# Exploring Syntactic and Lexical Complexity in Narrative Tasks

Roz Hirch

## I. Introduction

(1) Darwin thought that species gradually became more complex.
(2) Darwin's gradual rise to mounting complexity . . .
(Haliday and Martin, 1993, as quoted in Norris & Ortega 2009)

This paper is interested primarily in describing complexity, developing methods of analyzing it for diagnostic purposes, and in finding ways that it is (or is not) elicited in various types of tasks. Numerous attempts have been made to explain what complexity is. The sentence and clause at the top are an example; they have the same meaning but different complexity, showing what Systemic Functional Linguistics calls "a change from a dynamic style to a synoptic style of expression," (Norris & Ortega 2009). Such a change involves the gradual transformation of text, condensing it by changing sentences into clauses (as above), clauses into phrases, and phrases into words. Although this description of complexity could be problematic—perhaps, it is a conflation of content density and syntactic complexity—still, it is productive to begin with this definition and test how effective methods of measuring complexity through Systemic Functional Linguistics are.

There are numerous rating scales that purport to measure complexity in some way, the most well-known being those of the Common European Framework of Reference (CEFR) and the American Council on the Teaching of Foreign Languages (ACTFL). These offer qualitative analyses based largely on experiential research (Pallotti 2009); yet these are problematic, particularly for instructional purposes, for two specific reasons. First, they are not theory-based, and therefore offer little salient information as regards

empirically-based methods for understanding elements such as development and complexity (Norris & Ortega 2009). By their very nature, then, they are not effective for diagnostic purposes of improving writer quality. Similarly, they are not fine-grained enough to offer information that will be effective for improvement; in order to be effective for diagnostic purposes, feedback must be sufficiently fine-grained to direct a writer on specific ways to improve (Jang 2009). Such a rating scale must, therefore, be able to measure development and complexity in a way that gives writers substantial information in order to direct progress to more advanced levels of complexity (Norris & Ortega 2009). As a part of this, such a scale must not only be able to describe complexity in general, but must be able to describe complexity that is specific to various levels of proficiency / development (Norris & Ortega 2009, Jang 2009).

On the surface, this research seems mostly about testing and improving complexity for test-takers; yet application as a diagnostic tool suggests a greater purpose for second language acquisition (SLA) generally. Writing ability is a general requirement for a wide variety of academic and professional areas; as such, research into writing complexity must be applicable across a broad range of disciplines, and must be readily adaptable (McNamara *et al* 2010). Even beyond describing how to improve writing quality, such research is important for defining and describing developmental levels (Pallotti 2009). Then again, the prevalence of tasks within SLA settings necessitates a profound understanding of how tasks can be used to increase complexity in learner output (Pallotti 2009). This understanding will allow teachers to "manipulate learner performance in predictable ways, thereby promoting learning opportunities and developing proficiency" (Tavakoli & Foster 2008). Specific examples include developing tasks that require incorporation of forms that use "think," "expect," and "know," which lead to more complex syntactic structures in writing (Robinson & Gilabert 2007). In this way, the properly understood effects of tasks give the fine-grained direction to instruction that is both required for improvement, and can be specific to the student (Jang 2009). Purposefully applied, tasks give direction to instruction, focusing student attention on those areas of writing (and also reading, speaking, and listening) that they need to work on, and to strategies they can employ in order to achieve

effective goals (McNamara *et al* 2010, Kormos 2011, Jang 2009).

As stated in the beginning, this paper examines the idea of complexity, first by looking at its relationship with two other, closely related factors—accuracy and fluency—and the complex interactions of the three. It then focuses on various measures of complexity, discussing both syntactic and lexical complexity and how these two may be combined to describe performance on a task, as well as the idea of using native English speaker output as a baseline. Finally, it offers a small, exploratory study that compares the output of three groups—one group of native English speakers (NES), and two groups of non-native English speakers (NNS) divided into advanced and intermediate—on two different types of short writing tasks, in order to see if some of the measures of complexity described in this paper are effective for predicting and describing the differences in performance among the three groups.

## II. Complexity, Accuracy and Fluency

The first step in understanding the interaction of complexity, accuracy and fluency (commonly referred to collectively as CAF, as it will be throughout this paper) is to define what is meant by them. CAF is viewed as the "dimensions for describing language performance, most frequently used as dependent variables to assess variation with respect to independent variables such as acquisitional level or task features" (Pallotti 2009). While some theorists argue against placing too much emphasis on delineating developmental levels (see, for example, discussions in Pallotti 2009 & Skehan 2009), still it is productive for observing and describing how language may develop in specific learners, as well as in general. Ultimately, it may help us to document "what parts of the interlanguage system change as acquisition unfolds, in what ways anticipated change proceeds, and perhaps why sometimes not much change seems to take place" (Norris & Ortega 2009). For these reasons, while accuracy and fluency are not the primary focus of this paper, they are taken into account when looking at measures of complexity.

Of the three dimensions of CAF, accuracy is the easiest to define and measure objectively. Accuracy represents the degree to which some output—

whether in the form of written, as will be explored in this paper, or spoken—conforms to a specific standard of measure (Pallotti 2009). However, one thing to bear in mind, as this paper questions for other aspects of CAF, is: what exactly is the standard of measure by which accuracy is measured? This speaks to the question of prescriptive and descriptive grammar, and which is the more appropriate for use in analysis. Though not a focus of this paper, it is worth considering whether using a NES standard—in other words, utilizing descriptive grammar as a standard—would not be more appropriate than prescriptive. This concerns the interface between accuracy and fluency, and shows that even something that seems as objective as accuracy may not, in fact, be so obvious.

Fluency is more frequently referenced in speaking assessment, rather than in writing assessment (Skehan 2009; Yu 2010; Robinson 2001; Robinson & Gilabert 2007). In speaking situations, fluency is generally not so much defined as it is described with reference to aspects such as rates of speech (speed), number and timing of pauses (breakdown fluency), repairs, and false starts (repair fluency) (Pallotti 2009; Skehan 2009; Norris & Ortega 2009). Measures of fluency are almost always done through comparison to a NES standard, not so much to encourage native-like production but to obviate "researchers' bias toward seeing learners as defective language users, who always need to 'do more'" (Pallotti 2009). Fluency in writing is also compared to NES standards, though almost exclusively in terms of total length of output (Norris & Ortega 2009). Other, somewhat controversial, measures include average lengths of T-units and clauses (Wolfe-Quintero et al 1998), which is not the purpose for which these measures were intended (Norris & Ortega 2009). Furthermore, none of these measures looks at sociopragmatic aspects of fluency, which are more difficult to quantify (Ong & Zhang 2010). However, these may be appropriate to examine, even with reference to grammar, as is this paper's focus; while a task may not produce very complex structures on the part of the NNS, a comparison with NES output could show the result is because of fluency and accuracy being favored over complexity.

Complexity is, perhaps, the most difficult of the three to measure. Most theorists begin by differentiating between two aspects of complexity: that of

the task that is given to the NNS, and the performance or output of that task (Pallotti 2009; Skehan 2009; Robinson 2001). This paper follows the differentiation made in Skehan (2009), in which complexity refers to subject output—whether NES or NNS—while difficulty refers to the task itself, and how it differentiates among learners. In defining complexity with reference to subject output, it can be described as the combination of multiple elements within a specific language production (Pallotti 2009). Put another way, complexity is the presence of several elements within one unit of measure. To achieve syntactic complexity, then, one can assume that the writer must provide a variety of grammatical forms, in combination, within one grammar unit. However, there is a caveat; "'more complex' does not necessarily mean 'better'" (Pallotti 2009). This is especially the case when one considers readability; the very elements that make a text readable are not always those which make it complex, and vice versa. In fact, numerous studies have found that the very elements that researchers assume to comprise complexity are frequently absent in measures of readability (McNamara et al 2010). For this reason, using NESs as a baseline may be more productive for measuring complexity in how it interacts with fluency; it seems unreasonable to assume that NNSs should produce texts that are more grammatically complex than those NESs produce. The most common measures of syntactic complexity are words per T-unit, words per clause, and clauses per T-unit (Wolfe-Quintero et al 1998; Hunt 1965; Kormos 2011; Norris & Ortega 2009). These are discussed further, below.

Another, important form of complexity explored in this paper is lexical complexity (Norris & Ortega 2009). Lexical complexity interacts with grammar to create both complexity and fluency, a relationship that results in what researchers call "sophisticated language" (McNamara et al 2010; Skehan 2009). Lexical complexity is frequently measured in two different ways: through diversity and frequency. Lexical diversity is the more frequently measured as it is considered "an important quality indicator of test performance" (Yu 2010). It is broadly defined as the number of different words used in a text, and is calculated through a type-token ratio (TTR), which may be corrected through calculating the D formula (Yu 2010; Skehan 2009). Lexical frequencies are measures of the types of words used in a language production, compared with

frequency lists from NES corpora analysis, to determine from what level the majority of words contained within a spoken or written sample come (Skehan 2009). However, this method of analysis is not specific to a task, and therefore ignores the language use requirements of that task; the result may be language use by NNSs that is inappropriate to the task, and therefore lacking fluency.

While CAF is widely studied for analyzing language output, the precise interaction of the three elements is unclear. The next section of this paper explores the two main theories regarding CAF and how the three factors interact: Skehan's Trade-off Hypothesis, and Robinson's Cognition Hypothesis.

## III. Interactions of CAF: Trade-off Hypothesis and Cognition Hypothesis

One aspect of task design that is highlighted by both Robinson (2001) and Skehan (2009), and that is necessary to understanding their hypotheses, is the allocation of writer resources, and whether a task has elements that are resource-dispersing or resource-directing (Robinson 2001). Resource-dispersing elements are non-linguistic factors that tend to affect performance, including things like planning time and prior knowledge (Ong & Zhang 2010). Creating task difficulty through resource-dispersal in undesirable for developing complexity; instead, resource-directing elements should create increased difficulty in order to accurately observe language performance. Resource-directing elements are linguistic ones, and generally make demands on the writer's linguistic capacity in areas such as syntax and lexis (Robinson 2001). An important element of discerning whether the task is performing as it should, as far as both resource-directing and -dispersing elements are concerned, is the performance of NES writers on the task, which "gives an insight into which characteristics of the writing tasks are inherent in the task itself" (Kormos 2011).

Skehan's Trade-off Hypothesis is based on psycholinguistic research into limited attentional capacity (Skehan 2009). The fundamental element of this theory is that human capacity for attention is limited; therefore, when the re-

quirements of attention exceed the capacity available, performance will be affected. So, if a person is required to do two tasks at once, the likely outcome is either that one will be done "better" than the other, or that both will be done slowly or poorly (Foster & Tavakoli 2009). Skehan applied this model to language performance on a task, in the area of CAF; if the difficulty of the task exceeds a NNS's attentional capacity, then the NNS will tend to privilege one or another aspect. This is the trade-off—the NNS chooses one or two elements to benefit, such as sacrificing accuracy to complexity, or complexity to fluency (Ong & Zhang 2010; Foster & Tavakoli 2009). As mentioned above, task difficulty triggers the trade-off; less difficult tasks, whether because of simpler resource-directing or resource-dispersing elements, allow CAF to be accessed in balance, while more difficult tasks require disadvantaging one or two elements. Task difficulty in Tradeoff Hypothesis can be determined by resource-dispersing elements such as familiarity with the topic or format, clarity of directions or information, as well as resource-directing elements (Foster & Tavakoli 2009).

Robinson's Cognition Hypothesis also sees the three elements of CAF interacting, but there is no trade-off. Instead, influenced by Givon (1995, quoted in Robinson & Gilabert 2007)'s assertion that structural and functional complexity proceed simultaneously and perform equally on a difficult text, Cognition Hypothesis predicts that CAF will improve in tandem as task difficulty increases (Robinson & Gilabert 2007). This is because a difficult task demands more complex syntax and lexis from the subject, so that the three elements do not perform independently, but concurrently (Tavakoli & Foster 2011). However, such an effect is only achieved if the resource-directing dimensions, which include the immediacy of the content, reasoning demands, and task structure, increase in difficulty (Ong & Zhang 2010). According to Cognition Hypothesis, then, increasing the difficulty of tasks along resource-directing dimensions will "push learners to greater accuracy and complexity of L2 production in order to meet the greater functional and conceptual communicative demands they place on the learner" (Robinson & Gilabert 2007).

There are a number of similarities between the Trade-off and Cognition Hypotheses. Both agree that there is a measurable effect of task difficulty

on CAF (Pallotti 2009). They are also similar with regard to the idea that resource-dispersing factors tend to have a detrimental effect on CAF (Ong & Zhang 2010). Where the two primarily differ is in the effect that increasing task difficulty through resource-directing elements will have on language production complexity. Trade-off Hypothesis predicts that subjects will either be forced to draw on one or more elements of CAF, to the detriment of the others, or that the performance in general will be less effective. Cognition Hypothesis, by contrast, predicts that CAF will work together and that the output will actually display increased proficiency in all three areas (Skehan 2009; Ong & Zhang 2010). Ultimately, the difference between the two is in the voluntary control of resources, as Trade-off "assumes that attention is subject to voluntary regulation" (Kormos 2011).

Narrative tasks, especially through pictures, are effective for studying complexity for a number of reasons. To begin with, narrative tasks require different language in that writers must "respond" to the pictures, which frequently elicit using less frequent lexis and structures, especially forms such as "I think" or "It seems" (Skehan 2009). At the same time, it can be difficult to avoid certain forms and lexis, as the tasks often require specificity; as an example, one task used in this study requires knowing the term "picnic basket," for which there is no synonym (Skehan 2009). Learners are therefore pressed to use specific lexis and grammar. For these reasons, story-telling tasks using picture sequences were used for this paper.

There are several ways in which tasks should predictably affect both syntactic and lexical complexity in resource-directing factors (Skehan 2009). In general, of course, we can predict that certain tasks will require more or less of writers: it is reasonable to expect, for example, that a task about astronauts will elicit more complex vocabulary than one about children at school (Pallotti 2009). More specifically, the two aspects of task difficulty explored in this paper are called storyline complexity and inherent structure by Foster & Tavakoli (2009); for the purposes of this paper, as well as to minimize confusion among different uses of the term "complexity," they will hereafter be referred to as "storyline multiplicity" and "structure fluidity." Storyline multiplicity refers to the number of activities that occur within a story, or put more simply,

whether the prompt presents one or more storylines simultaneously (Foster & Tavakoli 2009). The assumption is that dual storylines will require subordination, specifically with terms such as "while," "at the same time," "after," and in that way make greater demands on those resource-directing factors (Tavakoli & Foster 2008). Single storylines, however, will not require such subordination, and will more likely produce greater accuracy and fluency, but lower complexity (Foster & Tavakoli 2009). The other aspect of task difficulty being investigated here, structure fluidity, refers to the structure of the story itself. A task that presents a clear storyline is expected to be relatively simple, as it requires less cognitive effort to create a story; a task with vague relationships among the elements of the story, on the other hand, will require more cognitive effort (Skehan 2009). The difference between these two factors, then, especially regarding picture tasks, can be summarized as the load placed on resource-directing factors as pertains to relating elements within each picture of a sequence (storyline multiplicity) and between each picture of a sequence (inherent structure). Tasks can then be described as +/– multiplicity and +/– fluidity.

## IV. Measures of Complexity

Many aspects of complexity in writer output have been proposed for both writing and speaking, including syntactic, lexical, interactive, and propositional, among others (Pallotti 2009). This paper focuses on syntactic and lexical complexity, and how task difficulty specifically affects these areas of complexity.

T-unit analysis was introduced in Hunt (1965) as a method for measuring complexity in native-speaker written production. Since then, it has also been frequently applied to analyses of complexity in spoken production, and also for measuring fluency in both written and spoken production, though these applications can be problematic (Norris & Ortega 2009). A T-unit is defined as an independent clause and its subordinate clauses, and was devised as a unit of measure in order to obviate the problems of false readings of complexity caused by focusing on sentence-length units, as well as to aid in measuring

situations where the sentence structure is unclear (Hunt 1965). Complexity measures based on T-unit include clauses per T-unit, which shows the number of total clauses per independent clause, and is the most frequently applied measure (Wolfe-Quintero *et al*). Other often described measures using T-unit analysis that are incorporated into this paper's analysis include words per T-unit and words per clause, which are intended to show the relationship between the independent and dependent clauses in terms of length (Norris & Ortega 2009).

Another important area to investigate and measure resource-directing elements of a task is subordination and coordination, particularly in intermediate and advanced levels of NNS development. The expected progression of development is coordination at lower levels of development, with increasing subordination as language learners progress in their understanding and effective ability with the second language (Norris & Ortega 2009). However, subordination terms should reach rather a climax at the upper intermediate levels, and should then gradually subside again as writing improves further, and more complex syntactic forms are mastered, such as in the example at the beginning of this paper (Norris & Ortega 2009). Understanding syntactic complexity therefore requires finding a way to analyze these various levels of syntactic complexity appropriately; an overuse of subordination, for example, suggests that the language learner must focus on more complex syntactic forms in order to increase complexity. Qualitative analysis and counting the use of words that indicate coordination and subordination are currently the most effective methods of measurement (McNamara et al 2010). Since the study conducted in this paper focused on advanced and upper intermediate level NNSs, the analysis here is exclusively on measures of subordination.

Lexical sophistication is also an important element of complexity, and must be considered in any measures of complexity or of CAF (Skehan 2009). Just as resource-directing factors should increase syntactic complexity, so too should they increase lexical sophistication (Ong & Zhang 2010). Complexity in lexis is considered in two different ways: variation, and frequency (Norris & Ortega 2009, Yu 2010). Variation refers to the number of different words used by a writer, and is mainly calculated using a type-token ration, which may be

adjusted for length through methods such as the D-formula (Yu 2010). Frequency refers to the relative level of words that are used, based on analysis of an external corpus, such as Nations' frequency lists (Kormos 2011). Through a combination of these analyses, writers can be placed according to the level of vocabulary knowledge and use that is displayed in their output. However, such analyses do not address specificity in the task; although a NNS writer used a high-level word, according to standard frequency measures, it may not be appropriate to the task. This would essentially give a "false positive reading" of complexity, in the sense that it is complexity for the sake of complexity, and actually impedes fluency and readability. This issue has not been addressed in any other studies in this area that I have found.

To address the issues discussed in this paper, this study attempts to answer three research questions regarding the analysis of complexity:

How will output differ in complexity between the two tasks? Will this output support Cognitive Hypothesis or Trade-off Hypothesis, or neither?

How will NNS and NES output differ in both syntactic and lexical complexity within each task?

How effective will the methods for describing complexity be at differentiating among NES, advanced NNS and intermediate NNS?

## V. This Study

*V.A Subjects*

The subjects for this study were 16 Koreans and 3 NESs. The Koreans were subdivided into two groups: 7 Advanced (ave. TOEFL iBT score=112) and 9 Intermediate (ave. TOEFL iBT score=88). The advanced group was all female; the intermediate had 2 males and 7 females. All but 1 from each group were graduate students in the Developmental Education program at Seoul National University, which program is taught in English, and the remaining 2 participants both worked in ESL publishing, so all used English on a regular basis. All Korean participants had lived in English-speaking countries for anywhere from 1-11 years, and ranged in age from 28-47. Certain homogeneity in use of English was attempted in order to control for writing ability level,

in order to ensure that no participants were unfamiliar with writing regularly in English (Yu 2010). For comparison purposes, homogeneity also seemed an important goal in the NES group, in part because it has been so difficult to establish a baseline for NESs, given the diversity that naturally occurs in a language (Foster & Tavakoli 2009). Therefore, all NES participants in this study worked, either part- or full-time, in publishing, and all had at least a master's degree. This narrow selection attempts to ensure a higher level of proficiency that might be more in keeping with that of the Korean subjects (similar academic purposes for writing English) in CAF by which to compare them. The NESs ranged in age from 36-40.

*V.B Test Design*

The test was comprised of 2 parts, though this paper only analyzes the second part of the test; however, it is described here as its presence may have had an unintentional priming effect on Part 2. Part 1 was based on tasks designed by Norris and Chapelle, but instead of creating tasks that could only have one answer in the gap; these tasks were designed to have a broader range of answers. This could have been done in 2 different ways: either by making the gap longer, requiring one longer answer from the subjects; or by requesting multiple answers of the test-takers. Since this portion of the test was focussing solely on tense, the latter was chosen; the former would be more appropriate for more complex forms, such as subordinate clauses. Part 2 was based on Tavakoli & Foster (2011), which was two narrative tasks: the first was +multiplicity, +fluidity; the second, –multiplicity, –fluidity. Presumably, the first should result in greater subordination, but the second in greater complexity, at least according to Cognition Hypothesis; the results would be less predictable according to Trade-off Hypothesis, and would depend on how test-takers allocated resources. In either case, the difference in fluidity between the two should have made the second more complex. All tasks for the test are given in Appendix A. The items were put on PowerPoint, for several reasons. First of all, this allowed the tasks to be strictly timed, which more closely emulates the time constraints of a speaking task, and which is considered more authentic for testing purposes (Kormos 2009). At the same time, it allowed the

tests to be done at home or in a comfortable environment, rather than a high-stress situation, which has been found to affect output and emergence, particularly in that test-takers have been found less likely to take risks if they are in a high-stress testing situation (Robinson & Gilabert 2007, Kormos 2011). Responses were done on pen and paper, partly from necessity (it was impossible to find an easily accessible program that met the timing requirements and still allowed answers to by typed in), but also to prevent any interference from test-takers dealing with an unfamiliar computer program, which could tend to increase difficulty through resource-dispersing factors.

*V.C Analysis*

Since the groups involved are relatively small, this paper uses descriptive statistics and qualitative analysis. The analysis of Part 1 of the test is not presented here. For analysis of Part 2, quantitative comparisons for CAF include total words for fluency, words/T-unit, clauses/T-unit and words/clause for complexity, error-free clauses/T-unit for accuracy, frequency of the subordination indicators "while," "after," and "when" for subordination (as the test-takers were intermediate-advanced, coordination measures were not done), and type-token ratios for lexical variation (analyzed using AntConc 3.2). In all cases, these measures were calculated for each individual subject's output, and then the average of these calculated for each group. Qualitative comparisons of lexis, focusing on the words used by NNSs compared to a NES baseline, are also incorporated. Ideally, a word frequency comparison with an external corpus would have been done, but this requires special software that was unavailable at the time of writing, so such an analysis must be reserved for future research.

## VI. Results

Tables 3, 4 and 5 show the results for each of the three groups, using the measures described above for CAF.

**Table 3.** Native English Speaker results for CAF measures

| Native English | Task #1 | Task #2 |
|---|---|---|
| Words | 141.7 | 140.7 |
| Words/T-unit | 11.5 | 12.8 |
| Clauses/T-unit | 1.7 | 1.7 |
| Words/Clause | 6.8 | 7.7 |
| Word Types/Tokens | 0.63 | 0.63 |

**Table 4.** Advanced Speaker results for CAF measures

| Korean—Advanced | Task #1 | Task #2 |
|---|---|---|
| Words | 95.1 | 90.4 |
| Words/T-unit | 12.5 | 10.5 |
| Clauses/T-unit | 1.8 | 1.6 |
| Words/Clause | 7.1 | 6.8 |
| Error-free Clauses/T-unit | 1.3 | 1.2 |
| Word Types/Tokens | 0.59 | 0.71 |

**Table 5.** Intermediate Speaker results for CAF measures

| Korean—Intermediate | Task #1 | Task #2 |
|---|---|---|
| Words | 70.4 | 69.8 |
| Words/T-unit | 8.4 | 8.1 |
| Clauses/T-unit | 1.4 | 1.4 |
| Words/Clause | 5.9 | 5.7 |
| Error-free Clauses/T-unit | 0.9 | 0.9 |
| Word Types/Tokens | 0.67 | 0.71 |

To read these results: the higher number suggests a higher degree of accuracy (words), complexity (words/T-unit, clauses/T-unit and words/clause), accuracy (error-free clauses/T-unit), and lexical variety (type-token). These results show mixed effects for task difficulty within groups. In all cases, the more difficult task had fewer words in the output, which shows a similar influence on fluency for all groups. However, these results also suggest that, while the NESs did produce more complex (albeit shorter) texts for task #2, the opposite occurred for both NNS groups; in other words, increased task difficulty showed increased complexity for the NESs, but not for the NNSs. Accuracy hardly changed for either NNS group. Lexical variation only appeared to have a noticeable difference for the advanced group, with the first task show-

**Table 6.** Comparison of 3 groups for task 1

| Task 1 | NES | Advanced | Intermediate |
|---|---|---|---|
| Words | 141.7 | 95.1 | 70.4 |
| Words/T-unit | 11.5 | 12.5 | 8.4 |
| Clauses/T-unit | 1.7 | 1.8 | 1.4 |
| Words/Clause | 6.8 | 7.1 | 5.9 |
| Error-free Clauses/T-unit | N/A | 1.3 | 0.9 |
| Word Types/Tokens | 0.63 | 0.59 | 0.67 |

**Table 7.** Comparison of 3 groups for task 2

| Task 2 | NES | Advanced | Intermediate |
|---|---|---|---|
| Words | 140.7 | 90.4 | 69.8 |
| Words/T-unit | 12.8 | 10.5 | 8.1 |
| Clauses/T-unit | 1.7 | 1.6 | 1.4 |
| Words/Clause | 7.7 | 6.8 | 5.7 |
| Error-free Clauses/T-unit | N/A | 1.2 | 0.9 |
| Word Types/Tokens | 0.63 | 0.71 | 0.71 |

ing somewhat lower variation than the second.

Tables 6 & 7 compare the three groups on all CAF measures for each task.

The tasks performed appropriately, as far as syntactic complexity among different levels is concerned, in that the NES group displayed greater complexity and fluency than the other groups, and the advanced group showed greater complexity, fluency and accuracy than the intermediate group. However, lexical variation performed opposite to what would be expected, in that NESs showed slightly lower variation than NNSs.

Subordination analysis was of the use of “while,” “after,” and “when.” Below are tables (Tables 8 & 9) that give frequency ratios on the three subordination terms being examined for each group as a whole, and are therefore calculated as the total number used by the group, divided by the number in each group. This is not the best measure, but still proved useful, even given the small sample sizes.

As this chart suggests, Task 1 led to more frequent use of “while” and “when”, while Task 2 led to more use of “after.” A comparison among NES and NNSs suggests that, at least on Task 2, subordination occurred as predicted in Nor-

**Table 8.** Frequency of 3 subordination terms for Task 1

| Task 1 | NES | Advanced | Intermediate |
|---|---|---|---|
| While | 0.7 | 0.6 | 0.1 |
| After | 0.3 | 0.1 | 0 |
| When | 0.7 | 1 | 0.4 |

**Table 9.** Frequency of 3 subordination terms for Task 2

| Task 2 | NES | Advanced | Intermediate |
|---|---|---|---|
| While | 0 | 0.3 | 0.4 |
| After | 1 | 0.9 | 0.8 |
| When | 0 | 0.1 | 0 |

ris & Ortega (2009), in that NESs did not use "while" or "when," suggesting they found other ways to say things that required use of those words. Understanding the exact nature of what happened among the three groups would require closer inspection through qualitative analysis, which was not done for this study.

Since there was a discrepancy in type/token ratios, a quick comparison was made among the top 20 words of each group to see if this gave any insight into why NESs seemed to use fewer types. The results are given in Tables 10 & 11 below, represented as the ratio of each grammar part used per person in the group.

This analysis suggests that the NESs tended to use many more functional words, especially prepositions, than NNSs, which may account for why there were fewer types/tokens among NESs than NNSs. Furthermore, a closer analysis suggests that the NESs used not only a greater number but a wider variety of pronouns and prepositions, as well as using a broader range of nouns to describe one thing (e.g. both "kids," and "boy and girl" were used by NESs, but primarily "kids" by NNSs). It is also worth noting that all the advanced and most intermediate NNSs used the word "went" (this was the only verb (other) in Task 1, and the most frequent in Task 2), while none of the NESs used "went" specifically, and only 1 NES in each task used "go." This suggests that the NESs found another method to express the same idea; finding out how would require closer qualitative analysis, which was not done

**Table 10.** Ratio of grammar parts per person for Task 1

| Task 1 | NES | Advanced | Intermediate |
|---|---|---|---|
| Articles | 15.7 | 10.4 | 6.8 |
| Pronouns | 12 | 9.1 | 4.9 |
| Conjunctions | 5 | 2.9 | 2.6 |
| Prepositions | 12 | 6.2 | 4.7 |
| Nouns | 15 | 10.6 | 6.9 |
| Verbs (to be) | 4 | 4 | 2.3 |
| Verbs (other) | 0 | 1 | 0.8 |

**Table 11.** Ratio of grammar parts per person for Task 2

| Task 2 | NES | Advanced | Intermediate |
|---|---|---|---|
| Articles | 20 | 9.6 | 5.6 |
| Pronouns | 6.3 | 7.3 | 4.4 |
| Conjunctions | 5.7 | 3.4 | 2 |
| Prepositions | 18 | 6.7 | 4.2 |
| Nouns | 9 | 6.1 | 5.6 |
| Verbs (to be) | 1.7 | 0.9 | 2.1 |
| Verbs (other) | 0 | 2.1 | 0.8 |

here.

## VII. Discussion

This discussion addresses each of the research questions, one by one.

*VII.A How will output differ in complexity between the two tasks? Will this output support Cognitive Hypothesis or Trade-off Hypothesis, or neither?*

The results for this question are difficult to read, as the NES and NNS had opposite results. The NES results on syntactic complexity seem to support Cognitive Hypothesis, in that the task which should have been more complex from a –fluidity perspective did lead to greater complexity, while the +multiplicity task seemed to produce greater subordination. However, the NNSs showed more complexity on the +fluidity task. Subordination measures were somewhat difficult to understand, but suggest that both task types are capable of eliciting subordination, though in different ways; “while” and “when”

seem to be more indicative of +multiplicity, while "after" may have been more useful to help build cohesion in –fluidity. This is rather logical. Overall, it seems that NES results may not be reliable to establish task effect, possibly because of fluency. If this is the case, then the results seem to support Trade-off Hypothesis; however, a larger NES sample size would likely be needed to confirm this

*VII.B How will NNS and NES output differ in both syntactic and lexical complexity within each task?*

Within each task and according to T-unit analysis, the NESs definitely produced more syntactically complex texts than the advanced NNSs, whose texts were in turn more complex than those of the intermediate NNSs. The subordination analysis furthermore showed largely the same effect, which suggests that NES output is useful as a baseline for comparison with NNS output, and also as a developmental goal for NNS output. The same is also accurate for the subordination measures; in fact, according to the measures given here, the NES baseline was more effective at predicting the type of subordination that would occur than +/–multiplicity and +/–fluidity were. While the lexical complexity measures, especially type/token ratios, were initially a bit confusing, further analysis seemed to indicate that this field could also be useful, with the addition of detailed, qualitative analyses. These additional analyses were not suggested in other studies; more research needs to be done in this area in order to tease out the best methods for giving the most useful information possible. This is discussed further in the next question, below.

*VII.C How effective will the methods for describing complexity be at differentiating among NES, advanced NNS and intermediate NNS?*

The measures used were effective, but only up to a point. Certainly, the syntactic complexity measures proved effective; however, these were complicated by the lexical sophistication measures, which seemed to indicate that NESs did not perform as well as NNSs. This could have been caused by sample size, as was mentioned above, but there were attempts to mitigate this effect (after some experimentation) by taking the average per-person output, rather than

the average of the whole group output combined. Using a more precise method of calculation, such as D-formula, may also be useful, though even this has limitations, as it cannot give the *nature* of use (Yu 2010). Yet, the subsequent analysis suggests that there may be another dimension to complexity that is not explored in the description given by Systemic Functional Linguistics and discussed in the introduction above; namely, that complexity does not always equate with transformation. This is based on the greater number and range of functional words used by NESs on these tasks, especially prepositions. Through personal experience with making language tests, it is apparent that grammatical structures indicative of the transformations described above are rarely those which test-takers find most difficult; instead, the ones with greater difficulty tend to be those utilizing more functional words, especially prepositions, in both reading- and grammar-specific tasks. This is an aspect that needs to be explored more, which is the topic of the final section.

## VIII. Conclusion and Future Research

This study has suggested several areas for future research into complexity. Further investigation into using NESs as a baseline was predictably effective for comparison purposes (Foster & Tavakoli 2009), though it seemed to confuse the issue of Cognitive and Trade-off Hypotheses; it is therefore not entirely clear whether NESs should be a goal. However, this could be because of the nature of the measures used for task complexity, which tend to describe the amount of use of a form or word, but not the way it is used (Yu 2010). It would also be helpful to do a comparison of lexis based on a NES baseline that is task specific, and compare those results with an analysis on an external corpora, as was described above, to see how the word use may differ, and whether analyses based on corpora give false readings of complexity. More effective measures need to be found. Future analyses should also include larger sample sizes, especially for NESs. Incorporating different levels would likely be productive for further analysis of the effectiveness of these measures (Norris & Ortega 2009). This might also lead to more potential for in-depth error analysis, particularly regarding the types of errors that are made by different levels

of NNSs; for example, whether, in addition to using fewer prepositions, NNSs also tend to make more errors in those. Of course, such analyses require more precise methods; programs such as Coh-Metric or Celex may be beneficial (McNamara et al 2010, Yu 2010). Ultimately, the measures that are currently used seem sufficient for descriptive scoring purposes, but are not fine-grained enough either for diagnostic or developmental purposes. More work certainly needs to be done in this area.

## Bibliography

Foster, Pauline, and Parvanine Tavakoli. (2009). Native Speakers and Task Performance: Comparing Effects on Complexity, Fluency, and Lexical Diversity. *Language Learning*. 59/4, 866-896.

Jang, E., (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to *LanguEdge* assessment. *Language Testing* 26(31), 31-73.

Kormos, Judit. (2011). Task Complexity and Linguistic and Discourse Features of Narrative Writing Performance. *Journal of Second Language Writing*. 20/2, 148-61

McNamara, Danielle S., Crossley, Scott A., McCarthy, Philip M. (2010). Linguistic Features of Writing Quality. *Written Communication*. 27/1, 57-86.

Norris, John M. (2005). Using Developmental Sequences to Estimate Ability with English Grammar: Preliminary Design and Investigation of a Web-based Test. *Second Language Studies*. 24/1, 24-128.

Norris, John M., Ortega, Lourdes. (2009). Towards an Organic Approach to Investigating CAF in Instructed SLA: The Case of Complexity. *Applied Linguistics*. 3/4, 555-78.

Ong, Justina, Zhang, Lawrence J. (2010). Effects of Task Complexity on the Fluency and Lexical Complexity in EFL Students' Argumentative Writing. *Journal of Second Language Writing*. 19/4, 218-33.

Pallotti, Gabrielle. (2009) CAF: Defining, Refining and Differentiating Constructs. *Applied Linguistics*. 30/4, 590-601.

Robinson, Peter, Gilabert, Roger. (2007). Task Complexity, the Cognition Hy-

pothesis and Second Language Learning and Performance. *International Review of Applied Linguistics in Language Teaching*. 45/3, 161-76.

Robinson, Peter. (2001). Task Complexity, Task Difficulty, and Task Production: Exploring Interactions in a Componential Framework. *Applied Linguistics*. 22/1, 27-57.

Skehan, Peter. (2009). Modelling Second Language Performance: Integrating Complexity, Accuracy, Fluency, and Lexis. *Applied Linguistics*. 30/4, 510-32.

Tavakoli, Parvanine, Foster, Pauline. (2011). Task Design and Second Language Performance: The Effect of Narrative Type on Learner Output. *Language Learning*. 61/s1, 37-72.

Yu, Guoxing. (2010). Lexical Diversity in Writing and Speaking Tasks Performances. *Applied Linguistics*. 31/2, 236-59.

## Appendix A. Tasks Used for the Study

1. The woman _____ on Saturday.

2. The boy _____ pictures for his hobby.

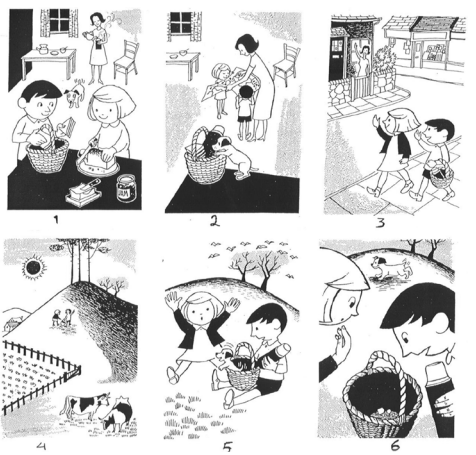
3. The boys _____ beach volleyball after school every day.

4. The girl _____ her bike on the weekend.

5. The dog _____ the man.

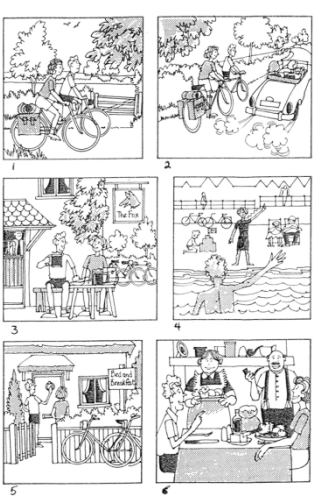
Part 2

1.



*Your response:*

2.



*Your response:*

**ABSTRACT**

# Exploring Syntactic and Lexical Complexity in Narrative Tasks

Roz Hirch

Improved complexity, accuracy, and fluency (CAF) is an often-stated goal of writing instruction, but there are few clear or well-established guidelines for describing and analyzing CAF. This paper looks at various measures of CAF as well as two theories for describing how they interact: Trade-off Hypothesis and Cognition Hypothesis. A study was conducted with 3 groups—a group of 3 native English speakers; one of 7 advanced non-native English speakers, and another of 9 intermediate non-native speakers—on two separate writing tasks developed in Tavakoli and Foster (2011). Analysis of the output was done utilizing a variety of measures discussed in the paper. The study offered support, albeit weak, for Trade-off Hypothesis. Suggestions for future research and improvements to analysis are suggested.

*Key Words* complexity, accuracy, fluency, CAF, writing assessment, Trade-off Hypothesis, Cognition Hypothesis, task difficulty