Lang. Teach. (2013), **46.2**, 256–271 © Cambridge University Press 2013
doi:10.1017/S0261444812000547

Thinking Allowed

Advancing research in second language writing through computational tools and machine learning techniques: A research agenda

Scott A. Crossley Georgia State University, Atlanta, GA, USA
scrossley@gsu.edu

This paper provides an agenda for replication studies focusing on second language (L2) writing and the use of natural language processing (NLP) tools and machine learning algorithms. Specifically, it introduces a range of the available NLP tools and machine learning algorithms and demonstrates how these could be used to replicate seminal studies in L2 writing that concentrate on longitudinal writing development, predicting essay quality, examining differences between L1 and L2 writers, the effects of writing topics, and the effects of writing tasks. The paper concludes with implications for the recommended replication studies in the field of L2 writing and the advantages of using NLP tools and machine learning algorithms.

1. Introduction

A key component of L2 proficiency is the ability to communicate ideas through writing. Writing in an L2 is an important skill for students interested in general language learning and professionals interested in English for specific purposes, such as business, science, and law. From a student perspective, writing at the sentence and discourse level is a key skill for conveying knowledge and ideas in the classroom. Writing skills are also important components of standardized assessments used for academic acceptance, placement, advancement, and graduation. For professionals, writing is an important instrument for effective business communication and professional development.

Communicative and pedagogical issues are at the root of L2 writing research. The pedagogical focus differs, depending on the role of culture, L1 literacy development, language planning and policy (Leki, Cumming & Silva 2005; Matsuda & Silva 2005), specific purposes (Horowitz 1986), and genre (Byrnes, Maxim & Norris 2010), but a vital element of pedagogy is still a focus on the written word and on the degree to which the combination of written words produces the intended effect on the audience: in other words, how well the text communicates its content. Thus, fundamentally, it is the quality of the text that learners produce – as judged by the reader – that is central. Of course, how the L2 writers arrive at these words and their combination (via the writing process and the sociocultural context in which they are learning) is also important; however, such considerations are probably unknown and in any

https://doi.org/10.1017/S0261444812000547 Published online by Cambridge University Press

case irrelevant to the reader, whose interest lies in developing a situational and propositional representation of an idea or a narrative from the text.

The situational model of the text develops through the use of linguistic cues related to the text's situational model, that is, its temporality, spatiality, and causality (Zwaan, Langston & Graesser 1995). Its propositional meaning is arrived at through the lexical, syntactic, and discourse units found within it (Just & Carpenter 1987; Rayner & Pollatsek 1994). Traditionally, L2 writing researchers have examined propositional meaning in students' writing for a variety of factors, including longitudinal writing development (Arnaud 1992; Laufer 1994), predicting essay quality (Connor 1990; Ferris 1994; Engber 1995), investigating differences between L1 and L2 writers (Connor 1984; Reid 1992; Grant & Ginther 2000), examining difference in writing topics (Carlman 1986; Hinkel 2002; Bonzo 2008; Hinkel 2009), and writing tasks (Reid 1990; Cumming et al. 2005, 2006). Fewer studies have investigated how situational models develop in L2 writing (cf. Crossley & McNamara 2009).

Many of the studies mentioned above provide a fundamental understanding of the linguistic development of L2 writers, how they differ linguistically from L1 writers, and how prompt and task influence written production. These studies are not only fundamental to our understanding of such things as L2 writing development, writing quality, and writing tasks, but are also prime candidates for replication (Language Teaching Review Panel 2008; Porte 2012; Porte & Richards 2012). Replication of these studies from a methodological standpoint is warranted because recent advances in computational linguistics allow for a wider range of linguistic features, measuring both situational and propositional knowledge, to be automatically assessed much more accurately. The output of these tools can also be analyzed using machine learning techniques to predict performance on L2 writing tasks and provide strong empirical evidence about writing development, proficiency, and variation. Such tools and techniques afford not only approximate replications of previous studies, but also constructive replications that query a wider range of linguistic features that are of interest in L2 writing research.[1] The purpose of this paper is to provide a research agenda that combines L2 writing research with newly available automated tools and machine learning techniques.

1.1 Natural language processing

Any computerized approach to analyzing texts falls within the field of natural language processing (NLP), which centers on the examination of how computers can be used to understand and manipulate natural language text (e.g., L2 writing texts) to do useful things (e.g., study L2 writing development). The principle aim of NLP is to gather information on how humans understand and use language through the development of computer programs created to process and understand language in a human-like manner.

[1] Approximate replications involve close duplication of methods, while constructive replications use new methods or designs to verify original findings. See Language Teaching Review Panel (2008), Porte (2012), and Porte & Richards (2012) for more information.

https://doi.org/10.1017/S0261444812000547 Published online by Cambridge University Press

A variety of NLP tools, recently developed for English, are available either free or for a minimal fee and require little to no computer programming skills. These tools include Coh-Metrix (Graesser et al. 2004; McNamara & Graesser 2012), Computerized Propositional Idea Density Rater (CPIDR; Brown et al. 2008), the Gramulator (McCarthy, Watenebe & Lamkin 2012), Lexical Complexity Analyzer (LCA: Lu in press), Linguistic Inquiry and Word Count (LIWC: Pennebaker, Francis & Booth 2001; Chung & Pennebaker 2012), L2 Syntactic Complexity Analyzer (Lu 2011), and VocabProfiler (Cobb & Horst 2011). These are discussed briefly below. For a complete summary of each tool, please see the references above.

1.1.1 Coh-Metrix

Coh-Metrix is a state-of-the-art computational tool originally developed to assess text readability, with a focus on cohesive devices in texts that might influence text processing and comprehension. Many of the linguistic indices reported by Coh-Metrix therefore measure cohesion features (e.g., incidence of pronouns, connectives, word overlap, semantic co-referentiality, temporal cohesion, spatial cohesion, and causality). In addition, Coh-Metrix reports on a variety of other linguistic features important in text processing. These include indices related to lexical sophistication (e.g., word frequency, lexical diversity, word concreteness, word familiarity, word imageability, word meaningfulness, word hypernymy, and word polysemy) and syntactic complexity (e.g., syntactic similarity, density of noun phrases, modifiers per noun phrase, higher-level constituents, words before main verb). An online version of Coh-Metrix is available free at http://cohmetrix.memphis.edu/cohmetrixpr/index.html.

1.1.2 Computerized Propositional Idea Density Rater (CPIDR)

CPIDR measures the number of ideas in a text by counting part-of-speech tags and then derives, using a set of readjustment rules, the number of ideas. CPIDR reports the number of ideas and the idea density (calculated by dividing the number of ideas by the number of words). CPIDR is available free at www.ai.uga.edu/caspr.

1.1.3 The Gramulator

The Gramulator reports on two key linguistic features of a text: the frequency of n-grams (contiguous sequences of two and three words) and lexical diversity. The Gramulator calculates the frequency of n-grams in two sister corpora in order to identify n-grams that differentiate between them. Specifically, the Gramulator identifies the most commonly occurring n-grams in contrasting corpora and retains those that are typical of one corpus but atypical of the other. The Gramulator also calculates a variety of sophisticated lexical diversity indices (e.g., MTLD, HD-D, M; see McCarthy & Jarvis (2010)) that do not correlate strongly with text length. The Gramulator is free to download at https://umdrive.memphis.edu/pmmccrth/public/software/software_index.htm.

https://doi.org/10.1017/S0261444812000547 Published online by Cambridge University Press

1.1.4 Lexical Complexity Analyzer (LCA)

LCA computes 25 different indices related to lexical richness (i.e., lexical sophistication and lexical density). These include sophisticated words (infrequent words), lexical words (content words), sophisticated lexical words (infrequent content words), verbs, sophisticated verbs, nouns, adjectives, and adverbs. The Lexical Complexity Analyzer is available free at http://aihaiyang.com/synlex/lexical/.

1.1.5 Linguistic Inquiry and Word Count (LIWC)

LIWC is a textual analysis program developed by clinical psychologists to investigate psychological dimensions expressed in language. LIWC reports on over 80 lexical categories that can be broadly classified as linguistic (pronouns, tense, and prepositions), psychological (social, affective, rhetorical, and cognitive processes) and personal concerns (leisure, work, religion, home, and achievement). LIWC is available for a modest fee at www.liwc.net.

1.1.6 L2 Syntactic Complexity Analyzer (L2SCA)

The L2SCA was developed to measure a range of syntactic features important in L2 writing research. The measures can be divided into four main types: length of production, sentence complexity, subordination, and coordination. The L2SCA is free to download at www.personal.psu.edu/xxl13/downloads/l2sca.html.

1.1.7 VocabProfiler

VocabProfiler is a computer tool that calculates the frequency of words in a text using Lexical Frequency Profiles (LFP) developed by Laufer & Nation (1995). VocabProfiler reports the frequency of words in a text using the first 20 bands of families found in the British National Corpus (the earlier version developed by Laufer & Nation reported on the first three bands only). An online version of VocabProfiler is available free at www.lextutor.ca.

1.2 Machine learning algorithms

As the size and variety of written corpora continue to grow, the amount of information available to the researcher becomes more difficult to analyze. Techniques are needed for automatically extracting meaningful information from these large and diverse corpora and discover the patterns that underlie the data. Replication research in L2 writing should therefore include the use of not only advanced computational tools, but also machine learning techniques that can acquire structural descriptions from corpora. These structural descriptions can be used to explicitly represent patterns in the data, in order to predict

https://doi.org/10.1017/S0261444812000547 Published online by Cambridge University Press

outcomes in new situations and explain how the predictions were derived (Witten, Frank & Hall 2011).

The output produced by the tools discussed above can be strengthened through the use of advanced statistical analyses that can model the human behavior found in the data. These models usually result from machine learning techniques that use probabilistic algorithms to predict behavior. The statistical package that best represents these advances and is the easiest to use is currently the Waikato Environment for Knowledge Analysis (WEKA: Witten, Frank & Hall 2011). WEKA software is available free from www.cs.waikato.ac.nz/ml/weka and allows the user to analyze the output from computational tools that use a variety of machine learning algorithms for both numeric predictions (e.g., linear regressions) and nominal classifications (e.g., rule-based classifiers, Bayesian classifiers, decision tree classifiers, and logistic regression). WEKA also allows users to create association and clustering models.

1.3 The intersections of L2 writing, NLP, and machine learning algorithms

We thus find ourselves at an interesting point in L2 writing research. We currently have available large corpora of L2 writing samples such as the International Corpus of Learner English (ICLE: Granger et al. 2009). We also have a variety of highly sophisticated computational tools, such as those mentioned above, with which to extract linguistic data from those corpora. Lastly, we also now have powerful machine learning techniques with which to explore this data. All of these advances allow us to replicate and expand a variety of studies that have proved to be important in our understanding of L2 writing processes and L2 writing development. In this paper, I will focus on a small number of influential studies related to assessing longitudinal growth in writing, modeling writing proficiency, comparing differences between fluent and developing L2 writers, and investigating the effects of prompt and task on L2 writing output. In each case, I will present earlier research on the topics and discuss the implications for recent technological advances in replicating and expanding these research areas.

2. Longitudinal studies of L2 writing

A variety of studies have attempted to investigate the development of linguistic features in L2 writing using longitudinal approaches (Arnaud 1992; Laufer 1994). Longitudinal approaches to understanding writing development are important because they allow researchers to follow a small group of writers over an extended period of time (generally around one year). While the power of the analysis is limited by the small sample size, longitudinal analyses provide an opportunity to analyze features that develop slowly, such as lexical networks (Crossley, Salsbury & McNamara 2009, 2010) or syntactic competence. Longitudinal studies also provide opportunities to examine growth patterns in more than one learner to see if they have any developmental trends in common.

One of the most frequently cited longitudinal studies of L2 writing is that of Laufer (1994), which investigated the development of two aspects of lexical richness in L2 writing: lexical

https://doi.org/10.1017/S0261444812000547 Published online by Cambridge University Press

diversity and lexical sophistication. Laufer's index of lexical diversity was a simple type/token ratio score, while her indices of lexical sophistication were early LFP bands (two bands that covered the first 2,000 word families in English), the university word list (UWL: Xue & Nation 1984), and words contained in neither the LFP bands nor the UWL. The data for the study came from 48 university students, who wrote free compositions at the beginning of the semester. These 48 students were broken into two roughly equal groups, one of which wrote free compositions at the end of the first semester and the other at the end of the second semester. Laufer then compared the essays written at the beginning of the semester to those at the end of the first and the second semester, using the selected lexical indices. Her primary interest lay in whether the writers showed differences in their lexical variation and lexical sophistication as a function of time. To assess these differences, she used simple *t*-test analyses.

The *t*-test analyses demonstrated that the lexical sophistication of the writers changed significantly after one semester of instruction and after two semesters of instruction: developing writers produced fewer basic words (words in the first 2,000 word families) and more advanced words (words beyond the first 2,000 word families). The findings for lexical diversity were not as clear, the students demonstrating significantly greater lexical diversity after one semester, but not after two. Laufer argued that the findings from the study demonstrated growth in L2 writing skills over time and indicated that greater emphasis should be placed on explicit lexical instruction in L2 writing classes.

Research task 1: Undertake an approximate or constructive replication of Laufer (1994)

Despite having been published some eight years ago, Laufer's study remains a solid representation of the basic methods and approaches used in longitudinal writing studies. It is also a prime candidate for approximate and constructive replication: first, because of the computational advances that have occurred in the last 20 years, and also because of problems with the lexical indices used by Laufer to assess lexical growth. The LFP bands she used are quite limited in scope (modern LFP bands as found in VocabProfiler assess 20 bands, each containing 1,000 word families) and may not be suitable for assessing word frequency production because of the information loss that comes with grouping words into families. Word families contain fewer distinctions than type counts and are naturally biased toward receptive rather than productive knowledge. Perhaps even more problematic was her use of simple type/token ratios to assess lexical variation: such indices are highly correlated with text length (McCarthy & Jarvis 2010), so it is possible that Laufer was measuring not lexical diversity, but the length of the students' writings.

New computational indices, available free, would be useful in an approximate replication of this study. For instance, the LFP bands reported by VocabProfiler provide greater coverage of the words in English and are based on much more representative corpora (i.e., the recent version of the BNC). However, these indices could be problematic because of the grouping approach, which diminishes lexical information and is more geared toward receptive vocabulary. Thus, the frequency indices reported by Coh-Metrix, which are count-

https://doi.org/10.1017/S0261444812000547 Published online by Cambridge University Press

based (i.e., not grouped into word families) may provide greater information on word frequency development in L2 writers. The lexical diversity indices reported by the Gramulator (MTLD, HD-D, and M) control for the text length effects found in simple type/token ratios and give a more accurate picture of the lexical diversity of text. An approximate replication study using these newer indices could provide additional support for the trends reported by Laufer (as they already have in spoken L2 production; see Crossley et al. 2009, 2010).

For a constructive replication, researchers may consider addressing other aspects of lexical richness and competence that were not considered in Laufer's study. For instance, the Lexical Complexity Analyzer reports on a range of indices of lexical density (an integral part of lexical richness) that could be used to assess the growth of L2 writers' lexical competence. Coh-Metrix also reports on other indices related to lexical richness (e.g., word familiarity) and lexical competence (e.g., word hypernymy, word polysemy, word meaningfulness, word concreteness, and word imageability). These indices could provide additional information about how the lexical aspects of writing skills develop.

3. L2 writing proficiency

Another important research area in L2 writing is the use of computational tools and machine learning algorithms to investigate L2 writing proficiency. In contrast to longitudinal studies, which investigate writing development, studies of writing proficiency assess the extent to which expert judgments of writing quality can be predicted by a variety of linguistic measures. Such investigations help to pinpoint which linguistic features are most likely to affect expert judgments of quality and provide a framework for capturing the proficiency of an L2 writer.

One of the more computationally robust studies to investigate writing proficiency is Grant & Ginther (2000), in which an automated tagging system was used to predict expert ratings of essay quality on a small L2 writing corpus. This study is a good candidate for replication because it looked at a wide variety of linguistic features, including text length, lexical specificity, cohesive devices, rhetorical features, grammatical structures, and syntactic structures. The data collection methods used were also sound, focusing on L2 writing samples produced in a standardized testing environment and holistically scored by independent raters. However, surprisingly, Grant & Ginther conducted no confirmatory statistical analysis of the findings, leaving interpretation of the study in doubt. This weakness, along with recent advances in computational tools, makes this study a prime candidate for replication.

In the Grant & Ginther study, 90 essays sampled from a larger corpus of essays written for the Test of Written English (TWE) were selected for analysis. The L2 writers taking the TWE were given 30 minutes to write an argumentative essay on a single prompt. The essays were scored on a holistic scale of 1 to 6 by two independent raters. Because not enough essays scored 1, 2, or 6, Grant & Ginther selected 30 essays that were scored 3, 30 that were scored 4, and 30 that were scored 5. The 90 selected essays were then automatically tagged by the Biber tagger (1988)[2] for length, lexical specificity (type/token ratio and word length), cohesive

[2] Because the Biber tagger is not available for public use, it was not selected as a potential tool for use in replication studies.

https://doi.org/10.1017/S0261444812000547 Published online by Cambridge University Press

devices (conjuncts and demonstratives), rhetorical features (hedges, amplifiers, emphatics, downtoners), grammatical structures (nouns, nominalizations, pronouns, verbs, models), and syntactic features (subordination, complementation, relative clauses).

Grant & Ginther conducted no confirmatory statistical analyses of the data reported by the Biber tagger and instead used only descriptive statistics (mean and standard deviation) to interpret the data. They interpreted the descriptive statistics as indicating that essays scored as higher quality by expert raters were longer and included more unique word choices, cohesive devices and rhetorical features, grammatical features, and complex syntactic features.

Research task 2: Undertake an approximate or constructive replication of Grant & Ginther (2000)

Without statistical analysis, such findings cannot be substantiated, so approximate and constructive replications of this study are needed. A major difficulty in such a replication is finding an appropriate corpus that consists of L2 essays holistically scored by expert raters. Often, researchers can approach the authors of the study to access the original corpora, but if this is not possible, similar corpora are available. For instance, the Educational Testing Service (ETS) will usually release a TOEFL iBT Public Use Dataset to qualified researchers. The dataset includes about 500 independent and integrated L2 writing samples that have been scored by expert raters. Alternatively, researchers can collect and score their own L2 writing corpora. Once the researcher has secured a database, an approximate replication study can be carried out, assessing linguistic features similar to those used by Grant & Ginther, such as text length (CPIDR), lexical specificity (the Gramulator and Coh-Metrix), cohesive devices (Coh-Metrix), rhetorical features (LIWC), grammatical structures (LIWC), and syntactic structures (Coh-Metrix and L2SCA). The data collected from these sources can then be analyzed using the machine learning algorithms in WEKA to provide confirmatory statistical analyses and assess the generalizability of the findings on a training set. The researcher could choose to investigate the expert scores as categorical functions (using machine learning algorithms such as logistic regression to classify the essays into groups based on scores) or as continuous variables (using machine learning algorithms such as linear regression to predict the expert scores).

A constructive replication would expand the number, variety, and type of linguistic features examined but focus on verifying the original findings. For instance, it would be interesting to know how the number of idea units contained in a text (as reported by CPIDR) relates to human judgments of quality. Likewise, it would be revealing to find out if the psychological properties of the words used in L2 writing (as reported by LIWC) influence such judgments. Lastly, Grant & Ginther examined a variety of linguistic features using a relatively simple POS tagger based on counts of surface-level features. Constructive replications could build on this original study by expanding the nature and depth of the features analyzed using the computational tools currently available. For instance, the lexical features reported by Coh-Metrix tap into much deeper attributes of lexical knowledge than simple type/token ratios and word length counts. Likewise, the syntactic structures reported by the L2SCA go beyond simple tags and began to examine clausal complexity, subordination, and coordination.

https://doi.org/10.1017/S0261444812000547 Published online by Cambridge University Press

4. Comparisons of L1 and L2 writing

A third research area that is important in L2 writing is comparisons of L1 and L2 writing, always bearing in mind that L1 written samples should be seen as a baseline for comparison and not an ideal. We cannot expect L2 writers to reach the fluency of L1 writers in most cases, but comparisons between L1 and L2 writers can give us a clearer understanding of the linguistic features of characterize L2 writing. These differences have important implications for writing assessment and instruction.

Numerous studies have been conducted on L1 and L2 writing differences in both the writing process and the writing product (see Silva 1993 for an early but thorough overview). Many of these have focused on differences in the use of cohesion devices (Connor 1984; Reid 1992) between L1 and L2 writers (and sometimes between L1 writers of different backgrounds). Perhaps the best study for replication is that of Reid (1992), in which she compared L1 and L2 writers' production of cohesion devices (pronouns, conjunctions, and subordinate conjunction openers) along with one indicator of syntactic maturity (prepositions). Reid collected the data for these linguistic features using an automated parser. Her corpus consisted of 768 essays, 540 of which were written by L2 learners from three different language backgrounds (Arabic, Chinese, and Spanish). L1 writers wrote the remaining essays (n = 228). The essays were written for two different task types (comparison/contrast/take a position and description/interpretation of a chart/graph) and on four different prompts, and were written under timed conditions (30 minutes). This experimental design allowed Reid to examine differences between L1 and L2 writing in general and between L2 writers from different language backgrounds. Here, I will focus on the former.

Using a series of ANOVAs, Reid reported that the L2 writers used significantly fewer pronouns than L1 writers and significantly fewer coordinating conjunctions. L1 writers also produced more prepositions than L2 writers, but no differences were noted in the production of subordinate conjunction openers. Reid argued that the increased use of pronouns by L2 writers may be symptomatic of the interactive prose that is common in oral language, indicating that L2 writers, unlike L1 writers, may lack awareness of their audience. Reid interpreted the differences in coordinating conjunctions as also demonstrating a reliance on oral communication on the part of L2 writers because natural conversation includes more coordinated structures, especially those using coordinating conjunctions. This is different from L1 writers, who are thought to alternate between cohesive devices. Lastly, Reid argued that differences in the use of prepositions probably signaled relationships between clausal constituents, and that the lesser use of prepositions by L2 writers indicated less clausal complexity.

Research task 3: Undertake an approximate or constructive replication of Reid (1992)

Reid's study is an excellent example of a comparison of L1 and L2 writing and one that demonstrates key differences between L1 and L2 writers. Within these differences there are important details that have proved useful for L2 writing assessment and instruction. However, the study requires replication because of the limited number of indices (four) available at the

https://doi.org/10.1017/S0261444812000547 Published online by Cambridge University Press

time. Current computational tools such as Coh-Metrix report on a wide variety of cohesion indices that go well beyond those reported by Reid, including lexical overlap, semantic co-referentiality, causality, spatial cohesion, temporal cohesion, anaphoric reference, lexical diversity, and a number of different connective and conjunctive indices. Very few of these have been tested for distinguishing L1 and L2 writing samples, with the exception of lexical overlap, semantic co-referentiality, causality, and spatiality (Crossley & McNamara 2009). In addition, Reid did not distinguish between personal pronouns and other types of pronouns, so she could not provide strong support for her arguments that L2 writers relied on interactive prose. These distinctions, however, are reported by LIWC. Thus both Coh-Metrix and LIWC enable approximate replications of this study.

Reid also investigated an index of semantic complexity, which she related to cohesion: studies have demonstrated that L1 writers first produce cohesive features (McCutchen 1986) and then move on to the production of more complex syntactic constructions (Haswell 2000). Such trends may demonstrate that the advanced writers began to use syntactic elements such as modification and embedding to draw implicit connections between ideas (Crossley et al. 2011), instead of explicit cohesive devices. Replication studies should therefore consider a broader range of syntactic indices not used by Reid, including the indices of syntactic complexity and syntactic similarity found in Coh-Metrix and L2SCA.

To address the basic research question about differences between L1 and L2 writers, constructive replication studies could investigate a wider range of linguistic features, including the psychological and rhetorical properties reported by LIWC, the lexical indices reported by Coh-Metrix, VocabProfiler, and LCA (see Crossley & McNamara 2009 for a constructive replication study using the Coh-Metrix lexical indices), and the n-gram indices reported by the Gramulator. Both constructive and approximate replication studies would benefit from the use of machine learning algorithms to detect and classify differences between L1 and L2 writers.

5. The effects of prompts on L2 writing

Another important research area in L2 writing is the effect of prompts on writing production. Hinkel (2002, 2003) demonstrated that writing prompts influence the linguistic output produced by L1 and L2 writers. We know that linguistic features are closely related to human judgments of essay quality, so it is important to understand the effects of prompts and how to control for prompt-based differences in writing assignments that may affect linguistic production and thus have implications for L2 assessment.

The study that best demonstrates prompt-based effects on linguistic production in L2 writing is probably that of Hinkel (2002), who investigated how different prompts influence the production of linguistic features for both L1 and L2 writers. Hinkel examined six different prompts related to parents, grades and learning, wealth, manner of instruction, opinion forming, and selecting a major. The prompts were modeled on those found in standardized tests such as the TWE, College Board, and the Scholastic Aptitude Test (SAT). Almost 1,500 students wrote essays in response to the prompts, including 242 native speakers and L2 learners from the following backgrounds: Chinese (n = 220), Japanese (n = 214), Korean

https://doi.org/10.1017/S0261444812000547 Published online by Cambridge University Press

(n = 196), Vietnamese (n = 188), Indonesian (n = 213), and Arabic (n = 184). All the L2 writers were of advanced proficiency.

In total, Hinkel examined 68 different lexical features: semantic and lexical classes of nouns (e.g., vague and enumerative nouns), personal pronouns, existential slot-fillers, indirect pronouns, verb tenses, verb aspects, semantic and lexical classes of verbs (e.g., public, private, and suasive verbs), modal verbs, participles, adjectives, semantic and lexical classes of adverbs (e.g., time, frequency, and place adverbs), noun and adjective clauses, adverb clauses, coordinating and logical conjunctions, and hedges. Hinkel used hundreds of Mann Whitney U tests to identify differences in the linguistic productions of L1 and L2 writers based on prompt difference. She reported that all the prompts led to significant differences in the linguistic features produced by the L1 writers and the L2 writers categorized by their L1. For instance, essays written on the manner of instruction prompt contained the fewest present tense verbs for the L1 writers and Korean and Vietnamese L2 writers. The opinion forming prompt led to a greater number of infinitives and, for Chinese writers, the fewest present tense verbs. Selecting a major verb elicited the most personal narratives and nominalizations. These prompts (manner of instruction, opinion forming, and selecting a major) also led to the highest rates of be-copulas for all but Arabic writers, phrase-level conjunctions for L1 and Chinese writers, and fixed strings for L1, Chinese, Japanese, and Indonesian writers.

Research task 4: Undertake an approximate or constructive replication of Hinkel (2002)

Hinkel's study demonstrates that prompts have a significant effect on writers' linguistic production. We know that linguistic features of writing influence human judgments of writing proficiency, so understanding prompt-based differences is an important area for replication. Specifically, in the case of Hinkel's study, replication studies should include a variety of linguistic indices that link directly to lexical sophistication, syntactic complexity, and text cohesion. While the features she selected are common in analyses of genre, many do not have strong links to writing theory, so from a linguistic perspective, studying features with stronger links to writing quality would help us understand how prompts affect writer production. Replication studies should include more standardized lexical features such as indices of lexical diversity and word frequency (as reported by Coh-Metrix, LCA, VocabProfiler). Syntactically, features related to complexity (t-units, clausal units, and length measure as reported by Coh-Metrix and L2SCA) should also be included in replication studies. Hinkel's study also reported on only a few indices of cohesion (e.g., coordinating and logical conjunctions). Replication studies should include more cohesion features with stronger links to textual cohesion, such as the semantic co-referentiality indices, lexical overlap indices, and anaphoric reference indices reported by Coh-Metrix.

Hinkel also failed to control for Type 1 errors (false positives) in her analysis: she conducted multiple statistical tests on her data but did not correct her alpha value to control for chance findings. There are several ways to control for Type 1 errors. Replication studies assessing prompt-based differences would do well to conduct Multiple Analyses of Variance (MANOVAs) with post-hoc analyses that include corrections for multiple comparisons. The indices that are the strongest predictors of prompt differences could then be fed into a machine learning algorithm (e.g., a logistic regression) that uses cross-validated training and test sets to

https://doi.org/10.1017/S0261444812000547 Published online by Cambridge University Press

develop a model to classify writing samples as belonging to one prompt or another, based on linguistic features alone. Such an analysis would provide strong evidence for prompt-based effects on linguistic production, and would be further bolstered by an analysis of how the prompt-based linguistic features in the writing samples affect human judgments of writing quality (see Research Task 2). A combination of these two areas would provide valuable information about how prompts can control the linguistic output of L2 students and how linguistic output can affect human judgments of essay quality.

6. Effects of task on L2 written production

The last important area of L2 writing that I will discuss is the influence of task on writing production. Successful writing tasks provide contextual and authentic opportunities for writers to use language to explain ideas, provide information, compare concepts, interact, or persuade. Perhaps the two most common writing tasks found in academic situations are independent writing (writing based on personal knowledge) and integrated writing (source-based writing). Studies have demonstrated that the writing task (whether it is independent or integrated) influences linguistic production (Reid 1990; Cumming et al. 2005, 2006). As with prompt-based differences (see Research Task 4), because linguistic features are predictive of human judgments of essay quality, it is important to understand how the writing task may affect the production of linguistic features.

While earlier studies on task-based differences focused on differences between compare/contrast essays and essays written to explain charts or graphs (Reid 1990), more recent research has begun to focus on differences between independent writing samples and integrated writing samples such as those found in the Test of English as a Foreign Language (TOEFL; Cumming et al. 2005, 2006). The Cumming et al. (2005) study is probably the best candidate for replication because it focused on differences between independent and integrated essays, using a number of linguistic features that have important links to writing quality.

Cumming et al. selected 216 essays written for the six tasks in the TOEFL-iBT (two independent writing tasks, two reading-to-write tasks, and two listening-to-write tasks). These 216 essays were written by 36 TOEFL examinees. All essays were scored from 3 to 5 on a six-level holistic scale of writing quality. Cumming et al. selected nine linguistic indices with which to analyze the essays in terms of differences between tasks. These nine indices all had theoretical links to assessments of writing quality. Six of the indices were generated automatically (text length, average word length, type/token ratio, number of clauses per t-unit, number of words per t-unit, and functional use of verb phrases taken from the source text), while human raters coded the other three features (grammatical accuracy, quality of argument structures, and orientation to source evidence). Cumming et al. used non-parametric MANOVAs to examine differences between the independent essays and the integrated essays. The analysis demonstrated that integrated writing tasks prompted shorter essays that contained longer words, a greater diversity of words, and more clauses that were also longer. Integrated essays were also less argumentatively oriented and contained more source material. Cumming et al. concluded that independent tasks produced essays

https://doi.org/10.1017/S0261444812000547 Published online by Cambridge University Press

containing extended written arguments and that integrated tasks produced essays responding to textual information.

Research task 5: Undertake an approximate or constructive replication of Cumming et al. (2005)

The Cumming et al. (2005) study provides a strong foundation for replication studies (both approximate and constructive). Recent advances in computational tools afford a greater understanding of differences between integrated and independent tasks in terms of their effects on lexical and syntactic features. Lexically, the Cumming et al. study needs approximate replication evidence because the two lexical indices used in the study are problematic. Word length is only a proxy for word frequency, so replication studies should focus on word frequency indices reported by Coh-Metrix or VocabProfiler. As discussed earlier, simple type/token ratio scores are highly correlated with text length. Therefore, it is possible that the type/token ratio scores were conflated with the text length measure. Replication studies should therefore consider using more advanced indices of lexical diversity, as found in the Gramulator, that control for text length. Lastly, approximate replication studies should consider some of the other syntactic complexity indices that are now available (as reported by L2SCA and Coh-Metrix). Constructive replications may consider testing a variety of linguistic features not considered by Cumming et al. but that still address their primary research questions. These would include indices related to cohesion features (as found in Coh-Metrix), idea units (as found in CPDIR), and psychological word properties (as found in LIWC). Both approximate and constructive replications should also take advantage of machine learning algorithms that could be used not only to assess differences between independent and integrated essays, but to also create models that could automatically categorize essays as belonging to one group or another.

While not related solely to replication studies, future research into linguistic differences between independent and integrated essays should also consider their effect on human ratings of essay quality. Unlike prompt-based writing analyses, some studies have investigated whether L2 essays demonstrate differences in holistic scores, depending on whether the writing task was independent or integrated. The results of these studies have been mixed, with some research reporting no differences in human scores for independent and integrated tasks (e.g., Gebril 2006) and other research reporting that L2 writers receive significantly higher scores for writing quality on integrated than on independent tasks (Esmaeili 2002). Future research should consider not only the human scores, but also how the linguistic features prompted by task influence human judgments of quality.

7. Conclusion

We are at an important intersection of language and technology where practical and accurate computational tools are readily available for advanced text analysis. Such tools can have an important impact on studies of L2 writing and provide us with the means to replicate studies from key areas of L2 writing research. Such replication studies can not only empirically support the original studies but also provide a deeper understanding of how L2 writing develops, how linguistic features in the text influence human judgments of writing quality,

https://doi.org/10.1017/S0261444812000547 Published online by Cambridge University Press

how, why, and where L2 writing diverges from L1 writing, how prompts affect text production and what this means for writing quality, and how different writing tasks incline L2 writers to produce different linguistic features. All of these areas of research provide us with fundamental knowledge about L2 writing and the effectiveness of L2 communication, which can help us develop instructional technology, pedagogical practices, and more sophisticated theories of the writing process and SLA in general.

My arguments for replication are linguistic in nature and computational at the core. However, linguistic and computer analyses can only provide answers to a defined range of research questions. Thus, while there are many tasks that computers are well placed to accomplish, there are others at which they are prone to fail. Many of these are important for a complete understanding of the L2 writing process. For instance, computational tools are fundamentally misaligned with research related to the writing process, literacy development, and language planning, all of which are important aspects of understanding L2 writing. In other areas, computational tools are still in their infancy and their application may be ill-advised (e.g., in sentiment analysis). However, from a linguistic perspective, computational tools offer many advantages over human analysis, including reduced costs (compared to human assessors), speed, flexibility, and reliability (Higgins, Xi, Zechner & Williamson 2011). These advantages enable a better understanding of the written word and how the choice of words in combination is indicative of writing development, writing quality, writing differences, and prompt and task-based effects.

Acknowledgments

The author would like to express his gratitude to the editor of *Language Teaching*, Graeme Porte, for the invitation to write this paper and for his close readings and suggestions during the writing and reviewing process. The author is also indebted to the anonymous reviewers of this paper and to the following people who have provided guidance in the development of the ideas expressed within the paper: Diane Belcher, Philip McCarthy, Danielle McNamara, and Sara Weigle.

References

Arnaud, P. J. L. (1992). Objective lexical and grammatical characteristics of L2 written compositions and the validity of separate-component tests. In P. J. L. Arnaud & H. Bejoint (eds.), *Vocabulary and applied linguistics*. London: Macmillan, 133–145.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Bonzo, J. D. (2008). To assign a topic or not: Observing fluency and complexity in intermediate foreign language writing. *Foreign Language Annals* 41, 722–735.

Brown, C., T. Snodgrass, S. J. Kemper, R. Herman & M. A. Covington (2008). Automatic measurement of propositional idea density from part-of-speech tagging. *Behavior Research Methods* 40.2, 540–545.

Bynes, H., H. H. Maxim & J. M. Norris (2010). Realizing advanced L2 writing development in a collegiate curriculum: Curricular design, pedagogy, assessment. *The Modern Language Journal* 94, Monograph Supplement.

Carlman, N. (1986). Topic differences on writing tests: How much do they matter? *English Quarterly* 19, 39–47.

https://doi.org/10.1017/S0261444812000547 Published online by Cambridge University Press

Chung, C. K. & J. W. Pennebaker (2012). Linguistic Inquiry and Word Count (LIWC): Pronounced 'Luke', … and other useful facts. In P. M. McCarthy & C. Boonthum (eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution*. Hershey, PA: IGI Global, 133–145.
Cobb, T. & M. Horst (2011). Does Word Coach coach words? *CALICO* 28.3, 639–661.
Connor, U. (1984). A study of cohesion and coherence in ESL students' writing. *Papers in Linguistics: International Journal of Human Communication* 17, 301–316.
Connor, U. (1990). Linguistic/rhetorical measures for international persuasive student writing. *Research in the Teaching of English* 24, 67–87.
Crossley, S. A., D. S. McNamara, J. Weston & S. T. McLain Sullivan (2011). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication* 28.3, 282–311.
Crossley, S. A. & D. S. McNamara (2009). Computationally assessing lexical differences in L2 writing. *Journal of Second Language Writing* 17.2, 119–135.
Crossley, S. A, T. Salsbury & D. S. McNamara (2009). Measuring second language lexical growth using hypernymic relationships. *Language Learning* 59.2, 307–334.
Crossley, S. A., T. Salsbury & D. S. McNamara (2010). The development of polysemy and frequency use in English second language speakers. *Language Learning* 60.3, 573–605.
Cumming, A., R. Kantor, K. Baba, U. Erdoosy, K. Eouanzoui & M. James (2005). Differences in written discourse in writing-only and reading-to-write prototype tasks for next generation TOEFL. *Assessing Writing* 10, 5–43.
Cumming, A., R. Kantor, K. Baba, U. Erdoosy, K. Eouanzoui & M. James (2006). *Analysis of discourse features and verification of scoring levels for independent and integrated tasks for the new TOEFL* (TOEFL Monograph No. MS-30). Princeton, NJ: ETS.
Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing* 4.2, 139–155.
Esmaeili, H. (2002). Integrated reading and writing tasks and ESL students' reading and writing performance in an English language test. *The Canadian Modern Language Review* 58.4, 599–622.
Ferris, D. R. (1994). Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. *TESOL Quarterly* 28.2, 414–420.
Gebril, A. (2006). *Writing-only and reading-to-write academic writing tasks: A study in generalizability and test method*. Unpublished doctoral dissertation, University of Iowa.
Graesser, A. C., D. S. McNamara, M. Louwerse & Z. Cai (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers* 36, 193–202.
Granger, S., E. Dagneaux, F. Meunier & M. Paquot (2009). *The International Corpus of Learner English. Handbook and CD-ROM*. Version 2. Louvain-la-Neuve: Presses Universitaires de Louvain.
Grant, L. & A. Ginther (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing* 9, 123–145.
Haswell, R. H. (2000). Documenting improvement in college writing: A longitudinal approach. *Written Communication* 17, 307–352.
Higgins, D., X. Xi, K. Zechner & D. Williamson (2011). A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech and Language* 25.2, 282–306.
Hinkel, E. (2002). *Second language writers' text*. Mahwah, NJ: Lawrence Erlbaum.
Hinkel, E. (2003). Simplicity without elegance: Features of sentences in L1 and L2 academic texts. *TESOL Quarterly* 37, 275–301.
Hinkel, E. (2009). The effects of essay topics on modal verb uses in L1 and L2 academic writing. *Journal of Pragmatics* 41, 667–683.
Horowitz, D. (1986). What professors actually require: Academic tasks for the ESL classroom. *TESOL Quarterly* 20, 445–462.
Just, M. A. & P. A. Carpenter (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review* 87, 329–354.
Language Teaching Review Panel (2008). Replication studies in language learning and teaching: Questions and answers, *Language Teaching* 41, 1–14.
Laufer, B. (1994). The lexical profile of second language writing: Does it change over time? *RELC Journal* 25.2, 21–33.
Laufer, B. & I. S. P. Nation (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics* 16, 307–322.

https://doi.org/10.1017/S0261444812000547 Published online by Cambridge University Press

Leki, I., A. Cumming & T. Silva (2008). *A synthesis of research on second language writing in English*. New York: Routledge.
Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly* 45.1, 36–62.
Lu, X. (in press). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*.
Matsuda, P. K. & T. J. Silva (2005). *Second language writing research: Perspective on the process of knowledge construction*. Mahwah, NJ: Lawrence Erlbaum.
McCarthy, P. M. & S. Jarvis (2010). MTLD, vocd-D, and HD-D: a validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods* 42, 381–392.
McCarthy, P. M., S. Watanabe & T. A. Lamkin (2012). The Gramulator: A tool to identify differential linguistic features of correlative text types. In P. M. McCarthy & C. Boonthum (eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution*. Hershey, PA: IGI Global, 312–333.
McCutchen, D. (1986). Domain knowledge and linguistic knowledge in the development of writing ability. *Journal of Memory and Language* 25, 431–444.
McNamara, D. S. & A. C. Graesser (2012). Coh-Metrix. In P. M. McCarthy & C. Boonthum (eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution*. Hershey, PA: IGI Global, 188–205.
Pennebaker, J. W., M. E. Francis & R. J. Booth (2001). *Linguistic Inquiry and Word Count (LIWC): LIWC2001*. Mahwah, NJ: Lawrence Erlbaum.
Porte, G. K. (2012) *Replication in applied linguistics research*. Cambridge: Cambridge University Press.
Porte, G. K. & K. Richards (2012). Replication in quantitative and qualitative research. *Journal of Second Language Writing* 21.3, 284–293.
Rayner, K. & A. Pollatsek (1994). *The psychology of reading*. Englewood Cliffs, NJ: Prentice Hall.
Reid, J. (1990). Responding to different topic types: A quantitative analysis from a contrastive rhetoric perspective. In B. Kroll (ed.), *Second language writing: Research insights for the classroom*. Cambridge: Cambridge University Press, 191–210.
Reid, J. R. (1992). A computer text analysis of four cohesion devices in English discourse by native and nonnative writers. *Journal of Second Language Writing* 1.2, 79–107.
Silva, T. (1993). Toward an understanding of the distinct nature of L2 writing: The ESL research and its implications. *TESOL Quarterly* 27.4, 657–675.
Witten, I. A., E. Frank & M. A. Hall (2011). *Data mining: Practical machine learning tools and techniques*. San Francisco, CA: Elsevier.
Xue, G. & I. S. P. Nation (1984). A university word list. *Language Learning and Communication* 3.2, 215–229.
Zwaan, R. A., M. C. Langston & A. C. Graesser (1995). The construction of situation models in narrative comprehension: An event-indexing model. *Psychological Science* 6, 292–297.

SCOTT CROSSLEY is an assistant professor at Georgia State University, Atlanta. His primary research focuses on corpus linguistics and the application of computational tools in L2 learning, writing, and text comprehensibility.

https://doi.org/10.1017/S0261444812000547 Published online by Cambridge University Press