Identifying and analyzing noisy spelling errors in a second language corpus

Alan Juffs
Department of Linguistics
University of Pittsburgh
juffs@pitt.edu

Ben Naismith
Duolingo and
University of Pittsburgh
ben.naismith@duolingo.com

Abstract

This paper addresses the problem of identifying and analyzing 'noisy' spelling errors in texts written by second language (L2) learners' texts in a written corpus. Using Python, spelling errors were identified in 5774 texts greater than or equal to 66 words (total=1,814,209 words), selected from a corpus of 4.2 million words (Juffs, Han, and Naismith 2020). The statistical analysis used hurdle() models in R, which are appropriate for non-normal, count data, with many zeros.

1 Introduction

The problem of 'noisy data' addressed in this paper is how to automatically identify and analyze spelling errors in texts written by speakers of English as a second language. This issue is important in automated scoring of written texts in high-stakes tests such as the internet based TOEFL (iBT) and Duolingo English Test (DET). Tests such as these use models that include numerous features, and these features may produce different values depending on whether they are considering correctly or incorrectly spelled tokens. Thus, this paper reports on one method of identifying the rate of spelling errors in the written output of learners of English as a second language in an Intensive English Program (IEP; Juffs 2020) over time and addresses the optimal statistical method for measuring those errors. The first languages (L1s) of these learners varied in their orthographies from abjads (Arabic), alphabets (Spanish, Korean), logographic characters (Chinese), and a mix of logographic and syllabaries (Japanese). At the time of data collection, the IEP had three levels of proficiency with approximate equivalent CEFR levels (Common European Framework of Reference; Council of Europe, 2001) as follows: level 3 low-intermediate, CEFR A2-B1; level 4 intermediate, CEFR B1+-B2; and advanced, C1 and above. Therefore, this corpus is representative of the population of IEP students across the USA at the time of data collection between approximately 2007-2012. (We note, however, that international student populations in US IEPs vary somewhat by region and have varied over time from the 1960s until present.)

English spelling poses special challenges because it uses a 'deep' orthography, meaning that the spoken sounds of English do not closely match their written forms and vice-versa. For example, the same sound /i/ is represented by different letters in 'ea' as in 'eat', 'e' as in the first 'e' in 'scene', 'ee' as in 'see', and 'y' as in 'quickly'.

Specifically, our research questions were the following. In terms of the rate of spelling errors in learners' written texts:

1. How can the spelling errors in a (typed) written corpus of 4.2 million words (Juffs, Han, and Naismith 2020) be automatically and accurately identified and calculated using Python?
2. Is there an effect for L1?
3. Is there an effect of proficiency level in the IEP?
4. Is there an interaction between L1 and IEP level?

2 Related Work

Spelling correction has been a long-standing challenge in natural language processing (NLP), with approaches ranging from traditional rule-based methods to modern deep learning models. Early spell checkers relied on edit-distance algorithms such as Damerau-Levenshtein (Damerau 1964, Mitton 1996), often combined with dictionary-based look-ups. However, these early methods

struggled with errors where a misspelling results in another valid word (e.g., 'form' instead of 'from'). Subsequently, statistical models leveraging n-grams (Brill and Moore 2000) and probabilistic approaches (Carlson and Fette 2007) were introduced, enabling some level of context-aware correction. More recently, deep learning methods have demonstrated superior accuracy by leveraging contextual embeddings (Devlin et al. 2018; Jayanthi, Pruthi, and Neubig 2020).

Among open-access models used for spell checking, *NeuSpell* is trained on diverse datasets and uses contextual embeddings such as BERT and ELMo (Jayanthi, Pruthi, and Neubig 2020). *SymSpell*, though often considered a rule-based system, incorporates bigram look-ups to enhance context awareness, allowing it to resolve some ambiguous cases where single-word spell checkers might fail. Similarly, *JamSpell* incorporates a 3-gram language model to refine corrections based on surrounding words (Ozinov 2019). Unlike deep learning models, which infer spelling corrections from large corpora, *SymSpell* and similar models use a pre-compiled frequency dictionary to determine valid words and generate correction candidates efficiently (Garbe 2021b). The Spell Checker Oriented Word List (SCOWL; Atkinson 2019) is one of the most widely used resources, providing a hierarchical lexicon of words categorized by frequency and linguistic validity. Other resources, such as Hunspell and Aspell, also use wordlist-based approaches, making them highly efficient for misspellings but limited when handling real-word errors (Näther 2020).

For L2 learner errors, the choice of a spelling correction system is particularly important. Rule-based systems offer a more conservative approach, as they avoid over-correcting errors that might be intentional learner choices or non-standard but comprehensible variants (Näther 2020). In contrast, deep learning models, while highly accurate, may introduce unwanted corrections that mistake the intended choices in learners' *interlanguage* (Selinker 1972), particularly when trained on L1-English corpora. Other proprietary systems, such as Google's spell checker and Microsoft's *BingSpell*, remain inaccessible for customization, though they benefit from large-scale user data and adaptive correction mechanisms. Therefore, in settings or applications focusing on learner data, a hybrid approach using open-source tools (e.g., using wordlist-based methods to avoid excessive intervention, supplemented by context-aware models for ambiguous cases) may be the most effective strategy (Bryant et al. 2019; Omelianchuk et al. 2020). In high-stakes English proficiency assessments that implement automated scoring of writing, spelling accuracy is explicitly listed as a dimension of the scoring models (e.g., TOEFL, DET, PTE). However, details about the spelling error identification methods are scarce.

Although the problem of correcting spelling with computers has a long history, as far as we are aware, spelling errors in a second language written corpus in L2 English with various L1s have not been addressed. The Pittsburgh English Language Institute Corpus (PELIC) is unusual in that it contains longitudinal data in addition to a variety of L1s. In addition, the appropriate statistical models for analyzing the rate of errors has not been determined. Applied linguists are not just interested in computational detection and correction, but also in the potential qualitative impact of spelling errors on human graders, along with pedagogical implications.

While the cited on-line spelling checking resources are coded in a variety of computer languages, for applied linguists who work with L2 data, Python is the main programming language, and so Python was used to provide accessibility to such researchers. A complete description of PELIC spelling error identification and correction is provided at a public GitHub repository and Jupyter Notebook (Naismith, Starr, and Bacas 2021), where links include the following resources which were used in this paper:

(1) SCOWL Lists (Atkinson 2019). This website contains English word lists that contain abbreviations, acronyms, British, American and Commonwealth spellings, contractions, and taboo words that can be used in spell checkers. The resource also contains scripts in perl for the creation of tailored lists.

(2) Symspell (Garbe 2021b). *Symspell* is a spelling correction algorithm that only deletes erroneous spellings according to limited specifications. Garbe 2021b claims that it is one million times faster than other models, for example, Norvig, which was 80-90% accurate. This program deals with single words, compounds, and word-segmentation. The website contains code in a variety of programming languages in addition to Python.

Related work in applied second language read-

ing and spelling research has noted that for L2, the challenges of English orthography are compounded by the influence of their L1 writing systems and limited vocabulary size (Hamada and Koda 2008, Humaidan and Martin 2019). An important construct in this domain is *lexical quality* Perfetti and Hart 2002, which established the importance of strong links in the mental lexicon among sound (phonology), orthography, and meaning. Poor links among sounds, graphemes, and semantics in any direction in lexical representations pose problems in both reading comprehension (Perfetti and Stafura 2013) and writing production (Dunlap 2012). Moreover, Baker and Hawn 2022 raise the problem that computer-automated grading may unfairly disadvantage some groups, known as 'algorithmic bias' in education.

Thus, this work is innovative because it is a rare(?) example of explicitly *interdisciplinary* work drawing on computational linguistics in automatic spell-checking and correction, applied statistics, with insights from applied linguistics research on literacy and instruction.

3 Spelling Identification

Spelling errors were identified using the following steps. First, SCOWL was consulted, and a SCOWL file was created and used to decide whether a word in the IEP texts was 'real' or not (SCOWL List for PELIC). All items were included from the lists except the abbreviations dictionary. Words that had previously been considered 'non-words' by dictionaries were added to our list, for example, 'southside', which is a neighborhood of Pittsburgh, 'frisbee', which is a toy/game, and 'onsen', which is a Japanese loanword for 'hot spring'. All hyphenated words were included as real words, forexample, 'prize-winning'. After running the revised dictionary, a list of misspellings with their adjacent words was created, followed by a dictionary of misspelled items. Examples in the dictionary of common misspellings included '*alot', '*becouse', '*sould', etc.

A Python module, Symspell (Garbe 2021a), was used that included the spelling errors and corrections for those errors. Examples, of corrections made are 'beccuase' -> 'because', 'nise' -> 'nice', 'friendlly' -> 'friendly'. Only word and bigram frequencies, but not full sentence context, were considered in resolving the appropriate target. This practice is consistent with other spellcheckers (Hunspell, pyspell, etc.). Thus, following this common practice the accuracy of corrected tokens will not be 100%. Nevertheless, the accuracy was inspected by random sampling and found that where the word is accurately spelled, the checker correctly does nothing 100% of the time (Naismith, Starr, and Bacas 2021), that is, there are no false positives.

An important caveat is that incorrect spellings that are actual words, for example, the pronoun 'him' misspelled as the real word 'hem', are not corrected. Such 'clang associates' (Schmitt and Meara 1997) are not counted as spelling errors in this automated process because it is difficult to automatically identify and correct misspellings that are real words. It might be possible to differentiate clang associates based on part of speech, for example, noun 'hem' vs. pronoun for 'him', or frequency of clang associate based on phonology, for example, 'ship' vs. 'sheep', but these possibilities have not yet been explored. Nevertheless, based on the corrections, it was possible to programmatically count and tally the misspellings in each text in the database. These text-based counts were the basis of the data in the study.

To control for number of errors by text length, the spelling errors were calculated by dividing the number of errors by the number of tokens in each text and multiplying by 100. Because the appropriate statistical analysis requires whole numbers (no decimals), 0.5 was added to the result of all calculations before the number was converted to an integer. Thus, 0.287 errors in a text remains 0, but a score of 0.847 became 1.347, and was converted to the integer 1.

4 Statistical Models

This section addresses the problem of the appropriate statistical model for non-normal count data with many zeros. Models that permit inferential quantitative investigation of count data include the Poisson family of analyses. Zeileis, Kleiber, and Jackman 2008 provide a detailed review of Poisson models that are both suitable and unsuitable for count data such as the spelling error data under consideration. Two points about our spelling error data are relevant here for Poisson analysis. First, standard Poisson analysis for count data is inappropriate for over-dispersed data, that is, data with very large numbers of outliers. Second, these data contain very large numbers of zeros, that is, texts without spelling errors – in fact over 50% for each

L1 and level. In this context, Crawley 2013 (Chapter 14) also raised the problem of high frequency of '0' in count data. Zeileis, Kleiber, and Jackman 2008 recommended the 'hurdle()' procedure for data with these characteristics. The hurdle() procedure is available in the (R package 'pscl') and is discussed in greater detail in the next section.

5 Results

In addition to L1 and level, other available student information that relates to the data includes standardized proficiency scores of a placement test and writing sample on entry to the IEP as well as self-reported biological gender. Neither the placement score nor the writing sample scores correlated at higher than r = -.07 to the number of errors and were therefore not included in any model even though these correlations were reliable at p<.0001 due to the large n sizes. Gender was also non-significant as a predictor.

The percentage of texts with 0 errors are displayed in Table 1 by L1 and level. It can be observed that over 50% of texts by each L1 at each level are error-free. Thus, the data are characterized by many scores of 0. In fact, 4554 of the total 5774 texts (78.7%) had 0 errors, not counting unknown clang associates. The numbers of students contributing data appear in Table 1. The majority of students at each level were Arabic speakers, while the fewest were Spanish speakers. Nevertheless, variability by L1 and level can be observed which makes the analysis important for proficiency assessment. Table 2 reports means and standard

Table 1: Percentage of Texts over 66 words with 0 Spelling Errors and Number of Texts by L1/Level.

	Level 3		Level 4		Level 5	
L1	%errors	Texts	%errors	Texts	%errors	Texts
Arabic	61.2	490	77	1126	85.2	709
Chinese	76.5	260	82.0	677	84.2	431
Japanese	70.0	60	79.1	249	90.3	134
Korean	67.0	276	80.7	685	87.9	404
Spanish	63.6	55	79.7	138	73.3	75

deviations of spelling errors, including texts with 0 errors. For example, the Arabic speakers at level 3 have an average of one error per text in their writing and also standard deviation of 1.52 errors. However, these means mask the fact that many texts by Arabic speaking learners have many more than just one error. The large number of texts reduces the mean but the visualization in Figure 1 shows

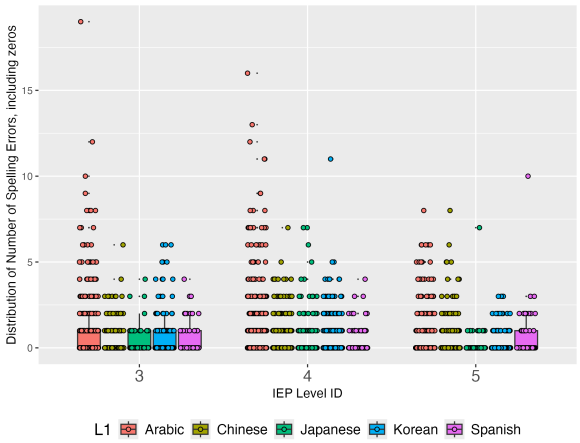

Figure 1: Box Plot Distribution of Errors (including zeros) in the count data by L1 and Level

the variability more clearly. As evident from Table

Table 2: Spelling Errors per text by L1 and Level

	Level 3		Level 4		Level 5	
L1	Mean	SD	Mean	SD	Mean	SD
Arabic	1.05	1.52	0.62	1.35	0.42	0.88
Chinese	0.43	0.55	0.28	0.46	0.33	0.56
Japanese	0.42	0.39	0.64	1.48	0.13	0.22
Korean	0.55	0.73	0.30	0.41	0.24	0.40
Spanish	0.76	0.87	0.27	0.33	0.39	0.51

2, which represents the raw mean errors per student in each language group, and Figure 1, which illustrates the *mean proportion of errors per 100 words for each language group*, there are large numbers of texts with zero errors. The errors that do occur are not normally distributed. Thus, in Figure 1, the red columns represent Arabic-speaking students who at level 3 and level 4 have many more spelling errors per 100 words than other students. The boxplot and outliers show that many more of the Arabic-speaking students' texts had errors, frequently over five in each text as indicated by the circles above the boxes. Based on Zeileis, Kleiber, and Jackman 2008, analyses showed that the data are overdispersed. As Crawley 2013 states, in standard Poisson analyses "it is assumed that residual deviance is equal to the residual degrees of freedom (because the variance and the mean should be the same)". In these spelling data, a standard Poisson model revealed that residual de-

viance was 8422.4 on 5759 degrees of freedom. Overdispersion can sometimes be dealt with using the quasi-Poisson technique. However, both Zeileis, Kleiber, and Jackman 2008 and Hoftstetter et al. 2016 show that a better method is the hurdle() technique. This approach provides regressions for the number of zeros and the count values separately by factor. Thus, one can determine effects of factors both on the number of zero counts and the count data in one model. Recall that all spelling error counts had been adjusted to 'count' integer data, that is whole numbers for analysis. Following Hoftstetter et al. 2016, we evaluated hurdle() and zeroinfl() negative binomial logistical regression models, concluding that the following hurdle() model was optimal: $hnb < -hurdle(PROP3\ L1 * level_id, data = L1FW3LNONA, na.action = na.exclude, dist = "negbin")$.

The results are provided in Table 3, with the statistics for the count data in the left half of the table and those for the texts with zero errors in the right part. In each half of Table 3, the intercept estimates are the log odds of spelling errors compared to zero errors, the other estimates are the log odds of those measures compared to the intercept.

Table 3: Hurdle Model Results

	Count Model (Truncated NegBin)		Zero Hurdle Model (Logit)	
Variable	Estimate	Std. Error	Estimate	Std. Error
Intercept	-1.151	0.891	-0.456	0.092***
L1Chinese	-0.904	0.291**	-0.726	0.173***
L1Japanese	-1.331	0.526*	-0.391	0.296
L1Korean	-0.917	0.251***	-0.253	0.158
L1Spanish	-0.683	0.457	-0.103	0.295
level_id4	-0.075	0.178	-0.752	0.116***
level_id5	-0.224	0.228	-1.293	0.140***
L1Chinese:level_id4	0.193	0.360	0.419	0.212*
L1Japanese:level_id4	0.929	0.600	0.267	0.342
L1Korean:level_id4	0.118	0.326	0.029	0.198
L1Spanish:level_id4	-0.525	0.620	-0.057	0.370
L1Chinese:level_id5	0.478	0.417	0.801	0.242***
L1Japanese:level_id5	0.265	0.813	-0.090	0.429
L1Korean:level_id5	-0.456	0.450	0.023	0.244
L1Spanish:level_id5	0.584	0.649	0.841	0.408*
Log(theta)	-2.376	1.014*		
Significance Codes: *** p<0.001, ** p<0.01, * p<0.05				
Log-likelihood		-4537		
N		5774		

Log odds can be converted to odds using the `exp()` function in R Levshina 2015. These converted odds are in Table 4, Appendix A, in the column Incidence Rate Ratios, produced using the `tab_model()` function. The intercept (reference level) was automatically selected (dummy coded) as Arabic level 3 in both models. This is why 'Arabic' appears nowhere in Table 3. The significance level of the intercept is an estimate of the outcome when the L1 and the level are at their reference levels.

A reliable chi-square statistic for the interaction of L1 and level_id in the model ($df = 16$, LRT = 30.37, $p = 0.016$) revealed that it contains frequencies of errors that are contingent on L1 and level. In addition, compared to an (inappropriate) standard Poisson model, the Akaike Information Criterion (AIC) confirmed that the `hurdle()` model with negative binomial was a better fit. The significant $\log(\theta) = -2.37$, $p = 0.019$ in the count data section in Table 3 also confirmed that these data were overdispersed and that therefore the negative binomial hurdle analytic technique was the appropriate one (cf. Hoftstetter et al. 2016, p. 524-525).

Unpacking these statistics, Table 3 can be interpreted following Levshina 2015, p. 257-266) and Hoftstetter et al. 2016. Visualization of the count data in boxplots appears in Figure 1, which included texts with zero errors. The count model (Intercept) shows the odds of a spelling error Arabic, $level3 = Exp(-1.15) = 0.32$, and is not significant compared to zero. This result makes sense because while 39% of texts do contain at least one error, 61% of level 3 Arabic texts are error free. However, the odds of a spelling error by Chinese learners at $level3 = Exp(-0.90449) = 0.40$ is reliably lower than the intercept (see Appendix A). Thus, the odds of Chinese speakers making a spelling error per 100 words at level 3 are reliably 0.4 times lower than Arabic at that level due to the negative estimate. The other L1 data can be similarly interpreted. The odds of a spelling error by Japanese, $level3 = Exp(1.33061) = 0.26$. Thus, the odds of Japanese speakers making an error at level 3 are 0.26 times lower than Arabic at that level. Note the higher variance and lower p value for the Japanese speakers, which reduces the odds compared to the Chinese speakers. The odds of a spelling error by Korean, level 3 speakers is also lower = $Exp(-0.9174) = 0.40$. The odds of a spelling error Spanish, $level3 = Exp(-0.68308) = 0.51$. The co-efficient is also negative. However, the higher variance, lower z, and non-significant p values mean the Spanish level 3 speakers are not statistically different from the Arabic level 3 learners in the count data.

The hurdle model, having selected the count data with a lower limit of 1, then proceeds to model the number of texts by L1 and level with 0 errors, that is, the zero data. For zero hurdle model coefficients,

Hoftstetter et al. 2016, p. 523) state that "the zero model represents the probability of observing a positive count". In this case, the Arabic level 3 intercept with errors is reliable: $Exp(-0.456) = 0.63$, which means the odds of Arabic speakers making a spelling error compared to zero is reliably negative 0.63, with Table 1 showing that 61% of their texts contain no errors. Only the Chinese speakers show a difference from the Arabic speakers' texts being even less likely to make a spelling error at level 3. In this case, Chinese speakers at level 3 are 0.48 times less likely than them to produce text with an error, consistent with the count data. In Table 1, the level 3 columns illustrate this result, showing that 76.5% of Chinese learner texts at level 3 have no spelling errors, which is higher than any other L1 by over 6%. However, the Japanese, Korean, and Spanish speakers are not different from Arabic speakers at level 3.

To the right part of Table 3, in the zero hurdle model, level-id is statistically significant at both level 4 and level 5. This result means that the odds of Arabic speakers' texts having an error decreased significantly at level 4 by 0.47 and at level 5 by 0.27 compared to Arabic level 3.

The Chinese speakers' estimates at level 4 and level 5 compared to the (Intercept) show a reliable difference, except this time in a positive direction. This result means that compared to texts with an error for Arabic levels 4 and 5, the odds of the Chinese level 4 and level 5 learner producing a text with even one error increases by odds of 1.52 and 2.23 respectively. At level 5, the Spanish speakers also reliably increase the odds of making an error by 2.32 compared to the intercept, with only 73% of their texts at level 5 being error-free. No other comparisons with Arabic-speakers' level 3 in the model are reliable.[1]

6 Discussion

The differences by L1 are statistically reliable according to the chi square test on the entire model. Thus, while spelling mistakes by all IEP learners with access to spell-checkers in word processing software are relatively low, they are noticeably and reliably different by L1. Moreover, it is important to note that overall the learners improved in their accuracy over time.

Returning to the research questions, the results first demonstrated an effect for L1. When a text contains errors, Arabic-speaking learners make more spelling errors than Chinese-speaking, Japanese-speaking, and Korean-speaking (but not Spanish-speaking) learners at level 3, but these differences disappear at levels 4 and 5. Regarding texts with 0 errors, the pattern is similar, but the odds of Chinese-speaking learners making an error remains somewhat higher at level 4 and level 5 compared to level 3 Arabic speakers. Thus, there is an interesting interaction and difference between the Arabic-speaking and Chinese-speaking learners such that Arabic speakers decrease their proportion of errors in texts, while Chinese speakers seem to be more stable compared to the Arabic-speaking learners. Taken together, a cautious interpretation of these results suggests the most reliable difference is between the Arabic-speaking in contrast to the Chinese-speaking learners as there are differences between these groups in both the count and zero hurdle models. While Japanese-speaking, and Korean-speaking learners at level 3 differ from Arabic speaking learners producing texts with fewer errors, the zero model showed no differences among these three L1s. The Spanish speakers make errors at a statistically similar rate to the Arabic speakers.

Second, as to the effect of level of IEP at 4 and 5, we can see that for texts with errors there is no effect. This means that the rate of errors in texts varies little across levels overall. However, numbers of texts with zero errors increases from level 3 to level 4 and remains steady at level 5. We may infer that use of the spell-checker with word processing skills improved along with knowledge of orthography, and especially for the Arabic-speaking learners.

Third, interactions exist in the rate of errors among Arabic-speaking learners. A decrease occurred from level 3 to level 4, but not for other L1s, indicated by the interaction of level for level 4 with numbers of zero error texts. In addition, for the zero-count data, Chinese speakers showed an interaction at levels 4 and 5, indicating that the number of zero error texts remained more constant compared to Arabic level 3. Figure 1 shows that

[1] It is possible to make multiple pairwise comparisons by changing the reference level from Arabic to other L1s. However, given the limited number of errors and the similar means and dispersion statistics in Table 2 and Figure 1, it is unlikely that other pairwise comparisons would be reliable. One possibility was the very low Japanese error rate at level 3 is different from L1s other than Arabic. Overall, Japanese speakers are reliably less likely make an error, but Arabic speakers' odds of errors increase consistent with the model in which they are the reference level.

errors by Chinese speakers at levels 4 and 5 remain higher, while other L1 error rates declined.

To some extent, this outcome is reassuring for automated scoring of many features, for example those related to lexical sophistication (vanHout and Vermeer 2007), because automated measures of lexical sophistication, for example, Advanced Guiraud (Daller, Turlik, and Weir 2013), based on word-processed texts will not unduly penalize one group at intermediate and high-intermediate levels, for example, Arabic-speaking learners, by excluding misspelled low-frequency words, that is, words with a frequency band higher than 2000 at a higher rate than other L1s. This possibility had been suggested by Naismith, Han, et al. 2018 but now seems to be less of a concern based on this analysis. This confidence is possible due to the low number of statistically significant differences among the groups and the low numbers of errors per text overall. The group most at risk would be the Arabic level 3 learners, who made the most errors. Specifically, automatic scoring of lexical sophistication measures derived from frequencies of lemmas in an external corpus will not be affected by learners losing credit for too many misspelled words above the 2k frequency band at intermediate levels and above.

However, Arabic-speaking learners' errors may be salient to human raters compared to other L1 groups. This impression arises from the visualization of the data, even if it is not statistically robust, because of the numbers of texts that contain outlier tallies of spelling mistakes. Although the L1 effect is only statistically reliable at level 3, the tendency is very noticeable on a qualitative level at levels 4 and 5 also. Such a pattern of errors could cause human raters to negatively perceive Arabic speakers' writing, when only 61.2% of their texts at level 3 are error free compared to Chinese speakers' 76.5%. Thus, in high stakes testing where *both* human and computer-based automatic scoring are deployed, spelling errors have spelling errors have the potential to create bias against one group, even though those learners 'know' the items in question.

Moreover, these spelling errors (even when using word processing software), and the evidence from the reading studies cited in the introduction, are indicative of wider problems with lexical quality Perfetti and Stafura 2013 at the low-intermediate level (level 3). Thus, these data support interventions with spelling for all L1s, perhaps especially at the low-intermediate stage at the early stages of learning. When spellcheckers highlight many words – including proper names not frequent in English – it may be difficult for learners to guess which words are misspelled and, perhaps more importantly, which ones are the correct replacements. In fact, due to the saliency of spelling errors reported previously by Dunlap 2012 in student transcriptions of their own speech, one IEP instituted a dictation component as part of its curriculum to address lexical quality. This decision is given additional support by these data and other studies such as those reported in Humaidan and Martin 2019.

7 Qualitative Review of 'Noisy' Errors

This section provides a qualitative review of the type and frequency of orthography mistakes in these word-processed data to complement the quantitative analysis based on automatic tagging in the previous section. This review provides additional insights into these 'noisy' data that vary by L1 and proficiency. The process through which this was done was that the first author, an experienced English as a second language teacher, reviewed all the spelling errors in the texts. Thus, the list is not exhaustive but provides some indication of the challenges that learners face. The mistakes fall generally into four categories: (i) mistakes many learners make with frequent words; (ii) errors across L1s with the use of English spelling conventions; (iii) those forms influenced by L1 morpho-phonology; (iv) forms flagged as errors even though they are correct, for example, the (now sadly outdated) blend 'Brangelina' or abbreviations, for example, NHK, CBS (Japanese and US TV stations), and RMB (= Renminbi, the Chinese currency).

In the first category, regardless of L1 and level, many learners made mistakes with some *frequent* words, for example, 'because' (the range of misspellings of this word is very large) and 'studying', with the 'y' plus 'ing' creating uncertainty. Errors flagged due to spelling conventions of English double consonants were also frequent across learners both at morphological boundaries, for example, *writting, *eightteen, *eatting, vs. *regreting, and within words, for example, *recomendation, *profesion vs. *bussiness. Unsurprisingly, given the different double consonant spelling rules in Spanish, Spanish-speaking learners made many errors with double consonants.

Second, errors influenced by L1 morpho-

phonology seemed especially frequent at level 3. Caution is in order as some could be simply 'typos', and others of these errors could be spacing problems that are influenced by English chunks (e.g., many learners made a mistake with *alot) or reduced stress on functors such as indefinite articles. However, Arabic speakers seemed to produce more with pronouns such as *iowe, and *idid, in addition to more frequent lack of spacing between indefinite articles and nouns (e.g., *anest and *abeach). In addition, trilled /r/ pronunciation could plausibly have produced *bearrd. Omitted vowels by Arabic speakers at the lower levels are also quite frequent, even in common words, but especially with liquids /l/ and /r/, for example, *evry and *evrybady, but also other words *amrica and *cmfortable. Such omission may be attributed to the influence of the abjad orthography, which only marks some vowels in Arabic and which also affects reading L2 reading in English (Martin and Juffs 2021). Because Arabic lacks the phoneme /p/, there is also an occasional, predictable voicing error in orthographic 'p' vs. 'b' (e.g., *laptob).

For Chinese speakers, it is possible to identify errors due to syllable structure constraints in Mandarin, which disallows consonant clusters (some possible with glides) in onsets and permits only alveolar and velar nasals in syllable final position. Potential examples of such influence include epenthesis (e.g., *samalled = 'smelled' and *sipricy = 'spicy') and what one could term metathesis *firiend 'friend' and *porblem 'problem'. In general, Chinese, Japanese (e.g., *toraditional 'traditional'), and Korean speakers (e.g., *zebara 'zebra') seemed more likely to insert a vowel compared to the Arabic speakers from the examples reviewed in the data. However, such qualitative observation would need to be quantitatively confirmed with inferential statistical analysis.

Finally, all L1s showed influence of vowel quality pronunciation on spelling, especially the [aɛ] as in 'cat' vs. [ɛ] in 'ketchup', for example, Korean level 3, *trevel = 'travel' and Korean level 4 *demage = 'damage' shows vowel raising from [aɛ] to [ɛ] in learners' phonological representations of these words. (We note that a merger between these two sounds may be occuring in some English varieties in Australia and in the northern cities of the USA near the Great Lakes.) Schwa [ə] in unstressed syllables also caused learners to have problems in identifying the correct grapheme, for example, Chinese level 4 *mechine = 'machine' and Arabic level 5 *sentance = 'sentence'. Some diphthongs (e.g., [ej] for Japanese *fervorite = 'favorite') also posed challenges, but less so.

The impression from this review of specific errors is that the influence of L1 morpho-phonology on spelling accuracy was more evident at level 3. This finding suggests that because learners do not control the pronunciation of the word, it is harder for them to evaluate choices provided by the spell-checker. This difficulty is due to their own representation of meaning to sound, being influenced by L1 phonology, is stronger than the link from meaning to orthography. Thus, this qualitative review suggests a possible developmental trajectory of orthographic accuracy from influence of L1 pronunciation on spelling accuracy at lower proficiency progressing to greater challenges based on proper nouns, abbreviations, and longer, less familiar technical words at the higher levels. Such an impression requires careful review of the whole data set to be supported quantitatively and faces the challenge of reconstructing the source of each error, which is a non-trivial task.

It is worth re-emphasizing that a limitation to this analysis is that we do not account for 'clang' effects in the spelling data (e.g., *sees, *case, and *scene for a target such a 'cease') which we found in responses to prompts in reading vocabulary test data (Heilman et al. 2010) or 'hem' vs. 'him', which actually occurred in the corpus. Automated spelling correction cannot correct words that are actually in the dictionary without further refinement of the correction algorithms. It is possible that were clang effects included in the analysis as spelling errors, the results would be different.

8 Conclusion

This paper considered the problem of identifying and measuring orthographic errors in a written IEP corpus by five different groups of L1 speakers across three levels of proficiency. Python coding enabled the identification and enumeration of errors. Using statistical models for overdispersed count data, the findings are that at the low-intermediate level, Arabic speakers make errors at a significantly higher rate per text than peers from some, but not all, L1 groups at the same level of proficiency. These L1 differences are reduced at intermediate and high-intermediate/advanced levels, with Chinese-speaking learners changing some-

what less than other groups in the proportion of texts with spelling errors. Gender was not found to be significant, perhaps because so many of the Arabic-speaking learners were male and hence it is difficult to tease apart this factor from L1 influence. The statistical models did not demonstrate important differences in the groups in spelling errors at higher proficiency levels, which should be seen as a positive outcome for automatic assessment. However, the large numbers of errors by Arabic L1 students in some texts could create a perception that they are worse than other L1 groups, when in fact they are not at levels 4 and 5.

9 Directions for Further Research

Future research might develop automatic coding to identify learner errors based on L1 phonological influence at lower levels of proficiency to confirm the qualitative examples identified in this paper and which are well known to English teachers, for example, confusion among 'ship', 'sheep', 'sip', and 'seep', which involves knowledge of contrasts between tense vs. lax vowels and alveolar vs. post-alveolar fricatives. The results also suggest that many low-intermediate students could benefit from targeted spelling instruction to improve lexical quality. Instructional interventions can be created to determine if instruction makes a difference in speeding up the progress and accuracy of learners. This instruction would not only improve spelling, but also reading comprehension through improved lexical access during text processing (e.g., Hopp 2016). It might also help students make the *correct choices* when choosing the appropriate form in spell-checking and potentially improve their grades for mechanics in tests that grade for that component. Therefore, these data support the call in Humaidan and Martin 2019 for an additional pedagogical intervention in orthographic skills that would improve not only writing but also reading competencies.

Finally, spell-checkers might be made more tolerant of common non-English words, acronyms, and abbreviations, which would further reduce false positives of 'errors' in students' writing.

Acknowledgments

None of this work could have been carried out without the help and guidance of Dr. Na-Rae Han. In addition, we are grateful to our undergraduate research assistants Eva Bacas, Guiseppe Livorno, John Starr, and Sean Steinle. This work was supported by the National Science Foundation for their grant via the Pittsburgh Science of Learning Center, funded award number SBE-0836012. (Previously NSF award number SBE-0354420) and the Social Sciences and Humanities Research Council of Canada.

References

Atkinson, Kevin (2019). *SCOWL: Spell Checker Oriented Word Lists*. Retrieved from http://wordlist.aspell.net.

Baker, Ryan and Aaron Hawn (2022). "Algorithmic Bias in Education". In: *International Journal of Algorithmic Bias in Education* 32, pp. 1052–1092. DOI: 10.1007/s40593-021-00285-9.

Brill, Eric and Robert C. Moore (2000). "An improved error model for noisy channel spelling correction". In: *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pp. 286–293. DOI: 10.3115/1075218.1075255.

Bryant, Christopher et al. (2019). "The BEA-2019 shared task on grammatical error correction". In: *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 52–75. DOI: 10.18653/v1/W19-4406.

Carlson, Andrew and Ian Fette (2007). "Memory-based context-sensitive spelling correction at web scale". In: *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*, pp. 166–171. DOI: 10.1109/ICMLA.2007.50.

Crawley, Michael J (2013). *The R Book*. Chichester, West Sussex: Wiley. DOI: https://onlinelibrary.wiley.com/doi/book/10.1002/9781118448908.

Daller, Michael, John Turlik, and Iain Weir (2013). "Vocabulary acquisition and the learning curve". In: *Vocabulary Knowledge: Human ratings and automated measures*. Ed. by Scott Jarvis and Michael Daller. Amsterdam: John Benjamins, pp. 185–218. DOI: https://doi.org/10.1075/sibil.47.

Damerau, Frederick J. (1964). "A technique for computer detection and correction of spelling errors". In: *Communications of the ACM* 7.3, pp. 171–176. DOI: https://doi.org/10.1145/363958.363994.

Devlin, Jacob et al. (2018). "BERT: Pre-training of deep bidirectional transformers for language understanding". In: *Proceedings of NAACL-HLT 2019*. DOI: 10.18653/v1/N19-1423.

Dunlap, Susan (2012). "Orthographic quality in English as a second language". Thesis. DOI: http://d-scholarship.pitt.edu/13614/.

Garbe, Wolfe (2021a). *https://github.com/wolfgarbe/symspell*. Computer Program. DOI: URL : https://github.com/wolfgarbe/symspell.

– (2021b). *SymSpell: 1 million times faster spelling correction & fuzzy search*. Retrieved from https://github.com/wolfgarbe/symspell.

Hamada, Megumi and Keiko Koda (2008). "Influence of first language orthographic experience on second language decoding and word learning". In: *Language Learning* 38.1, pp. 1–31. DOI: 10.1111/j.1467-9922.2007.00433.x.

Heilman, Michael et al. (2010). "Personalization of reading passages improves vocabulary acquisition". In: *International Journal in Artificial Intelligence in Education* 20.1, pp. 73–98. DOI: 10.3233/JAI-2010-0003.

Hoftstetter, Hedwig et al. (2016). "Modeling Caries Experience: Advantages of the Use of the Hurdle Model". In: *Caries Research* 50. DOI: 10.1159/000448197.

Hopp, Holger (2016). "The timing of lexical and syntactic processes in second language sentence comprehension". In: *Applied Psycholinguistics* 37.5, pp. 1253–1280. DOI: 10.1017/S0142716415000569.

Humaidan, Abdulsamad Y and Katherine I Martin (2019). "Instructor-generated Orthographic Assessments in Intensive English Classes". In: *Handbook of Research on Assessment Literacy and Teacher-Made Testing in the Language Classroom*. Ed. by Eddy White and Thomas Delaney. Hershey, PA: IGI Global, pp. 204–243. ISBN: 9781522569879. DOI: 10.4018/978-1-5225-6986-2.ch011.

Jayanthi, Shardul M., Danish Pruthi, and Graham Neubig (2020). "NeuSpell: A neural spelling correction toolkit". In: *Proceedings of the 2020 EMNLP (Systems Demonstrations)*, pp. 158–164. DOI: 10.18653/v1/2020.emnlp-demos.21].

Juffs, Alan (2020). *Aspects of Language Development in an Intensive English Program*. Routledge Studies in Applied Linguistics. New York: Taylor and Francis. DOI: 10.4324/9781315170190.

Juffs, Alan, Na-Rae Han, and Ben Naismith (2020). *The University of Pittsburgh English Language Institute Corpus (PELIC)*. Online Database. DOI: 10.5281/zenodo.3991977.

Levshina, Natalia (2015). *How to do Linguistics with R*. Amsterdam: Benjamins. DOI: https://doi.org/10.1075/z.195.

Martin, Katherine I and Alan Juffs (2021). "Eye-tracking as a window into assembled phonology in native and non-native reading". In: *Journal of Second Language Studies* 4.1, pp. 66–96. DOI: https://benjamins.com/catalog/jsls.19026.mar.

Mitton, Roger (1996). "Spellchecking by Computer". In: *Journal of the Simplified Spelling Society* 20.1, pp. 4–11. DOI: www.spellingsociety.org/uploaded_journals/j20-journal.pdf.

Naismith, Ben, Na-Rae Han, et al. (2018). "Accurate Measurement of Lexical Sophistication with Reference to ESL Learner Data". In: *Proceedings of the 11th International Conference on Educational Data Mining*. Ed. by Kristy Elizabeth Boyer and Michael Yudelson, pp. 259–265. DOI: http://educationaldatamining.org/EDM2018/.

Naismith, Ben, John Starr, and Eva Bacas (2021). *PELIC Spelling*. Online Database. URL: https://github.com/ELI-Data-Mining-Group/PELIC-spelling.

Näther, Marius (2020). "An in-depth comparison of 14 spelling correction tools on a common benchmark". In: *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pp. 1849–1857. DOI: https://aclanthology.org/2020.lrec-1.228/.

Omelianchuk, Kostiantyn et al. (2020). "GECToR – Grammatical Error Correction: Tag, not rewrite". In: *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 163–170. DOI: 10.18653/v1/2020.bea-1.16.

Ozinov, Filipp (2019). *JamSpell*. Retrieved from https://github.com/bakwc/JamSpell. Online Resource.

Perfetti, Charles and Lesley Hart (2002). "The lexical quality hypothesis". In: *Precursors of functional literacy*. Ed. by Ludo Verhoeven, Carsten Elbro, and Pieter Reitsma. 11th ed. Studies in

Language and Literacy. Amsterdam: John Benjamins, pp. 189–214. DOI: 10.1075/swll.11.14per.

Perfetti, Charles and Josepeh Stafura (2013). "Word knowledge in a theory of reading comprehension". In: *Scientific Studies of Reading* 18.1, pp. 22–37. DOI: 10.1080/10888438.2013.827687.

Schmitt, Norbert and Paul M Meara (1997). "Researching vocabulary through a word knowledge framework: word associations and verbal suffixes". In: *Studies in Second Language Acquisition* 19, pp. 17–36. DOI: https://doi.org/10.1017/S0272263197001022.

Selinker, Larry (1972). "Interlanguage". In: *International Review of Applied Linguistics* 10, pp. 209–231. DOI: https://doi.org/10.1515/iral.1972.10.1-4.209.

vanHout, Roland and Anne Vermeer (2007). "Comparing measures of lexical richness". In: *Modelling and Assessing Vocabulary Knowledge*. Ed. by Helmut Daller, James Milton, and Jeanine Treffers-Daller. Cambridge: Cambridge University Press, pp. 93–115. DOI: https://doi.org/10.1017/CBO9780511667268.

Zeileis, Achim, Christian Kleiber, and Simon Jackman (2008). "Regression Models for Count Data in R". In: *Journal of Statistical Software* 27.8, pp. 1–25. DOI: https://www.jstatsoft.org/article/view/v027i08.

A Appendix A. Table of Effects in hurdle() model[2]

Table 4: Effects in hurdle() model

Predictors	*Incidence Rate Ratios*	*CI*	*p*
Count Model			
(Intercept)	0.32	0.06 – 1.81	0.197
L1 [Chinese]	0.40	0.23 – 0.72	**0.002**
L1 [Japanese]	0.26	0.09 – 0.74	**0.011**
L1 [Korean]	0.40	0.24 – 0.65	**<0.001**
L1 [Spanish]	0.51	0.21 – 1.24	0.135
level_id [4]	0.93	0.65 – 1.32	0.675
level_id [5]	0.80	0.51 – 1.25	0.326
L1 [Chinese] * level_id [4]	1.21	0.60 – 2.46	0.591
L1 [Japanese] * level_id [4]	2.53	0.78 – 8.21	0.121
L1 [Korean] * level_id [4]	1.13	0.59 – 2.13	0.717
L1 [Spanish] * level_id [4]	0.59	0.18 – 1.99	0.397
L1 [Chinese] * level_id [5]	1.61	0.71 – 3.65	0.251
L1 [Japanese] * level_id [5]	1.30	0.27 – 6.41	0.744
L1 [Korean] * level_id [5]	0.63	0.26 – 1.53	0.311
L1 [Spanish] * level_id [5]	1.79	0.50 – 6.40	0.369
Zero-Inflated Model			
(Intercept)	0.63	0.53 – 0.76	**<0.001**
L1 [Chinese]	0.48	0.34 – 0.68	**<0.001**
L1 [Japanese]	0.68	0.38 – 1.21	0.187
L1 [Korean]	0.78	0.57 – 1.06	0.109
L1 [Spanish]	0.90	0.51 – 1.61	0.726
level_id [4]	0.47	0.38 – 0.59	**<0.001**
level_id [5]	0.27	0.21 – 0.36	**<0.001**
L1 [Chinese] * level_id [4]	1.52	1.00 – 2.30	**0.048**
L1 [Japanese] * level_id [4]	1.31	0.67 – 2.56	0.435
L1 [Korean] * level_id [4]	1.03	0.70 – 1.52	0.884
L1 [Spanish] * level_id [4]	0.94	0.46 – 1.95	0.878
L1 [Chinese] * level_id [5]	2.23	1.39 – 3.58	**0.001**
L1 [Japanese] * level_id [5]	0.91	0.39 – 2.12	0.834
L1 [Korean] * level_id [5]	1.02	0.63 – 1.65	0.926
L1 [Spanish] * level_id [5]	2.32	1.04 – 5.16	**0.039**
Observations	5774		
R^2 / R^2 adjusted	0.040 / 0.038		

[2]Variance explained in Poisson/hurdle() models, which are special types of logistic regression, is difficult to interpret. The table of results from the model in Appendix A includes an R^2 statistic that suggests that just 3.8% of the variance in the entire model (count and zero) is accounted for by the factors of L1 and level. This result is unsurprising given that 78.7% of texts in the entire sample are error free.