Estudios de
lingüística inglesa aplicada

elia

COMPARING NATIVE SPEAKER RATINGS AND QUANTITATIVE MEASURES OF ORAL PROFICIENCY IN IELTS INTERVIEWS

COMPARACIÓN DE LA EVALUACIÓN DE HABLANTES NATIVOS CON LOS INDICADORES CUANTITATIVOS EN LA ENTREVISTA DE IELTS

Katherine O'Donnell Christoffersen
University of Texas Rio Grande Valley, USA
katherine.christoffersen@utrgv.edu

DOI: http://dx.doi.org/10.12795/elia.2017.i17.10

Research on second language acquisition has used various quantitative and qualitative measures to assess oral proficiency, yet there is little empirical research comparing these measures. Comparisons between quantitative measures and native speaker ratings are especially rare. Four of the most common quantitative measurements applied in L2 research include the type-token ratio as a measure of lexical diversity; the T-unit as a measure of syntactic complexity; the error-free t-unit as a measure of grammatical accuracy; and average speech rate as a measure of fluency. The present study compares these four quantitative measures of oral proficiency and one qualitative measure of oral proficiency, i.e., native speaker ratings, based on the speech of three non-native English speakers during the International English Language Testing System (IELTS) oral *interview. The results indicate that measures of syntactic complexity and speed fluency correlate with native speaker ratings; however, the measure of lexical diversity does not correlate with the native speaker ratings. Interestingly, the measure of grammatical accuracy displays an inverse relationship to the native speaker ratings. These results are discussed in light of an accuracy-fluency continuum. This finding demonstrates the importance of careful consideration in determining which measure of oral proficiency is appropriate for a given research context.*

Key words: *oral proficiency, assessment, quantitative measures, native speaker ratings, Second Language Acquisition*

Las investigaciones sobre la adquisición de segundas lenguas (L2) han utilizado varios indicadores cuantitativos y cualitativos para medir la competencia oral. Sin embargo, hay poca investigación empírica que compare dichas medidas. De hecho, las comparaciones con los referentes de hablantes nativos son especialmente raras. Cuatro de los indicadores cuantitativos que se aplican con mayor frecuencia en las investigaciones de L2 incluyen la proporción type-token como medida de la diversidad léxica; el T-unit como medida de la complejidad sintáctica; el error-free t-unit como medida de la precisión gramatical; y la velocidad media de habla como medida de la fluidez. El presente estudio muestra una comparación de cuatro indicadores de competencia oral basadas en el habla de tres hablantes no nativos de inglés durante la entrevista oral del International English Language Testing System (IELTS). Los resultados indican que los indicadores de complejidad y fluidez sintácticas se correlacionan con los valores de hablantes nativos; sin embargo, las medidas de precisión gramatical y diversidad léxica no se correlacionan con los valores de hablantes nativos. De hecho, la medida de precisión gramatical muestra una relación inversa con respecto a los valores de hablantes nativos. Dichos resultados se analizan bajo la perspectiva del continuo precisión-fluidez. Este resultado demuestra la importancia de determinar cuidadosamente qué medida de competencia oral es apropriada para un contexto de investigación.

Palabras claves: *competencia oral, evaluación, medidas cuantitativas, valores de hablantes nativos, adquisición de segunda lengua*

1. Introduction

Researchers dealing with spoken second language (L2) data often seek measures of proficiency to understand the language use and development of L2 speakers in their studies. In particular, many L2 researchers have adopted measures from first language acquisition research that facilitate the quantitative analysis of spoken data, such as the T-unit (Hunt, 1965), the error-free T-unit (Larson-Freeman & Strom, 1977), and the type-token analysis (Johnson, 1944). Such quantitative analyses offer measures external to any particular data set and provide numerical comparisons between speakers (Gaies, 1980). However, while they may be internally reliable, it is questionable whether these specific measurements are generalizable to an overall oral proficiency rating. Extensive research in this area is based on the critical assumption that these measures are generalizable, but it is essential to evaluate the validity of this claim.

The research questions for the present study are as follows. First, what is the strength of the correlation between native-speaker ratings on L2 oral production and lexical diversity, syntactic complexity, and the number of lexical/ morphosyntactic errors in that production? And secondly, how do these measures compare with a qualitative analysis of the interview data?

2. Literature Review

Although quantitative measures of proficiency are commonly accepted in first language acquisition research (Hunt, 1970; Nippold et al., 2005; Scott, 1988), their place in second language acquisition has been a steady source of debate. Criticism over the years has led to the development of various alternatives, including the error-free T-unit (Larson-Freeman & Strom, 1977), sentence analysis (Bardovi-Harlig, 1992) and the AS-unit (Foster et al., 2000). Several researchers have simultaneously compared various quantitative measures of proficiency. For instance, Iwashita (2006) compared various measures of grammatical complexity including length of T-unit, number of clauses per T-unit, number of independent clauses per T-unit and number of dependent clauses per T-unit, and found the length of T-unit and number of clauses per T-unit to predict learner proficiency most reliably. Few studies have compared native speaker ratings to quantitative measures of proficiency. Such studies are necessary 1) in order to understand what aspects of speech native speakers consider in their ratings and 2) whether specific quantitative measures may be used as measures of overall oral proficiency in place of native speaker ratings. It is imperative that researchers carefully consider which type of oral proficiency measure is being chosen and what that measure does in fact reveal about the individual's speech. Thus, this study adds to thoughtful inquiry into oral proficiency assessment. In one notable exception, Iwashita et al. (2008) compared quantitative measures of grammatical accuracy and complexity, vocabulary and fluency to holistic scores by raters, and found that features from each category helped determine overall proficiency, with particular influence from vocabulary and fluency. To probe these findings, the present study compares four of the most commonly used quantitative measures (type-token analysis, T-unit, error-free T-unit, and fluency) to native speaker ratings.

The type-token ratio (TTR) analysis is a measure of lexical diversity determined by the ratio of different words (types) to total words (tokens). For instance, if a piece of discourse included 40 words and all 40 were different, the result would be 40/40, an 'ideal' TTR of 1.00. Many researchers have criticized the proposed baseline and the implications for its use (e.g., Covington, 2010; Hess et. al, 1989; Richards, 1987; Templin, 1957). Since, in fact, 40 different words in one segment of speech may result in awkward phrasing if an individual would need to repeat certain relevant terms more than once. For this reason, Templin (1957) proposed a baseline of 5 tokens of every type. Alternatively, Richards (1987) proposed a Verbal Diversity measure; Hess et. al (1989) provided evidence that TTR measures were not a reliable measure of performance for elementary school children; and Covington & McFall (2010) provided a mathematical argument for the unsatisfactory qualities of the TTR measurement. In essence, the problem is that the TTR is affected by sample size. The longer a text/sample goes on, the more likely one is to encounter a repeated word. Additionally, since it depends on sample size, it renders the measure ineffective when comparing across participants with texts of differing sample sizes, as many researchers have noted (Chotlos, 1944; Hess et. al, 1986; Richards, 1987;

Malvern & Richards, 2002). However, recent studies continue to use the TTR as a measure of lexical diversity and oral proficiency (Genesee et. al., 1995, Johnson, 2008; Nicoladis et. al., 2009). The continued use suggests that some researchers do consider the TTR a useful measure, despite its questionable validity.

Similar to the TTR, the T-unit was first developed as a measure of proficiency and linguistic maturity in first language acquisition (Hunt, 1965). Defined as "a main clause plus all subordinate clauses and non-clausal structures attached to or embedded in it" (Hunt, 1970, p. 189), the T-unit was originally used to measure the syntactic complexity of children's written language. Measures of syntactic complexity for spoken discourse range from semantic units (Sato, 1988; Pica et. al., 1989; Kroll 1977) to intonational units (Crystal & Davy, 1975; Chafe, 1980; Crookes & Rulon, 1985; Ellis et al., 1994; Foster & Skehan, 1996) to syntactic units (Quirk et al., 1985; Kroll, 1977; Hunt, 1970). However, upon comparing various measures of syntactic complexity, Iwashita et. al. (2006, p. 165) found the standard T-unit (Hunt, 1970) the "best way to predict learner proficiency", as Halleck (1995) found.

In tailoring the T-unit to the purposes of L2 research, Larsen-Freeman & Strom (1977) developed the error-free T-unit, since the errors characteristic of L2 acquisition are distinct from errors in first language acquisition. However, one of the most noticeable problems to this approach is determining what constitutes an *'error'*. Various scholars define the error-free T-unit as perfect in all respects (Larsen-Freeman & Strom, 1977), that is, free from morphological and syntactic errors (Scott & Tucker, 1974), or from morphosyntactic and lexical errors (Vann, 1978). While originally viewed as a measure of syntactic complexity, Polio (1997) explained that the error-free T-unit is more appropriately a measure of accuracy. Several studies have found the error-free T-unit a useful proficiency measure. For instance, Vann (1978) found that although the mean length of T-unit did not correlate with the Test of English as a Foreign Language (TOEFL) scores, the scores did correlate significantly with mean length of error-free T-unit and ratio of error-free T-units to total T-units.

A growing body of research explores quantitative measures of learner fluency, including such measurements as speech rate, number of hesitations, number and length of pauses, number and length of runs, and number of false starts. So far, the results are varied and conflicting. Some researchers argue that quantitative measure of fluency are reliable (Baker-Smemoe et al., 2014; Beigi, 2009; Cucchiarini et. al., 2000), while others find that L2 fluency and L2 accuracy are not highly correlated (Brand & Gotz, 2011), and that it was dependent on context and proficiency level (Garcia-Amaya, 2009). Tavakoli & Skehan (2005) propose three aspects of fluency: speed fluency, breakdown fluency, and repair fluency. The present study analyzes speed fluency, also referred to as speech rate, through the calculation of syllables per second (Hilton, 2009).

The adequacy and efficacy of measures of proficiency remain inconclusive, perhaps in part due to a lack of research comparing different types of measures. The current study addresses this gap in the literature by comparing four quantitative oral proficiency measures (TTR, T-unit, error-free T-unit, speech rate) to native speaker ratings of overall proficiency.

3. Methodology

The data was taken from the oral interview portion of IELTS assessment of three non-native speakers from different countries, which had been posted on the video-sharing website youtube.com. According to the Institutional Review Board, youtube videos are exempt from human subjects review protocol, since it is an open and public forum where all individuals are able to access the videos. However, the present study uses pseudonyms and does not disclose the video names or urls of the videos in order to protect the individuals in the videos, especially since practice for an oral proficiency test in a second language is a vulnerable experience. The first non-native speaking participant in the IELTS interview is Tiffany from Mexico. The second non-native speaking participant is Edgar from China. Although these two participants mentioned their home countries during the interview, the third participant, David, did not identify a nationality. The participants also did not state their age, although they all appear to be within the range of 15 to 25 years old. Therefore, speaker's L1 and speaker's age are both uncontrolled variables in the present study. These videos were chosen given that they were authentic samples of IELTS interviews rather than contrived samples presented for the sole purpose of the study. At the time of the study, very few authentic videos of IELTS interviews were available on youtube.

The IELTS interview is typically segmented into three parts. In Part 1, test takers answer questions about themselves and their family. In Part 2, test takers are given a prompt and are asked to speak about that topic for a few minutes. In Part 3, test takers have a longer discussion on that same topic with the interviewer. Due to availability constraints, different portions of the interview were available for each of the three non-native speakers. Only Part 3 was available for David. Parts 1 and 2 were available for Edgar, and Parts 2 and 3 were available for Tiffany. Although this may change the results of individual speaker evaluations, the main purpose of this study is not to accurately rate each non-native speaker, but to analyze to what degree the various measurements of proficiency correlate with one another. The difference in the parts of the oral interview available for each participant is another important limitation to take into consideration in the analysis of the data. However, in this case it was necessary due to the limited amount of authentic IELTS interviews available on youtube.com. Additionally, the case study of few participants allows for a close study of these types of differences, unlike larger datasets, although it is hoped that in the future more studies with larger datasets and controlled variables will follow.

The three native English-speaking raters who analyzed these videos were graduate students in an L2 Assessment course at the time. They had studied oral proficiency ratings, in particular, the American Council on the Teaching of Foreign Languages (ACTFL) oral proficiency scale. The raters used the ACTFL oral proficiency scale, since it is frequently used in a variety of contexts, and it is rather more intuitive and simplistic in nature. The main goal was to get a native speaker's general overall rating, without being distracted by technicalities, rather than to evaluate the effectiveness of the test itself. This rating scale was then converted to a 10-point scale, ranging from Novice Low (1) to Superior (10). The ratings were averaged, and the variances for each speaker were computed. The criterion for acceptable inter-rater reliability was set at a variance less than or equal to one. After choosing a score from 1-10, native English speaking raters were then prompted to comment on the reason that they chose that score for the individual IELTS interview video.

Each interview was transcribed in the CHAT program (MacWhinney, 2000). Then, the CLAN software was used to run an analysis of TTR for each non-native speaker interview (MacWhinney, 2000).To determine the mean length of T-unit, the researcher coded each main clause along with its subordinate clauses. The spoken nature of the data led the researcher to make several adjustments in the strict definition of the T-unit. Foster (2000) notes that "and" is frequently used in spoken discourse for different purposes. For this reason, all instances of "and" were excluded from the analysis. Additionally, all instances of "um" and "uh" were omitted in both the TTR and T-unit analysis. These words were tagged with '&' in order to exclude them from the quantitative measurements. The length of each T-unit was calculated manually, and then averaged to determine the mean T-unit length. For example, the segment of transcript below shows how these words were omitted as well as how this factored into the coding of main clauses and subordinate clauses for the purpose of TTR analysis.

*(1) *EDG: Because she &uh was really susessful at work [C] &and &uh she treat everyone equally [C] and &uh people thinks people think that she was a really nice person [C].*

The above segment of transcript shows how the "&" code was entered before each instance of "uh", "um" and "and" in order to omit these from the analysis of TTR and clause determination. The [C] code was entered for main clauses. This segment of speech also demonstrates how in spoken dialogue, "and" is used very frequently. Counting each instance as a subordinate clause instead of a main clause would yield an inaccurate representation of the speech production. This alteration is necessary for the TTR, since it was originally developed for written speech only.

On the transcript, the researcher coded and counted errors using the code "&e". The researcher's coding system was checked by a research assistant. Then, a command in the CLAN software was used to count the grammatical errors. Only syntactic, morphological and lexical errors were counted in the error-free T-units, as the audio quality of the videos are not sufficiently high for accurate phonological analysis. The number of clauses containing errors were subtracted from the total T-units. The ratio of error-free T-units over total T-units was then used for the measure of grammatical accuracy. Below is an example of the same excerpt, demonstrating the coding of errors with the code "&e".

*(2) *EDG:Because she was really susessful [sic] at work [C] and uh she treat &e everyone equally [C] and uh people thinks &e people think that she was a really nice person [C].*

In the above example, we see two syntactic errors, marked using "&e" in two different T-units. Since there are three T-units in this section of the interview, this section would have one error-free T-unit, or a ratio of 1:3 or .333.

Speed fluency was calculated through the measure of syllables per second, following Hilton (2009). Speed fluency was chosen instead of repair and breakdown fluency, since the latter two converge with grammatical accuracy. Instead, the goal was to select a quantitative measure based solely on fluency. The number of syllables were calculated for each turn, and divided by the amount of seconds per turn to calculate syllables per second.

4. Results

The present study analyzes which quantitative measures of L2 oral proficiency are effective predictors of native speaker ratings. Specifically, this analysis compares five measures of oral proficiency: a measure of overall oral proficiency (average native speaker rating), a measure of lexical diversity (type-token ratio), a measure of syntactic complexity (T-unit analysis), a measure of grammatical accuracy (error-free T-unit), and one measure of fluency (sillables per second). (See Table 1.)

Proficiency Category	*Proficiency Measure*
Overall Oral Proficiency	Average Native Speaker Rating
Lexical Diversity	Type-Token Ratio
Syntactic Complexity	T-unit
Grammatical Accuracy	Error-Free T-unit
Speed Fluency	Syllables per Second

Table 1: Proficiency Measures Used in the Present Study

Speaker	*Overall Oral Proficiency*	*Lexical Diversity*	*Syntactic Complexity*	*Grammatical Accuracy*	*Speed Fluency*
Edgar	7.6	.488	.331	.619	2.214
David	9	.386	.512	.594	3.267
Tiffany	5	.362	.299	.906	2.07

Table 2: Comparison of Oral Proficiency Measures by Speaker

Table 2 presents a comparison of the oral proficiency measures by speaker derived from the previously described analyses of the IELTS interviews by Edgar, David and Tiffany. For the measure of overall oral proficiency, the average native speaker ratings, David has the highest rating at 9, followed by Edgar at 7.6, and Tiffany at 5. However, Edgar has the most lexical diversity (.488) followed by David (.386) and then Tiffany (.362). In terms of syntactic complexity, David outperforms Edgar with average length of T-unit at .512 compared to the .331 for Edgar and .299 for Tiffany. In terms of speed fluency, David has 3.267 syllables per second, Edgar has 2.214 syllables per second, and Tiffany has 2.070 syllables per second. Yet, in grammatical accuracy, Tiffany scores above the others with a percentage of 90.6% error-free T-units compared to 61.9% for Edgar and just 59.4% for David.

A Pearson's *r* correlation between native speaker ratings and the measure of syntactic complexity (length of T-unit) found a correlation that is approaching significance ($r = .85$). Additionally, a Pearson's *r* correlation between native speaker ratings and speed fluency (syllables per second) found a correlation that is approaching significance (*r* = .83). The measure of lexical diversity does not clearly correspond to the other measures. Also, interestingly, there is a notable inverse relationship between measures of grammatical accuracy and speed fluency. A Pearson's r correlation found this correlation to approach significance as well (*r* = -.07). This may be due to the fact that to speak with great speed fluency, the speaker gives up precision in grammar.

4. Discussion

The findings suggest that native speaker raters may value fluency and syntactic complexity more highly than grammatical accuracy in oral proficiency exams. However, as this is a small scale case study with the intent to describe and analyze closely few participants, these statistics are meant for descriptive purposes only. Thus, the findings are not generalizable but merely intend to describe patterns for future analysis of larger datasets.

With regard to the correlation between syntactic complexity and native speaker ratings, speakers with greater lengths of T-units may be perceived as more proficient due to the overall structure of their conversations. If a speaker's turn is grammatically accurate or filled with diverse word choices but framed in short clauses, it may not sound as natural as longer, more complex sentence structures. For instance, compare the following clauses by Tiffany and David.

*(3) *DAV:Well, to be honest, uh personally I'm a huge sportsfan, [C] so for me there's not too many sports on tv [C] But of course for some people maybe, there might be uh a little less sports or too many sports, [C] but for me, to be honest, there's not too many sports [C].*

*(4) *TIF:Okay, my favorite place to eat is Las Antorchas [C] Uh it is located um downtown of the city [C] and this is on a street on what's the name, past xxx on Madero [C] And they serve Mexican food [C]. Um, it's delicious [C].*

The difference between the average length of David's T-units (marked by [C]) as well as variation in structure and style compared to

Tiffany's average T-unit length is clear. For instance, Tiffany's statement, "Um, it's delicious" is much shorter than many of David's T-units. There is, however, no correlation between lexical diversity (type-token Ratio) and overall language proficiency (native speaker ratings). The considerable previous discussion on Type-Token Ratio (TTR) in the literature review section notes that many scholars have questioned its reliability, especially as a measure of oral proficiency. The purpose of including the TTR in this analysis was, in fact, to verify whether the TTR correlated with other measures of oral proficiency. The findings suggest that in this case it may not be a reliable measure of overall oral proficiency. It may be of interest depending on the individual's purpose in using the measure. In particular, note that the subjects discussed in the IELTS interviews are very informal: favorite restaurant, someone you admire, and sports. In these informal conversational interviews, it may not be necessary to use a great variety of lexical terms. In fact, when individuals talk informally about such subjects, they may choose to use a select quantity of specific words. As long as the word choice is appropriate for the context and the discussion, lexical diversity may be judged as rather tangential and irrelevant. This may be compared with an academic article or formal essay wherein repetition may be viewed as less scholarly and academic. In a formal written situation, the TTR may be a more relevant measure of proficiency.

Also, as noted above, there is an inverse relationship between the percentage of error-free T-units and native speaker ratings. This may be evidence of an inverse relationship between accuracy and fluency. As accuracy decreases, the non-native speakers may increase the fluidity of their speech or their speech rate. For instance, David was consistently rated highest in overall oral proficiency, and native speaker raters all commented on the fluency of his speech in a comment section following the numerical rating. However, he rated lowest on error-free t-units, or grammatical accuracy. This reveals that even though the graders were using the ACTFL scale which would have taken into account accuracy, the native speaker raters also accounted for fluency. In fact, it suggests that they counted fluency as more important to overall oral proficiency than accuracy, perhaps even subconciously.

5. Conclusion

In the present study, speech samples from three non-native speakers were analyzed according to commonly used quantitative measures of proficiency, specifically the type-token ratio which derives a measure of lexical diversity, the t-unit as a measure of syntactic complexity, error-free t-unit as a measure of grammatical accuracy, and syllables per second as a measure of speed fluency. These quantitative measures were then compared to native speaker ratings. It must be noted, however, that this is a case study of three individual IELTS interviews. As such, the results are not broadly generalizable but merely tentative findings which are useful for the close analysis of individual interviews and also serve as a basis of further exploration into this topic.

The measures of syntactic complexity and fluency correlated with the native speaker ratings. There was no correlation between the measure of lexical diversity and native speaker ratings. Notably, there was an inverse relationship between the measure of grammatical accuracy and native speaker ratings. This may suggest that quantitative measures of speed fluency and syntactic complexity may accurately substitute for native speaker ratings. This is of considerable interest, given the time-consuming nature of native speaker ratings of oral proficiency.

Yet, it is slightly troubling to note that measures of grammatical accuracy were inversely related to the measures of speed fluency, syntactic complexity and native speaker ratings. So while quantitative measure of speed fluency and syntactic complexity may approximate native speaker ratings, it raises the important consideration of the role of grammatical accuracy in assessing oral proficiency. Additionally, this study may question the validity of native speaker ratings as the absolute best oral proficiency score. In fact, it may instead suggest that native speaker raters have difficulty attending to both grammatical accuracy and fluency while listening to an oral proficiency exam and judging both aspects equally. Instead, the most precise overall oral proficiency score is most likely a combination of factors including syntactic complexity, speed fluency, and grammatical accuracy. It is also quite possible that the individual speaker's "accent" or phonological characteristics of the oral interview impacted native speaker ratings; however, in this case the quality of the videos eliminated accurate phonological analysis as a possibility. It is also possible that native speaker raters may be influenced by such factors as the speaker's appearance, stereotypes about patterns and ways of speaking in addition to topics of conversation. Therefore, while time consuming, the best measure of oral proficiency may indeed prove to be a combination or triangulation of all measures in order to provide the language learner with an accurate assessment of different aspects of their speech.

Furthermore, the results question the ability of specific measures of lexical diversity, syntactic complexity and grammatical accuracy to generalize to an overall measure of oral proficiency, since grammatical accuracy does not align with native speaker ratings. Instead a combination of these measures may be more advisable, such as that proposed by Iwashita et. al. (2008). In a field where research is highly dependent on proficiency level and the effects of proficiency, this cannot be disregarded. Valuable research is being done in the field of second language acquisition, but it may be evaluated incorrectly if an unreliable measure of proficiency is misused or misappropriated. Recent research including the present study, however, has not upheld the reliability of using a singular quantitative measure of proficiency as an overall measure of oral proficiency, especially the types-token ratio and error-free T-unit. For this reason, further research on quantitative measures of proficiency among non-native speakers and comparisons with native speaker ratings are of the utmost importance in order to ensure continued high quality research in the field.

This research also holds important implications for the instruction of English as a second/foreign language. As students are preparing for oral proficiency exams, their instructors must prepare them not only in the areas of grammar and lexicon but also speed fluency and syntactic complexity which are highly influential to native English speaking raters.

References

Baker-Smemoe, W., Dewey, D., Brown, J., & Martinsen, R. (2014). Does measuring L2 utterance fluency equal measuring overall L2 proficiency? Evidence from five languages. *Foreign Language Annals, 47*(4): 707-728. https://doi.org/10.1111/flan.12110

Bardovi-Harlig, K. (1992). A second look at t-unit analysis: Reconsidering the sentence. *TESOL Quarterly, 26*(2): 390-395. https://doi.org/10.2307/3587016

Beigi, H. (2009). *Computer rating of oral test responses using verbosity*. Report No. RTI 20091211-01. Yorktown Heights, NY: Recognition Technologies.

Brand, C. & Gotz, S. (2011). Fluency versus accuracy in advanced spoken learner language: A multi-method approach. *International Journal of Corpus Linguistics, 16*: 255-275. https://doi.org/10.1075/ijcl.16.2.05bra

Chafe, W. (1980). The deployment of consciousness in the production of narrative. In W. Chafe (ed.): *The Pear Stories: Cognitive, Cultural and Linguistic Aspects of Narrative Production*. Norwood, NJ: Ablex, pp. 9-50.

Chotlos, J. (1944). Studies in language behavior: IV. A statistical and comparative analysis of individual written samples. *Psychological Monographs 56*: 75-111. https://doi.org/10.1037/h0093511

Covington, M. & McFall, J. (2010). Cutting the Gordian knot: The Moving-Average Type-Token Ratio. *Journal of Quantitative Linguistics, 17*(2). 94-100. https://doi.org/10.1080/09296171003643098

Crookes, G. & Rulon, K. (1985). Planning and interlanguage variation. *Studies in second language acquisition*, *11*: 367-83. https://doi.org/10.1017/S0272263100008391

Crystal, D. & Davy, D. (1975). *Advanced Conversational English*. London: Longman.

Cucchiarini, C., Strik, H., & Boves, L. (2000). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *Journal of the Acoustical Society of America, 107*(2): 989-999. https://doi.org/10.1121/1.428279

Foster, P. & Skehan, P. (1996). The influence of planning on performance in task based learning. *Studies in second language acquisition, 18*: 299-324. https://doi.org/10.1017/S0272263100015047

Foster, P., Tonkyn, A., & Wigglesworth, G. (2006). Measuring spoken language: A unit for all reasons. *Applied Linguistics, 21*(3): 354-375. https://doi.org/10.1093/applin/21.3.354

Gaies, S. (1980). T-unit analysis is second language research: Applications, problems and limitations. *TESOL Quarterly, 14*: 53-60. https://doi.org/10.2307/3586808

Garcia-Amaya, L. (2009). New findings on fluency measures across three different learning contexts. In J. Collentine (Ed.), *Selected Proceedings of the 11th Hispanic Linguistics Symposium* (pp. 68-80). Somerville, MA: Cascadilla Proceedings Project.

Halleck, G. (1995). Assessing oral proficiency: A comparison of holistic and objective measures. *The Modern Language Journal, 79*(2): 223-234. https://doi.org/10.1111/j.1540-4781.1995.tb05434.x

Hess, C., Haug, H., & Landry, R. (1989). The reliability of type-token ratios for the oral language of school age children*. Journal of Speech and Hearing Research, 32*: 536-540. https://doi.org/10.1044/jshr.3203.536

Hess, C., Sefton, K., & Landry, R. (1986). Sample size and type-token ratios for oral language of preschool children. *Journal of Speech and Hearing Research 29*: 129-134. https://doi.org/10.1044/jshr.2901.129

Hilton, H. (2009). Annotation and analyses of temporal aspects of spoken fluency. *CALICO Journal*, 26, 644–651.

Hunt, K. (1965). Grammatical structures written at three grade levels. Research Report. No. 3. Urbana, IL; National Council of Teachers of English.

Hunt, K. (1970). *Syntactic maturity in school children and adults*. Monograph and the Society of Research into Child Development. https://doi.org/10.2307/1165818

Iwashita, N. (2006). Syntactic complexity measures and their relation to oral proficiency in Japanese as a foreign language. *Language Assessment Quarterly, 3*(2): 151-169. https://doi.org/10.1207/s15434311laq0302_4

Iwashita, N., Brown, A., McNamara, T. & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics, 29*(1): 24-49. https://doi.org/10.1093/applin/amm017

Johnson, W. (1944). Studies in language behavior: I. A program of research. *Psychological Monographs, 56:* 1-15. https://doi.org/10.1037/h0093508

Kroll, B. (1977). Combining ideas in written and spoken English: a look at subordination and coordination. In E. Ochs & T. Bennett (eds.). *Discourse Across Time and Space*. Southern California Occasional Papers, 5.

Larsen-Freeman, D. & Strom, V. (1977). The construction of a second language acquisition index in development. *Language Learning, 27*: 123-34. https://doi.org/10.1111/j.1467-1770.1977.tb00296.x

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk.* 3rd edition. Mahwah, NJ: Lawrence Elbraun Associates.

Malvern, D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing, 19*(1): 85-104. https://doi.org/10.1191/0265532202lt221oa

Nippold, M.A., Hesketh, L.J., Duthie, J.K., & Mansfield, T.C. (2005). Conversational versus expository discourse. A study of syntactic development in children, adolescents, and adults. *Journal of Speech Language, and Hearing Research, 48*. 1048-1064. https://doi.org/10.1044/1092-4388(2005/073)

Pica, T., Halliday, L., Lewis, N. & Morgenthaler, L. (1989). Comprehensible outputs as an outcome of linguistic demands on the learner. *Studies in Second Language Acquisition, 11*(1): 63-90. https://doi.org/10.1017/S027226310000783X

Polio, C. (1997). Measures of linguistic accuracy in second language writing research. *Language learning, 47*: 101-143. https://doi.org/10.1111/0023-8333.31997003

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. Harlow: Longman.

Richards, B. (1987). Type/token ratios: What do they really tell us? *Journal of Child Language, 14*: 201-109. https://doi.org/10.1017/S0305000900012885

Sato, C. (1988). Origins of complex syntax. *Studies in Second Language Acquisition, 10*(3): 371-95. https://doi.org/10.1017/S027226310000749X

Scott, C.M. (1988). Spoken and written syntax. In M.A. Nippold (Ed.), *Later language development: Ages nine through nineteen* (pp. 49-95). Austin, TX: Pro-Ed.

Scott, M. & Tucker, G. (1974). Error analysis and English language strategies of Arab students. *Language Learning, 24*: 69-97. https://doi.

org/10.1111/j.1467-1770.1974.tb00236.x

Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239–276). Amsterdam: John Benjamins. https://doi.org/10.1075/lllt.11.15tav

Templin, M. (1957). *Certain language skills in children: their development and interrelationships*. Westport, CT: Greenwood.

Vann, R. (1978). A study of the oral and written English of adult Arabic speakers. Ph.D. Dissertation, Indiana University.

First version received: May 2017

Final version accepted: September 2017