Fluency Changes with General Progress in L2 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

Second language (L2) learners tend to speak slower at every level of linguistic analysis, often in an uneven tempo, with longer pauses at the start and before some words and constructions, than is typical of native speech. As noted by Zhang & Elder [1], native listeners focus on phonological fluency in making judgments about L2 proficiency. Improved understanding of how fluency grows with progress in overall oral proficiency may lead to measures of fluency that would be useful for measuring proficiency itself. Spontaneous speech sampled from populations of L2 speakers of English and Spanish showed orderly, seemingly linear increments in the rates at which words and larger constituents are spoken as a function of human-judged general proficiency level. Results suggest that unit/time fluency measures match native expert perception of oral proficiency, supporting the hypothesis that performance-in-time is a core attribute of speaking proficiency and efficient spoken communication.

Index Terms: L2 production, psycholinguistics, fluency measurement, proficiency test

1. Introduction

In a previous study [2], fluency and structural complexity were compared as predictors of L2 oral proficiency. Results suggested that timed measures that reflect fluency are better predictors of overall proficiency than are counts of units or measures of content complexity. Here we wish to find out just how fluency changes with progress in general L2 speaking proficiency.

Zhang & Elder [1] find that native listeners, including skilled language teachers, naturally integrate different kinds of evidence in the assessment of L2 spoken language proficiency. That is, judges are sometimes given specific rubrics to judge by (e.g. content, complexity, and form-accuracy) but they still give strong weight to performance aspects such as fluency. Segalowitz and Hulstijn [3] (Page 371) say that "automaticity refers to the absence of attentional control in the execution of a cognitive activity", and posit it to be the underlying source of measurable fluency. Conventional measures of fluency used with both spontaneous speech and oral reading include the rates (in time) of the production of spoken linguistic units, like words per minute.

To understand the growth of fluency in spoken language as general spoken language proficiency develops, we reanalyze the experimental data. Two sets of spoken responses are analyzed here: one is L2 Spanish (SPN) and the other is L2 English (ENG). We describe the materials and procedure quite briefly below, because details are available in [2].

2. Method

2.1. Overview of English and Spanish Experiments

Two data sets are analyzed - one has English L2 learners and one has Spanish L2 learners. Each learner took two different speaking tests within a 7 day period. One test was an automatically administered and scored Versant listening-speaking test that kept a recording of all candidate responses. 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 two separate scores (from two raters), and for the English test we have only one officially reported score. For each dataset, we generated: 1) oral proficiency test scores based on human judgments, 2) a set of phonological fluency measures (e.g. nominal-phonemes/time), and 3) a set of standardized Versant fluency scores. Differences in the datasets are discussed below in sec. 4.1, but evidence from two languages may suggest that the findings are not artifacts.

2.2. Design

The experimental design is shown in Figure 1, which relates the subjects, the instruments and the meausurements.

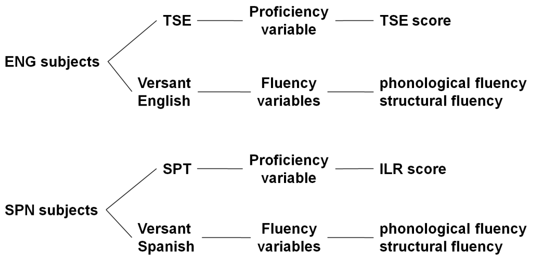

Figure 1: Representation of the experimental design

Each subject in either the ENG or SPN group 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), and the 38 SPN subjects took a Spoken Proficiency Test (SPT) in Spanish; both tests provided holistic proficiency levels. The TSE subjects were adult learners in Albany, NY with various first language backgrounds (e.g., Russian, Chinese, Spanish). The majority of the SPT subjects were English speakers (92%), recruited from different parts of the U.S. The learners also took Versant English and Versant Spanish tests [4, 5]. We analyzed relatively extended spontaneous spoken responses from the Versant tests. Each of two open questions in the Versant English Test provided a 40 second response window (80 seconds total per ENG learner). Each of three opinion questions and three passage retellings offered a 30-second response window (totaling 180 seconds) for each SPN learner to speak spontaneously. The total time available to a learner to speak (ENG: 80 seconds, and SPN: 180 sec) will be called "window time" or "wT". This is opposed to "speech time" or "sT", which is the sum of the durations (for all response windows) from the beginning of the first spoken word to the end of the last spoken word, including all interword silences, but excluding initial and trailing silence.

2.3. Human Analysis of Spoken Material

All extended speech samples were transcribed by a trained native and all transcriptions were then reviewed by a second, supervisory transcriber. The responses were transcribed in an augmented orthographic form of English or Spanish that also recorded the occurrence of the following linguistic units:

- Words: Response-relevant orthographic words
- Cohesives: Coordinators, logical connectors, and devices for semantic relation [6], e.g. *but, before, yet, because, so, even if, therefore, however*, etc.
- Clauses: Structures with a (usually finite) verb
- T-units: Main clauses with subordinate structures [7]

Usual speech rate measures (e.g. words/minute) are easily derived from ASR-aligned signals. Kormos [8] summarizes ten common temporal variables 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 [8] reports that speech rate and mean length of runs are the best predictors of L2 fluency judgments. We selected four phonological fluency measures that are easily measured with speech processing technology.

- Total Pause Time: duration of interword pauses (silent or filled)
- Mean Pause Time: average duration of interword pauses
- Articulation Rate (ART): phonemes/second of speech time (sT)
- Rate of Speech (ROS): words/minute of speech time (sT)

A second set of fluency variables represent the rate of production for larger structures (e.g. clauses). These rates are measured with respect to the time available for speaking (window time, wT), and not with reference to the trimmed speech time, sT. The four variables are:

- Words/minute: words/minute of window time (wT)
- Clauses/minute: clauses/minute of window time (wT)
- T-units/minute: T-units/minute of window time (wT)
- Cohesives/minute: cohesives/minute of window time (wT)

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

3. Results

The results of this study may serve as preliminary benchmarks for understanding the development of fluency in spontaneous spoken language as general spoken language proficiency develops. The mean score of the TSE was 39.1 with a standard deviation of 10.7 in the range from 20 to 60. The mean overall score of the Versant English test was 48.2 with a standard deviation of 18.4 in the range from 20 to 80. The mean score of the SPT was 2.1 with a standard deviation of 1.0 in the range from 0 to 5. The learners cover the whole range of the TSE and nearly the whole range of the SPT, as shown in Figure 2.

Figures 3 and 4 present the data for four fluency variables as a function of oral proficiency (x-axis). The Tukey boxes are drawn from five numbers: the smallest observation, lower quartile (Q1), median, upper quartile (Q3), and largest observation. Separate points shown as"+" are outliers.

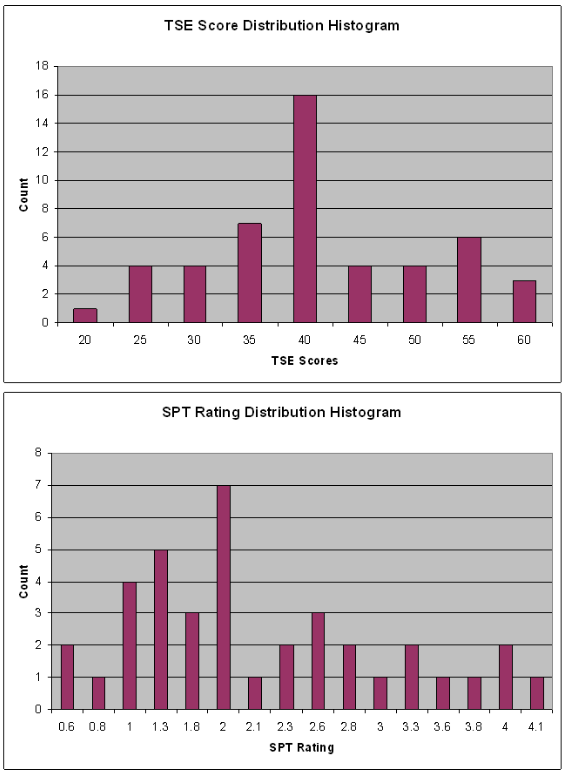

Figure 2: Histograms show distribution of subjects on TSE and SPT. The SPT shown is the average score from two raters.

Figure 3 shows the increasing trend of Rate of Speech, ROS, in words per minute and Articulation Rate (ART) in phonemes per second in Spanish as a function of increasing ILR (Interagency Language Roundtable) levels in spoken Spanish [9]. Learners judged to be at early levels of learning produce Spanish at about 60 words per minute (when leading and trailing silence is trimmed), and they almost triple in rate of speech until they are speaking at about 150 words per minute when they reach the ILR level 4. Articulation rate also increases steadily from the first stages of learning (about 7 phonemes/second) to the ILR level 4 (about 12 phonemes/second). These results for Spanish show a clear and consistent increase in fluency that likely reflects increased automaticity in L2 lexical retrieval and L2 articulation.

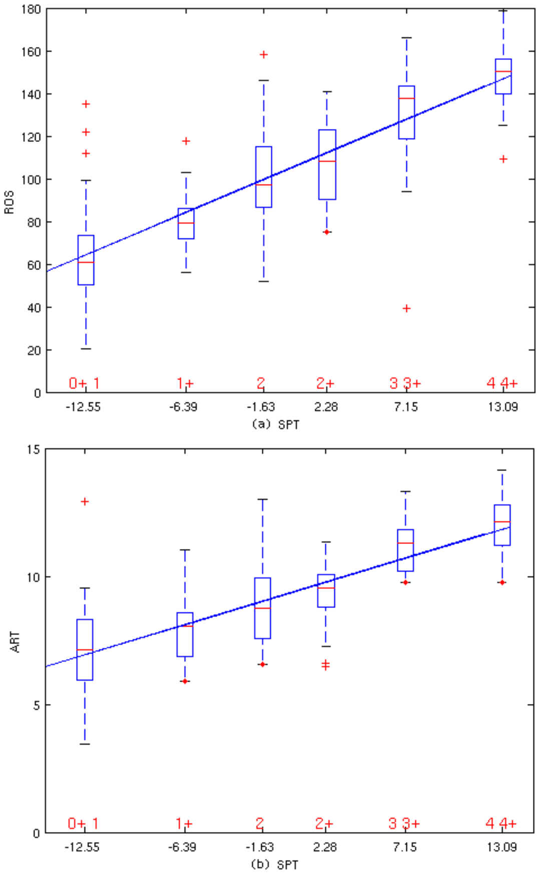

Figure 3: (a) Rate of Speech (ROS) and (b) Articulation Rate (ART) as a function of Spanish SPT test score grouped into six levels derived from listeners' overall proficiency judgments. Each of the 219 data points represent one 30-second response. Median scores and quartiles shown in the boxes. Data range is shown in the vertical extensions and outlier points. A linear regression line is superimposed on the graph.

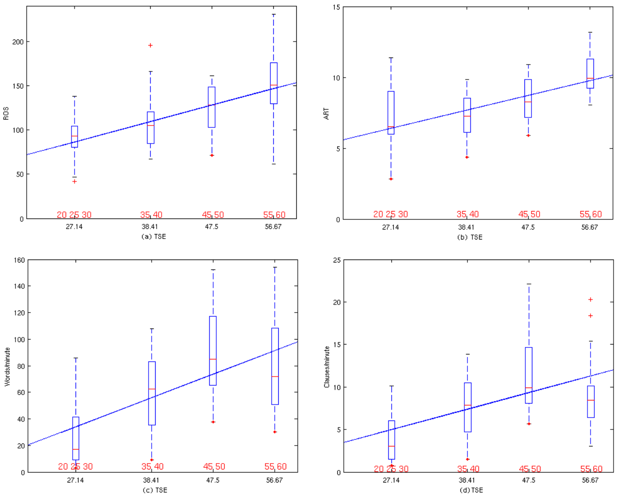

Figure 4: Four fluency variables shown as a function of TSE score in four score-level groups (20-30, 35-40, 45-50, 55-60). (a) Rate of speech (ROS), (b) Articulation Rate (ART), (c) words per minute, and (d) clauses per minute. The 92 data points are 40-second responses. Linear regression lines are superimposed on each graph.

The upper two displays (a & b) in Figure 4 show (a) the increasing trend of Rate of Speech, ROS, in words per minute and (b) Articulation Rate (ART) in phonemes per second in English as a function of increasing general proficiency in spoken English. Learners judged to be at early levels of learning produce English at about 90 words per minute (when leading and trailing silence is trimmed), and they almost double in rate of speech until they are speaking at about 150 words per minute when they reach the TSE scores of 55 or 60. Articulation rate also generally increases from the first stages of learning (about 7 phonemes/second) to TSE scores of 55 or 60 (about 10 phonemes/second). Thus, L2 English like L2 Spanish, shows a consistent increase in phonological fluency that may reflect increased automaticity in L2 lexical retrieval and L2 articulation.

The lower two displays (c & d) in Figure 4 show the trend of words per minute (wT) and clauses per minute (wT) in English as a function of increasing communication levels in spoken English. Learners judged to be at early levels of learning produce around 15-25 words per minute in a 40 second turn, and they are speaking at about 80 words per minute when they reach TSE scores above 45. Rate of clause production also triples from about 3 clauses per minute at the first stages of learning to about 8-10 clauses per minute when TSE scores reach 45 or above. Thus, the structural fluency variables for English also show large increases from low TSE levels up to TSE levels of 45 and above. These increases would be consistent with increased automaticity in the construction of phrases and clauses. Note that in both Figure 4 (c) and (d), the rate of word and clause production flattens or even seems to decrease slightly above the TSE 50 level. This pattern may be due to the nature of the task demand in relation to the measure, in that these structural fluency measures are calculated with reference to the elapsed time of the recording window; leading and trailing silence is not trimmed. Subjects are given a fixed 40-second recording window, and some proficient L2 English speakers (above TSE 50) may have felt that they answered the open question adequately and did not need to fill the allotted time.

	Lowest group TSE: 20-30, N = 9		Highest group TSE: 55-60, N = 9	
	Mean	Median	Mean	Median
Total Pause Time (ms)	2339	2365	4570	3660
Mean Pause Duration	148.4	128.3	96.7	49.8
Phones/sec	7.2	6.6	10.4	9.9
Words/min (trimmed)	92.8	89.9	152.9	154.8
Words/min	22.3	16.5	76.7	75.6
Clauses/min	3.6	3.0	9.3	9.2
T-units/min	2.5	1.9	5.7	4.9
Cohesive/min	1.1	1.5	3.5	4.3

Table 1: Average fluency values at highest and lowest English speaking levels.

Table 1 displays the mean and median values for the highest and the lowest proficiency cohorts as measured by the TSE scores. The values give an indication of how the 8 variables range from English learners who give evidence of little effective communication to L2 speakers whose communication is judged to be almost always effective. Learners who cannot communicate much according to TSE standards are speaking at about 90 words per minute when they speak, but they do not say much. From the data in Table 1, you might expect learners to double their words per T-unit and triple their clauses per minute over the course of learning spoken English to a high level.

4. Discussion

4.1. Differences in the Data Sets

These data sets are dependent on the accuracy and reliability of the human judges who assigned the criterion scores to these subjects, and there is no reason to suppose that these scores or levels are particularly inaccurate, and both the TSE scores and the ILR levels reported in this study have high reliabilities as measured in operation in these administrations. So the criterion measures of general speaking proficiency are accepted.

There are real differences in the two data sets: in the two criterion measures, in the two learner populations, and in the kind and length of spontaneous speech samples analyzed. Most importantly in the criterion measures in our data sets: the SPT scores are more reliable than the TSE scores (0.94 vs. 0.89). This difference in reliability index indicates that there is 12% more usable information in the SPT scores than in the TSE scores used in the analysis (and more than three times as much error variance in the TSE scores). On the assumption that the fluency variables are a key element in general spoken language proficiency, lower reliability of the criterion variable will be seen in lower correlations with it. Also, nearly one third of the ENG subjects are at the mid-point of the TSE scale, whereas the modal ILR level of the SPN subjects has only about one sixth of the subjects, making for a more even distribution. A steeply center-peaked distribution attenuates correlation coefficients, all else being equal.

Another difference in the criterion measures is that the oral proficiency construct for the SPT raters was more interactive, more time-sensitive, than the communicative ability construct that the TSE raters were using. The SPT interviews were real-time interactions over the telephone, with some social responsiveness expectations in force, whereas in the TSE administrations the test-takers had printed support for the tasks, time to prepare for the responses, and the responses were spoken into and scored from record/play devices. Some of these elements may have made the TSE scores less sensitive to timed measures.

4.2. Inferences from the Data

We can review what we have found out from these studies and consider what sort of interpretation to impose on the results. Among both Spanish and English learners, there are similar patterns of linear-seeming growth in measured fluency for almost any unit/time as general proficiency grows. This pattern in the Spanish and the English data is consistent with the hypothesis that interviewer/raters are quite sensitive to the rate at which information is conveyed and structures are produced, regardless of the holistic rubrics the raters are instructed to follow.

If speech communication is viewed in terms of a communication channel, the key quality sought is bandwidth, or information per time. In this view, noise (interpreted here as unintelligible segments, inappropriate words, and unstructured phrases) is interchangeable with time. Given enough time, one can transmit and receive messages even when there is noise in the line. In a one-on-one conversation, participants can negotiate meaning even in the early stages of L2 development, as long as neither participant is in a hurry. Satisfactory, proficient communication, however, assumes a reasonable rate of successful transmission. The data found in this study is compatible with the conclusion that communicative effectiveness and speaking proficiency are at least as evident in the timing of the spoken material as in the relative complexity of it (cf [2]).

This leaves more fundamental questions, such as: How can speaking proficiency be taught more effectively? What are the core components of a speaking proficiency construct? Have we learned anything about how we can induce fluency improvement in a language learner? What are the implications for assessment and testing if performance in time is at the core of speaking proficiency and spoken communication?

4.3. Future Directions

The best choice of unit per time is open. We have measured latencies, pauses, phones, words, runs, clauses, and T-units per time. One could try other structural units such as lexical dispersion (e.g. new-types/time) or subordinate clauses per time. Alternatively, semantic or functional measures like the rate at which propositions [10], pragmatic effects, or something else are successfully transmitted to the expected listener. For now, the data reported simply suggest that timed variables are carrying at least as much proficiency information as complexity variables and probably more.

Two questions are closely related: How can speaking proficiency be taught more effectively? How can speaking proficiency be measured efficiently? If we can build and validate speaking tests that are very efficient, convenient and accurate, such test instruments can serve a thousand experiments in instructional design. The current data may inform the design of 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. Better ability estimates from a fixed amount of assessment material can support more accurate assessments and take up less student time.

5. References

[1] Y. Zhang and C. Elder, "Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs?," *Language Testing*, vol. 28, no. 1, pp. 31–50, 2011.

[2] J. Bernstein, J. Cheng, and M. Suzuki, "Fluency and structural complexity as predictors of L2 oral proficiency," in *Interspeech 2010*, pp. 1241–1244.

[3] N. Segalowitz and J. Hulstijn, "Automaticity in bilingualism and second language learning," in *Handbook of bilingualism: Psycholinguistic approaches*, pp. 371–388. Oxford University Press, Oxford, 2005.

[4] J. Bernstein and J. Cheng, "Logic and validation of a fully automatic spoken English test," in *The Path of Speech Technologies in Computer Assisted Language Learning*, V. M. Holland and F. P. Fisher, Eds., pp. 174–194. Routledge, New York, 2007.

[5] J. Bernstein, A. Van Moere, and J. Cheng, "Validating automated speaking tests," *Language Testing*, vol. 27, no. 3, pp. 355–377, 2010.

[6] D. Crystal, *The Cambridge encyclopedia of language*, Cambridge University Press, Cambridge, 1997.

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

[8] J. Kormos, *Speech production and second language acquisition*, Lawrence Erlbaum Associates, Mahwah, New Jersy, 2006.

[9] Interagency Language Roundtable, "ILR speaking skill scale," `http://www.govtilr.org/Skills/ILRscale2.htm`, 2011.

[10] W. Kintsch, "The role of knowledge in discourse comprehension: A construction-integration model," *Psychological Review*, vol. 95, no. 2, pp. 163–182, 1988.