Fluency and Structural Complexity as Predictors of L2 Oral Proficiency

Jared Bernstein, Jian Cheng, Masanori Suzuki

Knowledge Technologies, Pearson
299 S. California Ave, Palo Alto, California 94306, USA
jared.bernstein@pearson.com, jian.cheng@pearson.com, masanori.suzuki@pearson.com

Abstract

Automaticity and real-time aspects of performance are directly relevant to L2 spoken language proficiency. This paper analyzes data from L2 speakers of English and Spanish spread over a range of proficiency levels as identified by traditional holistic, rubric-based human ratings. In spontaneous speech samples from these L2 populations, we studied timed measures of spoken fluency (linguistic units per time) that co-vary with proficiency level and compared the timed measures to indices of the linguistic complexity of the same spoken material. Results indicate that duration-based fluency measures yield as much or more information about proficiency as do structural complexity measures. These empirical findings suggest that expert perception of oral proficiency relate to automatic, real-time aspects of speaking and that the oral proficiency construct may be enriched by adding timing to its communicative/functional framework.
Index Terms: automaticity, psycholinguistics, fluency

1. Introduction

Skilled language teachers integrate linguistic and non-linguistic evidence in the assessment of L2 spoken language proficiency. That is, when listening to a functionally appropriate passage of speech to assess a learner's oral language, listeners seem to combine linguistic judgments of the content, complexity, and form-accuracy of a learner's speaking with a seemingly non-linguistic judgment of the apparent automaticity, fluency, or facility of speech production. Compared to native speakers and proficient non-native speakers, early language learners say less and they say it more slowly, and what they say is structurally simpler – even when the learner is successfully communicating.

In this paper, we will take "fluency" to represent the salient manifestations of automaticity in speaking. The basic measures of fluency are the rates (in time) of the production of spoken linguistic units, like words per minute, leaving aside for now the more complicated problem of how to measure smoothness or continuity in spoken language. But what is automaticity? Segalowitz and Hulstijn [1] say that "automaticity refers to the absence of attentional control in the execution of a cognitive activity", but they cannot pin it down to a single operational definition that all will agree to. Some psychologists classify a behavior as automatized when performance is asymptotically accurate without conscious attention to the mechanics of the task (e.g. decoding written words during highly skilled reading). Automaticity can be operationalized as asymptotic accurate performance that is indifferent to performance of another attention-needing task. For example, highly proficient speakers can talk quite fluently on light topics while steering a vehicle through traffic, because their speaking is automatized.

A current view in applied linguistics is that language use encompasses multiple communicative competencies: grammatical competencies related to vocabulary, syntax and so forth; 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 [2]. The claim is that oral proficiency is the ability to accurately produce certain structures appropriate to the context, using the elements of the competencies as they are ordered in increasing difficulty or complexity.

This paper analyzes two sets of data. We offer some estimates of the relation of the timed production of language structures (fluency measures) and the structural complexity observed in that same data to the growth of oral proficiency, as judged by expert raters.

2. Method

2.1. Overview of English and Spanish Experiments

The experiments reported in this study were conducted in the period from 1998 to 2007. They are based on two data sets, one collected in 1998-99 from non-native speakers of English (the ENG subjects) and one collected in 2002-3 from non-native speakers of Spanish (the SPN subjects). The two data sets will be considered two experiments; one on English learners and one on Spanish learners. In both experiments, each subject took two different speaking tests in close temporal proximity with less than 7 days elapsed between the test administrations. One test was an automatically administered and automatically scored listening/speaking test that took 12-18 minutes to complete. For the automatic tests, all the candidate responses were recorded and are available, along with several summary scores. In both the English and Spanish experiments, the other test was a human-scored oral proficiency assessment from which holistic speaking proficiency scores were obtained based on multiple human ratings. For the Spanish test we have the two separate scores from the two raters, and for the English test we have only one officially reported score. For each dataset, a corresponding set of measures were generated: 1) oral proficiency test scores based on human judgments, 2) a set of phonological fluency measures (e.g. phonemes/time), 3) a set of structural fluency measures (e.g. clauses/time), 4) a set of structural complexity measures (e.g. words/clause).

The purpose of the experiments in the current context is to understand the growth of fluency and structural complexity in spoken language as general spoken language proficiency develops. In particular, which aspects of spoken language production relate most closely and consistently with the more general construct of spoken language proficiency as judged by expert raters? The original spoken response data was collected as part of concurrent validation studies for the Versant English and Versant Spanish tests, and it has now been analyzed to ex-

tract phonological fluency, structural fluency, and complexity indices.

2.2. Design

In these studies, the independent variable is general proficiency in spoken language, and the dependent variables are eleven measures of production fluency and three measures of structural complexity as observed in learners' spoken responses. The experimental design is schematized in Figure 1, which shows how the subjects, the instruments and the independent and dependent variables are related.

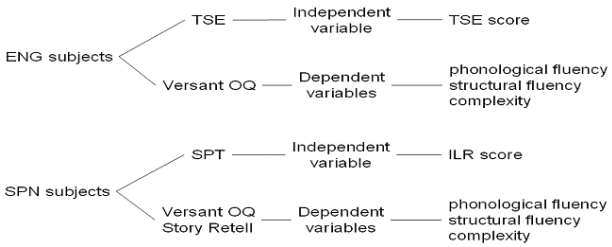

Figure 1: A graphical representation of the experimental design

In the following sections, the test instruments and scoring procedures are described. The independent criterion variables are rubric-defined scores that are based on expert human judgments of communicative effectiveness in speaking or based on oral proficiency. The dependent variables are measured from extended spoken turns in linguistic units/time or ratios of embedded linguistic units.

2.3. Independent Criterion Variables: Speaking Proficiency

Each of the subjects in the ENG and SPN data sets took a general oral proficiency test, with holistic scores assigned by expert human listeners. The 58 ENG subjects took the Test of Spoken English (TSE) offered by Educational Testing Service (ETS). The TSE was administered to the ENG subjects between October 1998 and April 1999. The TSE is a well accepted test that is used to measure the ability of nonnative speakers of English to communicate orally in English, particularly in post-secondary education and in professional settings. According to Educational Testing Service (as cited in [3]), the TSE measures "the ability to accomplish specific language tasks comprehensibly, accurately, coherently, and appropriately with respect to specific interlocutor/audience, topic, and purpose (1994, p.1)."

Each of the 38 SPN subjects took a Spoken Proficiency Test (SPT) in Spanish. This Spanish SPT was an oral proficiency interview test administered in one of the U.S. government organizations. The scores provided were holistic proficiency levels following the Interagency Language Roundtable (ILR) descriptors (0, 0+, 1, 1+, 2, 2+, 3, 3+, 4, 4+, 5). The ILR scale is presented on the ILR website http://www.govtilr.org/ILRscale2.htm.

2.4. Recorded Question Answers and Story Retellings

Subjects took Versant English and Versant Spanish tests. Pearson's Versant tests are automatically administered and scored using a computerized test delivery system and speech processing technology respectively. All the Versant tests in these two experiments were administered to the subjects over the telephone and had durations in the range of about 12-18 minutes per administration.

In the experiments reported here, we analyze only the last few responses from the Versant tests in which subjects were given a fixed length recording window to answer an open question or to re-tell a narrative. In the last section of the Versant English test, the English Open Questions, the system presented two items and provided a fixed 40-second recording window for each response. In the last sections of the Versant Spanish test, the Spanish Open Questions and Story Retellings, the system presented three items of each type and gave a fixed 30-second recording window for each item response. The two open questions in the last section of the Versant English provide subjects a total of 80 seconds to respond. The three open questions and three story retellings in the Spanish test provide subjects a total of 180 seconds to respond.

2.5. Human Analysis of Open Questions and Story Retellings

These responses (two per English test and six in each Spanish test) were all transcribed into an augmented orthographic form by an experienced team of transcribers (English or Spanish). All transcriptions were then reviewed by a second, supervisory transcriber. The transcription methods are applied to tens of thousands of responses each year as part of the usual test development procedure at Pearson. What we describe below is the hand analysis of these transcriptions to count the occurrence of the four kinds of linguistic units shown below:

- Words: Response-relevant orthographic words, not counting apparent disfluencies (e.g. self-corrections, false starts, fillers, or stutters)
- Cohesives: Coordinators, logical connectors, and devices indicating repetition, lexical relationship, reference, ellipsis, comparison, and conjunction [4], e.g. *but, before, yet, because, so, even if, therefore, however*, etc.
- Clauses: Structures with a verb (usually finite), including independent and various kinds of dependent clauses (subject, adverbial, relative, etc)
- T-units: "One main clause plus any subordinate clause or non-clausal structure that is attached to or embedded in it" [5]

From these counts, we derived measures of fluency with longer units (e.g. words/minute), and measures of unit complexity (words/clause or clauses/t-unit).

For the ENG data set, four content coders were recruited. First, the coders had a training session in which the goals of and procedures of the response analyses were discussed. Then, each coder independently analyzed training transcriptions for each syntactic measure. After the training session, each coder was given the same 103 transcribed Open Question responses and independently coded for each measure. For the SPN data set, four different content coders were recruited and trained. The Spanish coders followed the same rating procedures as the English coders. After the training session, each coder was given 220 Spanish transcriptions and independently coded them for each measure.

2.6. Dependent Variables

Fluency Measures: There are many speech parameters that have been identified with fluency. For example, Kormos [6]

summarizes ten most common temporal variables that are often used in L2 fluency research. Those variables are: Speech Rate, Articulation Rate, Phonation-Time Ratio, Mean length of Runs, Number of Silent Pauses per Minute, Mean Length of Pauses, Number of Filled Pauses per Minute, Number of Disfluencies per Minute, Pace, and Space. Kormos [6] reports that speech rate and mean length of runs are the best predictors of L2 fluency judgments.

In the present study, we selected four phonological fluency measures, as listed below, that could be measured easily from the response recordings using the speech processing technology.

- Total Pause Time: summed duration of silent or filled inter-word within-response pauses
- Mean Pause Time: average duration in milliseconds of inter-word pauses
- Phones/Second: number of phonemes per second (while speaking)
- Words/Minute (ROS): number of words per minute from beginning to end of speaking

The second set of fluency variables relate to the production of larger structures, like clauses, or special words that link larger structures. These structural fluency variables are also measured with reference to time – how many are produced per time or how long does it take to produce one. In this study, they are measured with respect to the time available for speaking, and not with reference to the trimmed speaking time. The four variables under study are:

- Words/minute: relevant words per minute over the available recording time
- Clauses/minute: relevant clauses per minute over the available recording time
- T-units/minute: relevant T-units per minute over the available recording time
- Cohesives/minute: cohesive words or phrases per minute over the available recording time

These variables represent the number of units that are produced per minute as part of a task-relevant response over the whole duration of a recording period. They are not the only possible structure/time variables that could be studied, but they may be representative of the class of such variables.

Complexity Measures: There have been many studies in the second language acquisition literature that investigated complexity of output made by L2 learners in either written format [7] or spoken format [8, 9, 10]. Most of the syntactic complexity measures used to analyze speech samples employed the same measures used in L2 writing analysis [10], although speech samples are often more complicated to analyze due to false starts, hesitations, and fillers. Foster, Tonkyn, and Wigglesworth [11] propose a new analysis unit called AS-unit specifically keeping characteristics of speech samples in mind; however, coding AS-units is a fairly complex process and "so few studies have used it" [10].

In the present study, some commonly used syntactic complexity measures (including T-units/minute, Clauses/minute, Cohesive Devices (CD), Words/minute, Words/T-unit (WPT), Words/Clause (WPC), Cohesives/word) were identified from previous studies [particularly 10, 12]. The parameters analyzed in this study are:

- Cohesives/Word: Average number of cohesives per word
- Words/Clause: Average number of words per clause
- Words/T-unit: Average number of words per T-unit
- Clauses/T-unit: Average number of clauses per T-unit

These four complexity variables are not determined with reference to time and do not represent the ability to produce units fluently. A single complex T-unit, with three clauses and many words counts the same as four such T-units produced in the response recording window. Again, these are not necessarily the best such variables to analyze, but they are probably similar in behavior to other such variables that might be tracked and analyzed.

3. Results

The results reports on the correlations between the variables and highlights the relative centrality of different aspects of fluency and complexity as they contribute to spoken language proficiency. A set of variables that can combine to account for most of the variance in proficiency may also inform instructional design or language testing.

A correlational analysis of the results indicates that most, but not all, of the dependent measures change in an orderly manner over the course of second language development. Table 1 displays the correlation of fluency parameters and complexity parameters with the over all spoken language proficiency measures.

type	variable	Corr1	Corr2
Phonological Flu.	Total Pause Time	0.22	-0.29
Phonological Flu.	Mean Pause Duration	-0.21	-0.68
Phonological Flu.	Phones/second (speaking time)	0.57	0.81
Phonological Flu.	Words/minute (utterance time)	0.59	0.86
Structural Flu.	Words/minute	0.62	0.90
Structural Flu.	Clauses/minute	0.56	0.85
Structural Flu.	T-Units/minute	0.47	0.85
Structural Flu.	Cohesives/minute	0.55	0.79
Complexity	Cohesives/word	0.02	0.26
Complexity	Words/T-Unit	0.67	0.70
Complexity	Words/Clause	0.50	0.55
Complexity	Clauses/T-Unit	0.39	0.51

Table 1: Correlations of individual fluency and complexity variables with proficiency measures. Corr1: Correlation with TSE (English) (80 sec. sample). Corr2: Correlation with Spanish SPT (180 sec. sample).

With these data set sizes, we can consider any correlation $\geq |0.50|$ to reflect a "meaningful" predictive relationship, as the dependent variable accounts for more than a quarter of the variance in the independent variable, proficiency. In general, the Spanish correlations are higher with only two correlations of magnitude less than 0.50. In the English data, seven of the twelve variables correlate with magnitudes of 0.50 or greater, including two or three variables from each of the variable groups.

For the English data, all the meaningful fluency and complexity variables are within a narrow range, between 0.50 and 0.67. No individual variable and no group of variables seems to stand out in having a particularly (or differentially) strong relation to proficiency.

In the Spanish data, the fluency and complexity variables are spread over a wider range of magnitudes, from 0.51 to 0.90. Also different is that the phonological and structural fluency variables seem to have a stronger relation to proficiency than do the complexity variables. This is born out in the analyses presented in Table 2, where a bootstrap cross-validation pro-

cedure [13] produces a fair estimate of how a combined variable based on a multiple linear regression of the four variables in each group is correlated with the corresponding proficiency measure. Table 2 presents the mean correlation value from 50 random partitions of the subject data set. Note that five of the six estimates of the expected correlations between the 4-variable multiple regressions and the oral proficiency values are less than the best single variable correlations with proficiency in the same macro-cell of Table 1. This is probably due to the errors caused by the limited data points.

Type	Variable	Corr1	Corr2
Phonological Flu.	Utterance time	0.56	0.88
Structural Flu.	Aavailable time	0.58	0.89
Complexity	Density, complexity	0.60	0.62

Table 2: Correlations of fluency and complexity variable groups with proficiency measures. Corr1: Correlation with TSE (English) (80 sec. sample). Corr2: Correlation with Spanish SPT (180 sec. sample). Variables are in Units/minute.

The results in Table 2 suggest that there is no evident difference among the measured variable types for the English speaker sample, but the Spanish data set suggests that the both the fluency variable types are much more strongly related to proficiency than are the complexity variables.

4. Discussion

First, consider why the Spanish performance data seems to show a more orderly relation to proficiency than the English data. There are four plausible explanations that come to mind: the aggregate spoken response time is longer; the re-telling task is more constrained; the independent variable is more reliable; and the independent variable comes from a more interactive task.

Although there are fewer SPN subjects, each SPN subject had more than twice as much time available to provide data for this comparative analysis. That is, there is more data per measurement at the first level of analysis, so the first-level counts are more stable and thus correlation coefficients are likely to be greater, even assuming the same underlying relations between the variables for the two languages.

The data may be better in Spanish because the Story Retelling task provides more constraints on response content. That is, purely personal or intellectual (non-language) differences in approaching the task may be more evident in the responses to a less-constrained task like the Open Questions. The Spanish data is based on three open question responses and three story retellings. Splitting the Spanish data to compare Story Retelling to Open Questions, all four of the structural fluency variables correlate better with proficiency when using the data from the retellings rather than from open question responses. The English data is all from Open Questions, which are less constrained and provide less information per time about proficiency.

For these reasons, phonological and structural fluency measures seem much more closely related than complexity measures to what interlocutors hear as proficiency, as is indicated by the L2 Spanish data. However, a more conservative interpretation, looking at the L2 English data, may be that each of the three types of variables is similar in its relation to perceived spoken proficiency. As presented in Table 2, the phonological fluency indicators (amount and length of pausing, speed at pronouncing segments, and rate of word production) can together predict about one third of the variance in oral proficiency. The structural fluency and the complexity indicators can do about as well. The question that remains is: is there evidence here to suggest that one or another of these variable types is more central to the construct? Perhaps not, but there is some evidence that producing a greater amount of acceptable lexical and structural information per time is a lot of what is heard as oral proficiency.

The current data may help lead to the design of more efficient and accurate spoken language tests. Combining linguistic analysis of spoken material with measures of the timing of the production of these linguistic units will yield better estimates of speaking proficiency.

These empirical findings suggest that the human perception of oral proficiency has a strong relation to automatic, real-time aspects of speaking and that the oral proficiency construct may need revision from an underspecified functional grounding that takes fluency to be just one of many equally important elements that make a spoken performance effective in communication.

5. References

[1] Segalowitz, N. and Hulstijn, J., "Automaticity in bilingualism and second language learning", In Handbook of bilingualism: Psycholinguistic approaches (pp.371-388), Oxford University Press, Oxford, 2005.

[2] Bachman, L. F., "Fundamental considerations in language testing", Oxford University Press, Oxford, 1990.

[3] Powers, D. E., Shedl, M. A., Wilson-Leung, S., and Butler, F.A., "Validating the revised TSE® against a criterion of communicative success", Research Report 99-05, Educational Testing Service, Princeton, NJ, 1999.

[4] Crystal, D., "The Cambridge encyclopedia of language", Cambridge University Press, Cambridge, 1997.

[5] Hunt, K. W., "Grammatical structures written at three grade levels", National Council of Teachers of English, Research Report No. 3, Urbana, IL, 1965.

[6] Kormos, J., "Speech production and second language acquisition", Lawrence Erlbaum, 2006.

[7] Wolfe-Quintero, K., Inagaki, S., and Kim, H., "Second language development in writing: Measures of fluency, accuracy, and complexity", University of Hawaii, Second Language Teaching and Curriculum Center, Honolulu, 1998.

[8] Halleck, G. B., "Assessing oral proficiency: A comparison of holistic and objective measures", The Modern Language Journal, 79, 223-234, 1995.

[9] Iwashita, N., McNamara, T., and Elder, C., "Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information processing approach to task design", Language learning, 51(3), 401-436, 2001.

[10] Iwashita, N., "Syntactic complexity measures and their relation to oral proficiency in Japanese as a foreign language", Language Assessment Quarterly, 3(2), 151-169, 2006.

[11] Foster, P., A Tonkyn, A., and G Wigglesworth, G., "Measuring spoken language: a unit for all reasons", Applied Linguistics, 21(3), 354-375, 2000.

[12] Farhady, H. and Farzanehnejad, A. R., "An objective measure for evaluating EFL compositions", In 25 years of living with applied linguistics: A collection of Articles (353-365), Rahnema Publications, 2006.

[13] Efron, B. and Tibshirani, R., "An introduction to the bootstrap", Chapman and Hall, New York, 1993.