All Theses and Dissertations

2013-03-18

# Toward a More Inclusive Construct of Native Chinese Speaker L2 Written Error Gravity

Steven K. Holland
*Brigham Young University - Provo*

Toward a More Inclusive Construct of Native Chinese Speaker L2 Written Error Gravity

Steven K. Holland

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Arts

Wendy Baker Smemoe, Chair
Dan P. Dewey
Norman W. Evans

Department of Linguistics and English Language

Brigham Young University

March 2013

ABSTRACT

Toward a More Inclusive Construct of Native Chinese Speaker L2 Written Error Gravity

Steven K. Holland
Department of Linguistics and English Language
Master of Arts

The purpose of this study is to determine two types of error gravity in a corpus of texts written by native Chinese learners of English (ELLs)—one that enriches the traditional construct of gravity found in error gravity research by including error frequency, or how often an error occurs in a text relative to others, as an intervening variable, and one that applies the new error gravity data in a practical way to help establish salient grammatical focal points for written corrective feedback (WCF). Previous error gravity research has suggested that the amount of irritation caused by error is determined by the extent to which an utterance departs from "native-like" speech. However, because these studies often neglect the role of frequency in determining gravity—relying on isolated sentences, pre-determined errors, and manipulated texts to define it—a more complete view of error gravity is needed. Forty-eight native English speakers without ESL teaching experience and 10 experienced ESL teachers evaluated a set of 18 timed, 30-minute essays written by high intermediate to advanced native-Chinese ELLs. Errors were identified, verified, tagged, and classified by the level of irritation they produced. Results show the most serious errors included *count/non-count (C/NC)*, *insert verb (INSERT V)*, *omit verb (OMIT V)*, and *subject-verb agreement (SV)*. The most frequent error type was *word choice (WC)*, followed by *singular/plural (S/PL)*, *awkward (AWK)*, and *word form (WF)*. When combined, *singular/plural (S/PL)*, *word form (WF)*, *word choice (WC)*, and *awkward (AWK)* errors were found to be the most critical. These findings support Burt and Kiparsky's (1972) global/local error distinction in which global errors, or those lexical, grammatical and syntactic errors that affect the overall organization or meaning of the sentence (Burt, 1975) are deemed more grievous than local ones, which affect only "single elements (constituents)" (Burt, 1975, p. 57). Implications are discussed in terms of future research and possible uses in the Dynamic Written Corrective Feedback classroom.

Keywords: error gravity, error frequency, irritation, global error, local error, L2 writing, written corrective feedback

# ACKNOWLEDGEMENTS

TABLE OF CONTENTS

# LIST OF TABLES

## LIST OF FIGURES

# CHAPTER 1: INTRODUCTION

It would be ideal to live in a world where second language (L2) instructors could teach so effectively and L2 students learn so completely that no error correction (EC) would ever be needed (Corder, 1967). Yet, L2 students obviously do make errors—in some instances, even grave ones—and instructors are therefore compelled to address them (Evans, Hartshorn, & Tuioti, 2010). In a survey of over 1,053 L2 English instructors in 69 different countries, the use of written corrective feedback (WCF), a subcategory of EC, was widely reported. Roughly 99% of respondents used at least some WCF in their classroom, and of these, 92% reported using WCF on a regular basis—on over 66% of the writing submitted by their L2 students (Evans, Hartshorn, & Tuioti, 2010). EC is clearly a common practice in writing pedagogy.

EC "concerns the way in which teachers (and other learners) respond to learners' errors," and it is usually "discussed in terms of whether errors should be corrected, when, how, and by whom" (Ellis, 2008, p. 961). According to James (1998), it can entail (1) *feedback*, "informing the learners that there is an error, and leaving them to discover it and repair it themselves," (2) *correction,* "providing treatment or information that leads to the revision and correction of the specific instance of error (the error token) without aiming to prevent the same error from recurring later," and (3), *remediation*, "providing learners with information that allows them to revise or reject the wrong rule they were operating with when they produced the error token" ( p. 237). Instructors may wish to employ any number of these in their pedagogy, and in some instances, an EC approach may involve all three, as is the case with a methodology developed by researchers at Brigham Young University called dynamic written corrective feedback, or DWCF (Evans, Hartshorn, McCollum, & Wolfersberger, 2010; Hartshorn, Evans, Merrill, Sudweeks, Strong-Krause, & Anderson, 2010; Evans, Hartshorn, & Strong-Krause, 2011).

DWCF is an adapted two-part form of WCF that focuses on "(a) feedback that reflects what the individual learners need most, as demonstrated by what the learner produces, and (b) a principled approach to pedagogy that ensures that writing tasks and feedback are *meaningful*, *timely*, *constant*, and *manageable*" (Hartshorn et al., 2010, p. 87). In brief, the process entails students writing 10-minute paragraphs 3-4 times a week and submitting multiple drafts to their instructor for written feedback. Additionally, students must track their progress through a set of forms, including an error tally sheet (see Appendix C), used as "the basis for explicit instruction essential to skill-acquisition theory" (Hartshorn et al, 2010, p. 88).

The method has seen proven success. Hartshorn et al (2010) provided DWCF to 47 advanced-low to advanced-mid English as a Second Language (ESL) students, and Lee (2009) to 53 intermediate-mid to intermediate-high ESL students studying at Brigham Young University's English Language Center (ELC). Both studies used a pre-test/post-test design to determine the linguistic accuracy gains resulting from 15 weeks of DWCF. Hartshorn et al report a "relatively large effect on improving the mean accuracy scores of those students in the treatment group compared with those in the contrast group" (p. 100), and Lee (2009) concludes DWCF is an effective means of improving linguistic accuracy because students prefer the method over traditional grammar instruction. Furthermore, at the collegiate level, Evans et al (2011) determined that university-matriculated English as a second language (ESL) students also can benefit from DWCF, as it again produced significant accuracy gains when the treatment group was compared to the control. DWCF is a valuable system that can accomplish valuable results.

Still, as Evans et al (2011) remind us, "the variables that influence the outcomes of WCF are many, and each must be carefully considered if we are ever going to clarify efficient and effective practices" (p. 231). WCF is a highly nuanced process that involves learner, situational,

and methodological variables (Evans et al, 2011), and to better understand how each contributes to WCF, each must be broken into its component parts, or subcategories, and included "in a full research agenda" (p. 231). Two subcategories that will be addressed in this present study are the theoretical construct of error gravity and its practical application to determining salient grammar focal points in DWCF.

First, stating that the error tally sheets form "the basis for explicit instruction" (Hartshorn et al, 2010, p. 88) implies a marked preference for error frequency, or how often an error occurs. It assumes that the seriousness of error is determined largely by its quantity, a view similar to that of Albrechtsen, Henriksen, and Faerch (1980), who hypothesize that irritation, or the affective response to error, "is directly predictable from the number of errors which an IL [interlanguage] text contains, regardless of error type or of other linguistic aspects of the text" (p. 394). Yet, the issue presented by this assumption is that it is incomplete. It accounts for only part of the rating context, as error is at least as much social as it is numerical (James, 1972).

Delisle (1982) states the problem in these words: "if our goal is to achieve absolute linguistic correctness, all errors are equally serious and will be rated accordingly. However, if we define our objectives in terms of communicative success, then we will probably use a different rating scale" (p. 39), meaning we will need to think of error in degrees if we are to address fully its social impact, or "the seriousness of an error" (Ellis, 2008, p. 961). Given the role of frequency in determining linguistic norms and the role of linguistic norms in deciding error, it would, of course, be foolish to completely discount the role of frequency in error evaluation, but other factors, i.e. the readers' ability (actual or perceived) to understand an utterance, the amount of attention drawn by certain errors, and their social acceptability, must also be accounted for.

Second, Robinson (1973, as cited in Johansson, 1978) notes "it may happen that the

teacher becomes obsessed with insignificant errors to the exclusion of others that are much more important" (p.1). If grave errors, those that are seen by society as unacceptable or that render the writing incomprehensible, are not also the most frequent (in its purest sense as an objective error count), the losses to both time and energy can be substantial (Khalil, 1985). In other words, a missing verb is a serious error because it affects the overall meaning of the sentence. Yet, because it occurs less frequently than, say, a serious word form error (such as using *constitution* when *constitutional* is called for) it would be irresponsible to devote the next lesson to verbs when the word form error is clearly more impactful. As Ferris (1999) argues, "it is vitally important for teachers to commit themselves to selective error feedback and to a strategy for building students' awareness and knowledge of their most serious and frequent grammar problems" (p. 7). A system that accounts for gravity *and* frequency will produce the best results.

Ferris (1999) continues by suggesting that "students can be successfully taught to self-edit their own texts if they are (a) focused on the importance of editing; (b) trained to identify and correct patterns of frequent and serious errors; and (c) given explicit teaching as needed about the rules governing these patterns of errors" (p. 5), Although she qualifies this statement by arguing that some errors (i.e. subject-verb agreement, run-ons, and comma splices) are more rule-governed than others (i.e. *some* missing words, unnecessary words, and word order problems) and therefore more amenable to correction than others, definite value can still come from identifying those L2 lexical, grammatical, and syntactic errors—rule-governed or not—that are endemic among a given language group or that cause the greatest amount of stigma within a target language community. In this manner, one empowers both students and instructors to not only set the most salient lexical, grammatical, and syntactic priorities but to also make the most effective use of their limited time and resources.

To account for both gravity and frequency in DWCF, one might look—with a minor addition—at Chan's (2010) taxonomical study of common lexicogrammatical ESL errors made by Hong Kong Cantonese speakers. Following a brief discussion of the value of both fields of inquiry (grievousness and prevalence), she adds "remedial efforts should of course be put on grievous errors, but prevalent errors, such as word class confusion, should also receive attention" (p. 314) and argues her taxonomy could be used to create "an error gravity scale and an error-prevalence scale" ranging from "the most grievous or prevalent to the least grievous or prevalent" (p. 314).

L2 instructors could use these scales to "sequence and prioritize their teaching focus according to the prevalence and gravity scales" (p. 315), a productive first step. Yet because gravity and frequency are not mutually exclusive constructs—error is determined, to some extent, by how often an utterance occurs in the target language (TL) relative to others—an additional scale that combines them is likely the better approach (James, 1998).

**Purpose of Study**

The purpose of this study is to assist instructors with finding the most salient lexical, grammatical, and syntactic focal points by (1) expanding the current construct of error gravity to account also for error frequency (a critical omission in previous protocols), (2) generating an error frequency for the errors made by native-Chinese speaking (NCS), pre-university, second language (L2) students, and (3) synthesizing the new error gravity and frequency data to determine those errors that are both grievous and frequent.

**Research Questions**

1. Which NCS L2 writing errors are most serious to NES university students when using authentic, essay-length discourse?

2. Which NCS L2 writing errors are most frequent to NES university students when using authentic, essay-length discourse?

3. Which errors are both highly serious and highly frequent to NES university students?

**Definitions**

Acceptability: "the degree to which a given L2 violates language norms" (Ludwig, 1982, p. 277), or as Khalil puts it, "the seriousness of errors" (p. 336). James (1998) argues "to decide on the acceptability of a piece of language we refer not to rules, but to contexts, trying to contextualize the utterance in question" (p. 67).

Comprehensibility: the "listeners' perceptions of difficulty in understanding particular utterances" (Munro & Derwing, 1995, p. 291).

Error evaluation: "the process whose duty is the systematic and objective determination of merit, worth or value" (Scriven, 1991, p.4).

Error frequency: a measure of how often certain errors occur in a text (Ellis, 2008). Frequency measures both the types and instances of error L2 learners produce (James, 1998).

Error gravity: the "seriousness of an error" (Ellis, 2008, p. 961). "Seriousness" is often understood to be the result of intelligibility, comprehensibility, irritation, and acceptability (Khalil, 1985).

Global Error: in error gravity research, lexical, grammatical, and syntactic errors that "affect overall sentence organization" (Burt, 1975, p. 56) and "cause the listener or reader to misinterpret the speaker or writer's message" (p. 57). These errors include verb tense, verb form, modal, conditional, sentence structure, word order, connectors, passive voice, and unclear meaning (Lane & Lange, 1993).

Intelligibility: the interlocutor's ability to understand "the meaning of the utterance" (Khalil, p. 344), "the extent to which an utterance is actually understood" (Munro & Derwing, 1995, p. 291)—i.e. how well the reader understands the text.

Irritation: "the result of the form of the message intruding upon the interlocutor's perception of the communication. For the listener or reader, the irritation continuum ranges from unconcerned, undistracted awareness of a communicative trait to a conscious, constant preoccupation with form, to the point that the message is totally obscured or lost" (Ludwig, 1982, p. 275). Irritation is also "a function of the speaker/writer's erroneous use of language measured against the characteristics and expectations of the interlocutor" (p. 275). Santos (1988) refers to irritation as the "bother" factor. Gynan (1985) and Khalil (1985) assert that it is the affective response to error, "native speakers' emotional reactions to deviant utterances" (Khalil, 1985, p. 336). It is the thing that causes people to stop and take notice of a textual feature.

Linguistic Norm: "what is agreed on by the speakers in the community concerned" (Milroy, 1992). Norms are democratic in nature, as each group has the power to determine what is or is not acceptable to the community on a local level regardless of where it fits on the scale of class, privilege, or prestige (James, 1998). If enough people start saying "could care less" instead of "couldn't care less," the latter would no longer carry the stigma it once had.

Local Error: an error affecting only "single elements (constituents) in a sentence" (p. 57), such as those errors involving word choice, spelling, and prepositions.

**Delimitation**

This study has several delimitations. First, this research centers on the evaluation of error, not the cause. It is difficult to say where an error comes from given the complexity of the issue. L2 acquisition is a highly idiosyncratic process, with each individual learner the sum of his or

her education, experience, environment, aptitude, motivation, and so on (e.g. Corder, 1967; Duskova, 1969; James, 1972). While it is possible to make educated guesses about where an error may originate, ultimately, one can never be entirely sure.

Second, it focuses on the evaluation of NCS L2 errors. Company (2012) includes L2 speakers of many different languages in her study. Yet, because of the need in this present study to obtain as many different NES viewpoints as possible on a finite set of errors, it proved to be an overwhelmingly large undertaking to focus on more than one L1 group. Fortunately, Company (2012) also divides her data by L1, so it remains possible to compare the frequency data obtained in this present study with hers.

Third, this evaluation relates to only certain narrowly defined errors. I use the error codes used in the DWCF course, the same ones used by Company (2012). Yet, these present only one way of describing error. Indeed, in the literature on the subject, there are several errors that appear, such as it-deletion and pronoun agreement that do not have a clear place in the DWCF codes. It-deletion could be classified as an insert error or an incomplete sentence. Pronoun agreement could be described as a word choice, awkward, or an unclear meaning issue. With this in mind, it must be recognized that there are inherent limitations in how the errors are identified.

Fourth, these limitations are unavoidable for the time being. For practical reasons, it is necessary to keep the number of codes used in DWCF to a minimum. Not only do the instructors need to be able to learn the concepts behind the symbols and resolve a method for applying them (i.e. whether to mark all errors, to mark only specific errors, to include any corrections) but the students must also be able to internalize them. Furthermore, this common language must be simple enough that it can be learned in a relatively short period of time. As such, a more detailed system would likely prove too cumbersome to the aims of DWCF and is therefore unadvisable.

**CHAPTER 2: LITERATURE REVIEW**

This chapter begins with a brief discussion on the underlying cause of error—deviance from accepted linguistic norms—and how the affective response to error (irritation) that stems from deviance affects error evaluation. Error has been described in error gravity research as being determined by the extent to which an utterance departs from "native-like" speech, a distinction which also appears to follow a trend from global errors, defined in the error gravity literature as errors that affect the overall organization of meaning of a sentence (Burt, 1975) to local ones, or those that affect only "single elements (constituents)" (Burt, 1975, p. 17). However, because these results are often obtained using a variety of flawed research practices, such as relying on isolated sentences, pre-determined errors, vaguely defined constructs of error, or neglecting the role of frequency in error gravity, additional study is required.

**A Social Constructivist View of Error**

Johansson (1978) asserts "the identification of errors presupposes a norm against which the learner's utterances can be judged" (p.1). Norms are the particular patterns of effective communication that are agreed upon by a given community (Milroy, 1992)—the codified expectations of discrete language groups, otherwise known as intelligibility nuclei or knowledge structures (Gergen, 1997). In linguistic terms, these nuclei are implicit in the constructs of *intelligibility* and *comprehensibility,* as both depend heavily on communities establishing "interrelated propositions that furnish a community of interlocutors with a sense of description and/or explanation within a given domain" (Gergen, 1997, p. 6).

Intelligibility is the interlocutor's ability to understand "the meaning of the utterance" (Khalil, p. 344), or "the extent to which an utterance is actually understood" (Munro & Derwing, 1995, p. 291). It refers to the act of determining that one understands the intended meaning of the

writer, and it is best measured through objective (i.e. restatements and rewrites) rather than subjective (i.e. surveys) means. In contrast, comprehensibility is defined as the communicative "accessibility of the content" (James, 1998, p.212), or the "listeners' perceptions of difficulty in understanding particular utterances" (Munro & Derwing, 1995, p. 291). It refers to how difficult they believed it was to understand the utterance's intended meaning, not an indication of the actual understanding itself. Piazza (1980) argues that the two are synonymous. However, given that using *comprehensibility* when *intelligibility* is warranted can be confusing, the subtle distinction between the two is worth maintaining.

Linguistic norms are often the means "of categorizing persons [and their language use] and the complement of attributes felt to be ordinary and natural for members of each of these categories" (Goffmann, 1963, p.2). To give an example, an English speaker who asks, "Have you eaten?" may expect a simple "Yes" instead of the full "Yes, I have eaten." That the speaker would use this abbreviated version of the language instead of the full form and that the questioner could understand the utterance, incomplete as it is, is a strong indication that the communication rests upon internal notions or patterns of acceptable linguistic use (Vygotsky, 1964). If someone were to condense English speech in any other way—using just *have* or *eaten*, for example—the pattern would be violated and the resulting utterance found to be unacceptable. By extension, whether speakers are aware of this contextual dependency may have implications for how they are perceived by native English speakers (NES), as the error could easily mark them as deviant, or not part of the English-speaking group.

From this categorization of "acceptable" and "unacceptable" language use, we might conclude that the primary issue, then, is not necessarily the error itself—after all, everyone makes errors—but how the errors affect one's ability to obtain membership within the dominant

community, membership that is often the byproduct of an errors' relative gravity or seriousness in the eyes of those who evaluate the texts. To determine what will likely be a grave error, however, let us first consider how errors are identified.

**Methods of Identifying Error**

Erroneous tokens must be detected and located (James, 1998), a process made possible by *irritation*, or the level of attention afforded to certain tokens. Ludwig (1982) defines irritation as "the result of the form of the message intruding upon the interlocutor's perception of the communication" (p. 275), a "function of the speaker/writer's erroneous use of language measured against the characteristics and expectations of the interlocutor" (p. 275). Santos (1988) refers to irritation as the "bother" factor. Gynan (1985) and Khalil (1985) assert that it is the affective response to error, "native speakers' (NS) emotional reactions to deviant utterances" (Khalil, 1985, p. 336). It is the thing that diverts reader attention "from the message to the code" (Johansson, 1975, as cited in Albrechtsen et al, 1980, p. 366).

Researchers generally speak of irritation in terms of degrees (Piazza, 1980) as in the continuation of the widely used definition by Ludwig (1982): "For the listener or reader, the irritation continuum ranges from unconcerned, undistracted awareness of a communicative trait to a conscious, constant preoccupation with form, to the point that the message is totally obscured or lost" (p. 275). For this reason, it may be best to think of it as a sort of friction, or slowing down of the mental faculties as the mind tries to process a deviation from the norm.

However, as to what, exactly, irritates someone is a complex question. James (1998) provides us with at least four different ways: grammaticality, acceptability, incorrectness, and strangeness. First, irritation may stem from ungrammaticality, such as when someone utters, "I am live in Utah." This criterion works fine for clear cut instances of deviant grammar, moments

when there is no possible way an utterance could pass as an example of standard English (as in the example given). Yet the further one gets from these definite cases, the more problematic the construct of grammaticality becomes as a viable method for determining irritation.

The notion of grammaticality, as well as that of error in general, relies heavily on the ideal NS, one who speaks the "code" perfectly (e.g. Chomsky, 1965; Duskova, 1969; Chastain, 1981; James, 1972, Corder, 1971). Yet, because no such speaker actually exists in real life, all grammaticality judgments are subject to the discrete code of those assessing the work (Hultfors, 1986). This definitional problem is aptly pointed out by Quirk and Svartvik (1966) when they state "'grammaticalness' involves two prime but interrelated difficulties—establishing what it is and determining native reaction in respect of it, the interrelation entering through the obvious fact that the second is dependent on the first, the linguist's categorical problem" (p. 1).

For instance, in the data for this present study, a common issue for NESs was the construct *help (someone) to*, as in the phrase *I helped my mother to clean the kitchen*. Some people were irritated by it and felt *to* should be omitted, while others hardly noticed. In this respect, the grammaticality of the construct in question is subject to divided opinions. Those who are irritated by it would likely say that it is grammatically incorrect. Yet, to those who were not, it may appear completely grammatical. It is these borderline cases that lead James (1998) to ultimately conclude grammaticality alone is a poor criterion for error because it fails to account for the community applying the standard. In other words, what is "grammatical" to one person may not be grammatical to another.

Yet, the grammaticality of an utterance is not always the same thing as its acceptability. For instance, a universal tag question is perfectly grammatical to speakers of Indian English (as in *She doesn't have enough money, isn't it?*) even if most speakers of American English believe

it is not. Similarly, the phrase *I have to go to the hospital* is standard in American English, while *I have to go to hospital* is more common in British English. An interlocutor who is familiar with the usage of these dialects might consider each one grammatical, even while still finding them to be unacceptable depending on the context in which they are used. As defined by Ludwig (1982), acceptability is "the degree to which a given L2 error violates language norms" (p. 277), or as Khalil puts it "the seriousness of errors" (p. 336). These "violations" and this "seriousness" can be assumed to be the direct result of utterances that deviate outside the standard usage of the TL community.

A third cause of irritation is incorrectness. Irritation of this type is dictated by textbooks and dictionaries, tokens that are only erroneous because a powerful individual said they are. A sentence ending with a preposition or using a split infinitive may seem perfectly grammatical and acceptable to society at large, yet to some, these "errors" still irritate. It is important to note that incorrectness is connected to notions of acceptability. To those who fully embrace the rules, a split infinitive error *is* unacceptable, and could cause significant amounts of irritation. However, to those who do not, the incorrectness of the utterance remains nothing more than an oddity. These individuals may know the rule. They might even have been required to follow it in their educational pursuits. Yet, if the rule does not make sense to them because no one in their immediate social environment seems to worry about it, they are left with only minimal irritation, if any at all. For instance, they may notice a violation of the rule and remember having learned something about it, but because they do not understand why it needs to be applied, they move on.

Finally, an error may irritate because it is strange or different from how a standard NES speaker would say it—i.e. using *white and black* instead of *black and white* or *she or he* instead of *he or she*. Strangeness, as with the other causes of irritation, is ultimately the result of

frequency, or how often a NES uses a word or a phrase in specific contexts in comparison to NNS L2 usage (James, 1974; Johansson, 1978; Hultfors, 1986).

Burt argues "in natural conversations, speakers constantly generate new sentences by applying the rules of the language they are speaking. Thus, even if a student has never heard a particular sentence, he can generate it if he has internalized (learned) the relevant grammar" (1975, p. 58). This generative ability implies that the more familiar a speaker or writer is with the standard forms of the target language, the better he or she will be in constructing new utterances that communicate. Obviously, even native speakers make frequent errors in their language. The difference is that the NES interlocutor's familiarity with the system can compensate more readily for gaps commonly made by native speakers because the errors are also familiar. Yet, when an error rarely, if ever, occurs among NES, it becomes more difficult for the interlocutor to interpret what is said, thereby resulting in higher levels of irritation.

**How Error Has Been Described**

This native/nonnative distinction is a prevalent finding of error gravity research (e.g. Vann et. al 1984; Janopoulos, 1992). Errors such as spelling, comma splices, and pronoun agreement were generally tolerated by native speakers because they are commonly made NES errors, while word order, it-deletion, tense, and relative clause errors were considered more irritating and less acceptable because they are made less frequently by NES, and thus considered "non-native" (Santos, 1988; Vann et. al, 1984; Janopoulos, 1992). Hyland and Anan (2006) also found word order and tense to be serious concerns for NES and add agreement and word form to the list. They posit that each of these error types vex NES raters because they impact the intelligibility (comprehensibility) of the sentence—which they measured through the occurrence

of NES statements "mentioning ambiguity, flow hindrance, confusion, fluency, etc." (p. 512)—and deviate most from rules of English grammar and syntax, making them "nonnative."

In some studies, attempts are made to categorize "native" and "nonnative" errors using a distinction between "global" and "local" errors (e.g. Tomiyana, 1980; Burt & Kiparsky, 1972, Burt, 1975; Albrechtsen et al, 1980). In error gravity terms, global errors "affect overall sentence organization" (Burt, 1975, p. 56) and "cause the listener or reader to misinterpret the speaker or writer's message" (p. 57). In contrast, a local error is one that affects only "single elements (constituents) in a sentence" (p. 57). Global errors are generally thought to be more serious than local ones because their effect on communication is greater (Burt & Kiparsky, 1972; Delisle, 1982). As Santos (1988) notes, "professors are willing to look beyond the deficiencies of language to the content in the writing of NNS students" (p. 84). Yet, when errors significantly interfere with meaning (global), they irritate and confuse NES more than local ones.

Burt (1975), in her study of 300 selected sentences obtained by Peace Corps volunteers serving in countries across the world, determined the most systematic global errors include

> wrong word order, missing, wrong, or misplaced sentence connectors, missing cues to signal obligatory exceptions to pervasive syntactic rules, and overgeneralizing pervasive syntactic rules to exceptions (in transformational terms, not observing selectional restrictions on certain lexical items. (p. 56-57)

Admittedly, her classification leaves substantial room for interpretation. For instance, what would constitute a "pervasive" syntactic rule and what are the "certain" lexical items? For this reason, it is useful to consider also the list of global errors contained in *Writing Clearly*, a student editing guide written by Janet Lane and Ellen Lange: verb tense, verb form, modal, conditional, sentence structure, word order, connectors, passive voice, and unclear meaning.

In this global/local error dichotomy, lexical errors, which are typically idiosyncratic to discrete contexts (i.e. *borrow* is a perfectly good word, even if the sentence "*He borrowed me the book*" is erroneous), are deemed less serious than grammatical ones. However, this claim runs counter to the findings of several gravity studies (e.g. Porte, 2008; Roberts & Cimasko, 2008; Johansson, 1978; Khalil, 1985), necessitating two important qualifications.

First, the conclusion that lexical errors are a definite concern for NES subsumes a certain way of describing error. It implies that lexical issues can be combined into one main category and that this category can be compared to other error types, such as verb tense and determiners (Ferris, 1999).Yet, lexical errors are tied to individual context to a much greater extent than others. They are far less rule-governed (Ferris, 1999), which makes them more difficult both to categorize and to address pedagogically—and suggesting, as do several researchers (e.g. James, 1998; Johansson, 1978; Hultfors, 1986), that gravity is also a function of the generalizability of the error to the norms and standards of the target language.

Second, Khalil (1985) asserts that semantically deviant utterances—which he defines as word choice and collocates—make sentences less intelligible and less comprehensible than grammatically deviant utterances (word order, concord, verb, and pronoun retention). Santos (1988) agrees, arguing that these errors are considered more serious than grammatical ones because they impinge content, which she discovered is rated more severely than grammar and syntax by university professors. However, because these results were obtained using isolated sentences, several researchers, such as Ludwig (1982) and Rifkin and Roberts (1995) advance the proposition that lexical errors are only irritating in these studies because word issues are more likely to stand out in isolated sentences than in paragraphs or essays. In other words, the issue may have more to do with protocol than it does gravity.

**Criticisms of Error Gravity Protocols**

*Problems in the Protocols*

Common criticisms of typical error analysis studies include the following. First, they tend to rely on isolated sentences to represent authentic discourse (e.g. Albrechtsen et al, 1980; Rifkin & Roberts, 1995; Chastain, 1980; Ellis, 2008). The concern raised by this practice is, as Chastain (1980) argues, "an isolated sentence often may not supply the sufficient information" (p. 212). Ellis (2008) asserts, "error evaluation is influenced by the context in which the errors occurred. Thus, the same error may be evaluated very differently depending on who made it and when, where, and how it was made" (p.60). For example, in the sentence *Yesterday, my sister will tell me about her new job*, the use of the simple future tense may be deemed erroneous. Yet, if the sentences surrounding it were also in the future tense, the use of *yesterday* would likely be the erroneous part, not the verb tense. In the absence of adequate context, it is difficult to say what an error actually means to NES response as, ostensibly, intelligibility decreases as context decreases (Albrechtsen et al, 1980). For this reason, Chastain (1980) concludes, "only by establishing a satisfactory universe of discourse can an investigator determine for sure whether or not comprehension is possible" (p. 212).

Second, attendant with using isolated sentences is the practice of including only one or two errors per utterance. Not only do researchers in error gravity studies frequently limit the context of the tokens, but they also predetermine which errors will be considered. Santos (1988) notes this approach does not "allow the NS judges to decide for themselves which errors are most glaring" and that "selectively inserted errors give equal weight to each error type by representing them only once each, an unrealistic condition that ignores the frequent recurrence of certain error types and the relatively infrequent occurrence of others" (p. 74). In an actual rating

context, raters must determine not just the gravity of certain errors, but also the gravity of certain errors as they relate to others in the text. For instance, a preposition or determiner error may be insignificant in the presence of relative clause or lexical errors, yet in their absence, using *the* when context requires an *a* could become more serious. Studies involving both limited discourse and pre-selected errors include Burt (1975), Janopoulos (1992), Tomiyana (1980), McCretton and Rider (1993), and Vann, Meyer, and Lorenz (1984).

Not all researchers agree with this assessment, however. Nor do they want to include additional context for practical reasons. Beason (2001) posits that using a naturalistic design "requiring subjects to locate the errors for themselves could provide useful results" (p. 38), but still chooses to boldface errors, in part, out of concern that respondents may feel uncomfortable or threatened if they are asked to identify errors, perhaps believing that it was their own language proficiency being tested and not the students; Even still, in keeping with the concerns of Santos (1988) and Rifkin and Roberts (1995), essay-level discourse is employed in this current study with no errors consciously pre-selected.

Third, these studies frequently substitute manipulated texts for authentic ones (Ellis, 2008; Khalil, 1985; Rifkin & Roberts, 1995). Khalil (1985), for instance, laments the many studies in which "linguistic context has been ignored" (p.336), while at the same time, ironically, committing this same error by revising sentences to include only one error per utterance (Mahoney, 2011). Chastain (1980) compiled a list of common errors seen by Spanish instructors at the University of Virginia and wrote his own list of 35 sentences, each with one error. Similarly, Tomiyana (1980) "mutilated" two 200-word paragraphs by inserting a variety of different "non-native" errors. This sort of manipulation results in the formation of an R-text, described by Rifkin and Roberts (1995) as the practice of using "a student sample to create

'simulated learner discourse'" (p. 516). In an R-text, NESs both react to and evaluate the text. In other words, they both create the "NNS" text and rate it. Of the 28 studies reviewed by Rifkin and Roberts (1995), half included this type of control in their protocols.

The issue created by R-texts is that they may cause raters to consider the errors lightly, "as a mere slip of the tongue (or pen) that does not reflect the author's true ability" (Mahoney, 2011, p. 110), may misrepresent the author's intended meaning, may result in sentences that are entirely bereft of any link to the learner, or may lead to findings that "cannot be necessarily applied to actual communicative events" (p. 110). Santos (1988) while acknowledging the benefits of using "artificially prepared" texts in terms of the control it accords to researchers, also criticizes the practice because "they also sacrifice the natural quality of unaltered connected discourse" (p. 74). Rifkin and Roberts (1995) assert that "at this essential level of research, namely selection of stimuli, we are faced with a basic challenge: if we are concerned with NS reaction to NNS error, then we must design studies that come as close as possible to using authentic language" (p. 517). In brief, the main concern posed by researchers is if we intend to study NNS written errors, we should use authentic, full-length, NNS texts.

*Problems in the Construct*

Fourth, error gravity study typically limits error evaluation to intelligibility, comprehensibility, irritation, and acceptability (Khalil, 1985). However, the inherent flaw in this approach is that these concepts are impossible to separate entirely (e.g. Ludwig, 1982; Johansson, 1975; Santos, 1988). Roberts and Rifkin (1995) point out that one critical omission in error gravity research has been a clearly defined concept of irritation. In effect, researchers are not sure where the separation between comprehensibility and irritation or acceptability and irritation lies. Ludwig (1982), for instance, concludes, while it is possible for a textual feature to

be understood yet still irritating, the two are "inextricably linked," arguing, "in general, higher comprehensibility implies lower irritation" (p. 275). Piazza (1980) makes a similar claim when she asserts "the more comprehensible an error type, the less irritating it was" (p. 424).

Additionally, the dividing line between irritation and acceptability is also elusive. Gynan (1995) posits that an error stems from an attitude, or "a learned predisposition to respond in a consistently favorable or unfavorable way with regard to an object, entity, person, or state (p. 161). He asserts that "attitude toward language may be measured by determining whether the listener feels good or bad about the language or whether the listener thinks the language is good or bad" (p.161), which indicates that defining an error likely includes both the evaluative and affective responses coexisting simultaneously. Interlocutors not only notice error because of the emotional response it creates, but they also make implicit judgments about it in the process. They do not just feel bad—they also decide whether feeling bad is acceptable to them. And, to the extent that these responses co-occur, irritation and acceptability defy complete separation.

In sum, while one may discuss intelligibility, comprehensibility, acceptability, and irritation as discrete constructs, the gap between them is often blurred in terms of data collection, as irritation will always co-occur to some extent in the intelligibility, comprehensibility and acceptability judgments individuals make. Indeed, it appears more a matter of how much irritation is generated by the token than it does whether or not the two concepts are mutually exclusive, and when researchers espouse to measure these concepts separately, they essentially make a false claim. Therefore, the construct used in this current study is irritation, as defined broadly by James (1998), with the other constructs assumed to be implicit.

Albrechtsen et al (1980) assert that "one should not expect to establish a hierarchy of errors with respect to irritation: All errors are equally irritating, provided they are in fact errors,

i.e., violations of a target language norm" (p. 395). Yet, as has been argued here, the notion of a universal "target language norm" is problematic. Norms change alongside the communities that create them, making it impossible to state conclusively whether a particular utterance will still be erroneous when detached from its native community (Hultfors, 1986). Subsequently, an argument that assumes a definitive accounting of these norms is achievable is inherently flawed.

*Problems in the Scope*

Finally, these studies often fail to account for the role of frequency in error gravity. Error frequency study is typically regarded as nothing more than a descriptive tool existing for three main purposes. The first is to "make explicit what would otherwise would be tacit and on the level of intuition" (p. 96). In other words, it allows us the opportunity to check our assumptions of common L2 errors against tangible data. The second is to account for the quantity of certain errors, helping us to see which errors are most endemic. And the third is to enable us to categorize L2 errors, which facilitates discussion of them. Yet, the impact of error frequency on error gravity is an issue that is far less discussed. Although an error that occurs only once may be irritating to a NS, when paired with several similar errors, the level of irritation increases. In this respect, an error occurring multiple times in a text is potentially more serious than one occurring only once or twice (James, 1998; 1974).

Error frequency also contributes to irritation on the macro level through the density of error, or "how many different errors occur per unit of text" (James, 1998, p. 211). This idea differs from simply determining quantities of errors by type or production frequencies because it includes the cumulative effect of all errors on irritation, intelligibility, comprehensibility, and acceptability. A likely process is as follows: when the same error is repeated, the reader eventually learns to interpret it by making adjustments to his or her reading. If the student always

omits articles, the reader will start to fill in the gaps. As Chastain (1980) notes, "given the amount of redundancy in the linguistic system, native speakers can obviously understand much nonnative speech, perhaps even a majority, even though the communications are not identical to those which they themselves would utilize in the same situation" (p. 210). As intelligibility nuclei become more and more ingrained, it is possible for listeners to fill in gaps in the code—much as they do in condensed speech—even when presented with abnormal usage. Conversely, in the presence of multiple error types occurring in tandem, this process is postponed, increasing NES irritation levels, and presumably, decreasing the grades NES raters assign.

**Toward a More Inclusive Definition of Error**

James (1998, p. 205) argues "the main reason for evaluation…is to get our priorities straight. We do not seek to hone the analytical scalpel so as to lay bare the tiniest error, but the opposite: to prevent obsession with trivial errors and give priority to the ones that really matter." Error evaluation does not seek to devalue L2 learners. It seeks to "assign relative values to errors" (p. 205) and thereby empower L2 learners to take greater control of their L2 production. If students and instructors know where to focus their attention, they will be better suited to reduce stigmas attached to their L2 language use and gain acceptance among their NES peers.

In an effort to help get the "priorities straight" (p. 205), this present study evaluates error gravity and frequency in two ways. The first is as discrete constructs, with a data set produced for each. This move is to determine (1) which errors are most serious to NES university students and (2) which errors are most frequent in NCS L2 writing.

A few error gravity studies have attended to certain language groups, such as Japanese (e.g. Hyland & Anan, 2006; Mahoney, 2012), Arabic (Khalil, 1985), Malaysian (McCretton & Rider, 1993), and Spanish (Porte, 1999), but there are, to my knowledge, currently no studies

specifically addressing the seriousness of errors produced by native-Chinese speakers (NCS). This is a critical omission in the literature given the ever-increasing Chinese-speaking student enrollments at US universities. According to the Institute of International Education, 194,029 Chinese and 23,250 Taiwanese enrolled at US universities in the 2011/2012 academic year. For the Chinese, this represented a 23.1% increase from the previous year. And, as the largest percentage of international students, NCS are certainly deserving of our time and attention.

Furthermore, little research has been done on the frequency of errors that NCS make. Liu (2011) collected a random sampling of 30 English majors, 40 non-English majors in Grade Two, and 35 non-English majors in Grade One at Beihai College of Beihang University in China. Verb errors were by far the most prevalent error type, accounting for 34.4%, 40.2%, and 46.9% of the total errors respectively, a result Liu (2011) suggests stems from the relative unimportance placed on word forms in Mandarin Chinese.

An additional finding by Liu (2011) is that preposition errors are also a large issue for NCS. Mandarin Chinese almost shares the same general word classifications as English (e.g. nouns, pronouns, verbs, prepositions, and so forth), but their application varies. In the case of prepositions, the Chinese use them less often than the English, and subsequently, "there are not so many strict rules" (p. 1063) in Mandarin Chinese. For instance, one character (在) can translate into *in, at,* or *on* depending on the sentence in which it is used. The implication is that Chinese frequently omit or misuse prepositions when writing in English (about 5.3% of the English major errors and 8% of the non-English majors).

A portion of Company's (2012) research also focused on NCS. Using the error tally sheets from the DWCF course at the ELC, she determined the five most common errors NCS make are word choice, spelling, singular/plural, word form, and determiner. However, because

these categories make a finer error distinction than that found in Liu (2011), it is ultimately impossible to fully compare the two. Liu (2011) combines all noun-related issues under the category "nouns," all verb-related issues as "verbs" and so forth, obscuring individual error types. For example, the category "nouns" might include any number of singular/plural, word choice, or count/non-count errors. As such, the additional frequency data provided in this present study would be quite useful to our awareness of prevalent NCS L2 errors.

The second way in which error gravity and error frequency is evaluated in this study is as a combined construct, in which absolute error gravity is assumed to be the composite of error gravity and prevalence—as traditionally defined. This step allows for a look at (3) the errors that are both highly serious and highly frequent, and follows a similar approach to the one used in *Writing Clearly*, by Janet Lane and Ellen Lange. An error may be serious and infrequent, less serious and frequent, serious and frequent, or less serious and infrequent. And for the reasons previously laid out, it is only after we have added the gravity and frequency data together that we will be able to determine what are, truly, the most serious errors.

In sum, this study will contribute to the literature in the following ways: the new gravity scale will offer an expanded view of error gravity that includes intelligibility, comprehensibility, acceptability, and irritation as influenced by error frequency. The frequency scale will provide information on the most endemic L2 English errors NCS make, and the third will provide salient focal points for grammatical and syntactic errors based on both grievousness and prevalence.

**CHAPTER 3: METHODOLOGY**

This chapter sets out the methodology used to conduct this study and to generate the necessary data to answer these questions: (1) what are the most grievous errors?, (2) what are the most frequent errors?, and (3) what are the most grievous and frequent errors made by native-Chinese speakers (NCS)? It begins by talking about the materials used—a set of 18 timed, 30-minute essays written by high intermediate to advanced native-Chinese ELLs as part of their coursework at Brigham Young University's English Language Center (ELC). It then discusses the two discrete groups of participants who identified and tagged the errors. Finally, it concludes with a description of the procedure used to identify, verify, tag, and classify the errors by the level of irritation they produced.

**Materials**

This study involves a set of 18 timed, 30-minute essays written by high intermediate to advanced native-Chinese English language learners (see Appendix D for the essays used) as part of their coursework at Brigham Young University's English Language Center (ELC). These students are Mandarin-speaking Chinese and Taiwanese. At the time the essays were written, their ages likely ranged from 18-25, the typical demographic of students at this school. However, because of privacy issues enacted at the time the essay test was administered, it was impossible to determine the writers' exact age and gender. As will be explained later, these essays were first collected, and then narrowed down from 169 essays to 18 using a series of evaluative filters

The ELC, though not ostensibly dedicated to preparing students for the TOEFL (Test of English as a Foreign Language) exam, requires students in its academic track to schedule a time in the computer lab five times each semester to be evaluated on their ability to write essays similar to those on the TOEFL—one integrated and one independent—with testing conditions

that largely mimic those of the actual exam. Integrated essays require students to listen to a short audio clip and read a short passage. The task is then to compare and contrast the two in their essay. In contrast, an independent essay entails students responding to one of various open-ended prompts on a subject of general interest (e.g. government, leadership, and technology). Successful completion of the essay required them to develop and support an opinion on the prompt. In this study, only independent essays are used.

Once gathered, 169 essays were methodically pared down to 18 through a series of evaluative filters. This decision to limit the samples size came as the result of several logistical concerns. First, to obtain the target 10 ratings per essay, as per Santos (1988), each of the raters would need to have highlighted at least 35 texts. Second, because each essay took around twenty minutes to complete, a total of 560 hours of labor would have been required. However, rather than simply taking a random sampling of the texts, I felt it would be more useful in terms of the data collection to focus my efforts on what I believed to be the most advanced texts, and which I defined as "university-level." This choice afforded me considerable control over both the global content (i.e. rhetoric, organization, complexity), minimizing the effect of these variables on the data, and several anticipated rater concerns, such as the effect of topic on the participants.

First, the essays were sorted by general writing ability. In reading through the essays, it quickly became apparent that not all of the samples were at a level appropriate for first-year university work, so each essay was evaluated for overall rhetorical content (i.e. thesis statements, topic sentences, supporting details, conclusions), word count (300+), and grammatical and lexical complexity (i.e. sentence variety, clauses, word choice). It should be observed, however, that while I do not believe that the criteria used for at least two of the categories, rhetorical content and word count, were so far removed from mainstream writing pedagogy that they

invalidate the results—it is, after all, plausible to expect a university-level writing instructor to mark essays that include a thesis statement, topic sentences, supporting details or that are of a certain length higher than those that do not, and thus, that by possessing these attributes the perceived quality of the essays increases—the fact that I am the sole creator and sole applicant of these standards should be taken into consideration. I did my best to use criteria that I felt would be commonplace among writing instructors, but the results could still be somewhat subjective.

Even still, as this research was focused on lexical, grammatical, and syntactic errors—not rhetorical content or word count—it was essential that these latter variables be controlled for. NES use thesis statements, organization, and topic sentences to focus their reading. The absence of such would have likely interfered with the interlocutor's understanding (intelligibility) and his or her perceived difficulty in understanding (comprehensibility), resulting in variable irritation levels. If they had had no idea what the passage was about, they may have either (1) overlooked features that would have been irritating in the presence of adequate context or (2) been excessively irritated by the fact that they did not know what the writer was attempting to say. In both cases, the end result would have been the same: data influenced by non-lexical, non-grammatical, and non-syntactic errors to a greater extent than after the evaluative filters.

Finally, on grammatical and lexical complexity, Wolfe-Quintero, Inagaki, and Kim (1998) present a carefully nuanced view that could have helped eliminate some of the subjectivity in this study. These complexities are manifest primarily through variation and sophistication within texts and are typically measured via ratios of features to units—i.e. how many past tense endings or academic words occur per T-unit (independent plus dependent clauses), sentence, or clause, and how varied the grammar or lexis is overall. Yet, while it would have been ideal to include such measures as part of the present protocol, the reality is that

following their recommendations could easily be the subject of multiple studies, as indeed they are in Wolfe-Quintero, Inagaki, and Kim (1998). Their chapter on grammatical complexity alone cites more than 10 unique studies, each assessing the complexities of various grammatical features ranging from adverbial clauses to prepositional phrases. Simply put, there just was not enough time to conduct more than a preliminary evaluation of the texts used in this study.

The second filter was the topic. Essays dealing with personal issues, such as friendship will likely rely on different vocabulary, different grammar, and different syntax than those that deal with business, science, or technology (Hinkel, 2009). For example, topics treating personal issues are best suited for the humanities or social sciences, and subsequently, one might expect to see fewer instances of the passive voice in these essays (largely because it is frowned upon in these communities) than were someone to write on a topic more typical to, say, the sciences, such as photosynthesis or the life cycle of a star. In this respect, we might apply an argument made by Duskova (1969) about the unpredictability of content when writers are free to express themselves in any way they see fit to the unreliability of the content when diverse topics are present: "while some grammatical points (such as articles, the past tense, the plural) were bound to occur in all papers, others (adverbs, relative pronouns, the future tense) appeared only in some" (p. 15). Thus, to maintain consistency in the specific vocabulary, grammar, and syntax used in the essays, the topic was limited to just business, leadership, and technology.

The third filter considered the affective response certain topics might have on the participants. BYU is a religious institution owned and operated by the Church of Jesus Christ of Latter-day Saints (LDS), so it was thought that topics dealing with religious or controversial subjects would unduly bias raters to approach the texts from a religious standpoint rather than an academic one (i.e. rate nicer than they normally would). Several studies in both social

psychology and second language acquisition conclude that bias is determined, to some extent, on the perceived role of the assessor towards the stereotyped population (e.g. Barden, Maddux, Petty & Brewer, 2004; Hosoda, 2006). This filter was therefore deemed necessary to prevent such bias from occurring. After applying these filters to the essays, it was possible to narrow the set to those that could represent at least entry-level work at a university to the NES participants.

Each selected essay was corrected minimally for spelling and punctuation. Roberts and Cimasko (2008) argue these error types "have not been shown to substantially affect native reader evaluation relative to other error types" (p. 131), and as evidence, they cite two studies, Vann, Meyer, and Lorenz (1984) and Sheorey (1986). Vann et. al (1984) includes two types of spelling errors: those based in dialect differences (i.e. color vs. colour) and those containing deletions (i.e. colr) and substitutions (i.e. calor). Their one punctuation error is a comma splice. After totaling the data, they determine, as Roberts and Cimasko (2008) claim, that neither spelling nor punctuation is a serious concern to NES when compared to other error types, such as word order and it-deletion. Indeed, in my personal observations of the texts involved here, so many other errors attracted my attention that the punctuation errors were hardly noticed. However, whether this lack of attention justifies editing out punctuation errors entirely, as Roberts and Cimasko (2008) claim it does, is debatable and begs the question, if these errors are not a big issue for raters, why not just leave them in? Unfortunately, I failed to realize the fallacy of this assumption until after I had edited them out of my own set of texts, something which I now regret given the negative impact it had on my ability to collect data on several error types included in this study (i.e. capitalization, incomplete sentence, run on).

Additionally, these findings assume that both punctuation and spelling errors occurred at a low enough frequency to allow rater attention to shift to other error types. In my personal

observation of the texts here, punctuation errors did not seem to be a big issue. They could have been kept in with minimal to no impact on the final results. In contrast, spelling errors were widespread, demanding so much of my attention that I felt something needed to be done to prevent them from overwhelming the attention of the participants. Thus, if spelling errors are not likely to irritate NES (Roberts & Cimasko, 2008; Vann, Meyer, & Lorenz, 1984), this result seems contingent upon their frequency.

NES can auto-correct many spelling errors in a text because context will generally determine the most appropriate interpretation. Yet, after a critical mass of spelling errors has been reached, NES begin to make assumptions about the education and background of the writers. "They should know better" or "they are just being careless," they posit, which causes them to score the essays lower (Beason, 2001). Furthermore, the ability to use context to understand the meaning of the passage becomes more limited as they find themselves increasingly having to use misspelled words to interpret other misspelled words. As such, to prevent this negative effect from happening, the spelling errors were edited out. In essence, unlike Roberts and Cimasko (2008), who eliminate spelling errors on the grounds that they are not a significant problem for NES, my own decision to edit them derives from the opposite concern—that they would attract too much attention and skew the results.

Finally, although several researchers have criticized the use of timed written assessments on the grounds that timed, high-pressure testing situations fail to give students an adequate opportunity to demonstrate their proficiency (e.g. Silva, 1997) or that the scoring of timed assessments is intrinsically flawed (e.g. McNamara, 1996), they are included here for three main reasons: (1) while better forms of assessment may exist, timed writing assessment remains unparalleled in terms of its ease of administration and grading efficiency, and subsequently, its

use is quite common in high-stakes assessment, (2) as the focus of this study is rater irritation caused by authentic features of nonnative written texts, the use of timed writing is ideal in that it prevents outside factors, such as NES peers or online resources, from interfering with their performance, much as would be the case in a genuine high-stakes testing situation, and (3) because the NCS L2 samples were taken from actual formative assessments, other factors relevant to high-stakes testing (i.e. motivation, anxiety, testing environment) have already been controlled for. In other words, because the texts used in this study were constructed genuinely in a high-stakes testing situation, they are appropriate for use in this study. See Appendix D, Native-Chinese Speaker Essays, for each of the 18 essays used in this analysis.

**Participants**

Two distinct groups participated in this study. The first consisted of 48 native-English speakers between the ages of 19 and 30, and included undergraduate, graduate, and post-graduate students from a range of academic disciplines. Gender was fairly evenly divided, with 22 females and 26 males. Roughly a quarter (23%) had lived abroad for at least one year. However, most of these had not had extensive exposure to nonnative academic writing, and only 6% had had any previous TESOL experience. Because this study complements Company (2012), who uses instructors at BYU's ELC, some might argue the decision to use non-ELC participants is inappropriate. Yet, as Beason (2001) notes in her study on the NES errors that most negatively affect writer ethos, it is possible for instructors to occasionally favor errors that are not irritating to the general public, and if we rely solely on data collected in the linguistic accuracy course, we neglect the larger social context in which L2 learners will eventually find themselves. Indeed, "our effectiveness, perhaps our ethos, can be impeded if we stress matters that other professionals see as trivial—or if we trivialize points they deem consequential" (p. 34), and there

is little or no value in using a narrow group of instructors if they do not represent everyday social judgments. Only by extending data collection outside of the ESL classroom could the broad, social view of irritation sought in this study be obtained.

Yet, that is not to say that ELC instructors had no part in this research. Although the initial decisions regarding whether a token was erroneous came from the general public, ten trained editors at the ELC assisted with tagging the errors, providing a level of consistency with Company's (2012) tagging. Not only did the taggers use the same set of codes and same set of materials to learn the codes (see Appendix A for a list of the codes used in the linguistic accuracy course at the ELC and see Appendix B for examples of how each of these codes are applied), but given that both studies drew from the same source, it is also entirely plausible that several of them are common to both studies. The taggers included here are predominantly female (80%), 26-35 years-old, and have an average of two semesters of experience using the DWCF codes.

**Procedure**

**Stage 1.** The 18 essays were distributed at random to the NES until a minimum of 10 raters per text was obtained. The average number of essays highlighted by each participant was six, although that number includes a larger set of essays sent out before the second and third filters were applied. Following Ferris' (2009) assertion that native speakers can often tell, even if they cannot put a name to it, if something is "not quite right" about some texts, students were asked to highlight the linguistic features that seemed strange or awkward to them. No attempt was made to identify for raters what does or does not constitute an error. Additionally, the construct of irritation elicited was broadened to include all items that bothered them. In fact, participants were told that an error did not have to be an "error," in the official, textbook sense, for them to mark it. This move was important because, at this stage, I was more interested in

irritation than I was in separating it into its component comprehensibility and acceptability judgments. The goal was to cast the net as wide as possible and leave the sorting for later.

Previous studies have examined error gravity using isolated sentences containing one error per sentence, but this study considers gravity using full-length essays, an approach that provides for a more naturalistic rating context, and therefore, more authentic data (Khalil, 1985; Rifkin & Roberts, 1995; Roberts & Cimasko, 2008). Errors not only occur in L2 student writing, but also simultaneously with other errors. Having multiple errors within the same text allows for competition between error types and subsequently, results in variable rater attention. A preposition error may seem more serious if there are no verb errors, for instance. Yet, in the presence of verb errors, the preposition may go unnoticed. In this respect, certain errors may seem more or less serious depending on the written context of the essay in which they exist. For this reason, Roberts and Cimasko (2008) assert, "rather than asking native writers [sic] to evaluate L2 errors in terms of abstract concepts such as acceptability or comprehensibility, a naturalistic performance by the evaluator (i.e. marking and editing any error they perceive) holds the potential to produce a more realistic accounting of response to L2 writing" (p. 126).

After the data had been collected, I set to work compiling the highlighted portions onto a single document for each essay by transferring the highlighting from each copy onto a single master copy via a system of circles and tallies. If an error contained different words from another, it was deemed original and, therefore, circled on the master draft. Repeated errors— those for which two or more participants had highlighted the exact same words—were marked first with a circle, then with tally marks indicating the number of people who had reported them.

The irritating features were entered onto a spreadsheet and a count made of how many individuals had reported each original feature, which implicitly may also be considered a

measure of each features' acceptability. Santos (1988) would likely disagree with this theory—her study used a similar protocol to identify salient features only—but preliminary data suggests a connection between salience and its communal acceptability. Because group norms and standards are used to establish error—what we might call "error by common consent"—certain tokens will be clearly more erroneous than others. Therefore, if an error is marked by the majority it can be assumed to be one of the least acceptable tokens, while those that have only been highlighted by one or two participants is more acceptable to the general community (Duskova, 1969).

Next, due to the open nature of the highlighting (participants could highlight as many or as few words as they wanted) it was further necessary to remove any redundant features from the data set. Redundancies were defined by identifying the overlapping portions of the highlighting. Participants followed certain behavioral patterns in which one or two participants would highlight an entire sentence, some a short phrase or two within the highlighted sentence, and others just a word or two. And under the assumption that the most highlighted portions of each sentence or phrase were the erroneous parts of the sentence, all other potential errors related to them were deleted from the list. Once the highlighting and data compilation were completed, the remaining 1123 "errors" represented a socially constructed corpus, as determined by over 11,000 individual judgments.

**Stage 2.** The final step required a group of trained ESL raters at the ELC to tag the errors using a marking system created for the ELC's applied version of DWCF (see Appendix B for examples of each in use). The rationale for using these raters, as opposed to the NES participants in the first stage was three-fold: (1) asking the NES to tag the errors themselves would have required additional training and a meta-linguistic awareness that was simply too unrealistic in the

time allowed, (2) even if they had received training, asking them to put a name to what irritated them could have skewed the results toward more textbook errors (i.e. grammaticality, incorrectness) by making them feel they had to identify the error in order for it to be considered one, and (3) instructors at the ELC regularly tag student errors using the classification system, so including them in this study made sense because they were the best qualified for the task overall. Additionally, as previously noted, it allows for comparison to a study by Company (2012) on the most common NCS textual errors made by students enrolled in the linguistic accuracy class.

To ensure reliability in the tagging, each essay was tagged by two raters. If one or both raters failed to mark an error because they did not feel it was erroneous, or if they otherwise disagreed on an error because the codes were simply too ambiguous (i.e. one or more could be applied to the same error), a third rater was consulted. The codes used were as follows (see also Appendix A for the list used in the ELC course):

Table 1

*Error Categories and Codes Used in Teacher Marking and Analysis*

| Error Type | Code |
|---|---|
| Determiner | D |
| Subject-verb agreement | SV |
| Verb form | VF |
| Run-on sentence | SS ro |
| Incomplete sentence | SS inc |
| Verb tense | VT |
| Preposition | PP |
| Word form | WF |
| Word choice | WC |
| Singular/Plural | S/PL |
| Countable/Non-count | C/NC |
| Unclear meaning | ? |
| Awkward | AWK |
| Word order | WO |
| Omit | ~ |
| Insert | ^ |
| Capitalization | C |

It should be mentioned that these codes are slightly different from those used by Company (2012). In her study, she includes all of the above marks, with the exceptions of *unclear meaning (?), awkward (AWK), omit (~), and insert (^)*. She argues these errors "could encompass several things" (p. 17), which is true. An *awkward (AWK)* phrase could be the result of poor word choice, abnormal syntax, or even just momentary misreading. However, given the above discussion about what irritates NES and how this complicates our understanding of how error is perceived, I felt it would be informative to leave these codes in to see if there are noticeable patterns to them as well.

*Awkward (AWK)* and *unclear meaning (?)* are the most problematic of the new codes because they involve asking the reader to determine something that is often undeterminable— what is awkward or unclear about a phrase. It is not always possible to pinpoint the exact part of an utterance that causes communication to breakdown. An interlocutor may be able to say what bothers him or her, even if he or she does not know why, so no attempt was made to separate these error types into smaller parts in this present study. The data are reported here as is with no claim regarding what, exactly, they mean to error correction.

In contrast, *insert (^)* and *omit (~)* errors proved far easier to work with. Even though these categories also include "several things" (Company, 2012, p. 17), they are based on regular patterns of English syntax. Unlike *awkward (AWK)* and *unclear meaning (?)*, which deal primarily with content, *insert (^)* and *omit (~)* errors involve structure, something any NES could easily piece together. Therefore, in the data collection, omit and insert errors were subdivided by the part of speech of the word needing to be inserted or omitted, much as they were in Hultfors (1986). These error types can contribute much to our understanding of error gravity.

For the purposes of this research, *insert (^)* errors are defined not as errors of insertion, but as ones involving words that have been left out of the utterance, or missing words that are noticeably absent. This error type is labeled by what needs to be done, not by what has been done. For example, the erroneous sentence *She must appreciate* the *nature* involves an unnecessary *the*. This is an *omit (~)* error, meaning the writer can correct the sentence by omitting *the*. In contrast, the sentence *Only by being rational can * lead his company* contains an *insert (^)* error. The pronoun *he* should be added before the verb in order for the utterance to make sense.

To obtain data for the *insert (^)* errors data a two-part method was used. First, the existence of the error was determined through the tagging done by the ELC instructors. Second, the part of speech was identified. In many instances, the taggers provided the word or series of words that needed to be inserted, which left only the task of determining the part of speech to the researcher. However, in instances where the *insert (^)* error was tagged, but the missing word not identified, the researcher needed to determine both the missing word and the part of speech, based on what it could likely be.

**CHAPTER 4: ANALYSIS/RESULTS**

This chapter discusses the methods and results of analyzing the gravity, frequency, and gravity + frequency data. To answer the first two research questions—which errors are most serious and which are most frequent?—the data were separated into scales for both gravity and frequency. To answer the third—which errors are both serious and frequent?—the scales were added together.

**Gravity**

The data resulted in two fields, and because of these, it was necessary to devise an equation that synthesized them. The first field was the number of raters who highlighted each particular error. Because the number of raters for each essay was 10, there were ten different categories, one for each number of raters who could have noticed the error (see QI in Table 2). This field was closely tied to the second, which was the number of each discrete errors noticed by a specific number of raters (QE in Table 2). For example, determiners yielded the following data in Table 2.

**Table 2**

*Determiner Irritation Levels by Quantity*

| QI | 10 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 |
|----|----|----|----|----|----|----|----|----|----|----|
| QE | 1 | 0 | 3 | 7 | 7 | 10 | 11 | 7 | 11 | 10 |

Here, the top row (QI) is the number of participants who tagged a given error (a single occurrence, not the entire error type). The second row (QE) is the number of errors rated at that level, so for the data given above, only one determiner error was marked by all ten raters, while three determiner errors were marked by eight people, and so forth.

I multiplied the quantity of each error (QE) by the quantity of participants irritated by it (QI) for each discrete point on the irritation scale to obtain a sum irritation (SI) value for each error type and for each irritation level. This first step may be summarized by the equation

(QE*QI) = SI

The SI values were then added together for each error type, resulting in total sum irritation score (tSI).

SI1+SI2+SI3+SI4+SI5…+SI8+SI9+SI10 = tSI

Applying this equation to the determiner data in Table 2, we have

(1*10) + (0*9) + (3*8) + (7*7) + (7*6) + (10*5) + (11*4) + (7*3) + (11*2) + (10*1) = 272

Next, this sum (tSI) was averaged by the number of total errors made by type.

tSI / (QE1+QE2+QE3+QE4+QE5+QE6+QE7+QE8+QE9+QE10) = AVG

In the case of determiners, the tSI, 272, was divided by 67, which was the total number of determiner errors, to find the average amount of irritation caused. This number was 4, meaning that, of all determiner errors that occurred in the essays, an average of 4 raters out of 10 found them sufficiently irritating to warrant highlighting. This is in contrast to other error types, such as *count/non-count (C/NC)*—i.e. *transporting the* mails *for people*—that irritated an average of 6 out of 10 raters or *word choice (WC)*—i.e. *in my opinion, a good leader should be rational, and have a high* EQ—that saw only 3 out of 10 raters irritated. This process was followed for all remaining error types, and because the values from this equation ranged naturally from 3 to 6, the data were used to create the following 6-point scale in Table 3.

Table 3

*Error Gravity Scale with Undivided Insert and Omit Errors*

| 6 | 5 | 4 | 3 | 2 | 1 |
|---|---|---|---|---|---|
| C/NC | VT | AWK | ? | | |
| SV | WF | D | C | | |
| | | INC | PP | | |
| | | ^ | WC | | |
| | | ~ | | | |
| | | S/PL | | | |
| | | VF | | | |
| | | WO | | | |

Count/non-count (C/NC)—i.e. *minimize the* corruptions—and *subject-verb agreement (SV)*—i.e. *he is the one who* need—are among the most irritating errors a NCS could make. Both error types classify under Burt's (1975) description of global errors, and so it is not surprising to find them, along with *verb tense (VT)* and *word form (WF)*, in the top third of the gravity scale. Each indicates a failure to include "the cues needed to signal obligatory exceptions to pervasive syntactic rules" (Burt, 1975). In the case of *subject-verb agreement (SV)*, the pervasive rule for the simple present is to exclude any markers that indicate number. Obviously, for the third-person singular, though, one does, making it an exception to an otherwise standard rule. We see further evidence in support of Burt's (1975) global/local dichotomy if we divide the *insert (^)* error and *omit (~)* error categories into their subcategories as shown in Table 4.

Table 4

*Error Gravity Scale with Insert and Omit Errors Subdivided*

| 6 | 5 | 4 | 3 | 2 | 1 |
|---|---|---|---|---|---|
| C/NC | VT | AWK | ? | ~CONJ | |
| ^V | WF | D | C | | |
| ~V | | INC | ^ADV | | |
| SV | | ^D | ^CONJ | | |
| | | ^N | ^ADJ | | |
| | | ^PP | ~ADJ | | |
| | | ^PRO | ~N | | |
| | | ~ADV | PP | | |
| | | ~D | WC | | |
| | | ~PP | | | |
| | | ~PRO | | | |
| | | S/PL | | | |
| | | VF | | | |
| | | WO | | | |

Here, verb-related issues place in the top three irritation categories. Insert and omit errors involving structural words (i.e. determiners and prepositions) are all included in category 4, with the exception of conjunctions. Insert and omit errors involving content words (nouns, pronouns, adjective, and adverbs) are more divided, with *insert noun* (^N) a bigger concern than *omit noun* (~N).

**Frequency**

Error gravity was operationalized as the sum irritation caused by each error type divided by the total number of each error type. It differs from frequency—what was measured to answer the second research question—in that, for frequency, each error type was measured against the total number of errors (1123) and then divided by the total number of NCS writers (18), whereas, to determine gravity, error types were measured only against themselves using the formula given above. Perhaps an easier way to think about it is that frequency is the prevalence of certain error types compared to all errors made, while gravity is the prevalence of irritation by error type.

Frequency was calculated using the Quantity of Error (QE) sum for each error type and dividing it by the total number of errors, which in this case was 1123. This number was then divided by the total number of students (18) to obtain a measure of the frequency of the error type per person, similar to Company (2012). For example, determiners had a QE of 67, meaning of the 1123 tokens in this study, 67 were determiner errors. This number was then divided by the total number of errors (1123) to yield a frequency of 6%. Next, after converting this number into a decimal (and moving the decimal one place to the right to yield .6) the value was divided by 18. The end result was .33, meaning that each of the 18 writers produced an average of 3.3 determiner errors per essay.

*Word choice (WC)* comprised a significant majority of the errors, accounting for 321, or 28.5%, of the total tokens—a frequency of 1.6. This frequency was roughly double that of the next highest error type, *omit (~)* at.7. Hence, in the interest of visual clarity, *word choice (WC)* was removed from the chart below and the remaining data adjusted by subtracting 321 from 1123 (802) to reflect the change. *Omit (~)* registered in at a frequency of .98, followed by *singular/plural (S/PL)*—i.e. *controlled by different group of people*—at .82, *insert (^)* at .62, *awkward (AWK)* at .58 and *word form (WF)*—i.e.*to be success*—at .57. For these data, as well as those for the remaining error types, consult Figure 1 below.

Figure 1. Error Frequency
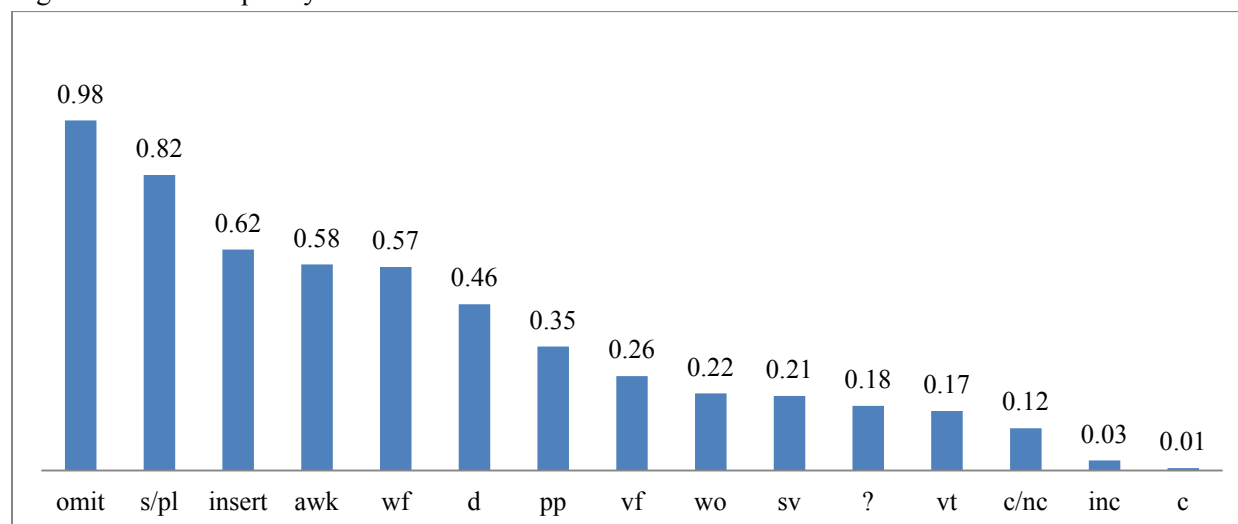


A significant number of the errors involved *awkward (AWK)*, *insert (^)*, and *omit (~)*, which, together, comprise nearly 40% of the remaining data. There are at least two preliminary conclusions one can make on this point. First, it shows that these errors are endemic in NCS L2 writing, thereby justifying their inclusion in the study. Second, even though they are endemic, the fact that they "could encompass several things" (Company, 2012, p. 17) suggests that they should be subdivided into smaller parts, if possible.

As noted, *insert (^)* and *omit (~)* errors are significantly easier to work with than *awkward (AWK)* because they relate to sentence structure, making it possible for native speakers to identify what needs to be included or omitted based on their familiarity with standard English syntax. Therefore, the *awkward (AWK)* errors were left alone, while the other two categories were subdivided according to the part of speech of the word needing to be inserted or omitted.

*Insert (^) Errors*

Omitted prepositions (i.e. *Only being rational can he lead his company*) were the most common type of *insert (^)* error, a .16 when included alongside all other error types in the study, followed by determiners (.11), pronouns (.10), conjunctions (.10), and verbs (.10). The remaining

types—nouns, adjectives, and adverbs—were either not a serious problem for NCS or not sufficiently irritating for NES to warrant much attention, as evidenced by their low frequencies (see Figure 2).

Figure 2. Insert Error Frequency
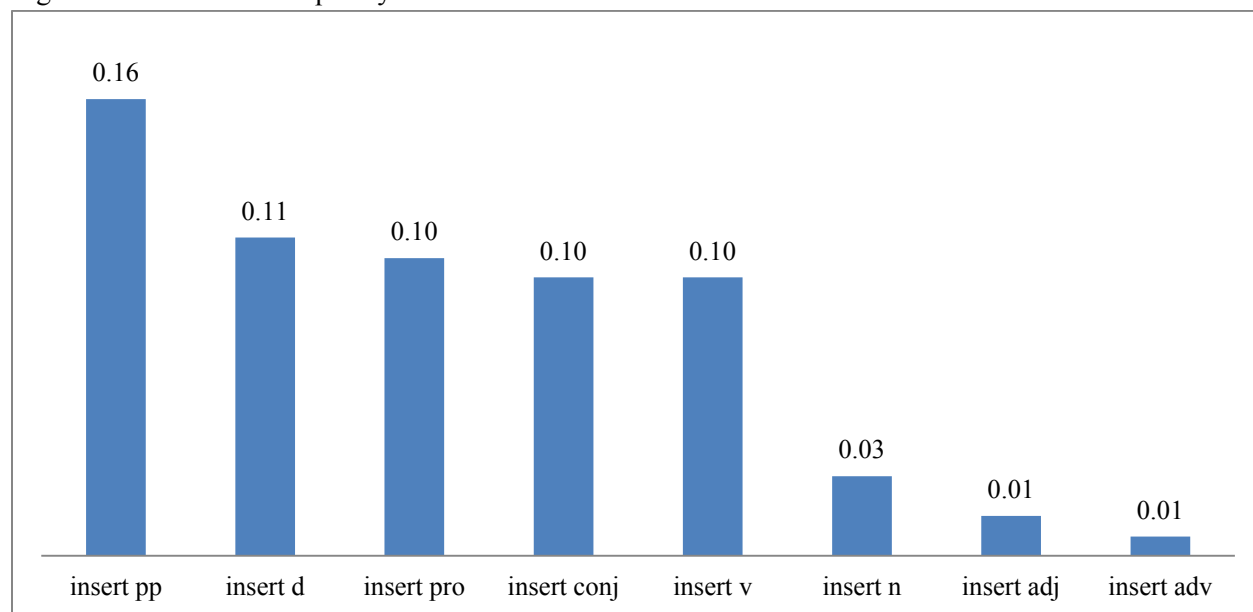


The data exist at two noticeable plateaus, which will be defined as three or more error types with frequencies within .05 of each other. The first includes *determiners (^D)*, *pronouns (^PRO)*, *conjunctions (^CONJ)*, and *verbs (^V)*, or structural and higher priority content words (higher priority because they need to be inserted before nouns *(^N)*, adjectives *(^ADJ)*, and adverbs *(^ADV)*, or lower priority content words). Additionally, when divided into smaller categories, none of the errors made it past the lower third of the data set. Each was relatively infrequent.

*Omit (~) Errors*

An *omit (~)* error involves words that were used in a sentence but need to be taken out because they are unnecessary. This error type was the second most frequent error type after word choice, with determiners (i.e. *She must appreciate* the *nature*) commanding a clear lead over the

remaining types. These data seem to imply that the NCS in this study are over-generalizing the omit rules of, most noticeably, determiners and prepositions. Mandarin lacks determiners, and as has been previously noted, the rules of preposition use are more relaxed than they are in English, so it is unsurprising that these, of all the other parts of speech, would give them the most trouble in both insert errors and omit errors. These and the remaining data are as follows in Figure 3.

Figure 3. Omit Error Frequency
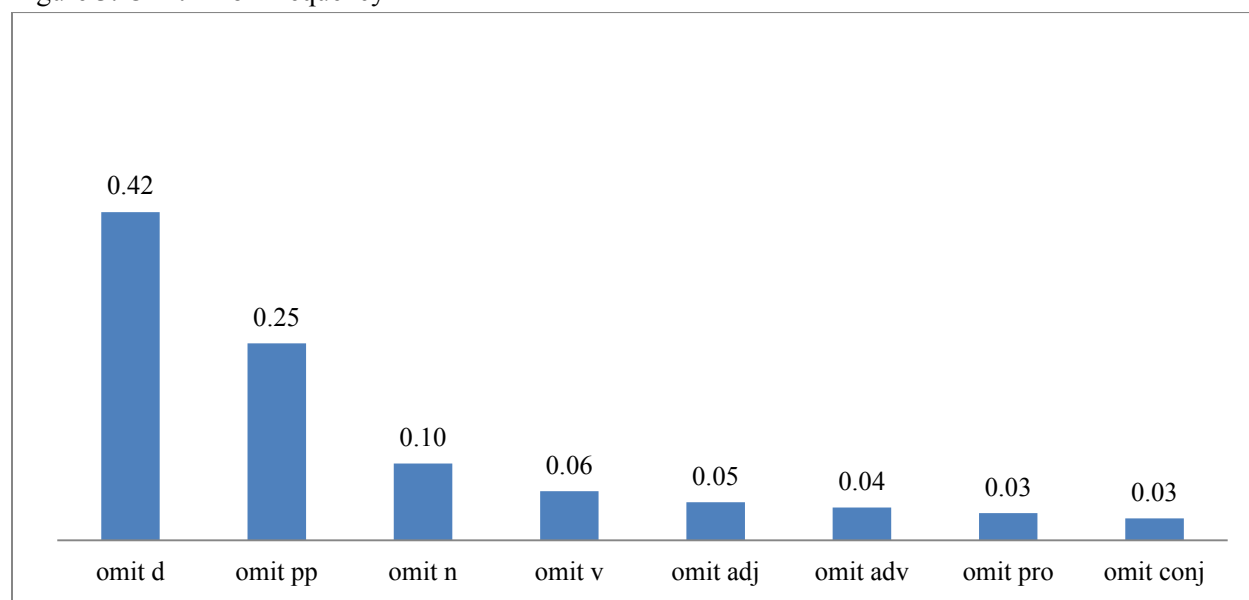


Unlike the insert errors, two of the omit error categories, *omit determiner (~D)* and *omit preposition (~PP)* placed above the lower third. However, the remaining errors were among the lowest in the data, signifying that most omit errors are also infrequent.

*Scale*

Following completion of the initial error frequency analysis, it was necessary to class the errors on a scale of 1 to 6. This step was included to facilitate comparison of the frequency and gravity data later on. Because I wanted to isolate each error type, as best as possible, to just one error, the subcategories for *insert (^)* and *omit (~)* errors were included alongside the overall data, resulting in Table 5.

Table 5

*Combined Error Frequency*

| TYPE | FRE | TYPE2 | FRE2 | TYPE3 | FRE3 |
|------|-----|-------|------|-------|------|
| WC | 1.59 | SV | 0.15 | ~V | 0.04 |
| S/PL | 0.59 | ? | 0.13 | ~ADJ | 0.03 |
| AWK | 0.41 | VT | 0.12 | ~ADV | 0.03 |
| WF | 0.41 | ^PP | 0.11 | ~PRO | 0.03 |
| D | 0.33 | C/NC | 0.08 | INC | 0.02 |
| ~D | 0.30 | ^D | 0.08 | ^N | 0.02 |
| PP | 0.25 | ^PRO | 0.07 | ~CONJ | 0.02 |
| VF | 0.19 | ^CONJ | 0.07 | ^ADJ | 0.01 |
| ~PP | 0.18 | ^V | 0.07 | C | 0.01 |
| WO | 0.15 | ~N | 0.07 | ^ADV | 0.01 |

*Word choice (WC)* errors were classed immediately as category 6 and removed from the remaining data set in an effort to prevent error types from clustering together. Because this error type was such an overwhelmingly large part of the findings, including it would have minimized any differences between the other data. Had I done so, *word choice (WC)* would have been classed far above the other types, meaning that the majority of the errors would have been category 1 or category 2. Additionally, for reasons explained later, the data for *word choice (WC)* is simply not amenable to comparison with the others in light of the wide range of possible *word choice (WC)* errors. With this assessment in mind, these errors were set aside.

The remaining data were divided by a class width of .14. This width was identified after subtracting the maximum frequency (after *word choice (WC)* had been omitted) from the minimum to obtain the range (.82). This range was then divided by . The final scale is found in Table 6 (please consult Table 5 for specific frequencies).

Table 6

*Error Frequency Scale*

| 6 | 5 | 4 | 3 | 2 | 1 |
|---|---|---|---|---|---|
| WC | WF | ~D | PP | VF | C/NC |
| S/PL | AWK | D | | ~PP | ^D |
| | | | | WO | ^PRO |
| | | | | SV | ^CONJ |
| | | | | ? | ^V |
| | | | | VT | ~N |
| | | | | ^PP | ~V |
| | | | | | ~ADJ |
| | | | | | ~ADV |
| | | | | | ~PRO |
| | | | | | INC |
| | | | | | ^N |
| | | | | | ~CONJ |
| | | | | | ^ADJ |
| | | | | | C |
| | | | | | ^ADV |

Although the data still cluster, especially in the category 1 errors, it made little sense to divide the data any further given the need for six final categories. The error gravity data discussed in the previous section resulted in six categories. To keep the two constructs at an equal weight, I decided to limit the frequency categories to six as well. Otherwise, I would have included an additional element to the research—the relative weights of gravity and frequency in the gravity + frequency paradigm—an idea beyond the scope of this present research. Subsequently, no additional attempt was made to separate the frequency data.

**Gravity + Frequency**

Finally, to determine which errors were both serious and frequent, and subsequently those that deserve the greatest attention in DWCF (the third research question), the 6-point frequency and gravity scales were simply added together (see Table 7)

Table 7

*Error Gravity and Error Frequency Combined*

| TYPE | FRE | GRAV | TOTAL | TYPE | FRE | GRAV | TOTAL |
|------|-----|------|-------|------|-----|------|-------|
| S/PL | 6 | 4 | 10 | ^PP | 2 | 4 | 6 |
| WF | 5 | 5 | 10 | ? | 2 | 3 | 5 |
| WC | 6 | 3 | 9 | ^D | 1 | 4 | 5 |
| AWK | 5 | 4 | 9 | ^PRO | 1 | 4 | 5 |
| D | 4 | 4 | 8 | ~PRO | 1 | 4 | 5 |
| ^D | 4 | 4 | 8 | INC | 1 | 4 | 5 |
| SV | 2 | 6 | 8 | ^N | 1 | 4 | 5 |
| VT | 2 | 5 | 7 | ^CNJ | 1 | 3 | 4 |
| C/NC | 1 | 6 | 7 | ~N | 1 | 3 | 4 |
| ^V | 1 | 6 | 7 | ~ADJ | 1 | 3 | 4 |
| ~V | 1 | 6 | 7 | ~ADV | 1 | 3 | 4 |
| PP | 3 | 3 | 6 | ^ADJ | 1 | 3 | 4 |
| VF | 2 | 4 | 6 | C | 1 | 3 | 4 |
| ~PP | 2 | 4 | 6 | ^ADV | 1 | 3 | 4 |
| WO | 2 | 4 | 6 | ~CNJ | 1 | 2 | 3 |

The final scale is as follows (Table 8).

Table 8

*Combined Gravity and Frequency Scale*

| 10 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 |
|----|---|---|---|---|---|---|---|---|---|
| S/PL | WC | D | VT | PP | ? | ^CONJ | ~CJ | | |
| WF | AWK | ~D | C/NC | VF | ^D | ~N | | | |
| | | SV | ^V | ~PP | ^PRO | ~ADJ | | | |
| | | | ~V | WO | ~PRO | ~ADV | | | |
| | | | | ^PP | INC | ^ADJ | | | |
| | | | | | ^N | C | | | |
| | | | | | | ^ADV | | | |

In summary, the most serious (gravity) errors included *count/non-count (C/NC)*, *insert verb (^V)*, *omit verb (~V)*, and *subject-verb agreement (SV)*. The most frequent error was *word choice (WC)*, followed by *singular/plural (S/PL)*, *awkward (AWK)*, and *word form (WF)*. Yet, when combined, *singular/plural (S/PL)*, *word form (WF)*, *word choice (WC)*, and *awkward*

*(AWK)* errors were the most critical. These error types are followed by *determiners (D), omit determiners (~D)*, and *subject-verb agreement (SV)*. The least serious errors (according to the gravity + frequency scheme) include *omit conjunction (~CONJ), insert conjunction (^CONJ), omit noun (~N)*, *omit adjective (~ADJ)*, *omit adverb (~ADV)*, *insert adjective (^ADJ)*, *capitalization (C)*, and *insert adverb (^ ADV)*.

**CHAPTER 5: DISCUSSION AND CONCLUSION**

This section discusses the results from the previous chapter and draws conclusions from the findings. Gravity is considered in terms of Burt and Kiparsky's (1972) global/local error distinction. Frequency is discussed in comparison to the new gravity data, as well as that of Company (2012) and Liu (2011). And because, to my knowledge, no other gravity + frequency studies for NCS exists, this data is presented in conjunction with its pedagogical implications.

**Gravity**

The final gravity rankings place *count/non-count (C/NC)*, verb-related issues (*subject-verb agreement (SV), verb tense (VT), insert verb (^V)*, and *omit verb (~V)* errors) and *word form (WF)* at a category 6 or 5, the most grievous errors as described on the scale discussed above. Except for conjunctions, insert and omit errors involving structural words (i.e. determiners and prepositions) are all included in category 4, making them slightly less grave than the category 5 and 6 errors. Also included in the category 4 errors are *awkward (AWK), determiner (D), incomplete sentence (INC), singular/plural (S/PL), verb form (VF), word order (WO)* and insert and omit errors involving certain content words (*insert noun (^N), insert pronoun (^PRO), omit adverb (~ADV),* and *omit pronoun (~PRO)*). The remaining errors ranked predominantly as category 3 errors, with *omit conjunction (~CONJ)* being the lone category 2 error. The implications of these data will be addressed by ranking and by word class in this section.

Burt and Kiparsky's (1972) global/local error dichotomy is strongly supported by the results. After separating the insert and omit errors by part of speech, *count/non-count (C/NC), insert verb (^V), omit verb (~V),* and *subject-verb agreement (SV)* were more frequently irritating to NES than any other error type. And although Hultfors (1986) argues the first type, c*ount/non-count (C/NC)* errors, is a relatively insignificant error, the fact that it is ranked serious in this

present study substantiates the need to consider the effect of protocol when considering error. Hultfors (1986) uses a survey in which discrete sentences have been revised to include only one token each, practices that have come under attack from more recent studies (i.e. Rifkin & Roberts, 1995; Ellis, 2008). In contrast, this current study neither isolates sentences, nor pre-determines errors, therefore offering a more authentic view of error evaluation.

Furthermore, the high degree of gravity seen with the *count/non-count (C/NC)* errors may be explained in terms of their generalizability to the rest of the text (James, 1998; Johansson, 1978). When an English speaker attaches number to a noun, it holds implications, not just for its intrinsic meaning (and subsequently, the irritation caused by using the word incorrectly), but also for referents found later in the text. For instance, if a writer refers to *rice* as *rices*, English grammar requires changes to both the pronouns used to refer to it and its subject-verb agreement. This type of error has the proclivity to cause other errors to occur, making it highly serious.

The next three category 6 errors, *omit verb (~V)*, *insert verb (^V)*, and *subject-verb agreement (SV)* are part of a cluster of verb-related issues that appear at the top of the scale. This result is unsurprising for two reasons: first, verbs play a key role in determining which subjects and objects can be used in each sentence. Omitting a verb or including one unnecessarily may confuse the reader, making him or her unsure of how the subjects, verbs, and objects link together. In other words, it seriously impedes communication between participants. Subsequently, this finding is, second, typical of other studies (i.e. McCretton & Rider, 1993; Roberts & Cimasko, 2008; Vann et al., 1984).

Collectively, verb-related issues span the top three error rankings (*verb tense (VT)* is a category 5 error and *verb form (VF)* is a category 4). This fact implies that even these errors may be considered within Burt and Kiparsky's (1972) global to local error distinction. The first

concern is the use of the verb in relation to the sentence's overall structure and meaning. In other words, the first questions the NES interlocutor appears to ask are whether the verb is missing (^V) and whether there are too many of them(~V). Once resolved, the concern then moves to tense, which primarily affects sentence meaning. If one were to use the simple past tense when the simple present is called for, it could cause the reader to be unsure of when the event happened. The structure may be sound, but the meaning is unclear. Finally, the gravity shifts to verb form, an error type that involves an isolated set of meanings (i.e. the nature of the action). These findings imply a general sequence that NES follow when interpreting the textual gravity of verbs—structure, tense, and form. The data for verb errors are included in Table 9.

Table 9
*Verb Error Gravity Scale*

| 6 | 5 | 4 | 3 | 2 | 1 |
|---|---|---|---|---|---|
| INSERT V<br>OMIT V<br>SV | VT | VF | | | |

Other error types appear to follow a modified pattern, while still favoring a global to local interpretation of gravity. If we subdivide the insert and omit errors by structural (i.e. prepositions, determiners) and content words (i.e. nouns, pronouns, adjectives, adverbs), we see a marked preference for structure over content—structure appears to be the gravest concern for this particular group of NES. In fact, all of the insert and omit errors for structural words are category 4 errors, with the exception of those involving conjunctions.

The decision to separate insert and omit errors into discrete subcategories is significant because, to my knowledge, only two other studies, Hulfors (1986) and Tomiyana (1980), have ever done so. Hultfors (1986) uses highly specific error categories, such as *insertion of definite*

*article* and *omission of do –periphrasis*. Yet, once again, his data is quite different from my own. In Hultfors (1986), *insert determiner (^D)* and *insert preposition (^PP)* were ranked as the 13th and 3rd most grievous errors out of 17 errors, respectively. *Omit determiner (~D),* and *omit preposition (~PP)* errors were ranked as the 8th most and 2nd most grievous errors, respectively. That difference amounts to a 35 percent divide between the insert errors and a nearly 60 percent divide in the omit errors. But in my study, there was no recognizable difference between them.

This inconsistency can perhaps be explained by again considering the protocols. Hultfors (1986) includes a range of different insert and omit errors, but the specificity with which he labels each, ironically, also limits the breadth of his research. For example, he provides data on two different types of article errors, writing these specific types into his research, but says nothing about the use of other determiners. So, while they have been included in this study under the category *determiner*, his favors articles while mine includes several other errors, such as possessive pronouns and demonstrative pronouns. In this respect, his categories are smaller in scope than my own, which possibly accounts for the differences in our data.

Tomiyana's (1980) study presents a similar comparison problem. It focuses only on article and connectors and divides each, as was done in the present study, by wrong choice, insertion, and omission. However, his use of *articles* when I use *determiner* implies the same issue seen in trying to compare Hultfors (1986). His *determiner* is narrower in scope than my own. Yet, at the same time, his *connector* is only included minimally in this study as an insert or omit conjunction error. So while I may have the broader category for *determiner*, his *connector* category is broader than mine. In essence, the categorical issues posed in both Hultfors (1986) and Tomiyana (1980) not only point out that several of my categories encompass several things, but also that significant comparison of insert and omit errors across studies is difficult.

Indeed, a major issue presented by the findings is that comparing them to other research is highly problematic. A quick review of over 16 different studies shows an array of different protocols, error categories, and error selections. For instance, Johansson (1975) and Hultfors (1986) talk about determiners, but Burt (1975), James (1998), and McCretton and Rider (1993) do not. Johansson (1975), Burt (1975), and Roberts and Cimasko (2008) omit prepositions, while Vann et al (1984), McCretton and Rider (1993), and Hultfors (1986) leave them in. The result is a mass of data that defy systematic comparison because, at their core, they are dissimilar.

As noted, many gravity studies rely on inherently insufficient research methods, such as using isolated sentences, pre-determined errors, and other manipulated texts (i.e. Burt, 1975; Janopoulos, 1992; Tomiyana, 1980; McCretton & Rider, 1993; Vann, Meyer, & Lorenz, 1984). Yet, the extent to which the texts are manipulated differs widely across studies. Some, like Vann et al (1984) and Janopoulos (1992), rely on purely manufactured texts and sentence-level context. Others, such as Khalil (1985), include context that is more than the sentence-level, but not quite paragraph-level. And others still (e.g. Roberts & Cimasko, 2008; Hyland & Anan, 2006) use authentic, essay-length discourse to investigate gravity. This inconsistency begs the question of, if some studies used flawed methods—and used them to varying degrees—while others used improved ones, can comparisons between them be justified?

Additionally, the attempted comparison with Hultfors (1986) and Tomiyana (1980) revealed that some errors, such as determiner errors, can have multiple meanings. Both studies include only article errors out of the several others they could have chosen, pointing out a now obvious flaw in several of the categories used here. If error categories remain too broad, different compositions of errors within error types might skew the results. This result is implied by Hultfors (1986), who ranks *insertion of the definite article*—16 of 39 errors—as a more serious

concern for NES than *insertion of the indefinite article*—24 or 39 errors. The resulting problem is, if multiple errors can exist within the same general error category, and if each of these errors can have different weights, then comparison with other studies that use these categories ultimately rests on a great deal of speculation when the compositions are unknown. A determiner error's gravity will depend on the number of definite article errors versus the indefinite article errors. If there are more definite articles, the weight will be greater. Yet, if there are more indefinite articles, the weight will be less. The same can be said of other categories as well (i.e. *word choice*, *awkward*, *unclear meaning*, and *word form*).

Finally, issues inherent in the process of ranking itself must also be considered. If five boxers are entered into a tournament, one will become the winner (unless there is a tie) while the others will take the remaining spots. If the winner is prevented from competing the following year, a new one will be crowned and the others will be ranked higher overall. There might even be a different ranking because they would not have to box the previous champion and the effect on the rankings from the absent boxer would be neutralized. Error gravity rankings work the same way. Tomiyana (1980) ranks determiner errors as the most serious while Vann et al (1984) place them among the least serious, but they also look at different sets of errors. Thus, comparing rankings from different studies is akin to comparing Mike Tyson and Mohammed Ali. They may seem similar—both, after all, were highly successful—but their competitive fields were completely different, thereby making comparisons between the two difficult to make.

McCretton and Rider (1993) encounter both the definitional and field limitations when trying to compare their results to four other studies: James (1977), Hughes and Lascaratou (1982), Davies (1983), and Sheorey (1986). They argue that the first, James (1977), is "not directly comparable to those of our own study: by 'case' we understand him to subsume our

category 'prepositions,' and 'tense' includes 'verb forms'....He also subdivided lexical errors into noun, verb, and adjective." (p. 184). The second, Hughes and Lascaratou (1982), included categories that did not overlap with their study, such as negation. And the remaining studies, Davies (1983) and Sheorey (1986), are omitted entirely on the grounds that they are too dissimilar "to enable us to usefully compare them" (p. 184). These differences lead McCretton and Rider (1993) to conclude, as do I, "because of the differences described above, it is unwise to make anything more than a tentative comparison between the results of these separate studies" (p. 184). Further examples are provided in Table 10.

Table 10

*Selected Error Gravity Results*

| | Johansson (1975) | Hultfors (1986) | Porte (1999) | Roberts & Cimasko (2008) | Santos (1988) | Holland (2013) |
|---|---|---|---|---|---|---|
| C/NC | _ | 13 | _ | _ | _ | 1 |
| D | 6 | 16 | 6 | 3 | 2 | 3 |
| PP | _ | 9 | 5 | _ | _ | 4 |
| S/PL | _ | 11 | _ | _ | _ | 3 |
| SV | _ | _ | 1 | _ | 4 | 1 |
| VF | 5 | 7 | _ | 2 | 5 | 3 |
| VT | _ | 4 | 4 | 2 | 7 | 2 |
| WC | 1 | 6 | 3 | 1 | 1 | 4 |
| WF | 5 | 5 | _ | _ | 8 | 2 |
| WO | 3 | 14 | _ | _ | _ | 3 |
| | | | | | | |
| SPELLING | 7 | _ | 2 | _ | _ | _ |
| NEGATION | _ | 15 | _ | _ | _ | _ |
| CONCORD | 5 | 12 | _ | _ | _ | _ |
| CONNECTOR | _ | _ | _ | _ | 3 | _ |
| COHESION | _ | _ | _ | _ | 6 | _ |

*Note*: Errors are ranked from most serious (1) to least serious. The first section is error types included in this present study. The second is error types found in other studies, but not found directly in this one.

Although it may be tempting to think of a few of these data as being similar, such as the two fives for *word form (WF)*—one for Johansson (1975) and one for Hultfors (1986)—or the three ones given to *word choice (WC)* by Santos (1988), Roberts and Cimasko (2008), and Johansson (1975), each implies different things about the gravity of those specific errors. Just as in boxing, where a win over seven opponents is more significant than a win over four, the relative weight of an error takes on greater meaning the more errors it is compared against. In other words, a grave error measured against five errors, tells us less about gravity than if it had been compared against 15 or 20 (to clarify, the different quantities of error types involved in each study required that gravity be ranked from one on in Table 10, instead from 6 to 1 as is the case in this present study). Therefore, this present study represents a marked improvement in error gravity research because, with the exception of Hultfors (1986), who includes 39 error categories, it considers more errors than any other study to date (30).

And so we move on in our discussion of the insert and omit error gravities found in this particular study and in this particular field of competition. In contrast to the structural words, which are all ranked as category 4 errors (again, my error gravity scale ranges from 6, the most serious, to 1, the least serious), there appears to be a divide in the content words between what could be called *strong content words* and *weak content words*, a distinction indicating that the severity of a content word error hinges, to some extent, on the sum meaning it provides. In the global/local error dichotomy, errors that interfere with overall structure and meaning are considered more serious than those that affect only isolated parts of the discourse. Inherent in this construct is the idea that certain errors carry more weight than others, and if this is true for different error types, which are essentially error that has been subdivided to allow greater ease in evaluation and treatment, then why not for error types subdivided even further? In this view, a

complete sentence, one that contains all of the requisite meaning-making parts (i.e. nouns and pronouns) is valued over ones that do not greatly affect the overall meaning (i.e. adverbs). In other words, nouns and pronouns will need to be inserted in the sentence before an adverb because they add more meaning to a sentence than the adverb. Therefore, they classify as strong content words, and adverbs weak. See Table 11 for the actual breakdown of the categories.

Table 11
*Insert Error Gravity Scale*

| 6 | 5 | 4 | 3 | 2 | 1 |
|---|---|---|---|---|---|
|   |   | ^D | ^ADV |   |   |
|   |   | ^N | ^CONJ |   |   |
|   |   | ^PP |   |   |   |
|   |   | ^PRO |   |   |   |

In terms of the omit errors, we may draw several conclusions regarding the relative importance of discrete word classes. First, adverb and pronoun errors are considered more grievous than nouns and adjectives. Adverbs affect the meaning of the verb. Saying that someone walks slowly or quickly provides us, when compared with the function of an adjective, with valuable information regarding how we are to interpret the utterance. By extension, a misused adverb negatively affects the overall meaning of the sentence to a greater extent than an adjective. If one were to write that a man walked fluently, the dissonance caused would be greater than saying that a Communist man walked quickly. The difference between the adverb and the adjective is that a misused adverb confuses the reader more than a misused adjective. The gravity of the error increases the more it affects the verb.

Second, pronouns are also not always essential, as in the case of *the girl wearing the red dress* and *the girl who is wearing the red dress*. In terms of structure, they are more prone to

personal preferences and style than those found in category 3, nouns and adjectives, with perhaps a higher frequency of raters marking them, not on the basis of grammaticality or correctness, but rather their acceptability or strangeness.

In contrast, nouns can contribute to major structural elements in a sentence, making them more valuable to leave in, even in cases when they need not remain. The primary concern with nouns appears to be whether the sentence structure is complete, as *insert noun (^N)* is a category 4 error and *omit noun (~N)* is only a category 3. From this data, one might infer that once the structure of a sentence is complete, omitting an unnecessary noun is less of a priority than other omissions.

Finally, with respect to adjectives, participants marked those that needed to be taken out far more often than ones that needed to be inserted. Indeed, adjective use appears highly idiosyncratic and leaving one in seems less of a concern to NES than leaving in an adverb or a pronoun (Strunk & White, 1999). The *omit (~)* errors are reported in Table 11.

Table 12
*Omit Error Gravity Scale*

| 6 | 5 | 4 | 3 | 2 | 1 |
|---|---|---|---|---|---|
| | | ~ADV | ~ADJ | ~CONJ | |
| | | ~D | ~N | | |
| | | ~PP | | | |
| | | ~PRO | | | |

One of the more surprising results was that both *insert conjunction (^CONJ)* and *omit conjunction (~CONJ)* errors were found on the low end of the gravity spectrum. Burt and Kiparsky (1972) and Lane and Lange (1993) both classify conjunction errors as global because

they interfere with the overall meaning of the sentence, so it is strange that it was not ranked above more local errors, such as *capitalization (C)* or *determiner (D)*.

A possible explanation for this inconsistency is that certain conjunctions may impact sentence structure and meaning more seriously than others. Chastain (1980) notes how NES can often fill in the gaps when they read nonstandard English writing. Based on this claim, one might assume that coordinating conjunctions would be easier to fill in because there are only a few of them. When the choice is between seven different words, with several occurring far more frequently than others (i.e. *and*, *but*, and *or*), it is plausible that insert *conjunction (^CONJ)* errors involving coordinating conjunctions would be less irritating than those involving subordinating conjunctions, for which there are far more options to choose from. Thus, as indicated above, the global/error dichotomy seems mitigated by a feature's generalizability to standard NES texts.

To conclude, the gravity data from this study substantiates Burt's (1975) separation of global and local errors. Those errors affecting the overall meaning of the sentence (i.e. verbs) are considered more serious than ones affecting smaller portions of the sentence (i.e. nouns and adjectives). Global meaning also appears closely connected to how the sentence is structured, as gravity follows a trend from inserting necessary structural words (i.e. prepositions and determiners) and some content words (nouns and pronouns) to inserting adverbs and adjectives. The trend also dictates omitting those words affecting the meaning of the verb (adverbs) first, and then moving to nouns and adjectives.

**Frequency**

The second goal of the study was to determine the relative frequency of the errors. Company (2012) provided data on error frequency for NCS at the ELC, so the question may arise as to the need for additional data. The simple answer is that it was necessary to maintain

consistency in the error gravity and error frequency data because they were to be added together. Company (2012) gathered her data by relying on the error tally sheets mentioned earlier (see Appendix C). But because of how the sheets are used in DWCF, the errors represent only those noticed by a handful of raters, ones that may have had biases in terms of what they noticed (e.g. Beason, 2001) by virtue of their status as L2 instructors. Depending on the week, different errors may have been noticed more because they were considered more serious (i.e. the instructor had discussed them recently or an error had stood out and subsequently drew more attention).

One must be cautious when analyzing error frequency data obtained through subjective measures because gravity will always be implicit. It would be problematic to claim that the data elicited through such measures represents an absolute error frequency count when all it does is show how frequently certain errors were noticed by a given population at a given time. This distinction is perhaps a more fruitful means of considering frequency, but one that is rarely if ever clarified as many studies just assume that their "error" is universal "error," as if that ideal, native-speaker really did exist (e.g. Lunsford & Lunsford, 2008; Janopoulos, 1992). Gravity and frequency are tied to the evaluators. Company (2012) used a different group of raters from those found in this present study, and since frequency is related to gravity in extended discourse (Janopoulos, 1992), it could not be assumed that the frequency data in Company (2012) would match that of this present study. New frequency data was therefore deemed essential.

The results of the new frequency study are as follows: *Word choice (WC)* was clearly the most frequent, but the higher degree of personal preference intrinsic in labeling *word choice (WC)* errors makes this finding unsurprising. To give a simple example, in Wisconsin, some people call a *bubbler* what others call a *drinking* or *water fountain*. Both are used similarly in the sentence, but one person may prefer the first word over the second or third.

Several errors placed similar in both frequency and gravity. *Word form (WF)* errors were not only frequent, but their rankings for gravity and frequency were also the same (category 5). This result implies that the error type is a definite concern for NES, so it should be addressed when it is seen in NCS L2 writing. *Determiners (D)* were ranked as category 4 in both data sets, as were the closely related error type *omit determiner (~ D)*. *Preposition (PP)* errors were also consistent across both scales (category 3). The implication is that the gravity of these error types appears closely tied to their frequency. It is perhaps even probable that, were these errors to occur more frequently, gravity would also increase.

Yet, most errors had noticeably different rankings from the gravity data. For instance, the second most frequent error type was *singular/plural (S/PL)*. Its category 6 frequency ranking is higher than its gravity ranking (4), meaning that it occurs often, but it is not among the most serious concerns of NES. This finding is perhaps for the reasons stated above by Chastain (1980), that NES can easily fill in some corrections because of grammatical regularity in English. If NES know that a word should be plural instead of singular, it is not much effort to auto-correct. A side-by-side comparison of the gravity and frequency data is seen in Figure 4.
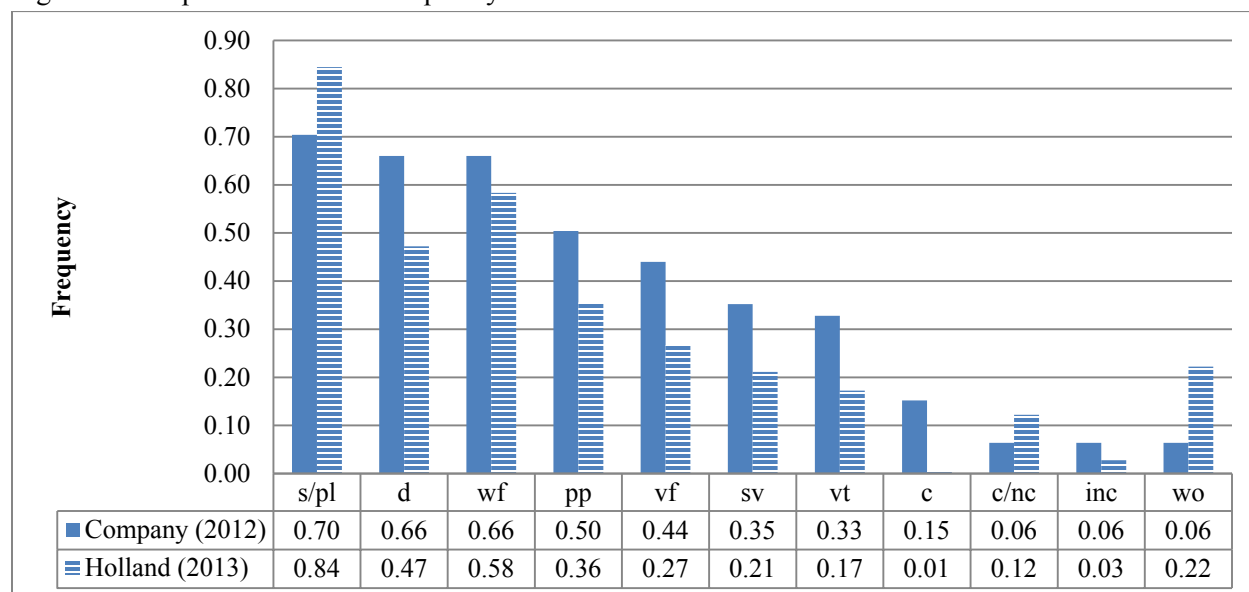
Figure 4. Error Gravity and Error Frequency Compared
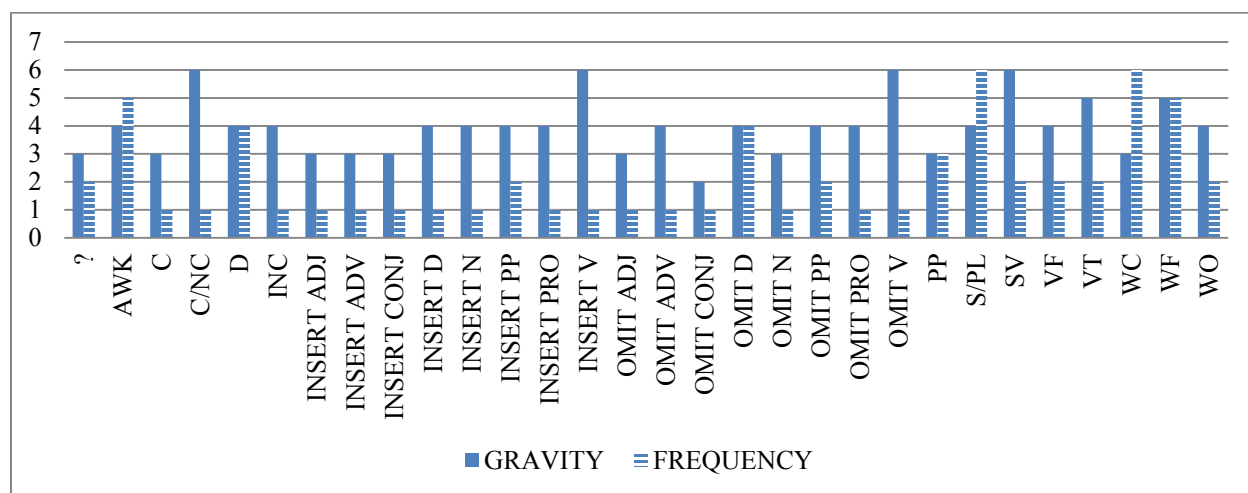
There are several key similarities between Company (2012) and this study that make it possible to draw, at least preliminary, comparisons. For instance, she collected her data from 2007 to 2011 at BYU's ELC, and her participants were enrolled in the linguistic accuracy class as part of their English coursework. The data set in this current study spans a ten year period at the ELC from 2000 to 2010 and uses essays produced as part of their linguistic accuracy coursework. Although not a perfect fit, it is entirely plausible that at least some of the students are similar to both studies. In any case, the students who attend the ELC at this level are similar from year to year. They are roughly the same age (~18-25), share roughly the same goals (to eventually attend a US university), and are roughly at the same level of proficiency (intermediate to advanced), making it reasonable to compare the two.

This process required that I adjust the categories to reflect only those errors that were common to both. In the case of this current study, I omitted *awkward (AWK)*, *unclear meaning (?)*, *insert (^)*, and *omit (~)*. From Company (2012), I took out *sentence structure—run on, (SS RO)* and *spelling (SPG)*. Finally, because of the different methods used to determine *word choice (WC)* errors—Company (2012) relies on individual rater judgment, while this present study uses collective judgment—I chose to take these errors out as well. The results are in Figure 5.

Figure 5. Comparison of Error Frequency Data



| | s/pl | d | wf | pp | vf | sv | vt | c | c/nc | inc | wo |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ■ Company (2012) | 0.70 | 0.66 | 0.66 | 0.50 | 0.44 | 0.35 | 0.33 | 0.15 | 0.06 | 0.06 | 0.06 |
| ≡ Holland (2013) | 0.84 | 0.47 | 0.58 | 0.36 | 0.27 | 0.21 | 0.17 | 0.01 | 0.12 | 0.03 | 0.22 |

The general trend from highest to lowest frequency appears the same, with a few noticeable exceptions. Both studies identified *singular/plural (S/PL)* as the most frequent error type—after *word choice (WC). determiner (D),* and *word form (WF)* errors had the same frequency in Company (2012). Yet, in the data presented here, there is greater separation between the two types, which is perhaps the result of having more tagged errors to work with (463) than Company (182).

Of more concern are the wider disparities in the frequency data that appear at the low end of the scale. A possible explanation for this change could be the inclusion of more data in the present study, as already explained for the high end of the scale. Another is the source texts in my study were more heavily filtered and spelling and punctuation errors edited out, meaning it is plausible that the frequencies of certain error types (i.e. capitalization and sentence structure) may have also been reduced in the process. And a third is that the source texts in this current study include more context, which may have affected the errors noticed. In any case, if we add the data from both studies together, we arrive at Figure 5.

Figure 6. Combined Error Frequency Data
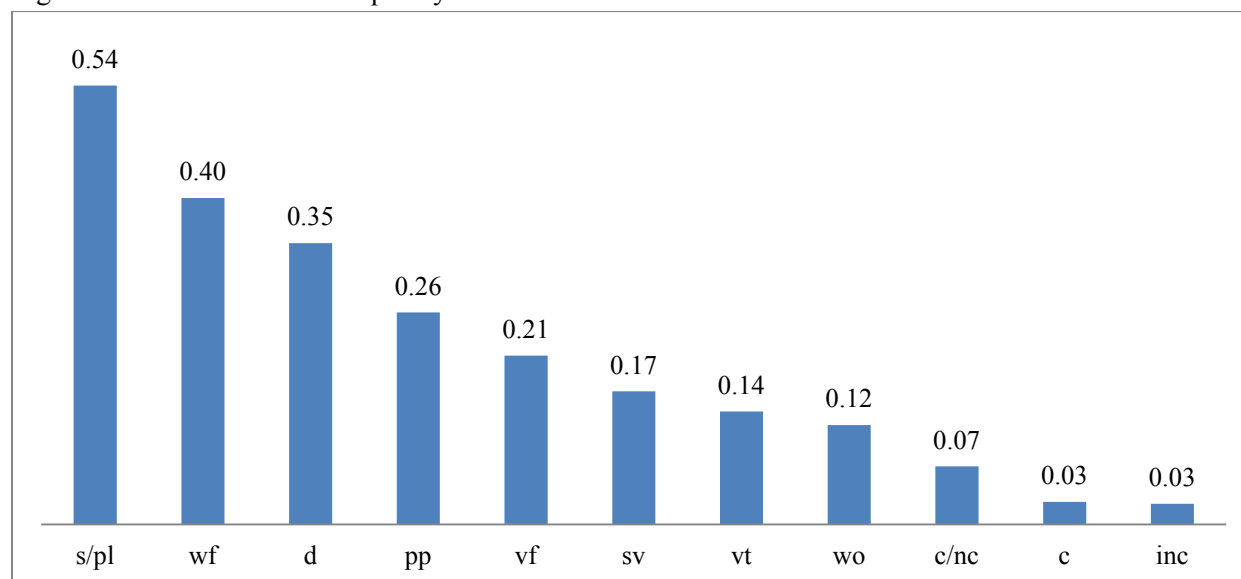


Here, *word form (WF)* overtakes *determiner (D)* for the second highest spot, and *word order (WO)*, *count/non-count (C/NC)*, *capitalization (C),* and *incomplete (INC)* shift places from their original spots in Company (2012). Given the increased data (645 tokens), broader pool of texts (43) and added diversity in the types of discourse used (paragraph *and* essay), I believe this chart more accurately represents the frequency of errors made by NCS—with the exception of the final two errors—and may prove to be a useful addition to the frequency data presented by Company (2012).

Yet the similarities do not imply that my initial assumption that different audiences would find different error frequencies is false. As noted, Company (2012) used a small number of ELC instructors to find the errors. Each text was considered only once and by only one person. In contrast, ten participants per essay were used in my study—a total of 180 discrete evaluations. There were more ratings per text in this present study, meaning that the similarity not only speaks well of the individual abilities of the ELC Linguistic Accuracy faculty, as fewer respondents, expectedly, produced similar results to that of a larger community of non-teachers,

but it also argues for the collective intelligence of both undergraduate and graduate students. As such, whether a comparable number of ELC teachers would produce the same frequency data as a group of non-teachers remains to be seen.

Finally, it is unfortunate that Liu (2011), another study that relies on teachers to generate the frequency data, uses different categories in his classification because it prevents productive comparison between his data and that of Company (2012) and my own. Liu (2011) combined errors by part of speech, meaning that all errors related to, say, nouns, are put together. The finer distinction I make in the insert and omit errors moves us a step closer to being able to compare the studies, which is yet another argument for including them in error frequency counts, but because he also separates word form by part of speech, and this present study does not, comparison between the two is ultimately impossible. In this respect, the more exact we are with error description the better suited we will be to make comparisons.

The take home message from the error frequency portion of the research is the following: Company (2012) removed *omit (~)*, *insert (^)*, *awkward (AWK)*, and *unclear meaning (?)* from her data on the grounds that they could "encompass several things" (p. 17). Yet, from the problems encountered when trying to draw comparisons to the Liu (2011) study—in addition to those encountered in the gravity data—it appears that other error types also include several things. Thus, there is a significant issue with error description in the current DWCF codes, at least as they relate to error gravity and error frequency study.

**Gravity + Frequency: Implications for Error Correction**

Frequency was included in the gravity data, but the several variations in the data from the previous sections argue for maintaining the distinction between the two. With the exception of *omit determiner (~D)* errors, omit and insert errors were relatively infrequent. All of them failed

to get above a category 2 ranking, most did not even get past a category 1. Similarly, verb-related issues—*verb form (VF), verb tense (VT), insert verb (^V), omit verb (~V),* and *subject-verb agreement (SV)*—were ranked far lower than they had been for gravity. While they are considered among the most socially grave errors a NCS could make, the fact that they do not occur as frequently as other errors, such as preposition or determiner, means they may not be as high of a priority as the gravity data suggest.

Hence, there are at least two different types of frequency in this study—one is the effect of frequency on gravity, the other is the number of errors noticed. While the obvious theoretical flaw in this separation is as old as the proverbial chicken and the egg—i.e. to what extent does gravity cause frequency and frequency gravity?—it is clear from the results that both are still viable and necessary types of error analysis. Gravity may be related to frequency, but it is not purely frequency, as the "error-as-frequency" fallacy would suggest, and although frequency is included in the gravity data, the considerable difference between the two data sets implies that the *quality* of frequency is not necessarily the same thing as its *quantity*.

The role of frequency in gravity is extended even further in this final stage by adding the two scales together. The contribution made by the first stage was that it ameliorated several common complaints regarding error gravity research—i.e. overreliance on isolated sentences, pre-determined errors, manipulated texts, and insufficient constructs (e.g. Albrechtsen et al, 1980; Rifkin & Roberts, 1995; Chastain, 1980; Ellis, 2008). The hole that had been left largely unfilled was that they had very rarely said anything about how the prevalence of specific error types shifts opinions of seriousness. In a standard written assessment, error rarely occurs on its own. It is more often than not found alongside a wide variety of other error types, including

multiple occurrences of the same error type. Subsequently, an error that occurs in isolation is likely to be rated differently than one that occurs as part of a larger corpus of error.

Combining all errors into the same assessment, not only by error type, but also every other co-existing error in the text and its frequency, allowed for a highly naturalistic competition to occur, one that accounted for the relative nature of error gravity. As previously noted, a preposition error may seem significant on its own, yet in the presence of other errors, it may be deemed less important. In this respect, the error gravity data in this study presents a perhaps more accurate accounting of gravity than most studies have typically given. It also works to correct several issues previously associated with the underlying theory of error gravity research.

But to apply this theory to EC, there must also be a more practical side to the data. James (1998) and Lane and Lange (1993) provide a supplemental theory to error gravity involving both the new gravity and frequency (in its purest, objective, error-count sense). The rationale is that a serious error that occurs infrequently is much less productive to focus on than a serious error that occurs frequently. Furthermore, there is the issue of which errors are so serious that they surmount some of the more frequently made, yet less serious errors. Based on these observations, including frequency as both a theoretical and a practical construct was warranted.

The new gravity + frequency data set *singular/plural (S/PL)* and *word form (WF)* errors as the most important to EC (category 10). *Word choice (WC)* and *awkward (AWK)* were the second most important, followed by *determiner (D)*, *omit determiner (~D)*, and *subject-verb agreement (SV)*, then *verb tense (VT)*, *count/non-count (C/NC)*, *insert verb (^V)*, and *omit verb (~V)*, and so on down the list (see Table 8). These data differ from the gravity rankings by shifting errors such as count/non-count, verb-related, and structural word to lower priority and by raising the importance of word form, awkward, and singular/plural errors. In contrast, the data

departs from the frequency rankings, most noticeably, by breaking the category 1 cluster, and dividing errors more evenly across the board. For instance, *count/non-count (C/NC)* goes from category 1 in the frequency data to a category 7 in the new gravity + frequency data. *Insert pronoun (^PRO)* errors were also category 1 but only make it to a category 5 in the new data.

If instructors rely solely on the gravity or frequency scales to address NCS L2 written errors, they might easily devote time to low priority items at the expense of others. Thus, as the representation of both the theoretical and practical application of both gravity and frequency, the new scale not only surmounts several potential time-wasters in EC, such as focusing on *subject-verb agreement (SV)* before *singular/plural (S/PL),* but it also presents instructors with a more structured, empirically driven approach. Instead of having to guess whether an error is serious or not based on personal observation or by relying on solely the new gravity or frequency data, L2 instructors may now use the chart above to determine how best to prioritize their EC efforts.

For instance, DWCF instructors can include these findings in their classes in one of several ways. First, they could memorize or refer to the codes as they mark, highlighting only those errors that fall within the top third of the gravity + frequency scale. Second, they could mark all errors, and then highlight the top-third errors to draw greater attention to them. Or third, they could mark all errors, or at least the ones that have been discussed in class, and refer to both the texts themselves and the data provided in this research to select the most salient grammar and usage points to be addressed each day. In this way, they would be empowered to respond more effectively to their students' work, as it would scale the errors in such a way that students would also no longer have to guess which errors are most important.

Ferris (1999) would here insert the caveat that not all errors are equally treatable, and thus that not all errors should be addressed in the same way. Ferris (1999) believes, as do I, that

"students can be successfully taught to self-edit their own texts if they are…(b) trained to identify and correct patterns of frequent and serious errors; and (c) given explicit teaching as needed about the rules governing these patterns of errors" (p. 5). Yet, as she was reflecting on her students' "needs and progress (or lack thereof, in some cases)" (p.6), she became "increasingly aware that [her] suggestions are only applicable when students' errors occur in a patterned, rule-governed way" (p.6). The implication is that even when students are trained to *identify* patterns of frequent and serious errors, some errors (i.e. subject-verb agreement, run-ons and comma splices, missing articles, and verb form errors) remain more ostensibly rule-governed than others—i.e. "a wide variety of lexical errors and problems with sentence structure, including missing words, unnecessary words and word order problems" (p. 6). If the pattern is not readily apparent to either the instructor or student, an alternative approach must be taken (Ferris, 1999).

Ferris' (1999) own approach to correcting "untreatable" error includes "a combination of strategy training and direct correction" (p. 6), but regardless of the method used to treat these errors, "ESL writing instructors would do well to give much more thought to how they provide error feedback regarding these different types of language forms and structures" (p. 6). This sort of ongoing reflection, of carefully monitoring and evaluating student success in self-editing, will not only contribute to L2 instructors being better prepared to address their students' needs, but it will also acknowledge that linguistic homogeneity can never be entirely assumed in the L2 classroom (Matsuda, 2006). Individual intelligibility nuclei are defined by their commonalities, much as are patterns of error frequency and error gravity. These prototypical views are "inevitable and even necessary" (Matsuda, 2006, p. 639), as "without those images, discussing pedagogical issues across institutions would be impossible" (p. 639). Yet, when over-applied, they also cease to be useful. Thus, while the information provided in this study is valuable to

finding the most efficient and effective WCF practices, immediate conditions in the classroom should always dictate both how and when it is applied.

**Limitations and Suggestions for Further Research**

As this research deals primarily with evaluating error, the new gravity + frequency scale will be better used if it is also paired with studies involving the cause, or diagnosis, of error. Focusing on just this list alone would be akin to a doctor prescribing medication for a condition that he or she does not fully understand. Although it might happen that the prescribed medicine is exactly what the patient needs, a far better approach would be obviously to first observe the patient's symptoms before applying treatment. The information provided here is nothing more than error triage. It ranks error types by their severity, allowing L2 instructors to devote their efforts to the most threatening, but says nothing absolute regarding their cause. This is one of several limitations that must be kept in mind when applying the results of this study.

Much debate in second language writing has centered on whether NES are a suitable standard for all English use (e.g. Silva, Leki, & Carson, 1997; Spack, 1997; Kubota, 2001). The research presented here has very deliberately limited the "standard" to that of American NES—not to imply that this is the only or best standard, but rather, that this is the standard L2 students studying in the United States are most likely to encounter. For this reason, one should proceed with caution when trying to apply the new gravity + frequency scale to populations outside of the US. Indeed, it would be interesting to see in subsequent research how other communities define and evaluate English errors in comparison to the data given here.

Furthermore, it would be unwarranted to assume that all error types made by NCS are included. Raaijmakers, (2003) reminds us of the "language-as-fixed-effect" fallacy, which is, in essence, the inaccurate conclusion that different texts can be relied upon to produce the same

range of errors. In other words, the error frequencies and error gravities discovered by this data set are self-supporting only, and subsequent studies must be conducted to determine whether the results will continue to hold. As Duskova (1969) claims,

> It should be borne in mind that the subjects were free to choose how to express the given content. As a result, while some grammatical points (such as articles, the past tense, the plural) were bound to occur in all papers, others (adverbs, relative pronouns, the future tense) appeared only in some. Thus lower frequency of an error need not necessarily mean that the point in question is less difficult, but simply that the point itself occurred only in some (not in all) papers. (p. 15)

Because it would be highly difficult, perhaps even impossible, to collect all of the possible error types and include them in error gravity research, there will always be some errors left out (Johansson, 1978). Additional research on the errors NCS make will therefore add greatly to our ability to determine whether the results found here are typical of NCS errors in general.

Unfortunately, two error types were consciously omitted, and it would be interesting to see how the addition of spelling and punctuation errors would affect the overall error ranking. As previously noted, these error types were edited out of the original essays, and this exclusion prevented me from obtaining data on related errors, such as *sentence structure, run-on (SS RO* and *incomplete sentence (INC)*. In retrospect, this editing was foolish because of the classification of sentence structure issues as global errors (Burt, 1975). Although I do not feel that the omission of these errors invalidates the data in any way—there is, after all, still a general ranking of error that is useful to EC—replication studies would be well-advised to include them.

Furthermore, due to the subjectivity involved in filtering the initial 169 essays, there is no doubt that other errors were unconsciously omitted as well. As previously noted, I do not believe

that the criteria used to assess rhetorical content (i.e. thesis statement, topic sentences, supporting details) and word count (300+) were so far outside the realm of mainstream writing pedagogy that they invalidate the results in any way, but I start to get nervous when I consider the more vaguely defined linguistic complexity category I used, especially given that this study looks at the relative gravities and frequencies of discrete errors. Were the time and resources available, the ideal would have been to follow the more objective approach championed by Wolfe-Quintero, Inagaki, and Kim (1998) since, in applying what I felt to be the marks of a linguistically advanced text (i.e. sentence variety, word choice, clauses), it is entirely plausible that I skewed the results to my personal preferences, leaving out several errors that would have been present had I been able to find willing participants to help with the sorting.

In addition to including a more accurate representation of "university-level" texts, replication studies would also benefit by including different evaluator subpopulations. In the data found here, a relatively narrow group (i.e. 20-30 year-olds, university-educated) took part, thereby warranting the question of whether this group is actually fit to stand in for the most significant "real world" contexts L2 students come across, such as their education or employment (Beason, 2001). As Johansson (1978) and others remind us, error type is just one part of the evaluation process (Johansson, 1978). Other factors, such as receiver characteristics (e.g. Santos, 1988; Weigle, Boldt, & Valsecchi, 2003; Barkaoui, 2010a; Barkaoui, 2010b), the type of language situation (Johansson, 1978), and the role of the sender (e.g. Barden, Maddux, Petty & Brewer, 2004; Hosoda, 2006) are also involved.

Research should continue into the error codes themselves. In the delimitations for this study, I say that some error types seen in other studies do not have a clear place in the DWCF codes, and at least for the time being, this situation was unavoidable. In the interest of better

training instructors of DWCF, as well as facilitating further studies, it would helpful to have a detailed classification that describes the exact error types included in each error code. Not only would this clarification improve the ability of DWCF instructors to mark errors more consistently and more accurately across classes, perhaps even reducing the marking of *awkward (AWK)* or *unclear meaning (?)* errors, but it would also provide a reference point for future research. As long as studies continue to describe error in their own way, it will be difficult—as indeed, it proved to be in this study—to compare the results. However, if there were to be a common language of error codes understood and spoken by all researchers involved in error gravity study, it would greatly aid our understanding of error.

Additionally, despite efforts to limit each code to one error apiece, there remain some categories that still "encompass several things" (Company, 2012, p. 17). To give an analogy, the atom was considered for a long time to be the smallest possible particle in the universe. After electrons, positrons, and neutrons were discovered, the categorization was revised, so that these particles were now considered the smallest. As time goes on, and research progresses, smaller and smaller particles are found, making what was once considered to be the smallest particle, the entire atom, a far more complex thing than originally thought. Error description is much the same way. For instance, a *verb form (VF)* error might be subdivided into simple, perfect, progressive, and so on, yet each aspect could also be separated even further. Perfect aspect could become, among other things, perfect to indicate experience, perfect to indicate change over time, or perfect to indicate accomplishments. Gravity and frequency data could be obtained for discrete error types (the electron) in each general category (the atom) to determine the individual compositions of each (Hultfors, 1986; Tomiyana, 1980). The finer the distinction, the more detailed our understanding of nonnative error will be.

Specifically, the following errors would be ideal for further study

*Word Choice (WC)*

Initially, it was surprising to see how many word choice errors were found in the texts, especially in comparison to the other types. However, given that word usage is, in many cases, a matter of personal preference, stemming from variable sensitivities to nuances in the connotations and denotations of certain words to the need to make a finer distinction in meaning in the presence or absence of context, a large pool of *word choice (WC)* errors fits (James, 1998). Yet, this mass of word choice errors is highly problematic to error correction. In spite of their high frequency, they are actually quite low (category 3) in terms of their overall seriousness to a NES population. This finding implies that *word choice (WC)* is a "wild card" error. It may irritate one rater, as a result of regional differences, education, the books one reads, and so on, which may result in a momentary decision to score the essay less, but that is just one person. To another individual, the same error may not seem nearly as serious. This inconsistency makes these errors difficult to determine. English has thousands of words. To expect everyone to think the same way about them would be highly unrealistic.

Even so, *word choice (WC)* deserves substantial research into the frequencies and gravities of specific word classes (i.e. nouns, adjectives, and verbs) because L2 learners sometimes associate a TL with its vocabulary (James, 1998) under the assumption that knowing a language means nothing more than knowing a lot of words. Thus, this research might profitably consider why NES think words are misused and whether there is an irritation continuum for word use. It might also consider how the denotations and connotations of words relate to error determination. For instance, to what extent are NES prone to favor certain words

in certain contexts over others? The answer to this question could be quite useful in developing lesson plans and curricula intended to correct *word choice (WC)* errors.

*Awkward (AWK)*

*Awkward (AWK)* errors were considered category 4. These errors are ones in which the content is accessible to some extent. In other words, they are intelligible, albeit minimally. The reader has a general idea of what the writer wishes to express, yet something about the sentence—a strange word choice or order, for instance—prevents the reader from easily grasping the meaning of the utterance.

The problem with this error type is that it is too easy to apply, and "we can all hear the ring of truth in the assertion that teachers are inconsistent in their ability and willingness to recognize and correct errors and to provide adequate grammar explanations to their students" (Ferris, 1999). Thus, an *awkward (AWK)* error could also mean the rater was irritated by a particular token, yet in the interest of time or energy, the cause of the irritation was left unexamined. This decision to "take the easy way out" could likely be avoided with more definite error categories. For example, instead of having to decide whether an it-deletion error classifies as an insert error or as an incomplete sentence, it could just be called an "it-deletion" error. Of course, *awkward (AWK)* errors could also include those errors that, even after deliberate analysis, still leave the rater baffled, but the point is, *awkward (AWK),* as a category, needs to be subdivided if it, and other error categories, are to be better understood.

*Unclear Meaning (?)*

That *unclear meaning (?)* is considered less serious (category 3) than *awkward (AWK)* possibly indicates a connection between the findings of Johnson and VanBrackle (2010) and Janopoulos (1995), who determine that perceived group membership affects rating decisions. In

the case of *awkward (AWK)* errors, a NES reading the text may consider the writer of such errors part of the English-speaking community. Perhaps after some threshold of meaning has been attained, the reader accepts the potential inclusion of the writer. But with *unclear meaning (?)*, it is possible the threshold has yet to be reached and it is therefore easier for the reader to discount the writer as a viable part of his or her community. Research can be conducted into exactly how many and what type of errors leads a NES to consider the writer "native" or "nonnative."

This finer distinction, of course, would require a shift away from describing errors generally, as is done here, to describing a discrete set of errors related to just one main error type. One study may focus only on verb errors, for example, while another on preposition errors. If the codes used to describe the errors and the protocols are similar enough, information about errors could essentially take a bottom-up approach. Small bits of information could be gradually added to a large corpus of error frequencies and gravities, much as is currently done in corpuses, such as the Corpus of Contemporary American English (COCA) or the British National Corpus (BNC) with general language use. To continue the atom analogy, we could first identify the specific weights of discrete error types and then use these data to analyze the "reactions" caused as these errors are compounded with others. This corpus could then be accessed by L2 instructors to sharpen their ability to structure lessons, materials, and syllabi that focus on key errors.

**Conclusion**

This study has sought to establish an error gravity index that accounts for both the prevalence of error and its gravity—traditionally a composite of comprehensibility, irritation, and acceptability. Given the differences in the frequency and gravity scales, we may conclude that frequency is not gravity, nor is gravity frequency. An error may be frequent but not serious,

just as a serious error may occur only infrequently, and it is for this very reason that EC should seek to combine the data. If one were to focus solely on frequent errors, one would miss a large portion of error's impact, or its social severity. Yet, by focusing only on gravity data, as it is typically considered in error gravity research, one would largely miss the prevalence of error. As such, a combined approach is not only more complete, but it also makes sense.

Even still, if the long list of limitations and suggestions for further research is any indication, there remains much in the field of error gravity research that is yet unknown. For example, how would different populations, different texts, and clearer error taxonomies increase our understanding of the highly nuanced subject of error? The results presented here are just one small piece of the larger picture, one that is yet becoming more complex as our understanding of error increases. This study reports on the gravity of error within the context of NCS L2 writers at the ELC and their NES peers at an American university. Whether these results can be generalized to a larger or related population remains to be seen in future research.

# References

Albrechtsen, D., B, H., & Faerch, C. (1980). Native speaker reactions to learners' spoken interlanguage. *Language Learning, 30*, 394.

Barden, J., Maddux, W., Petty, R., & Brewer, M. (2004). Contextual moderation of racial bias: The impact of social roles on controlled and automatically activated attitudes. *Journal of Personality and Social Pyschology, 87*(1), 5-22.

Barkaoui, K. (2010). Do ESL essay raters' evaluation criteria change with experience? A mixed-methods, cross-sectional study. *TESOL Quarterly, 44*(1), 31-55.

Barkaoui, K. (2010). Variability in ESL rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly, 7*, 54-74.

Beason, L. (2001). Ethos and error: How business people react to errors. *College Composition and Communication, 53*(1), 33-64.

Burt, M. K. (1975). Error analysis in the adult EFL classroom. *TESOL Quarterly, 9*(1), 53-63.

Burt, M. K., & Kiparsky, C. (1972). *The gooficon: A repair manual for English.* Rowley, Massachuesetts: Newbury House Publishers.

Chan, A. Y. (2010). Toward a taxonomy of written errors: Investigation into the written errors of Hong Kong Cantonese ESL learners. *TESOL Quarterly, 44*(2), 295-319.

Chastain, K. (1980). Native speaker reaction to instructor-identified student second-language errors. *The Modern Language Journal, 64*(2), 210-215.

Chastain, K. (1981). Native speaker evaluation of student composition errors. *The Modern Language Journal, 65*(3), 288-294.

Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge, MA: The M.I.T. Press.

Company, M. (2012). *Error frequencies among ESL writers: A resource guide.* (Unpublished Master's Thesis). Brigham Young University, Provo, UT.

Corder, S. P. (1967). The significance of learners' errors. *International Review of Applied Linguistics, 5*(4), 161-169.

Corder, S. P. (1971). Idiosyncratic dialects and error analysis. *International Review of Applied Linguistics, 9*(2), 14-25.

Davies, E. (1983). Error evaluation: the importance of viewpoint. *ELT Journal, 37*(4), 304-311.

Delisle, H. (1982). Native speaker judgment and the evaluation of errors in German. *The Modern Language Journal, 66*(1), 39-48.

Duskova, L. (1969). On sources of errors in foreign language learning. *7*(1), 11-36.

Earline Shairer, K. (1992). Native speaker reaction to non-native speech. *The Modern Language Journal, 76*(3), 309-319.

Ellis, R. (2008). *The study of second language acquisition* (2nd ed.). New York: Oxford University Press.

Ensz, K. Y. (1982). French attitudes toward typical speech errors of American speakers of French. *The Modern Language Journal, 66*(2), 133-139.

Evans, N. W., Hartshorn, K. J., & Strong-Krause, D. (2011). The efficacy of dynamic written corrective feedback for university-matriculated ESL learners. *System, 39*, 229-239.

Evans, N. W., Hartshorn, K. J., McCollum, R. M., & Wolfersberger, M. (2010). Contextualizing corrective feedback in second language writing. *Language Teaching Research, 14*(4), 445-463.

Evans, N., Hartshorn, K. J., & Tuioti, E. (2010). Written corrective feedback: Practitioners' perspectives. *International Journal of English Studies, 10*(2), 47-77.

Falk, G. (2001). *Stigma: How we treat outsiders.* Amherst, NY: Prometheus Books.

Ferris, D. (1999). The case for grammar correction in L2 writing classes: A response to Truscott (1996). *Journal of Second Language Writing, 8*(1), 1-11.

Ferris, D. R. (2009). *Teaching College Writing to Diverse Student Populations.* (P. Byrd, J. M. Reid, & C. M. Schuemann, Eds.) Ann Arbor: University of Michigan Press.

Galloway, V. B. (1980). Perceptions of the communicative efforts of American students of Spanish. *The Modern Language Journal, 64*(4), 428-433.

Gergen, K. (1997). *Realities and relationships: Soundings in social construction.* Boston: Harvard University Press.

Goffman, I. (1963). *Stigma: Notes on the management of spoiled identity.* Englewood Cliffs, NJ: Prentice-Hall.

Gynan, S. N. (1985). Comprehension, irritation, and error hierarchies. *Hispania, 68*(1), 160-165.

Hartshorn, K. J., Evans, N. W., Merrill, P. F., Sudweeks, R. R., Strong-Krause, D., & Anderson, N. J. (2010). Effects of dynamic corrective feedback on ESL writing accuracy. *TESOL Quarterly, 44*(1), 84-109.

Hinkel, E. (2009). The effect of essay topics on modal verb uses in L1 and L2 academic writing. *Journal of Pragmatics, 41*(4), 667-683.

Hosoda, Y. (2006). Repair and relevance of differential language expertise in second language conversations. *Applied Linguistics, 27*, 25-50.

Hughes, A., & Lascaratou, C. (1982). Competing criteria for error gravity. *ELT Journal, 36*(3), 175-182.

Hultfors, P. (1986). *Reactions to non-native English: Native English-speakers' assessments of errors in the use of English made by non-native users of the language.* Stockholm: Almqvist & Wiksell International.

Hyland, K., & Anan, E. (2006). Teachers' perceptions of error: The effects of first language and experience. *System, 34*, 509-519.

Institute of International Education. (2012). *Fast facts*. Retrieved from Institute of International Education: http://www.iie.org/Research-and-Publications/Open-Doors

James, C. (1972). The diagnosis and analysis of error--some insights from linguistics. *Audio-visual Language Journal: Organ of the British Association for Language Teaching, 10*(2), 75-79.

James, C. (1974). Linguistic measures of error gravity. *Audio-visual Language Journal, 12*(1), 3-9.

James, C. (1998). *Errors in language learning and use.* New York: Longman.

Janopoulos, M. (1992). University faculty tolerance of NS and NNS writing errors: A comparison. *Journal of second language writing, 1*(2), 109-121.

Janopoulos, M. (1995). Writing across the curriculum, writing proficiency exams, and the NNS college student. *Journal of Second Language Writing, 4*(1), 43-50.

Johansson, S. (1978). *Studies of error gravity: Native reactions to errors produced by Swedish learners of English.* Goteborg, Sweden: Acta Universitatis Gothobergensis.

Johnson, D., & VanBrackle, L. (2012). Linguistic discrimination in writing assessment: How raters react to African American "errors," ESL errors, and standard English errors on a state-mandated writing exam. *Assessing Writing, 17*, 35-54.

Khalil, A. (1985). Communicative error evaluation: Native speakers' evaluation and

    interpretation of written errors of Arab EFL learners. *TESOL Quarterly, 19*(2), 335-351.

King, M. L. (1959, November 11). *The future of integration.* Retrieved from The King Center:

    http://www.thekingcenter.org/archive/document/future-integration-1#

Kubota, R. (2001). Discursive construction of the images of U.S. classrooms. *TESOL Quarterly,*

    *35*(1), 9-38.

Lane, J., & Lange, E. (1993). *Writing clearly: An editing guide.* New York: Heinle & Heinle

    Publishers.

Lee, S. (2009). *The efficacy of dynamic written corrective feedback on intermediate-high ESL*

    *learners' writing accuracy*. (Unpublished Master's Thesis). Brigham Young University,

    Provo, UT.

Liu, Z. (2011). Negative transfer of Chinese to college students' English writing. *Journal of*

    *Language Teaching and Research, 2*(5), 1061-1068.

Ludwig, J. (1982). Native-speaker judgments of second-language learners' efforts at

    communication: A review. *The Modern Language Journal, 66*(3), 274-283.

Mahoney, S. (2011). Exploring gaps in teacher and student EFL error evaluation. *Japan*

    *Association of Language Teachers Journal, 33*(2), 107-130.

Matsuda, P. K. (2006). The myth of linguistic homogeneity in U.S. college composition. *College*

    *English*, *68*(6), 637–651.

McCretton, E., & Rider, N. (1993). Error gravity and error hierarchies. *International Review of*

    *Applied Linguistics in Language Testing, 31*(3), 177-188.

McNamara, T. F. (1996). *Measuring second language performance.* London: Longman.

Milroy, J. (1992). *Linguistic variation and change.* Oxford: Blackwell.

Munro, M. J., & Derwing, T. (1995). Processing time, accent, and comprehensibility in the

   perception of native and foreign-accented speech. *Language and speech, 38*(3), 289-306.

Okamura, A. (1995). Teachers' and nonteachers' perception of elementary learners' spoken

   Japanese. *The Modern Language Journal, 79*(1), 29-40.

Piazza, L. G. (1980). French tolerance for grammatical errors made by Americans. *Modern

   Language Journal, 64*, 425.

Politzer, R. L. (1978). Errors of English Speakers of German as Perceived and Evaluated by

   German Natives. *The Modern Language Journal, 62*(5/6), 253-261.

Porte, G. (2010). Where to draw the red line: Error toleration of native and nonnative EFL

   faculty. *Foreign Language Annals, 32*(4).

Quirk, R., & Svartvik, J. (1966). *Investigating linguistic acceptability.* London: Mouton & Co.

Raaijmakers, J. G. (2003). A further look at the "language-as-fixed-effect-fallacy". *Canadian

   Journal of Experimental Pyschology, 57*(3), 141-151.

Rifkin, B., & Roberts, F. (1995). Error gravity: A critical review of research design. *Language

   Learning, 45*(3), 511-537.

Roberts, F., & Cimasko, T. (2008). Evaluating ESL: Making sense of university professors'

   responses to second language writing. *Journal of Second Language Writing, 17*, 125-143.

Santos, T. (1988). Professors' reactions to the academic writing of nonnative-speaking students.

   *TESOL Quarterly, 22*(1), 69-90.

Scriven, M. (1991). *Evaluation thesaurus.* London: Sage.

Sheorey, R. (1986). Error perceptions of native-speaking and non-native-speaking teachers of

   ESL. *ELT Journal, 40*, 306-312.

Silva, T. (1997). On the ethical treatment of ESL writers. *TESOL Quarterly, 31*(2), 359-363.

Silva, T., Leki, I., & Carson, J. (1997). Broadening the perspectives of mainstream composition

 studies: Some thoughts from the disciplinary margins. *Written Communication, 14*(2),

 398-428.

Spack, R. (1997). The rhetorical construction of multilingual students. *TESOL Quarterly, 31*,

 765-764.

Strunk, W., & White, E. B. (1999). *The elements of style* (4th ed.). New York: Longman.

Tomiyana, M. (1980). Grammatical errors communication breakdown. *TESOL Quarterly, 14*(1),

 71-79.

Vann, R. J., Meyer, D. E., & Lorenz, F. O. (1984). Error gravity: A study of faculty opinion of

 ESL errors. *TESOL Quarterly, 18*(3), 427-440.

Vygotsky, L. S. (1962). *Thought and language.* Cambridge, MA: MIT Press.

Weigle, S. C., Boldt, H., & Valsecchi, M. I. (2003). Effects of task and rater background on the

 evaluation of ESL student writing: A pilot study. *TESOL Quarterly, 37*(2), 345-354.

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing:*

 *Measures of fluency, accuracy & complexity.* Honolulu: Second Language Teaching &

 Curriculum Center.

## Appendix A: Error Correction Symbols

*Indirect Coding Symbols Used to Mark Student Writing in Linguistic Accuracy Classes*

| | | | | | |
|---|---|---|---|---|---|
| 1. | D | = Determiner | 11. | S/PL | = Singular/Plural |
| 2. | SV | = Subject Verb Agreement | 12. | C/NC | = Count/Noncount |
| 3. | VF | = Verb Form | 13. | ? | = Meaning is not clear |
| 4. | Ro | = Run-on Sentence | 14. | AWK | = Awkward Wording |
| 5. | Inc | = Incomplete sentence | 15. | ⌒ | = Word Order |
| 6. | VT | = Verb Tense | 16. | C | = Capitalization |
| 7. | PP | = Preposition | 17. | P | = Punctuation |
| 8. | SPG | = Spelling | 18. | ℓ | = Omit |
| 9. | WF | = Word Form | 19. | ^ | = Something is missing |
| 10. | WC | = Word Choice | 20. | ¶ | = New Paragraph |

See next page for symbols used in context

## Appendix B: Error Correction Symbols Applied

| Error Samples | Correction |
|---|---|
| 1. The climber slowly ascended to ᴰtop. | *A determiner is needed before* top. |
| 2. She ˢⱽthink he will win the race. | *She thinks he will win the race.* |
| 3. Eat ⱽᶠpizza at parties is fun for us. | *Eating pizza at parties is fun for us.* |
| 4. He bought pizza ʳᵒshe came by they ate it. | *These independent clauses need to be separated or combined properly.* |
| 5. Because inflation ⁱⁿᶜhad risen so sharply. | *An independent clause is required.* |
| 6. Yesterday she ⱽᵀdive to Provo. | *Yesterday she drove to Provo.* |
| 7. He was always studying in ᴾᴾ7:00 AM. | *He was always studying at 7:00 AM* |
| 8. She was exceptional at mathoᴿᴾᴳmatics. | *She was exceptional at mathematics.* |
| 9. He truly was a very diligenᵂᶠce student. | *He truly was a very diligent student.* |
| 10. She typed the paper on her calcuᵂᶜlator. | *She typed the paper on her computer.* |
| 11. He bought five ˢ/ᴾᴸapple with the money. | *He bought five apples…* |
| 12. She breathed in the fresh aᶜ/ᴺᶜirs. | *She breathed in the fresh air.* |
| 13. The desk walked to the eat ? door. ) | *(requires clarification)* |
| 14. My family has 1 bᴬᵂᴷother and 1 sister. | *I have one brother and one sister.* |
| 15. She ran two times the marathon. | *She ran the marathon two times.* |
| 16. ᶜthen ᶜmr. ᶜwhite came home. | *Then Mr. White came home* |
| 17. She said ᴾI am so happy ᴾ | *She said, "I am so happy."* |
| 18. I will very study very hard. | *I will study very hard.* |
| 19. After class did all my homework. | *After class I did all my homework.* |

**Appendix C: Error Tally Sheet**

| | | | | | | | | | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| D | | | | | | | | | | | |
| SV | | | | | | | | | | | |
| VF | | | | | | | | | | | |
| VFger. | | | | | | | | | | | |
| VFinf. | | | | | | | | | | | |
| SS | | | | | | | | | | | |
| ro | | | | | | | | | | | |
| SS inc | | | | | | | | | | | |
| VT | | | | | | | | | | | |
| PP | | | | | | | | | | | |
| SPG | | | | | | | | | | | |
| WF | | | | | | | | | | | |
| WC | | | | | | | | | | | |
| adv. | | | | | | | | | | | |
| S/PL | | | | | | | | | | | |
| C/NC | | | | | | | | | | | |
| ? | | | | | | | | | | | |
| AWK | | | | | | | | | | | |
| W O | | | | | | | | | | | |
| C | | | | | | | | | | | |
| P | | | | | | | | | | | |
| omit | | | | | | | | | | | |
| ^ | | | | | | | | | | | |
| ¶ | | | | | | | | | | | |
| **Total** | | | | | | | | | | | |

**Appendix D: Native-Chinese Speaker Essays**

**Text 1**

How to become a good leader? Different people have their different answers. But in my opinion, a good leader should have courage, intelligence, and the sense of responsibility.

First, courage is essential to a leader. A good leader must be calm and confident whenever he is. For example, when a leader meets something difficult, he could not give up. Because he represents the whole image of a team. If he lacks the courage and give up, his partners will feel inconfident. What worse, the whole team will do not know what to do when their leader loses the courage to solve problem. At the end, the leader with his mates will fail together. Because of this reason, a leader must have enough courage to deal with anything.

Second, intelligence is also an important characteristic of a good leader. A person who just knows do the one thing again and again is not a leader at all. A good leader must know how to use different ways to maximize the benefit. Just like when a leader meets a good deal. Most time, an ordinary leader will just finish it. But a good leader will not do this. Conversely, he will think how to make more profits from this deal. He will bargain with the other side or he will arrange his mates to do research about the deal. If he does all of these things, he will benefit his team more with his intelligence. Obviously, intelligence is important to a good leader.

Finally, a good leader cannot lack the sense of responsibility. He, as the leader of others, ought to undertake more responsibility than others. Whenever his team meets troubles, he must take action firstly. He also has the responsibility to protect his mates if he wants to get support from them. In a word, he must to have a sense of responsibilities which include what he ought to do and what he should do.

In conclusion, it is not easy to be a good leader. If you are interested in this, you have to do something now. Have a good courage, intelligence, and sense of responsibility, you will make it.

**Text 2**

A good leader can lead an organization to success. He can bring encourage to other people. Therefore, a good leader is very important. In my opinion, a good leader should be a rational person with high EQ. Also he should be a smart person. Meanwhile, he must be a brave person who can face to the difficulties bravely.

As a leader, he usually needs to deal with many different kinds of problems. Sometimes, there are many difficulties coming together, the leader should deal with them patiently. In this situation, EQ is very important that can help you keep rational. Only be rational, people can make a good decision. Nowadays, the society is complicated. If people want to lead an organization to success, they should be calm to face everything. On the other hand, when the people see their leader is calm down, they will be confident to face the series of things.

As a leader, he should be smart. Wisdom can help people to figure out difficulties. Sometimes, the problems people facing cannot be settled by EQ. People need to think about a new way to settle these problems. At this moment, a leader should come forward and use his wisdom help his group to overcome those difficulties.

A leader should be a brave person because he stand in front of his group. When storm is coming, he is the first person to face storm. A brave person can burden more responsibilities. The organization needs the leader to take the responsibilities. Therefore, leaders cannot only enjoy the successful feeling, but also do more for their groups.

In a word, a good leader is the core of an organization. He should be excellent among the people. High EQ, high IQ can help him to be a leader. However, he should also be a brave person. When the leader does not fear anything, he can use his EQ and IQ freely. That is the characteristics of a good leader.

**Text 3**

For most people, to be a good leader is their dream when they were young. Some people devote themselves to this dream, and they work hard, get higher education, and even spend more time to get special training. However, it is not enough for someone who eager to be a good leader to study more. In general, a good leader needs to have excellent characteristics and abilities.

First, one of the most important characteristics of a good leader is communication skills. Obtaining a good communication skill is not as easily as talking to employees. Different people have different personality because of their own backgrounds, experiences, and emotional concern. Therefore, they may have distinctive thoughts and behaviors. A good communicator can use various methods to communicate with different people based on their personality. Usually, everyone in a team has different responsibility. For a good team leader, it is wise for him to be familiar with each person's role, so that he can control the communication. What's more, a leader can show his careness of his group members. By doing this, he will gain the reputation.

Second, not only good communication skill can help a person to be a good leader, but also the ability and the eager to learn. Only when a person can take all the works he assigns to the others, he has the ability to lead the whole team. As known to all, changes are happening every day in the world. Hence, it is significant for a leader to update his knowledge very often. If

a leader does not keep in touch with the development of the society, it is impossible for his team to make progress.

In conclusion, the process to be a good leader is not easy. Only a person with patience and strong faith can achieve this goal. During this period, different skills and ability play an important role. Even for those people who are leaders, it is necessary for them to update their thoughts and ways to operate the whole team.

**Text 4**

Recently, with the development of the economy and technology, there are more and more new discoveries. Telephones and computer are two of these new discoveries. Nowadays, it is common to see people use telephones and email to keep touch with their families and friends. Some people said that telephones and email have made communication between people less personal. And in my opinion, I totally agree with the thought. My view will be greatly confirmed by the following discussion.

In the first place, it is know to all that in the past days when there was no modern technology, people had to use animals such as horses, birds to send mails to families or friends who lived far away from them. The mails like these were only for the people who receive them. There even were some organizations for transporting the mails for people. People could not communicate with their families or friends directly; but there are rules and laws to protect the benefits of people who send the letters, nobody can know the contest of their letters. This is the situation that people faced about communication in the past.

In the second place, even though we can write letters to families or friends, the problem is that we cannot send the letters by ourselves. So we have to ask for help to the post office. We buy the stamps and then the postmen will send our letters to the person we want to. It is no use if

we try to finish the communication without help. This is not a personal thing and there are lots of people working on your communication.

However, since the telephones and email were discovered, the communication between people is more convenience. People can talk to each other at anytime, anywhere. When we talk to each other on the phone, we don't know whether people around us are listening. When we use email to communicate, we don't know whether there are virus on the internet which can steal our emails. We cannot talk about secrets about ourselves on the phone or email.

In a word, the telephones and emails are not as safe as we expected. Compare with the old communication ways, they are less personal; because we are not sure about people around us.

**Text 5**

Since the technology revolution, a great amount of new inventions appeared, such as automobile, telephone, and the Internet. With the help of these new advances, people's lives have experienced abundant changes. Recently, the topic that whether the new technology advances bring people closer or pushing them apart attracts more attention. This is a controversial topic. Some people may believe that these inventions bring convenience to their daily lives; meanwhile, others may insist that the telephone and the Internet decrease communication between people. Actually, by using these innovations, people are able to communicate with each other more easily than before.

First, it is known that the main transportation was coaches before the invention of automobile. It was not convenient for people to travel around. When their families or friends lived far away from them, it was not possible for them to visit them frequently. Without the Internet and telephone, the only way for people to keep in touch with each other was letters. However, it usually took several weeks or even months to mail letters and receive the responds.

Therefore, there were not a lot of chances for people who lived in the old time to communicate. However, after the technology revolution, people have more choices to contact. They can use telephone, the Internet to chat with each other. In addition, they are able to use modern transportations such as car and airplane to travel around the world. People live with new technology advances are closer.

Second, technology inventions also brings benefits to business world. Generally, businessmen do business around the world were not able to see each other in the past. They were strangers who were not similar with each other. Therefore, it was possible that these businessmen canceled their cooperation because of distrust. The main reason was they lacked of communication. In contrast, in modern society, the advanced transportations and communication tools make it possible to talk and see each other. Businessmen get the chance to negotiate, and it increases the opportunities to work together. Thus, the new technology plays an important role in the development of business relationship.

In conclusion, these technology inventions bring revolution to whole society. No matter families and friends or business cooperators obtain more chances to contact with each other directly. They are able to share their opinions and suggestions face to face. By doing this, their relationships are improved, and people are closer together instead of apart.

**Text 6**

What makes a good boss? Different people may come up with various answers. For me, a good boss should have three qualities. He should make right decisions for the whole team; he should be patient; he should know how to solve conflicts between employees.

First of all, a good boss should make right decisions for the whole team. This is priority of all the qualities because a boss is like a head of a whole body. He or she controls every part of

the "body". In a company or team, he controls every department or employee. When situations are complicated, the boss should keep a clear mind and make a wise decision. This is easier said than done. If a boss cannot keep a cool mind, the company will probably collapse.

The second quality a good boss should have is patient. Sometimes, employees are not able to understand what their boss wants them to do. If the boss just shout at them, things will only become worse. Wise boss always explain their ideas and intentions again and again until the employees aware the boss's whole plan. On the contrary, if the boss just criticize their work, the feelings of employees to the boss will turn to be extremely negative.

Third, good bosses should know how to solve conflicts between employees. When a huge number of employees work in a company, there are always numerous conflicts. If those employee come from different countries or have different background, their business ethics vary. In that situation, the boss should know how to solve these problems and unite them together. If the boss cannot deal with this kind of problems, the company cannot run on well.

In conclusion, a good boss should at least have three qualities: making right decisions, being patient; knowing how to solve conflicts between employees. There might be other qualities that a good boss should have; however, these three quality are the most important. If someone can come across a boss with all three qualities, he is very lucky.

**Text 7**

In my opinion a good leadership, a good communicate skill and a loving heart are very important for being a good boss.

First at all, a good boss needs to know how to lead his group. As the leader of the team, he need to perceive the right direction for team and know how to get everyone in the team work together in one to achieve the goal. He is the one need to know what are the special talents and

potential ability of each individual in his team and what is the strength and weakness of his whole team. Hence, a good leadership is one of the most important attribute of a good boss.

Second, a good boss knows how to communicate with his team. Having a great goal is not enough. A good boss needs know how to help everyone to understand the goal and how to do it together. A good communicate skill helps him to encourage his team, so that people will feel confident and satisfy with their career. And also, it will help him to know how to talk to his team, so that people will understand what they need to improve.

Last, having a loving heart is very important for being a good boss as well. A boss who has a loving heart really care of the people he work with. The team will be full of love because this. As a result, people who work in his team will feel degraded and happy. It makes them want to work even harder. It makes them feel that the team is their family. The satisfaction from this sometimes can be more meaningful than money for them.

Therefore, A good leadership, a good communicate skill and a loving heart make a good boss.

**Text 8**

Government is the most powerful department of a country. It has the power to decide almost everything. They decide the common people's daily lives and the trade business at home and all over the world. The government's decision influence the develop of a country. So it is very important to build a good government system.

First of all, a good government system should benefit the common citizens. Every country has its common citizens and they are the crucial part of a country. Government should provide the basic insurance of every individuals such as medical and education. For instance, if people are sick, they should have the abilities to see a doctor and cure the disease. That is

because people only have a good health, they can do anything they want. Moreover, education is very important to people. If everyone can get a good education, it will enhance the quality of a country. Therefore, government only benefits their citizens, so they can work for their countries and contribute to their countries.

Second, a good government system should bring interests to the country. Government people should know how to do business with others, but I do not just mean to earn money. The most important thing is to keep a good relationship with others. For instance, everyone may meet difficulties. At that time, a country need the help from other countries and people. Also when the other countries get into troubles, government should provide help to them. It can help to keep a long-distance relationship. In addition, they can exchange communication and develop together. Thus, government should bring interests to develop the country.

All in all, a good government should try its best to satisfy everyone and develop the country.

**Text 9**

Many parents allow their children to watch television, but many doctors argue that children should never be allowed to watch television. However, there are no definitely right answer on this point. Television has some benefits to children, at the same time, it also exists some drawbacks.

On the one hand, television has some advantages to the growth of children. For instance, children can learn some knowledge from T.V., which they cannot study from books. Some T.V. show can broaden their horizon and create their imagination. Moreover, they can watch television with their friends and families. They can talk about what they watched and what they

learned. Also, they can enjoy the feeling of staying with families and friends. Therefore, television brings a lot of fun to children.

On the other hand, television still have some disadvantages to children. For instance, television can influence the health of children. If children watch T.V. for a long time, they may feel headache and it will lower their sights. Also television can influence their study. Some children spend too much time on the television, then they start hating doing homework. In addition, some T.V. program is not good for children to watch such as violent and porn. Although the government are trying their best to reduce those kinds of program, they are still exiting inevitably. Thus, television sometimes is not good for the growth of children.

All in all, people cannot deny the benefits of the television and they also cannot avoid the drawbacks of the television. So what they can do is to reduce the time of watching television and prevent them from watching the unhelpful programs. With the efforts of parents and doctors, they can provide a good environment for children and let the children make good use of television.

**Text 10**

In 21st century, with the fast development of computer science, people's life style has been significantly changed. Computer is used everywhere in our life such as school, office, home and so on. Some people have the opinion that computers have made life more complex and stressful. However, others insist that computers have made life easier and more convenient. As far as I am concerned, the advantages of computer science are overweight disadvantages because computers can make science develop faster and bring us entertainment.

First of all, the more computer science develops, the faster science can be improved. As is known to all, some science research require extremely complex calculate, which is impossible

for people to do. Scientists have to rely on the computer to do the calculating. For example, the first computer in the world is used to calculate the science problem. In fact, at that time it is not called "computer". It is so-called "super calculator". After around 40 years' development, computers are gradually used as personal computer. Therefore, the development of computer can make the science develop faster.

Second, the development of computer science make our life more enjoyable. Recently, a survey show that more and more people using computer in their leisure hour instead of using TV or reading. Besides that, the age of people start to play computer is becoming younger. Many people think computer can bring more entertainment than others. For example, through the computer, people can enjoy the music, watch movie, and reading book. In addition, they can surf the internet with the help of computer. Therefore, computers like super play stations.

In conclusion, computers have made our life easier and more convenient. Without computers, it is hard to imagine how to develop science. Without computers, it is difficult to enjoy the internet charting at home. Without computers, our life will changed a lot.

**Text 11**

Nowadays there are many companies. Some companies are success, some are not. It is really important to have a success company if it has a good boss. Being a good boss, there are several ways: kindness, responsibility, and good working ability.

The most important quality of being a good boss is kindness. Kindness can affect employees positively that they will be willing and happy to work. For example, if the boss is always smiling and saying kind words to the employees, he makes the employees feel happy and not feeling nervous. With the happy feeling they can work very well and they also can be willing to work for the company.

Another important quality of being a good boss is responsibility. With good responsibility, the employees can work without worrys. They can trust the boss therefore the qualities of their work can be improved positively.

The last quality of being a good boss is a good working ability. To be a leader of a company is as important as a core of a tree. A good ability boss can help the employees to work as well as they can. Therefore with all the employees work well skill used, the company will be success.

Kindness, responsibility and good working ability are the important quality of a good boss. People work with a boss how has all these qualities can help the company work as good as it can be. Therefore, if you want to be a success boss, learn those qualities first!

Text 12

People live in a society; therefore, we need a government to organize our society. There are hundred of governments in this world. However, not all of them are good governments. What is a good government? In my opinion, a good government requires three main qualities, which are providing people good education, taking care of people, and developing the essential buildings.

First, a good government should provide people a good education system. A good education system will help people to get education and training, which they need. In fact, a education system includes kindergartens, elementaries, high schools, colleges and universities. Schools help people to learn what they are interested in and develop their skills, so that in future they could earn by what they learn. In sum, it is government's responsibility to have a good education system.

Second, a good government will take care of people in several ways. For example, a good government will provide them a good financial environment so that people can have jobs.

Furthermore, a good government will have a good medical system, even for poor people. The society is unfair; however, the government can play a role to balance this situation. In additional, a good government will provide a good environment for people to live in by reducing criminals.

Third, a good government will revolate in developing the essential buildings. These establishments need the power of the government. They need to have a scheme and lead people to a better level. For example, the transportation system, such as freeways, trains, subways, buses, and airplane all need to be planned under governments.

In summary, a good government should be a reliable government and have a good regime by educating people, protecting people, and establishing structures. Japan and Germany are good governments because people are happier.

**Text 13**

When people use "poor", "isolated" and "terrify" to describe a country, do you think which country it is? Most people's answers will be North Korea. As is known to all, the terrible situation in North Korea is due to its government system. Then, the question becomes: how can we avoid this and what are the characteristics of a good government system? Indeed, with different education background, different people might come up with various answers. However, when we review the history of human society, we will discover that three main factors build a healthy government system.

To start with, a healthy government system must be democratic. The democracy is not perfect, yet compares with other political systems, it has the fewest disadvantages. In other words, it is the least worst system so far. Democratic system gives every citizen the right to vote and choose the best candidates, which represents the willing of the majority.

Furthermore, powers have to be separated in a good government system. The three main

powers, which are administrative power, legislative power and jurisdictive power should be controlled in different group of people. The separation of powers is the most effective way so far to prevent a political leader becoming a dictator because all his decisions have to be checked by other officials.

Finally, a good government system requires no one is above the law, which means everything must be ruled by law, not by personal will. Every official, especially the high level official, has to obey the rule. In addition, if she or he breaks the law, no matter what her or his status is, he will receive a fairly conviction from the court.

In conclusion, a good government system must be democratic, the three kinds of power should be separated, and no one is above the law. If a government can fulfill all these requirements, it will not be too negative. However, if a government can not achieve any one of them, like the North Korea government, it is or will definitely be an unhealthy one.

**Text 14**

There are many great inventions in twenty-first century; the television is one of them. A quote said, "A television is like a magic box, which can take you to everywhere you want to." Therefore, many people love watching television. However, there are many problems increasing due to watching television. Some people think television helps them to relax; also they can learn something from the educational channel. On the other hand, I think television has bad influences for children for three reasons: providing common value, disturbing interactional behavior, ruining a life route.

First, the television creates a common value and trade. It is a very bad influence for children because they do not know how to evalue and judge, but just follow the evaluation which was provided from the television. For example, a channel spreads ideas, which are that tattoo is

cool, homosexual is acceptable, and sexual relationship before marriage is fine. If fact, it is not right, but after television spreads this kind of value, people think that it is nothing.

Second, the television makes people do not know how to interact well. People who watch too much television are like to image. However, imagination is not real; we live in a real world. It is good for us to have some imagination, but if we live in imagination, it will be bad for us. Sometimes, children will act like actors on television.

Third, the other bad influence of the television is we watch too much television and do not organize time well. In fact, when I was a teenager, I used to watch television for five hours after school. Sometimes, I watched too much television, so that I forgot time and did not do my homework or prepare for the test. Therefore, I got a bad grade and made my parents angry.

In summary, television is a great invention in twenty-first century, but it has three main bad influences on children, which are providing a common evaluation, interrupting children's interaction, wasting time on watching television. There are still some advantages on television, such as educational channel, learning channel, discovery channel.

**Text 15**

Doctors have believed that children should never be allowed to watch television. However, are doctors' suggestions always correct? Is watching television so harmful that we should never allow children to watch it? The answer is definitely no. In fact, watching television can help children be aware of the culture and help them to learn some knowledge that never demonstrated on text books as long as the parents guide them.

First, watching television is a useful way to know the culture. In our daily lives or the academic researches, we always can hear someone's talking or find somebody's quote related to the television program or advertisements we watch on T.V. Watching the T.V is a necessary

method to acquire these information. We can imagine that if a person never watch TV, he will be isolated when his friends talks about any TV related cultural events with him.

Moreover, many excellent TV programs will help children to learn various kinds of knowledge. Unlike the tedious text books, TV programs are more interesting and attractive to children, so children can learn knowledge when they have fun. For example, children can form interests to natural science or social science from watching the National Geographic, Discovery Channel and different kinds of news channels. The key point is that the parents should suggest appropriate programs to them.

To sum up, watching T.V is necessary for children's grows. Indeed, we do not deny the fact that being a coach potato is definitely not a right choice for children, but the question is: should we completely abolish T.V because of its negative effects? Of course no! Doctors give these suggestions only based on the healthy concern, but as parents, we should keep a clear mind. Every coin has two sides. Watching TV is the same. How to use the double-edged sword is the key. Thus, guiding the children to choose the right programs is what a responsible parent should do. Do not be a lazy parent.

**Text 16**

2012 is coming; Most of the medias in the United States are busy for the election of presidential election of 2012. American voters are so zealous for their presidential election. Coincidentally, the presidential candidates are busy on their campaign. American government system is a good example in the world because it is democratic and monitorable.

The U.S. government is a democratic government. All the American voters who aged 18 and above have their right to vote for their government. In the United States, either the federal government or state governments are elected by their voters. The government has to be voted

every four years. Therefore, the government has to perform as it promised before getting voted. Otherwise, it will loose in the next election. Such kind of government mostly will perform their duty on the behalf of their voters---all Americans.

The U.S. government separates their power of legislation, judgement and execution. Those divided parts of power could monitor each other. This kind of government system can minimize the corruptions in a more effective away. Also, there are two dominant parties in the United States to compete each other in campaigns for their power. They always monitor each other. Either one of them tries to perform better for winning the elections. This government system is monitored by voters and other parties. Therefore, it can minimize the corruptions. It just that saying" a government without monitors, it will be a corrupted government", but Americans are doing well.

As a conclusion, American government is elected by all American voters and monitored by all the voters. Therefore it is a good government.

**Text 17**

A good government is very important for citizens. A good government can bring different benefits for citizens. They can protect their citizens if they can trusted by their citizens. A good government should for the people and by the people. There are some important characteristics for a good government system. Such as, the good government should give people different rights, also, the good government should consider people's situation, and they should stand by people's side.

Firstly, a good government should give citizens different rights. Such as, speech rights, human rights and religion rights. For example, some governments never give the speak right to their citizens. Citizens cannot judge their government or they cannot say their opinion in public

area. Speak right of citizens can help governments to change their mistakes. Also, there are many different rights. Religion right is also important for human. Everyone should have their own religion. They should believe somethings. There are some research about whether people should have religion or not. The research shows that human have to have a religion, because religion can balance people's life and religion can change people's behavior. However, many governments whose political system is community party forbid people to believe religion.

Secondly, the good government should consider about their citizens. The good government should stand by the citizens' side. They should consider about what citizens want to need. They should not change their systems around by themself. A good government should trusted by citizens. For example, Beijing hold Olympic game in 2008. The government paid amount of money to hold this game. Because the government want to other countries to saw a strong country. The government thought this was an opportunity to show China. However, there are many people who lose jobs and are homeless. The government can use this money to help their citizens. The stand of strong country is not how to show itself. It is how to help people to change their situation.

To sum up, a good government should not forbid citizens' rights. A good government should change citizens' situation. They should consider about what citizens need.

**Text 18**

How to be a good leader? How to lead your company to earn a lot of many? It's a dilemma question. Some successful man isn't a good leader. He is just a good manager. However, a boss has good leadership will lead his company set up a good program to success. In the following, I will provide some reasons and instances to tell you how to be a good leader.

First of all, he needs to have a good judgment. It will help him to overcome any troubles in different situation. For example, economy problem is a huge problem in recent year. When you are a boss, you need to find the method to solve it. However, it is too late when you encounter it. A good leader will predict this situation first and prepare well to challenge this problem. He will find the question first and do something when the problem really happen.

The second quality of a good leader is that he is also good speaker. When he wants to do some project, people always choose to believe him. For instance, in the basketball team, the team leader is very important role. He not only need to communicate between coach and players, but also lead this team to win the game. When they lost the game in the half time, a good team leader will say something to encourage his teammate. He also follow the coach's order. His speech is like a magic to make his teammate together. It will finally to lead his team to win the game.

In conclusion, there are two characteristics that how to be a good leader. He is a good speaker and he can predict the problem first. When you have this quality, you will have enough abilities to face any challenge.