what is standard for the language and how far a non-native speaker deviates from this standard.

The results from these validation experiments show that the scores from tests built on the workable models are reliable and distinguish native from non-native speakers. Moreover, the results indicate that the models can discriminate different levels of performance on a continuum and this discrimination of speakers' facility with the spoken language correlate well with traditional and completely independent measures of spoken language ability.

4 Adequacy of Models for Other Dialects

Now that we have a methodology for creating workable models of a language, the question arises as to whether or not the model generalizes. If the models are trained on data from one or more specific regions, does this representation of the standard language generalize to other dialects? As a specific example, for Latin American Spanish, the model of the language was trained on Spanish speakers from several Latin American countries, and not on speakers from Spain. To assess the performance of the model with

speakers from a different dialect, an experiment was conducted with Iberian Spanish speakers. The hypothesis was that the model would fit the Iberian Spanish speakers just as well as the native Spanish speakers from Latin American countries since the model represents the spoken language in general terms and is not tied to a specific region. Although pronunciation is recognizably different between the two groups and some vocabulary is characteristic of only one group or the other, the model construction method suggests that these differences would not constitute a large enough disparity to cause significantly different scores. The metric used to quantify how well the model fit the speech of the native speakers was the Overall score on the SST.

For the experiment, 153 Iberian Spanish speakers took the SST. The scores were then compared to speakers from the Latin American countries. Figure 6 shows the cumulative distribution functions of the scores from native Spanish speakers from five different countries.

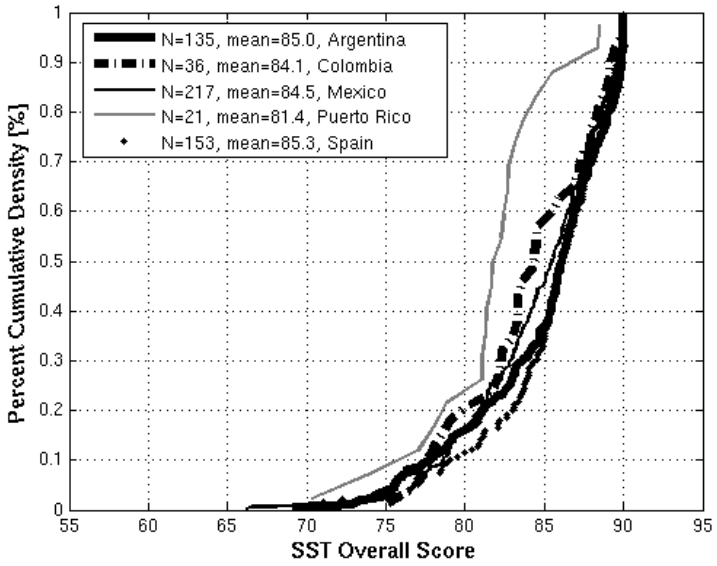

Figure 6. Cumulative distribution functions of SST scores for native Spanish speakers from five different countries.

As seen in Figure 6, the Iberian Spanish speakers (depicted by the light thin line) scored just as well as speakers from the Latin American countries. Further analyses showed that the Iberian Spanish speakers also scored just as high as, if not higher than, many native Spanish speakers from Latin America on Pronunciation (Balogh, Bernstein, Barbier & Rosenfeld, 2005). The results are particularly remarkable for Pronunciation since dialectal differences in lexical form had been avoided in the spoken probes, leaving phonetic patterns as the most salient remaining differences between dialects of Spanish. Similar results were observed for the other three subscores. The findings support the notion that the models of spoken Spanish reflect a standard of native performance that spans across dialects.

5 Language Standards

Given that the models can distinguish native versus non-native speakers, a potential topic of research is a more thorough understanding of the differences between these two groups. The method for building workable models of the language produces response distributions than can be mined for information about what constitutes native-like language use. For example, when an English native speaker hears the prompt, "What is frozen water called?" about 99 percent of the native speakers respond with the single word *ice*. A few offer less frequent lexical items such as *slush*. The patterns are significantly different for non-native speakers who often respond with loose semantic associations with the word, such as *cold* or *winter*. These types of answers do not occur in the native speakers' responses.

Response distributions of about 50 short answer questions in the SET-10 were analyzed to extract patterns in the speech of native versus non-native speakers. For short-answer questions designed to elicit a predictable one or two-word answer, natives almost always respond in a consistent way. A small amount of variation was observed with the following characteristics:

Enhanced vocabulary - Natives sometimes used vocabulary that appropriately addressed the question, but that occurs less frequently in the language compared to the majority response. For example, some natives say *aircraft* instead of the more common word *airplane,* or *verse* instead of *poetry*.

Mastery of phrasal structure – Instead of saying a single word as a response to a question, native speakers sometimes couch the word in a larger

phrase structure, for example by adding a determiner. When presented with the question, "Who is more likely to be older – John or his grandmother?" natives often say *his grandmother* instead of *grandmother*. When non-native speakers attempt to do this, they often use an unexpected determiner such as *her* or *my*. These patterns reveal a possible lack of knowledge about the semantics of the spoken probe or an inability to track the constituents in a discourse due to taxed cognitive resources. Natives, in contrast, naturally construct an appropriate phrase structure for the answer.

Structural agility – Native speakers sometimes embed a response in a complete sentence that answers the question. For example, if asked "What do tadpoles grow into?" natives will sometimes say "Tadpoles grow into frogs." The ability to answer the question in a complete sentence based on the question shows agility with the syntactic structure of the language. When repeating sentences, native speakers are amazingly adept at parroting entire sentences verbatim. When deviating from the prompt, substitutions tend to retain the meaning of the sentence. For example, native speakers sometimes replace *might* for *may* and *which* for *that*.

Extended cognitive capacity – Sometimes native speakers answer the question but add comments that describe exceptions to the question. They can do this because they are using less of their cognitive capacity to respond appropriately to the question compared to non-native speakers. These responses may also be attributed to cultural differences and personal style.

The patterns for non-native speakers are quite different. Aside from having trouble with basic meaning of questions that almost every native speaker answers with ease, the non-native speakers sometimes have difficulty in the following areas:

Parsing – Non-native speakers sometimes misinterpret a word for another word that sounds similar, for example, *ride* for *write, mouth* for *mouse, queen* for *clean,* and *store* for *story*. Because of these lexical recognition errors are not sufficiently compensated for by structural or semantic constraints in larger constituents, the speaker has difficulty making sense of the question or sentence and sometimes constructs a reasonable meaning and then hears other subsequent words incorrectly. Often, the meaning of an entire question can hinge on a single word or part of a word. For example, with the previous example, "What does a tadpole grow into?" some non-native speakers miss the *to* of *into* and answer, "water." Non-native speakers may also miss the type of information the question is asking for and often respond with *who* when the question is asking *where*.

Syntax – In contrast to natives who have a strong sense of what sounds grammatical and ungrammatical, non-native speakers have much weaker intuitions and must learn syntactic rules including correct agreement, mor-

phology, and sentence structure. Because syntactic information is still being induced or learned by many non-natives, they often produce unexpected forms in their responses, for example, "a trunks," "an eyes," "two duck," "by my eyes," "most the people." Semantic errors can happen when the speaker tries to re-situate the answer in a complete sentence based on the question. For example, a non-native speaker might say, "I will …" to a question starting, "Would you …"

Cognitive Load – For many non-native speakers, formulating responses to prompts is much more challenging than for native speakers and often taxes an individual's cognitive resources. This becomes evident when the speaker cannot remember the constituents mentioned in the sentence, and the speaker says, "the first one," or "not Susan" in response to simple questions that give two choices. These types of responses do not appear in the native speaker's response distributions.

Bringing the patterns of the two groups together, it seems that there are several attributes of spoken language that are common to native speakers and distinguish them from non-natives. Natives share an ability to parse the language effortlessly, have access to a basic lexicon of at least 8,000 words, and have mastered performance with the various levels of structure. Natives perform automatically with the language, which allows them to track the entities in a discourse and follow complex trains of thought beyond the mechanics of using the language. Automaticity is the ability to access and retrieve lexical items, to build phrases and clause structures and to articulate responses without conscious attention to the linguistic code (Cutler, 2003; Jescheniak, Hahne, and Schriefers, 2003; Levelt, 2001). Automaticity is required for the speaker/listener to be able to devote attention to what needs to be said rather than to how the code is to be structured. Automaticity provides a framework for the test construct: *facility with the spoken language*. The patterns from the response distributions clearly show that natives have extra capacity to focus on other thoughts aside from the language, while non-native speakers who are not proficient in the language do not. The fact that the two groups differ in this way is consistent with the construct of the test and also offers validation of the way the models were designed.

6 Conclusions

When considering the diversity of language, one challenge is to be able to define what the standard of a language should be in order to assess the performance of an individual speaker in that language. The method proposed here produces a workable model that captures some aspects of a language that are uniform across cultures and geography in order to encapsulate what is common (and thus standard) for almost all native speakers. The method

has been employed in the development of tests of spoken American English and Latin American Spanish and is currently being used to develop tests of Japanese, Dutch, French, and German. The method starts by developing a representation of the sounds of the language and continues to the level of the sentence by modeling responses to spoken probes posed in colloquial forms. The development process steps through how to model the content of what a speaker says in addition to the way the speaker says it with regard to pronunciation and fluency. The model may be useful in that it can distinguish native versus non-native speakers in the context of a language test.
The evidence supporting the effectiveness of the models are high reliability, clear separation of native versus non-native scores of spoken performances, and high correlation with other human-rated measures of spoken language proficiency. Even when presented with a group of native speakers from a dialect outside those dialects used in constructing the models, the models accurately characterize these speakers' high level of proficiency with the language. By discriminating native versus non-native speakers, the model embodies information about what constitutes a standard native performance for that language.

A by-product of the method is a set of data for how natives and non-natives respond differently when using the language. These patterns can contribute to our basic understanding of which characteristics are core traits of the language, as they are manifest in the speech of native speakers, but also the diversity introduced by learners of the language. Our understanding of non-native performances can enrich not only our notions of what should or should not be considered standard, but also our pedagogy and our perspectives on how languages change in contact.

References

Bachman, L.F. 1990. *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Balogh, J., J. Bernstein, I. Barbier & E. Rosenfeld. 2005. Impact of Native Accent on a Computerized Assessment of Pronunciation. Poster presented at the *1st Acoustical Society of America (ASA) Workshop on Second Language Speech Learning,* Vancouver, Canada.

Canale, M. & M. Swain (1980) Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics* vol. 1: pp. 1-47.

Carroll, J. B. 1961. Fundamental Considerations in Testing for English Language Proficiency of Foreign Students. *Testing*. Washington, DC: Center for Applied Linguistics.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Cutler, A. 2003. Lexical Access. In L. Nadel (Ed.), *Encyclopedia of Cognitive Science. Vol. 2, Epilepsy – Mental Imagery, Philosophical Issues About*. London: Nature Publishing Group, 858-864.

Eysenck, M. & M. Keane 1995. *Cognitive Psychology*. Hove, UK: Psychology Press.

Hymes, D. 1972. "Models of interaction of language and social life," in J. Gumperz & D. Hymes (eds.): *Directions in Sociolinguistics: The Ethnography of Communication*. New York: Holt Rinehart and Winston: pp. 35-71.

Jescheniak, J. D., A. Hahne & J. J. Schriefers. 2003. Information Flow in the Mental Lexicon during Speech Planning: Evidence from Event-Related Brain Potentials. *Cognitive Brain Research,* 15(3), 261-276.

Lado, R. 1961. *Language Testing*. New York: McGraw-Hill.

Levelt, W. J. M. 2001. Spoken word production: A theory of lexical access. *PNAS,* 98(23), 13464-13471.

Linacre, J. M. 2003. *Facets Rasch Measurement Computer Program*. Chicago: Winsteps.com.

Linacre, J. M., B. D. Wright & M. E. Lunz. 1990. A Facets Model of Judgmental Scoring. *Memo 61*. MESA Psychometric Laboratory. University of Chicago. www.rasch.org/memo61.html.

Miller, G. A. & S. Isard. 1963. Some Perceptual Consequences of Linguistic Rules. *Journal of Verbal Learning and Verbal Behavior, 2,* 217-228.

Young, S., D. Kershaw, J. Odell, D. Ollason, V. Valtchev and P. Woodland. 2000. *The HTK Book Version 3.0*. Cambridge, England: Cambridge University.