Learner Corpus Research: LCR2013 Conference Proceedings

BeLLS 2015 Vol. 6, ISBN 978-82-998587-7-9, s. 107–132 BeLLS.uib.no

Patterns of misspellings in L2 and L1 English: a view from the ETS Spelling Corpus[1]

*Michael Flor, Yoko Futagi, Melissa Lopez, and Matthew Mulholland**
Educational Testing Service, Princeton, NJ, USA

Abstract

This paper presents a study of misspellings, based on annotated data from the ETS Spelling corpus. The corpus consists of 3000 essays written by examinees, native (NS) and non-native speakers (NNS) of English, on the writing sections of GRE® and TOEFL® examinations. We find that the rate of misspellings decreases as writing proficiency (essay score) increases, both in TOEFL and in GRE. Severity of misspellings depends on writing proficiency and not on NS/NNS distinction. Word-length and word-frequency have strong influences on production of misspellings, showing patterns associated with proficiency. For word-frequency, there is also a clear effect of NS/NNS distinction.

Keywords: misspellings; learner corpus; annotation; writing proficiency; word length; word frequency

* *Principle contact:*
Michael Flor, Research Scientist
NLP group, R&D Division, Educational Testing Service, Princeton, NJ, USA
Tel.: 609-734-1591
E-mail: mflor@ets.org

[1] Many thanks to Beata Beigman Klebanov, Michael Heilman, and Swapna Somasundaran, for valuable comments during preparation of the manuscript. This article has also benefited from the comments of anonymous reviewers and the editors of the LCR2013 proceedings.

Bergen Language and Linguistic Studies (BeLLS), May 30th 2015. © Flor, Futagi, Lopez and Mulholland
DOI: 10.15845/bells.v6i0.811

1. Introduction

The ability to write text with adequate spelling is an important aspect of writing proficiency, both for native speakers of a language (Lunsford and Lunsford 2008) and for foreign language learners (Bestgen and Granger 2011). Effective spelling is important for clarity of communication, and also because of its social overtones (Cook 1997). Nowadays, good spelling is also important for adequate automated processing of texts – for example, misspellings can distort various automated text metrics (Granger and Wynne 1999), and pose problems for automatic content scoring systems (Sukkarieh and Blackmore 2009; Leacock and Chodorow 2003). Bestgen and Granger (2011) have demonstrated that spelling errors may serve as a reliable predictor of the quality of L2 texts.

Corpus-based studies of spelling errors are traditionally focused on identifying the causes of spelling errors. Such studies are considered important for pedagogy (Botley and Dillah 2007; Cook 1997; Bebout 1985) and for development of spellcheckers (Pollock and Zamora 1984; Turba 1981). For typology of potential causes of misspellings, a classical distinction is between typographic errors, cognitive errors and phonetic errors (Kukich 1992). When it is assumed that the writer/typist knows the correct spelling but simply makes a motor coordination slip (e.g. *the→teh*, *spell→speel*), such errors are considered typographical (typos). Cognitive errors (e.g. *receive→recieve*, *conspiracy→ conspiricy*) are presumed to stem from misconceptions or a lack of knowledge on the part of the writer/typist. Phonetic errors are those where the writer substitutes a phonetically similar sequence of letters for the intended word (e.g. abyss→abiss). However, reliable classification of spelling errors in any given corpus is problematic. This is succinctly illustrated by Kukich:

> It is frequently impossible to ascribe a single category to a given error. Is 'recieve' necessarily a cognitive error, for example, or might it simply be a typographic transposition error? Similarly, is 'abiss' a phonetic or typographic error? Fortunately, it is often unnecessary to categorize errors in order to develop a useful spelling correction technique because many correction techniques handle typographic and cognitive misspellings equally well (Kukich 1992, 387).

Notably, with the addition of phonetic algorithms, automatic spellcheckers can handle many phonetic misspellings as well (Flor 2012; Pollock and Zamora 1984).

The question whether analysis of patterns of spelling errors is necessary for improving spellcheckers is still open. Recent research indicates that generic spellcheckers, that are developed for native language speakers, are not well suited for language learners' needs, and thus studies of L2 spelling errors are considered important for improving spellcheckers (Hovermale 2010; Rimrott and Heift 2008; Mitton and Okada 2007). On the other hand, Flor and Futagi (2012) have demonstrated that a contextually driven spellchecker corrects spelling errors, generated by native and non-native English writers, with almost the same rate of success. In a related study, Flor (2012) has demonstrated that the error correction performance of an automatic spellchecker is influenced by the overall quality of a text (i.e. a holistic writing-proficiency score assigned by a human scorer). The error correction rate was higher for better quality essays and lower for lesser quality essays. This leads to an intriguing question – to what extent patterns of misspellings reflect the native/non-native distinction, and to what extent they reflect overall writing proficiency levels.

In this paper, we present a large-scale study of patterns of misspellings in essays written by native and non-native speakers of English to the writing prompts of TOEFL and GRE examinations. We utilize the large annotated corpus of misspellings that was developed for evaluating the performance of an automatic spell-checking system (Flor and Futagi 2013, 2012; Flor, 2012). An important feature of this corpus is that

misspellings were annotated in full context of the essay and corrections were supplied. We specifically refrain from guessing the sources of individual spelling errors. We use several objective attributes of misspellings: severity of error (as approximated by edit distance), length of the intended correct word and language-frequency of the intended word.

The paper is structured as follows. First, we present details about the corpus, and explain the annotation process (Sections 2-3). Section 4 provides some descriptive statistics on the spelling errors in this dataset. Section 5 introduces the breakdown for native and non-native speakers. Section 6 describes an analysis of fusion errors. Section 7 presents analysis of misspellings by edit distance (to correct form). Section 8 looks at the relation between misspellings and the length of the intended word. Section 9 looks at the relation between word frequency and misspellings.

2. The corpus

The ETS Spelling Corpus is a collection of essays systematically annotated for misspellings. It was produced at the Educational Testing Service in 2011–2013, for the task of developing and evaluating new spell-checking software. The spellchecker development research has produced an advanced contextually-aware spellchecker, Conspel, as described by Flor (2012), Flor and Futagi (2012). This system is already included in automated essay-analysis systems at ETS.

The corpus comprises essays written by examinees on the writing sections of GRE® (Graduate Record Examinations) and TOEFL® (Test of English as a Foreign Language) (ETS 2011a,b). The TOEFL test includes two different writing tasks. For the Independent task, examinees receive a predefined topic and have to write a short opinion essay on that topic. On the Integrated task, examinees receive two different sources presenting conflicting arguments about some issue. Examinees' task is to write a summary essay comparing the arguments from the two sources. The GRE Analytical Writing Section also includes two different writing tasks. On the GRE Issue task, participants write a short argumentative essay by taking a position on an assigned topic. On the GRE Argument task, test takers are presented with a short argument text (the prompt) and then write an essay evaluating those arguments. The writing tasks of both TOEFL and GRE tests are delivered (via internet) on computers at test centers around the world. The setting makes mandatory use of the standard English language computer keyboard (QWERTY). Editing tools, such as a spell checker, are not provided in the test-delivery environment (ETS 2011a). All writing tasks have time constraints.

To illustrate the kinds of spelling errors encountered, the excerpt presented below was taken from a low scoring essay. In addition to spelling errors, it also involves grammar errors and anomalous word order.

> the person who is going to be take a movie to saw the film is to *takn* to pass the star heroes movies . *iam* suppose to *takn* that is not *valied* is to *distroy* to take all *heroneos*

Currently, the corpus includes 3000 essays, for a total of 963K words. All essays come from operational test administrations conducted between the years 2007 and 2009. The essays were selected equally from the two testing programs (4 tasks, 10 prompts per task, 75 essays per prompt). The corpus essays cover the full range of essay scores (as a proxy for English proficiency levels) for each task. Scores range from 1 to 6 on GRE, and from 1 to 5 on TOEFL, with higher score indicating better proficiency. For each prompt, we sampled an approximately equal number of essays for each score level.[2] The majority of

[2] While the typical distribution of essay scores is 'normal'-like, i.e. most essays get scores in the middle of the rating scale and

essays in this collection were written by examinees for whom English is not the first language. Out of the 1500 TOEFL essays, 1481 were written by non-native speakers of English (98.7%).[3] Out of 1500 GRE essays, 866 were written by non-native speakers of English (57.7%).

3. Annotation

The annotation procedure and tools used for this project were described in detail by Flor and Futagi (2013). Annotators were required to 1) find and mark misspellings in texts, and, 2) provide the correct word(s) for each misspelling. Since human annotation of misspellings is tedious and also error-prone (annotators often miss some misspellings), we used a semi-automated technique, where a system auto-detects many misspellings (by consulting dictionary files) and highlights misspelled words. However, the annotator must not only accept/reject automated suggestions, but also check if any misspellings were missed by the system.

The annotation scheme for this project defines four classes of misspellings (see Table 1). The primary distinctions are 1) whether the words (strings) are found in the dictionary, and 2) whether an error spans one or multiple word tokens. Single-token non-word misspellings are spelling errors where an intended word was misspelled so that the result is a string that is not in a dictionary, e.g. "*businees*" where "*business*" was intended. This category also includes fusion errors (e.g. "*taketo*" where "*take to*" was intended). Another category involves single-token errors where the resulting string happens to be found in the dictionary, but is not the proper or 'intended' word in the context. For example, typing "*they*" for "*then*". Such errors are known under a variety of labels, such as malapropisms, real-word errors, confusion errors or contextual errors (because the error is apparent only in context). The third category are non-word misspellings that span more than a single word. Note that these are not simply cases of adjacent single-word misspellings. Multi-token misspelling is defined as a case where a correction involves simultaneously more than one token (e.g. "*mor efun*" when "*more fun*" was intended). For this category, at least one of the tokens in a sequence is not a dictionary word. The complementary fourth category are multi-token real-word misspellings where all component strings are dictionary words (e.g. "*with out*" for "*without*"). This categorization schema was motivated by the efforts of the spell-checking software development. Single-token non-word errors (NW) are the most abundant type of errors (see Table 1), while dictionary lookup is the most reliable error-detection approach (Kukich 1992).

In the annotated corpus, different spelling variants (alternative spellings of same word) were acceptable. This consideration stems from the international nature of TOEFL and GRE examinations – the examinees come from all around the world, being accustomed to either British, American, or other English spelling standards; so, it is only fair to accept all of them.

In annotation we deliberately ignored repeated words (e.g. "*the the*"), missing spaces around punctuation (e.g. "*...home.Tomorrow...* ") and improper capitalization (e.g. "*BAnk*"). Many of the essays in our corpus have inconsistent capitalization, while some essays are written fully in capital letters. Although issues of proper capitalization fall under the general umbrella of orthographic errors, we do not consider them 'spelling errors'. Another issue involved is the boundary between spelling and other grammar errors, especially cases bordering on the real-word error category. In annotation we did not consider improper

fewer essays get extreme scores, for this study we selected a stratified sample – for each prompt we took almost equal numbers of essays from each scoring level, so as to provide a good amount of data for extreme ends of the rating scale. For example, for one of the GRE prompts, for scores 1,2,3,4,5,6 we had 9,14,14,14,12,12 essays respectively.

[3] The native language of a test-taker is self-reported. Writers of our 1500 GRE essays came from 51 countries, of them 583 from USA and 517 from India. Writers of our 1500 TOEFL essays came from 55 countries, the largest subgroups being from China (393), Korea (351) and Japan (209).

inflectional variants as misspellings (such as may arise when subject-verb agreement is violated, e.g. "*the kids plays in the park*", etc.). Also, preposition errors (such as "*in/at*") and article errors ("*a/an*") were not considered as misspellings and were not annotated.

Table 1. Classification of annotated misspellings in the ETS Spelling Corpus

Type	Description	Count in corpus
1	(NW) single token Non-Word misspelling (e.g. "businees") also includes fusion errors (e.g. "niceday" for "nice day")	21,142 (80.05%)
1b	NW misspelling for which no plausible correction was found in context[4]	52 (0.20%)
2	multi-token non-word misspelling (e.g. "mor efun" for "more fun")	574 (2.17%)
3	(RW) single token Real-Word misspelling (e.g. "they" for "then")	3,393 (12.85%)
4	multi-token Real-Word misspelling (e.g. "with out" for "without")	1,251 (4.73%)
	Total	26,412 (100%)

The dictionaries used for this project include about 360,000 entries. The core set includes about 130,000 single token entries and 110,000 multi-word entries, providing a comprehensive coverage of modern English vocabulary. This lexicon includes all inflectional variants for a given word (e.g. 'love', 'loved', 'loves', 'loving'), and international spelling variants (e.g. American and British English). Additional dictionaries include about 120,000 entries for international surnames and first names, and names for geographical places, brand names and many additional names. Inclusion of a wide variety of names is particularly important for an international testing setting, such as TOEFL and GRE examinations – essays written on these tests often include names of famous people, places and brands from all over the world. Inclusion of names-dictionaries provides for a certain shift in categorization of spelling errors. For example, "*hince*" is a misspelling of 'hence', but 'Hince' is also a common surname, so in our corpus "*hince*" is classified as a real-word misspelling.

4. Annotation process

Corpus annotation was carried out in two stages. This paper presents the results of the first stage, where the exhaustive annotation effort focused on non-word misspellings. An in-house annotation software was developed for the project (Flor and Futagi 2013). It automatically highlighted all non-words in a given text and provided candidate corrections. Each text was independently reviewed by two annotators. They were required to check all highlighted strings, accept/reject flagged words and candidate corrections, and could also provide their own corrections.[5] They also marked real-word misspellings.[6] Classification of annotated strings was automatic: an annotated string was auto-marked as non-word if it was not found in the system dictionaries, and as a real-word misspelling if it was found in the system dictionaries. Annotators also

[4] For example: "In agriculture side, a lot of crops >exploish<. And the pollution is growing."

[5] Annotators could provide out-of-dictionary corrections, or accept non-dictionary words as correctly spelled. Such new words were later vetted and added to the dictionary.

[6] In the first stage, the annotators were instructed to mark real-word misspellings when they saw them, but they were not instructed to look for such errors exhaustively. In the second stage, we focused specifically on real-word errors. Annotators used the same software, and saw the adjudicated annotations from the first stage. The task was to scan each essay and look for additional (yet undiscovered) cases of real-word misspellings. This task is currently in adjudication, not yet fully completed.

marked multi-token errors, and the annotation software automatically tagged them as 'multi-token with non-word' (if at least one of the tokens was a non-word) or 'multi-token real-words'. Inter-annotator agreement was calculated in two steps. First, we considered agreement on marking of misspellings (with two categories: misspelled or not misspelled). Agreement was 99.3%, Cohen's Kappa=0.85, p<.001 (notably, most words in the whole corpus are not misspelled). A strict criterion was applied for calculating agreement on corrections: correction of a misspelling was considered 'agreed' only if two annotations both marked a word as misspelled and provided exactly same correction. Among all cases initially marked by annotators, they strictly agreed on correction in 82.6% of the cases. For all cases that were not in strict agreement, all differences and difficulties were resolved by an adjudicator.

5. Descriptive Statistics

This section provides general descriptive statistics about the texts in our corpus and misspellings found in them. The annotated corpus has 3000 essays. Average essay length is 321 words (the range is 28-798 words). 130 essays turned out to have no misspellings at all. Total spelling error counts are given in Table 1. The average error rate was 2.74% for all spelling errors in general, 2.2% for single-token non-words. Notably, both TOEFL and GRE scoring guides do not require penalizing essays for spelling errors (ETS 2011a,b). In general, lower quality essays often involve many spelling, mechanics and grammar errors, though their holistic scores also take into account their 'narrative' and topical/argumentative quality (Ramineni et al. 2012a,b). In this study, we use the holistic essay scores, assigned by human scorers, as estimators of writing proficiency.

Table 2. Summary statistics for the ETS Spelling Corpus

	GRE Argument	GRE Issue	TOEFL Independent	TOEFL Integrated	TOTAL
Total essays	750	750	750	750	3,000
Essays without misspellings	60	21	18	21	120
Total Word Count	263,578	336,301	212,930	151,031	963,840
Average Word Count	351	448	284	201	321
Total count of Misspellings	5,935	7,962	7,285	5,230	26,412
Misspellings as % of all words	2.25%	2.37%	3.42%	3.46%	2.74%

Next we examine whether there are differences in the rate of misspellings between the testing programs. Table 2 provides the breakdown of essays and spelling error counts by program/task. Essays written for the GRE tasks are, on average, longer than those written for TOEFL. The proportions of misspellings are also unequal. Overall, essays written for TOEFL have a larger proportion of spelling errors than essays written for GRE. We used a test for Difference Between Two Independent Proportions to compare the error rates.[7] The average percent of misspellings for GRE Issue essays (2.37%) is significantly larger than the average percent of errors for GRE Argument essays (2.25%), z=2.96, p<0.002. The average percent of misspellings for TOEFL Independent essays (3.46%) is significantly larger than the average percent of errors for GRE Issue essays (2.25%), z=23.161, p<.0001. For the two TOEFL tasks, the difference in average error rates is not significant (p=0.25).

In sections 6-11, we focus on the single-token non-word errors, keeping the distinction between

[7] Two comparisons are within testing program (GRE or TOEFL). The third comparison is between GRE Issue and TOEFL Independent – those have similar writing assignments: write and support an opinion on a given topic.

GRE and TOEFL essays, but dropping the task distinction. Figure 1 shows that such errors are the largest category for every program/task. TOEFL Integrated essays have the largest proportion of single-token non-word errors (85%), which is significantly larger than that for TOEFL Independent (80.9%), z=5.918, p<.0001. The proportion for TOEFL Independent is, in turn, significantly larger than for GRE Issue (78.4%), z=3.829, p<.0001. The proportion for GRE Issue is significantly larger than for GRE Argument (76.7%), z=2.428, p<0.008. Despite those differences, it is evident that single-token non-word errors are the most prevalent type of error in all programs/tasks by a very large margin, and unification into two groups (TOEFL vs. GRE) is reasonable.

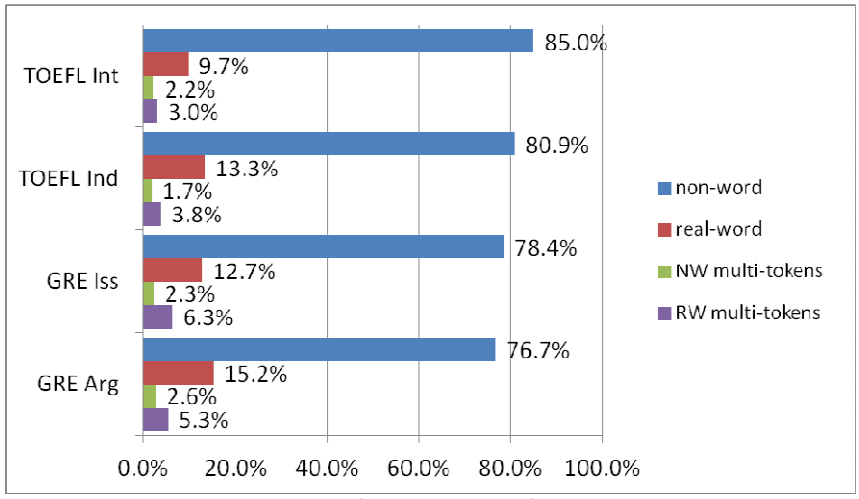

Figure 1. Relative proportions of various types of misspellings by program/task

6. Native and non-native speakers

In this section we examine some differences in the number of misspellings between native (NS) and non-native (NNS) speakers of English. Following Cook (1997), we start with a broad question: which population produced more error-less essays? The data is presented in Table 3. For NS writers, 10.7% of essays have no spelling errors, while for NNS writers, only 2.3% of essays have no errors. This may suggest that NNS writers are more prone to making spelling errors. Such a suggestion may be supported with data from Table 2 in the previous section, showing that the error rate (misspellings as % of all words) in TOEFL is larger than in GRE; in TOEFL 98.7% of essays in our corpus are NNS, whereas NS are well represented in GRE.[8]

Table 3. Essay counts in ETS spelling corpus, by test program and NS/NNS status

Group	TOEFL	GRE	Total count	Essays without misspellings
NS	19	634	653	67 (10.7%)
NNS	1481	866	2347	53 (2.3%)

[8] In the analyses in this paper, we use only TOEFL NNS data and exclude the 19 TOEFL NS essays.

A closer look at GRE essays, by NS/NNS status and essay score, reveals a more complex picture (see Figure 2). Most of the NS essays in our corpus have high scores (4-6), while most of the essays from NNS have low scores (1-3). Thus, we should consider the extent to which the quantity of spelling errors reflects NS/NNS status or writing proficiency (as represented by essay score). This consideration is presented in Figure 3, consisting of two parallel bar charts – one for GRE and one for TOEFL. For each population, the average percent of misspelled words (per essay) decreases with higher proficiency. There is a gap between NS and NNS at lower proficiencies (NS make fewer misspellings, on average), but it closes quickly as the scores go up. Analysis of variance found a significant main effect of score, $F(5,1500)=65.173$, $p<.0001$, a significant effect of NS/NNS status, $F(1,1500)==48.86$, $p<0.0001$, and significant interaction $F(5,1500)=5.447$, $p<0.0001$.The data for TOEFL NNS essays show a similar trend: the average percent of misspellings (per essay) decreases with higher writing proficiency.

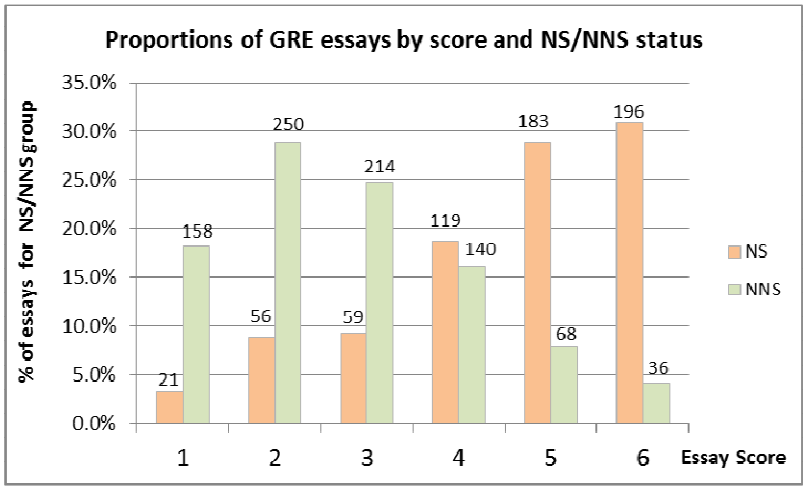

Figure 2. Percent of GRE essays by score and NS/NNS-status. The total for each color is 100%. Note: Essay counts are provided, but the bars are scaled by relative percentage.

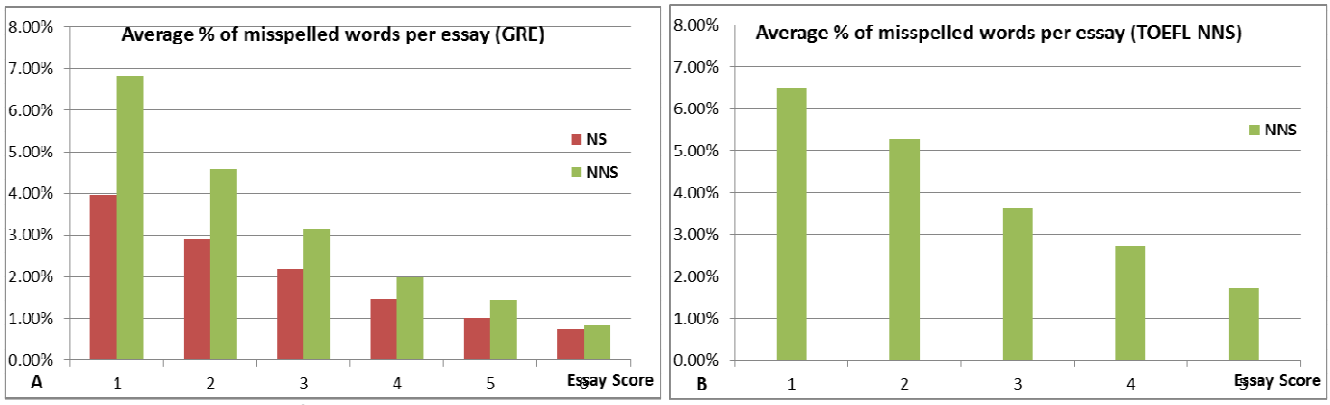

Figure 3. Average percent of misspelled words per essay, by NS/NNS and score (panel A – GRE data, panel B – TOEFL data)

7. Fusion errors

Fusion errors (also known as run-on errors) are single token non-word misspellings that result from fusion of two intended words, often due to omission of the space or hyphen between words. Examples from our corpus: *schoolstudents* (school students), *allthe* (all the), *doesnot* (does not), *inour* (in our), *selfconfidence* (self-confidence). Our research question in this section is whether the rate of fusion errors (among all single-token non-word misspellings) is related to writing proficiency and whether it is influenced by NS/NNS status. For GRE data, we compared how many of the single-token non-word errors are fusions, by NS/NNS status and essay score level. The data are presented in the two panels of Figure 4. NNS writers produce more fusion errors than NS writers. A Mann–Whitney U test indicates that the distribution of fusions is not the same for GRE NS and GRE NNS groups ($U=4$, $n1=n2=6$, $p<0.02$, one-tailed). The left-hand chart indicates that for NS writers only data for score level 1 seems to be different (6.8%), while the proportions of fusions for other score levels tend to stabilize around 3%. The improvement comes 'early' and stabilizes. Indeed, for GRE NS data, the proportion of fusions at score level 1 is significantly higher than at score level 2 ($z=1.898$, $p<0.03$), but none of the pair-wise comparisons of proportions for other score levels are significant. For the GRE NNS data, the tendency is different, continued reduction in proportion of fusions as writing proficiency improves. The difference of proportions between GRE NNS groups of score 2 and 3 is significant ($z=1.892$, $p<0.03$), and so is the difference between groups 3 and 4 ($z=1.945$, $p<0.03$), but the differences between groups 4, 5, and 6 are not statistically significant. For TOEFL NNS data, the proportions of fusions are similar for all five score levels, fluctuating around 5%.

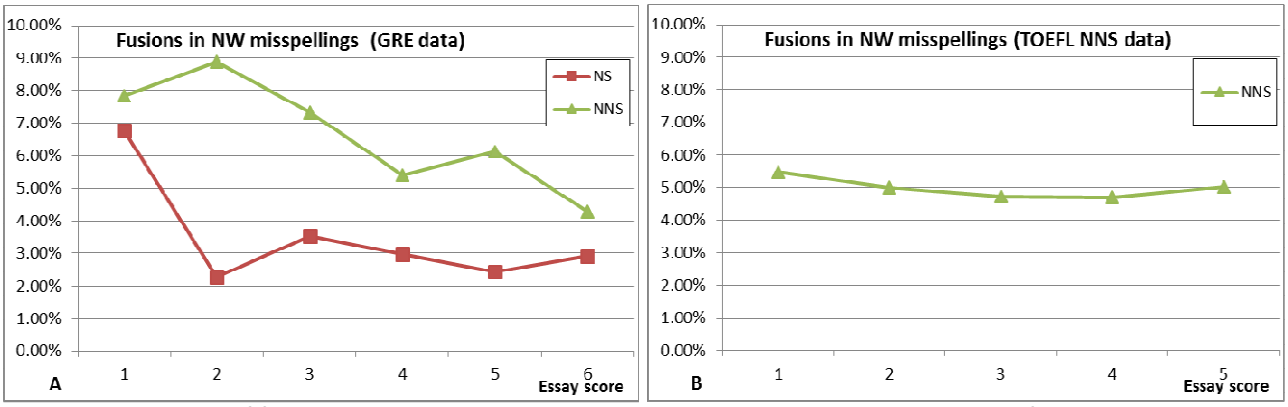

Figure 4. Proportion of fusion errors among single-token non-word misspellings, by essay score, for NS and NNS populations in GRE data (panel A) and NNS TOEFL data (panel B). Total counts are 80, 609, and 509, respectively.

Fusion errors may arise randomly when writers type rapidly on the keyboard. If this were the only cause for fusions, one might expect the rate of fusions to be roughly equal for all proficiency levels. However, fusion errors (or a writer's failure to correct them) may also be influenced by lack of knowledge on how to correctly write certain English word combinations. In such case, one might conjecture that the rate of fusions would decrease as the overall writing proficiency improves. The pattern of results obtained for GRE NS and GRE NNS data is consistent with the knowledge hypothesis: the rate of fusions decreases as essay scores increase. For native English speakers the knowledge transition is rather abrupt – from score level 1 to score level 2, and rather steady after that. On the other hand, the decrease in the rate of fusions for GRE NNS data continues at least until score level 4, which might reflect continued acquisition of orthographic knowledge for word combinations. Contrary to GRE NNS data, the TOEFL NNS data is not

consistent with the knowledge-influence hypothesis. The rate of fusions in the latter is almost flat across all score points. This discrepancy between the two NNS populations is puzzling, leaving an open question for further research.

8. Error severity

Edit distance is the minimal number of characters that need to be changed in order to transform one string into another (Levenstein 1966, Damerau 1964). Edit distance between the correct word-form and the error can be used as a rough indicator of the severity of a spelling error.[9] A misspelling that differs from the correct form by one character is a rather slight misspelling, while a misspelling that differs by 4 characters is a rather strong distortion of a word. For example, *sucessful→successful* (e.d.=1), *voultaneer→volunteer* (e.d.=4), *naiberhouad→neighborhood* (e.d.=6). In this study we use simple edit distance, where insertion, omission, and substitution of one character count for 1 point, and transposition of two letters also counts for 1 point.

Our research question is whether severity of errors is related to writing proficiency and to the NS/NNS distinction. Table 4 presents the breakdown of single-token non-word misspellings by edit distance, for NS and NNS populations in our corpus. Overall, NNS writers produce larger proportions of severe errors. The proportion of least severe errors (e.d.=1) is larger for NS writers (83.15%) than for NNS writers (79.47%). This difference is statistically significant, $z=4.436$, $p < 0.0001$. We may be tempted to infer that NNS writers produce more severe errors. However, a breakdown by proficiency level (essay score) reveals a different picture (see Figure 5). On the left panel of Figure 5, there seems to be no effect of NS/NNS status, and the average error severity declines as writing proficiency improves. With the GRE data (all single-token NW errors), we conducted a two-way ANOVA by essay score[10] and NS/NNS status, with edit distance as dependent variable. There is a significant effect of score, $F(5, 10799)=9.954$, $p<0.0001$, while the effect of NS/NNS status is decisively not significant ($F(1, 10799)=0.772$, $p=0.38$, and the interaction is also not significant ($F(5, 10799)=1.606$, $p=0.155$. For the TOEFL data (only NNS), a one-way ANOVA showed a strong effect of proficiency level, $F(4, 10260)=20.699$, $p<0.0001$. Our present finding is that 'severity of spelling errors' (as measured by edit distance) is directly related to writing proficiency. There is no evidence that the NS/NNS distinction has any role.

Table 4. Sample misspellings and their edit distances

Edit Distance	Total NW errors (tokens)	NS data		NNS data	
		Count	%	Count	%
1	16908	2393	83.15%	14515	79.47%
2	2957	372	12.93%	2585	14.15%
3	827	88	3.06%	739	4.05%
4	296	22	0.76%	274	1.50%
5	100	2	0.07%	98	0.54%
6	41	1	0.03%	40	0.22%
7	7			7	0.04%
8	2			2	0.01%
9	4	4		4	0.02%
Totals:	21142	2878	100%	18264	100%

[9] Notably, edit distance plays an important role in automatic spelling correction algorithms, from early proposals in the 1960s, through the advent of advanced spell-checkers (Kukich 1992) and today with contextually driven spell-checking (Flor 2012).

[10] Each misspelling carries the score of the essay where it was found.

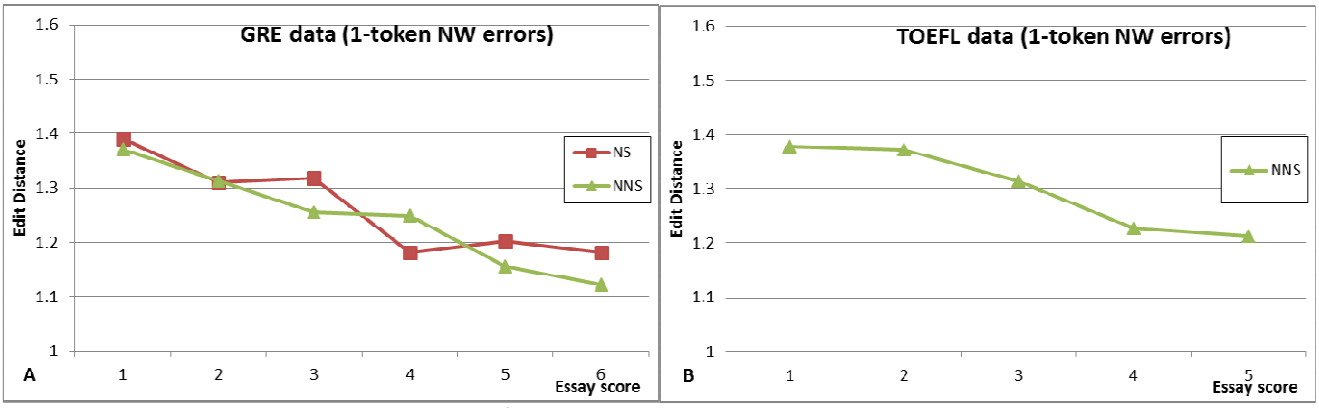

Figure 5. Average error severity (edit distance) for non-word misspellings, by essay score (panel A – GRE data, panel B – TOEFL data)

9. Word length and spelling

In this section we address the relations between word length and misspellings: whether short or long words are more likely to be misspelled, and whether there is a difference between NS and NNS populations on this issue. We consider only single-token non-word errors which are not fusions (i.e. we focus on cases where one intended word was misspelled and resulted in one non-word). There are 2715 such cases from GRE NS data, 7395 cases from GRE NNS data, and 9751 cases from TOEFL NNS data. For all such errors, we took the corrections (the intended words) and calculated their lengths (number of characters). Figure 6 presents the relative distribution of errors for each population, by length of the intended correct word. For all three groups, the proportion of errors increases as word length increases from 2 to about 7-8 characters, and then decreases, with a sharp drop after length 10. There are some obvious differences between NS and NNS.

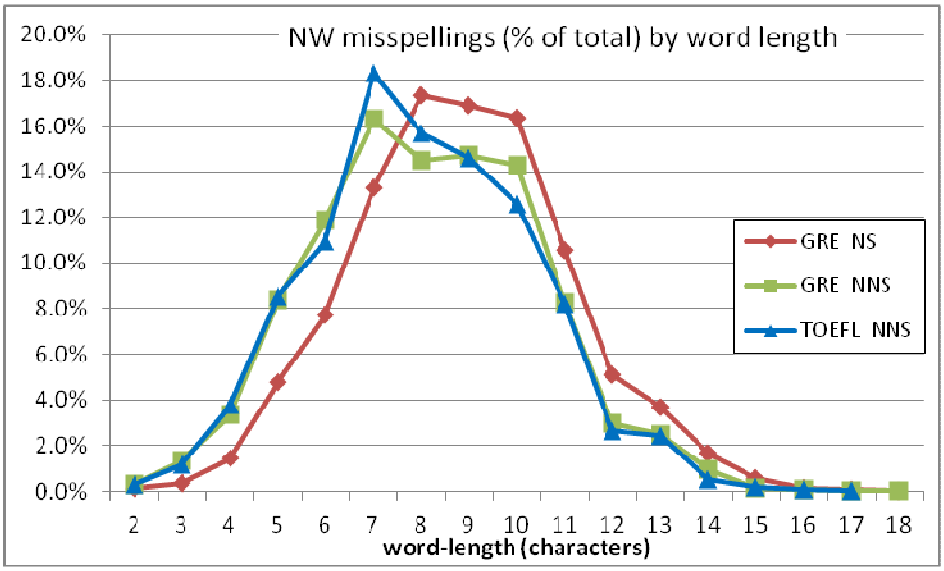

Figure 6. Relative proportion of single token NW spelling errors by length of the intended word. For each population group – distinct line on the chart – all errors sum to 100%

In GRE data, for words 3-7 characters long, NNS writers tend to produce more errors than NS writers (for each word length, the point on the NNS line is higher than the corresponding point on the NS line). For example, about 5% of all errors made by NS are words of length 5, but for NNS, such words account for more than 8% of errors. Words of length 6 constitute about 8% of all errors for NS, but 12% for NNS. This trend reverses for longer words (length 8-14 characters) – NS writers produce relatively more misspellings in such words than NNS writers do. The pattern for TOEFL NNS data resembles GRE NNS more than GRE NS data. The two NNS lines are almost overlapping for word lengths 2-6, and similarly peak at word length 7. They also overlap for word lengths 11 and longer. For words of length 7-10, TOEFL NNS data shows a rather steep drop in the proportion of errors, while GRE NNS data shows a leveled decrease. The average length of the intended word is 8.04 characters for TOEFL NNS, 8.13 for GRE NNS and 8.85 for GRE NS.

The data above suggest that NNS writers have more trouble with shorter words than NS writers. This is quite puzzling. One expects the non-native speakers to have more trouble spelling longer words, which are supposed to be more difficult. However, we should consider different interpretations for this finding. NNS writers may overuse short words and underuse long words, as compared to NS writers, and thus the opportunities for misspellings might be different between the groups. Alternatively, it could be an effect of proficiency levels, rather than NS/NNS status (in the GRE dataset, the distribution of NS essays is skewed toward higher scores, and the distribution of NNS essays is skewed towards lower scores, see Figure 2). In the following paragraphs we consider the distribution of words by word length, both for all words in the essays, and for the misspelled words.

One direct approach is to compare data for misspellings from GRE essays, by NS/NNS populations and by essay score. This dataset has 10110 single-token NW errors (excluding fusions). Panel A in Figure 7 presents average length of the intended word for the two groups, by six proficiency levels. It illustrates that there is indeed an effect of proficiency – average length of intended word (misspelled to NW) increases with higher proficiency. A two-way ANOVA indicated a significant main effect of essay score $F(5, 10110)=22.071$, $p<0.0001$, and a significant main effect of NS/NNS status, $F(1, 10110)=12.885$, $p<0.0001$. The interaction was not significant ($p=0.188$). On average, at almost each proficiency level, native English speakers misspell words that are slightly longer than those misspelled by non-native speakers. The small gap between NS and NNS closes at score=4, but then widens at scores 5 and 6. The data for TOEFL NNS population (n=9751 misspellings) is presented in panel B of Figure 7. A one-way ANOVA showed a significant main effect of score level, $F(4,9751)=56.319$, $p<0.0001$. Indeed, as the essays improve, the tendency for misspellings shifts to longer words, in all three populations.

If the finding presented in Figure 6 stems from 'opportunity' to misspell long words, we need to consider overall patterns of usage for words of different lengths. We expect that as essays get better, utilization of longer words also increases, and with that – the opportunity to misspell them.

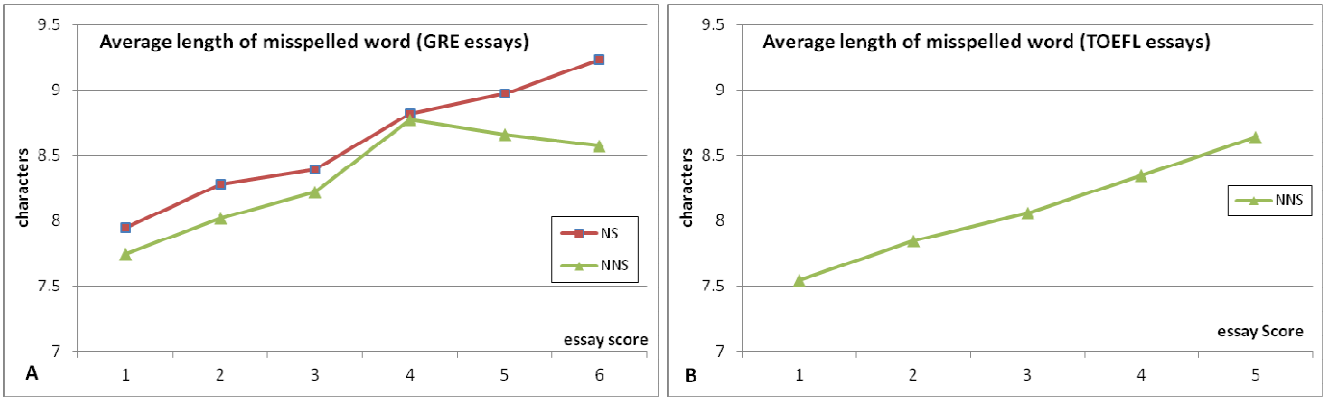

Figure 7. Average length of 'intended' words (that resulted in single token NW errors), by essay score, for three populations (panel A – GRE data, panel B – TOEFL data)

We computed average word length (for all words) per essay for the 1500 GRE essays. The data is presented in Figure 8, panel A. A two-way ANOVA revealed a significant main effect of proficiency (essay score): $F(5,1500)=18.667$, $p<0.0001$. The main effect of NS/NNS was marginally significant, $F(1,1500)=3.317$, $p=0.069$. The interaction was not significant ($p=0.711$). There is a slight tendency for NS writers to use longer words, however this tendency is not statistically significant. The gap closes as proficiency improves, and disappears at the high proficiency levels (scores 5 and 6). A similar analysis was performed for 1481 TOEFL NNS essays. A one-way ANOVA showed a significant effect of proficiency level (score): $F(4,1481)=24.181$, $p<0.0001$. More proficient writers use a higher proportion of long words than do less proficient writers. This provides some support to the idea of increasing opportunity to misspell longer words. However, although the average word length per essay increases with higher proficiency, note that the difference between averages of extreme score groups (1 vs. 6) is just about 0.3 characters.

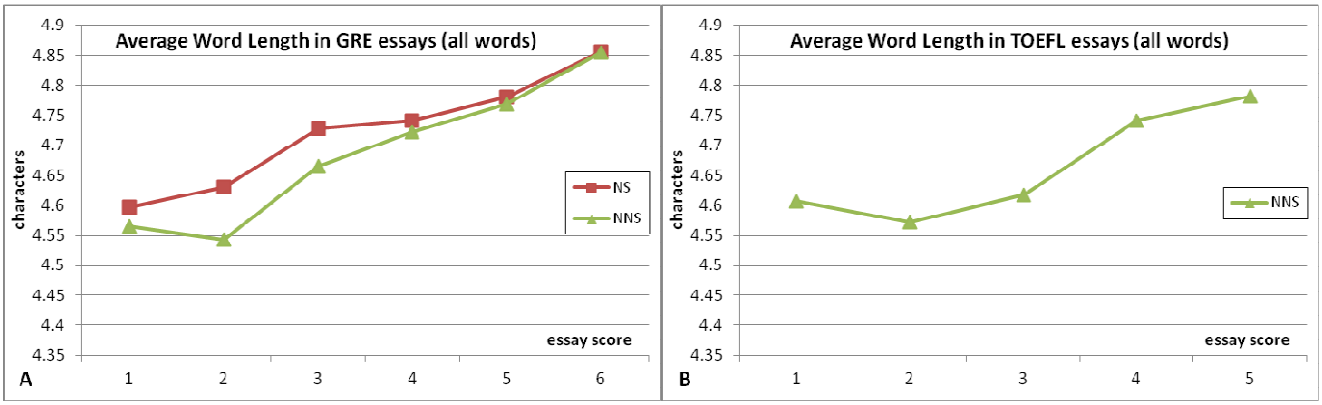

Figure 8. Average length of words per essay, by essay score, for three populations (panel A – GRE data, panel B – TOEFL data)

As noted above, for each group (GRE NS/NNS and TOEFL NNS), we tallied the number of all words by word length. We also tallied the number of single-token non-word misspellings for each group, tallying by the length of the correct intended word. For each group, we calculated the proportion of misspelled words out of all words, for each word length. The data is presented in Figure 9. This detailed breakdown reveals patterns that were not clear in the analyses above. Despite inevitable noise, and small

counter-examples, the general trend is common in all populations: for words of every length group, the proportion of NW errors made with words of that length tends to decrease as the writers' proficiency increases. For example, in GRE essays written by NNS, for words of length 8 characters, lowest-proficiency writers make a spelling error in 15% of their 8-character words, at proficiency level 4 they misspell 3.8% of their 8-character words, and at proficiency level 6 – just 2%. Thus, as writers of greater proficiency introduce more of the longer words (Figure 8), the proportion of misspellings they make with longer words actually decreases (Figure 9). This can be interpreted as improving lexical knowledge.

Another trend is also visible in Figure 9. For every population, for each proficiency level, the relative proportion of NW misspellings tends to increase as the word length increases. Within each proficiency level, longer words are more difficult to spell correctly than shorter words are. For example, for TOEFL NNS essays scored at level 3, writers misspell 4.3% of their 6-character words, 6% of their 7-character words, 9.4% of 8-character words, 10.3% of 9-character words, 14.4% of 10-character words and 19% of 11-character words. Note that Figure 6 indicates that most of the errors occur with shorter words – relative to total amount of misspellings. This is because shorter words are overall used more often. But when percent of misspellings is expressed as relative to the amount of words of given length, as in Figure 9, it shows that the proportions of misspellings are greater for longer words.

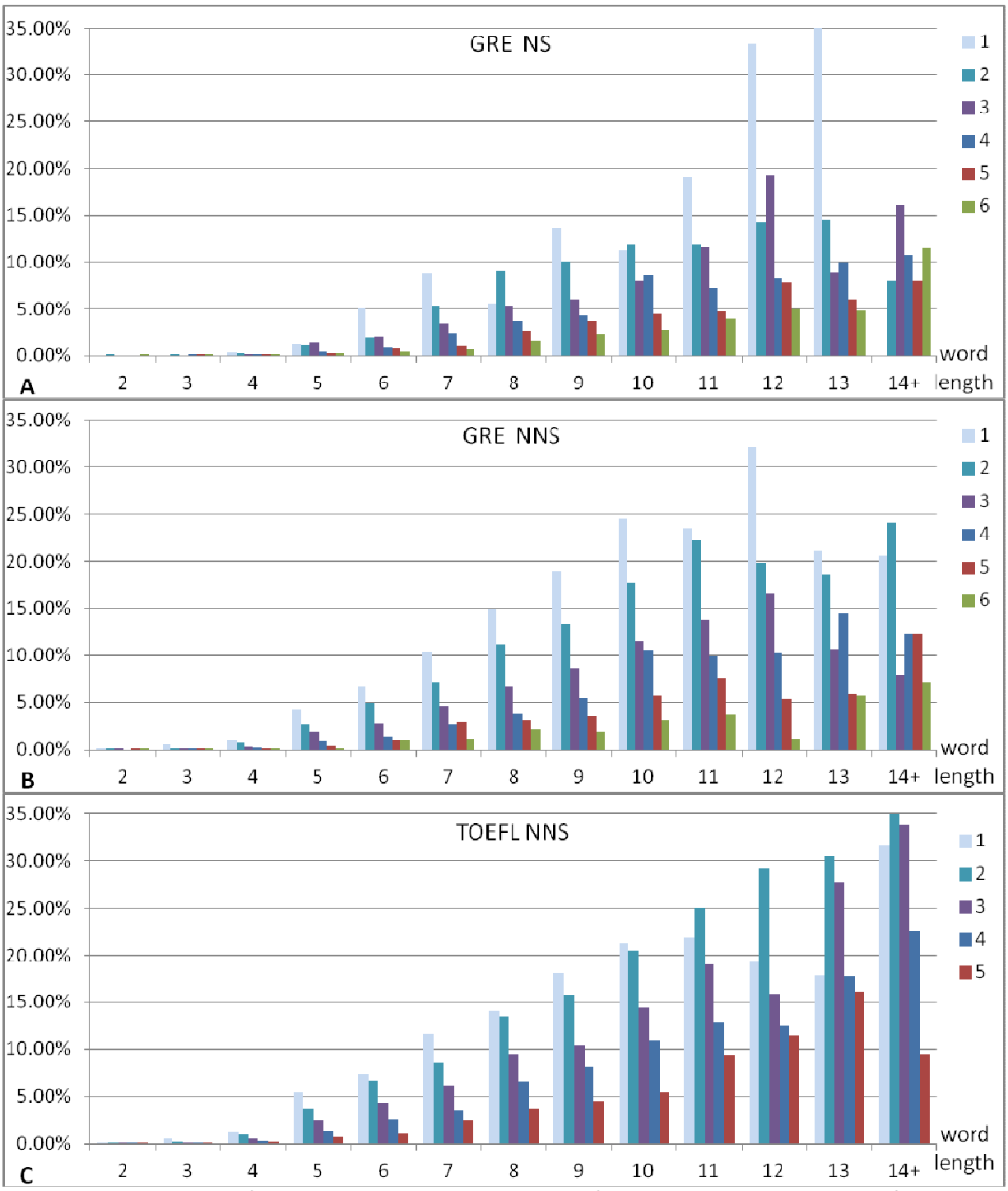

Figure 9. Relative proportions of NW-spelling-errors, by word length (of the correction) and essay score, for three populations Note that the sum of percents for each color (score group) does not add to 100%.

10. Word frequency and spelling errors

In this section we consider how word frequency may be related to spelling errors. Are misspellings committed more often with frequent or infrequent words? Infrequent words might be considered more difficult – the relative lack of familiarity may contribute to lack of knowledge about how to spell them correctly. We label such view a 'knowledge-based' hypothesis. This view predicts more errors with rare words than with more common words. The 'opportunity-based' hypothesis, by contrast, assumes that misspellings are random. Following this view, one would expect to find more spelling errors of the frequent words – because they are used more often and so provide greater chance for committing errors. We look at the interplay of word-frequency, NS/NNS status and overall writing proficiency, focusing the analysis on single-token non-word errors (excluding fusions).

For every word in every essay in our corpus, word frequency was obtained from a very large collection of English texts.[11] As our frequency measure we computed Standard Frequency Index (SFI; Carroll, Davies, and Richman. 1971), calculated as $10(\log_{10}(WF)+10)$, where *WF* is the relative frequency of a word form in the large collection. The SFI measure conveniently maps frequency values into a 0-100 range.[12] For each misspelling we retrieved the correct word from corpus annotations and obtained the SFI of the intended correct word.

We begin with a comparison of average SFI of misspelled intended words. The chart for GRE NS and NNS populations, broken down into 6 proficiency levels, is presented in panel A of Figure 10. The average SFI of the intended word decreases with higher proficiency, i.e. errors are committed with rarer words. A two-way ANOVA indicated a significant main effect of proficiency, $F(5, 10110)=21.008$, $p<0.0001$, and a significant main effect of NS/NNS status, $F(1, 10110)=12.371$, $p<0.0001$. The interaction is also significant, $F(5, 10110)=2.588$, $p<0.03$). The interaction is due to an unexpectedly high average SFI of misspelled words in the lowest-scoring essays from NS writers. If we consider only scores 2-6, there is an effect of score, $F(4, 8267)=17.734$, $p<0.0001$, and NS/NNS status, $F(1, 8267)=41.955$, $p<0.0001$, but no interaction. Words misspelled by NS writers are on average of lower frequency than those misspelled by NNS writers of comparable proficiency. The data for TOEFL NNS population is presented in panel B of Figure 10. A one-way ANOVA showed a significant main effect of score level, $F(4, 9751)=46$, $p<0.0001$. As proficiency increases, the average SFI of misspelled words becomes lower; same trend as the one seen in the GRE data.

Is there a relation between the trends for misspellings, presented in Figure 10, and overall word usage in essays in our corpus? One may expect that NS writers utilize more rare words than NNS writers, and that usage of lower frequency words increases with writing proficiency (leading to lower average word frequency). If this is the case, it might explain the trends observed for the misspelled words. To investigate this assumption, we consider the general trends of word usage by word-frequency. Using the all-spell-corrected versions of essays, we computed word SFI for every word in our corpus. Figure 11 presents average word frequency of all essay words, for each population, by proficiency level. A two-way ANOVA for GRE data showed a significant main effect of proficiency level, $F(5, 597442)=37.362$, $p<0.0001$, a significant main effect of NS/NNS status, $F(1, 597442)=8.441$, $p<0.005$, and even a significant interaction,

[11] For word frequency data, we used a combined corpus of 1.5 billion word tokens. It combines the GigaWord 2003 corpus (Graff and Cieri 2003) and an ETS internal corpus, consisting of popular science and fiction texts.

[12] Theoretically, a formula like $10(\log_{10}(WF)+K)$ maps frequency values from [0,1] range into a $(-\infty,100]$ range. Carroll, Davies, and Richman (1971) used K=4, but their WF was counts per million. With our WF, for a corpus up to a billion words, using K=10 ensures mapping into a [0,100] range. Since our reference corpus is larger, all negative SFI values were taken as zero.

F(1, 597442)=2.903, p<0.02. For TOEFL NNS data, an ANOVA showed a significant main effect of score level, F(4, 356181)=34.733, p<0.0001. Overall, the average word frequency decreases as the essay score increases, and average word frequency is lower in essays written by native speakers. Overall, these findings are similar to the finding for misspellings. However, we consider that this trend does not sufficiently explain the patterns for misspellings. For all-words data, the actual average frequency values are very close, between SFI 69 and 68 for all groups and levels (see Figure 11). For average SFI on all words, the range of variation is very small (about 1 SFI point), while the range of variation of average frequency of misspelled intended words is much larger – about 9 SFI points for GRE and 4 SFI points for TOEFL (Figure 10). The decrease in average frequency of misspelled words is much more dramatic than the overall decrease in average word frequency.

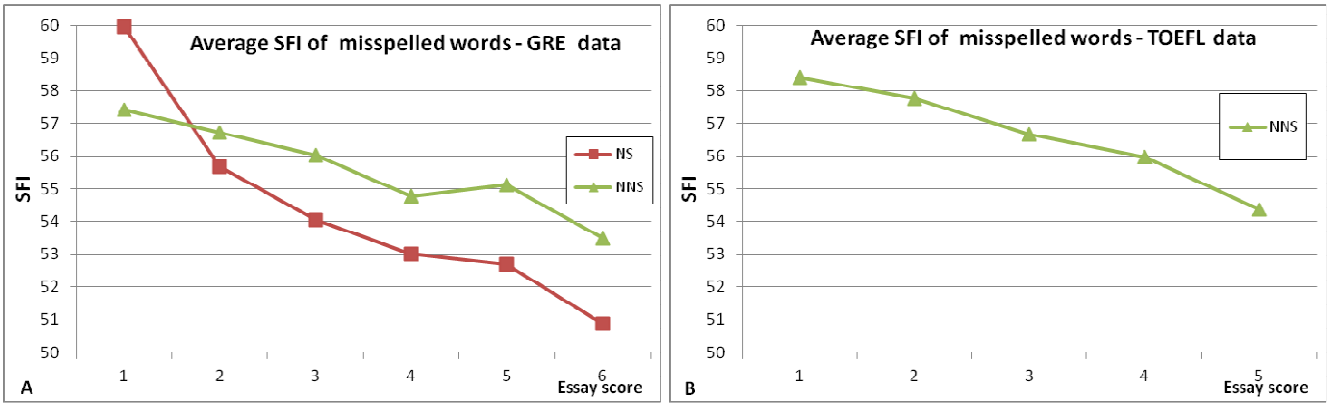

Figure 10. Average language-frequency of 'intended' words (that resulted in single token NW errors), by essay score, for three populations (panel A – GRE data, panel B – TOEFL data)

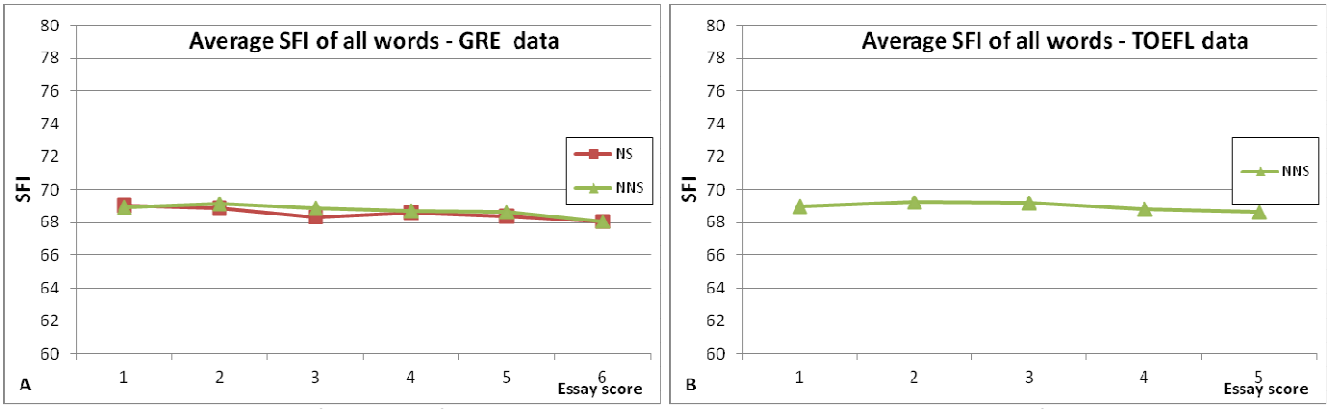

Figure 11. Average language-frequency of all essay words (spell-corrected essays), by essay score, for three populations (panel A – GRE data, panel B – TOEFL data)

We conducted another, different analysis. Instead of looking at average word frequency, we calculated how many words from different frequency levels are used by writers, and what are the relative proportions of misspellings. First we looked at overall word usage by word frequency. We defined 10 bins for word frequency bands, from very rare words (SFI<40), to the most frequent words (SFI=>80). For each population (GRE NS/NNS and TOEFL NNS), we tallied the number of all words by SFI values, using the

spell-corrected versions of essays. The data is presented in Figure 12. The tendencies for all three populations are quite similar. Words that are more frequent in English are also used more often in the essays. For frequency bands of SFI:40-45 to SFI:50-55 (rare to moderately-rare words), the relative percent of such words tends to increase with increased proficiency. For frequency bands SFI:65-70 and SFI:70-75, there is an inverse tendency – as proficiency increases, the relative percent of such words decreases. For frequency bands SFI:60-65 and SFI:75-80 the relative percent is rather stable across score levels. Some difference is apparent for bin SFI:55-60: the GRE NNS population has increase in relative percent of such words with increased proficiency, but for GRE NS and TOEFL NNS populations, such trend is barely apparent. For all three populations there is also a tendency to use more words of the highest frequency band (SFI:80+) as proficiency increases.[13]

Next, we looked at the distribution of misspellings by their word-frequency. We counted the number of single-token non-word misspellings (excluding fusions), binning by the SFI value of the intended word (the spelling correction). Figure 13 presents the proportion of NW errors in each bin – out of all NW errors, counting separately for each score level (color bars), and for each population (separate charts). For example, for TOEFL essays written by NNS writers (panel C), consider the highest scoring essays (score 5). Out of all NW misspellings in this score group, about 8% are with words of lowest frequency (SFI:<=40), 6.5% are with words of SFI:40-45, 15.7% of all misspellings are with words of SFI:45-50, 19.8% are with words of SFI:50-55, 23.9% are with words of SFI:55-60, and 14.7% are with words of SFI:60-65. However much fewer of their misspellings are with very frequent words: 6.6% with words of SFI:65-70, 3.8% with words of SFI:70-75, 0.7% with words of SFI:75-80 and just 0.3% with words of SFI:80+.

The chart shapes for the three populations (GRE NS/NNS and TOEFL NNS) are rather similar. For each chart in Figure 13 we distinguish a left part (bins SFI:<40 to SFI:55-60), where proportion of misspellings increases with word-frequency, and a right part (bins SFI:60-65 to SFI:80+), where proportion of misspellings decreases with word-frequency. There are similar patterns within every proficiency level (bars of same color): on the left side proportion of errors tends to increase with word-frequency; followed by a decreasing trend on the right side. This trend is not consistent with the hypothesis that misspellings happen most often with the frequent words (where the opportunities are). If misspellings happen more or less randomly, we would expect to see most misspellings for high-frequency groups of words. The charts indicate that for each population group, most misspellings happen with words of moderately-rare to medium-high frequency (SFI 45 to 65), whereas the high-frequency words (SFI>65) receive very low proportions of misspellings. Recall from Figure 12 that those high-frequency bins carry the largest portion of the essay words, and yet they have the lowest portion of misspellings. The right-side parts of the charts are consistent with a knowledge-based hypothesis – words that are more frequent are likely to be better known, and thus fewer misspellings there. The trends seen on the left side of the charts are consistent with the 'opportunity hypothesis': as word frequency increases from very rare to moderate, more such words are used in essays (left sides in Figure 12) and the proportion of misspellings also rises with word frequency (left sides in Figure 13). However, note that for all-words (Figure 12), the increase of usage reaches a local

[13] This may stem from increase of essay length with improving proficiency. The frequency band SFI:80+ includes just four words: *the, in, of, and*. As essays become longer, they have more clauses and phrases, and thus need more determiners and frequent prepositions. Those are often qualified with an article (*the*), a preposition (*in, on*), or coordinated with a connector (*and*). The absolute counts for those words should rise. Yet the relative increase in their usage in longer essays is intriguing and deserves a separate study.

peak in bin SFI:60-65, whereas for misspellings (Figure 13) the bin SFI:60-65 already shows a decrease in the proportion of misspellings.

There is another trend seen in charts of Figure 13. We consider how word frequency relates to writing proficiency (comparing different color bars within each SFI bin). In each population, within most word frequency bins on the left side of the charts: writers of higher proficiency produce more misspellings than writers of lower proficiency (e.g. bin SFI:40-45 in every panel). In the right side of the charts, the trend reverses: in most bins, as proficiency increases, the proportion of errors decreases (e.g. bin SFI:65-70 in every panel). More proficient writers are possibly better acquainted with medium-to-high-frequency words, which may explain that decreasing trend (within each bin) on the right sides of the charts.

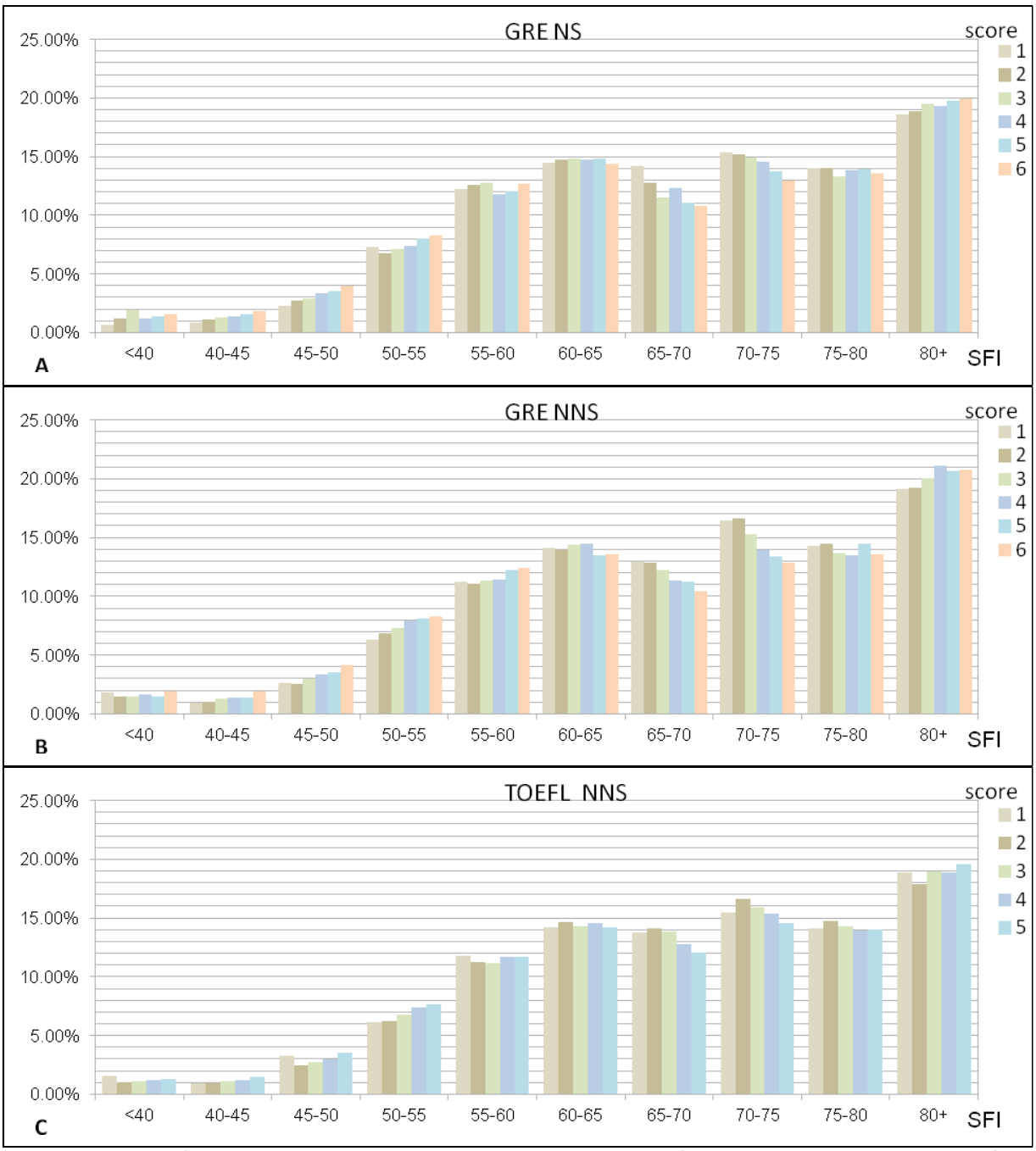

Figure 12. Rates of word-use in student essays, with breakdown by word-frequency (binned SFI) and by proficiency, for three groups of essays. Data reflects all words from spell-corrected essays. Each color in panels A-C represents a specific proficiency (score) level. The sum for each score level (color) on each panel is 100%

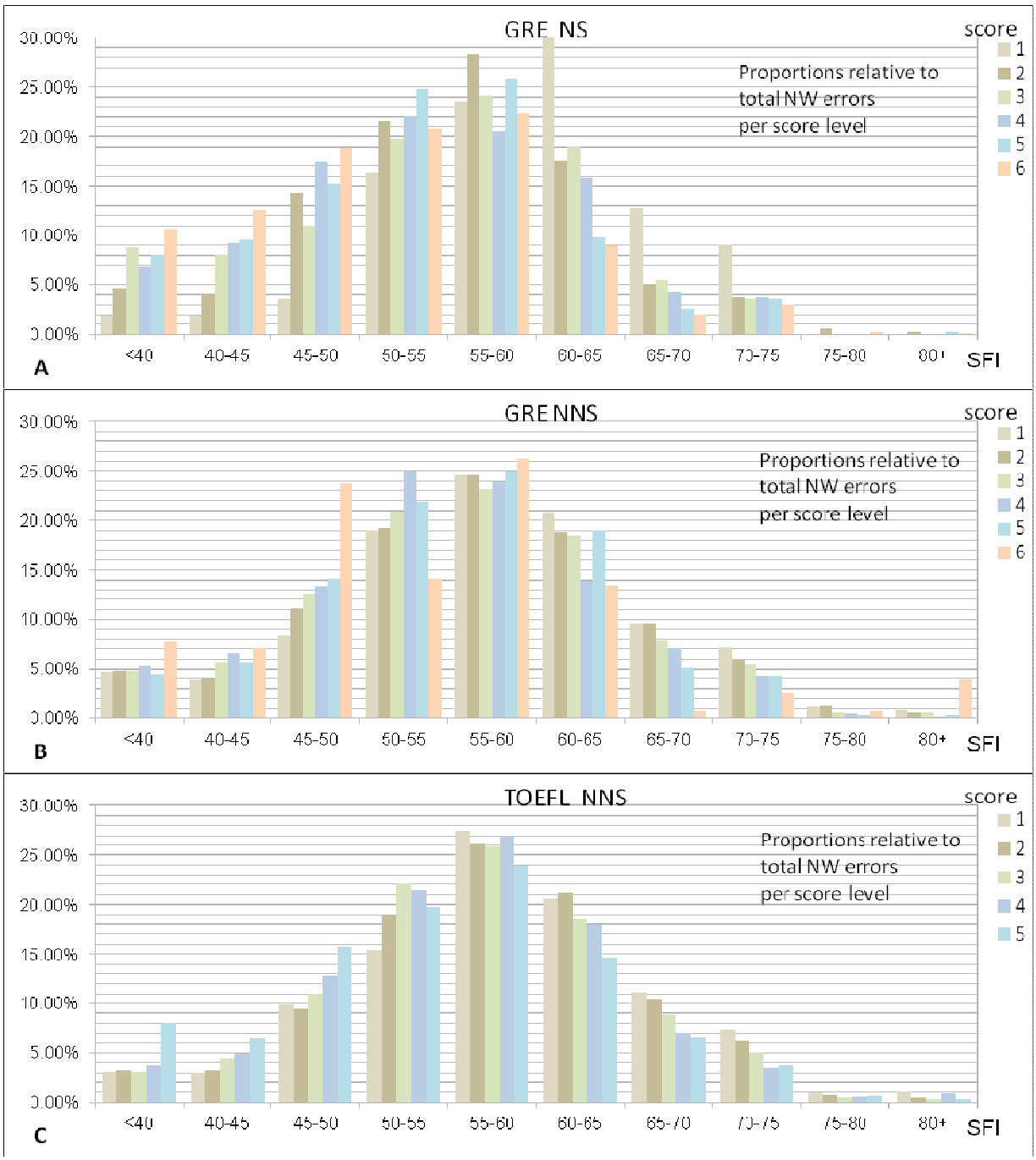

Figure 13. Rates of NW-spelling-errors in student essays, by word-frequency (binned SFI) and by proficiency, for three groups of essays. Each color in panels A-C represents a specific proficiency (score) level. The sum for each score level (color) on each panel is 100%

For the third analysis, we calculated the proportion of misspelled words out of all words, for each SFI bin, for each proficiency level, for each population. The data is presented in Figure 14. Given a group of essays (defined by population and score level), we find all words in those essays that belong to a given word frequency bin, and count how many of them were misspelled (as single-token non-words). For example, in GRE essays written by NNS writers (panel B), for words of SFI band 50-55, we compare the different-color bars in the bin: low-proficiency writers make a spelling error in 14% of such words; at proficiency level 2 writers misspell 9.3% of such words; at proficiency level 3 – 6.5%; and at proficiency level 6 – just 1.2%. Similar tendencies can be observed for GRE NS and for TOEFL NNS data.

There is a strong pattern. For each population, for words of every frequency band, as the writers' proficiency increases, the proportion of NW misspellings they make with words of that frequency-band decreases. This overall trend is remarkable, as it exists despite the fluctuations of overall use for words of different frequency bands (as seen in Figure 12). For words of SFI bands 40-45 to 50-55, the usage of such words increases with proficiency (writers are using more of infrequent words), yet the tendency to misspell such words decreases with proficiency. This trend can be interpreted as a clear sign of improving lexical knowledge – proficient writers may know more of the infrequent words, and use these words more often than the less proficient writers do, but proficient writers misspell such words less than the non-proficient writers do.

Another pattern in Figure 14 is apparent within each proficiency level: the proportion of misspelled words is higher for rarer words, and tends to decrease as word frequency increases (bars of same color get lower for higher-SFI bins, i.e. left to right). This trend may reflect the tendency of rarer words to be generally more difficult to spell, at any given level of writing proficiency. However, there is a noticeable exception: on the far left side of each panel, bars of same color tend to increase (get higher) in the first three bins, before they begin to decrease. This sub-trend is surprising and deserves a separate investigation.

The third pattern in Figure 14 relates to differences between native and non-native speakers. For rare to medium frequency bands, GRE NNS writers produce larger proportions of misspellings than GRE NS writers. For example, for the band SFI:45-50, in GRE NNS essays of score 2, 14.3% of such words are misspelled, while in GRE NS essays of same score the proportion is 12%. For the same SFI band, for essays of score 3, GRE NNS have 9.5% misspelled, while GRE NS have 6.6%. Generally, the bars in the GRE NNS chart are higher than in GRE NS chart. The proportions in TOEFL NNS data are also higher than in GRE NS data (although the essay score levels are not directly comparable, due to different scales). The differences between the three populations tend to disappear with increased proficiency levels (all bars diminish on the right sides of the charts).

Those three patterns can be seen as manifestations of improving lexical knowledge. As lexical knowledge increases, we expect greater use of rare and infrequent words, as demonstrated in Figure 12. As lexical knowledge improves, we also expect to see diminishing rates of misspellings for words of every frequency band, as demonstrated in Figure 14.

Several factors contribute to the steep decrease of average word-frequency of misspelled words, with improved writing proficiency (Figure 10). On one hand, writers (both NS and NNS) increase their use of rare and less-frequent words (Figure 12), and many of those words come out misspelled. Even if the rate of misspellings for those words were rather constant, it could contribute to the lowering of average SFI across proficiency levels. Actually, the rate of misspellings for such words increases with proficiency level (left sides of the charts in Figure 13), which lowers the average SFI even more. In addition, as proficiency improves, writers make relatively less misspellings with the common (frequent) words, which also contributes to the trend of lowering the average SFI of misspellings.

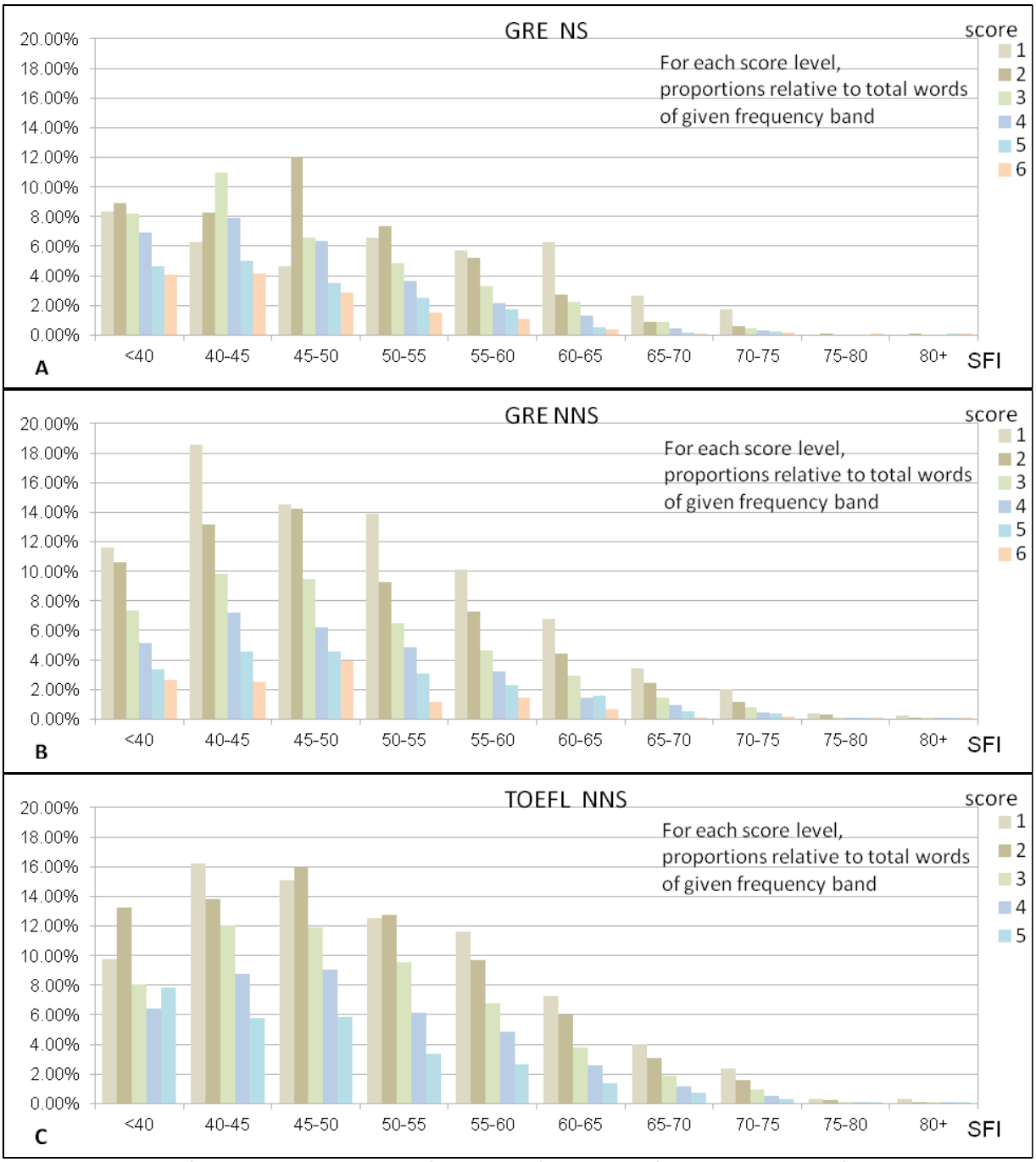

Figure 14. Percent of NW-spelling-errors (out of all words of given word-frequency bin, by proficiency level (essay score), for three populations. Note that the sum of percents for each color (score group) does not add to 100%

11. Conclusions

In this paper, we presented patterns of misspellings, based on data from the ETS Spelling corpus. The corpus comprises essays written by examinees on the writing sections of GRE and TOEFL examinations. The metadata of these essays includes scores assigned to the essays in operational testing, as well as background information indicating whether test takers were native (NS) or non-native speakers (NNS) of English. The corpus was manually annotated for misspellings of various types, and corrections were provided. The majority of misspellings (80%) were single-token non-word misspellings.

Analysis of average percentage of misspellings per essay showed that the rate of misspellings decreases as proficiency (essay score) increases, both in TOEFL and in GRE. While this finding is not surprising, the comparison between NS and NNS populations in GRE data has shown that there is a large difference at lower essay score levels (NNS writers produce more misspellings), the difference shrinks with increasing proficiency and disappears at the highest level of proficiency.

Using edit distance as an indicator of spelling error severity, we analyzed all single-token non-word misspellings. We found that severity of misspellings depends on writing proficiency. Writers of lesser proficiency produce more of the severe errors, and the average error severity decreases with better proficiency. We found no evidence that error severity is influenced by the NS/NNS distinction.

Analysis of single-token non-word misspellings by length of the correct (intended) word, revealed some general trends. With increased proficiency, essays become longer (more words). Writers of increased proficiency introduce more long words, but they also introduce more short words, and the relative proportions of words of different lengths remains roughly similar for all levels of proficiency, for GRE NS, GRE NNS and TOEFL NNS populations. When amount of errors (non-word misspellings) was expressed as relative to the total amount of words of given length, we observed two patterns. First, for every one of the three population groups, for each proficiency level, the relative proportion of non-word misspellings tends to increase with the increase of word length. Thus, word length is related to misspellings – longer words are more difficult to spell correctly. Another pattern, repeated in each population group, is that for words of every length group, as the writers' proficiency increases, the proportion of misspellings they make with words of that length decreases. So, while writers of greater proficiency introduce more long words (in absolute measure, such as average word length per essay), they are also less prone to misspell such words (as compared to writers of lesser proficiency). This finding may be taken as a sign of improving lexical knowledge.

Misspellings are also influenced by word frequency. The average frequency of words that are misspelled to non-words declines with writer proficiency, both in GRE and in TOEFL essays. GRE data show that for each proficiency level (except the lowest one), the average frequency of words where NS writers make misspellings is lower than that for NNS writers of same level. We also asked whether misspellings occur more often with infrequent words (the knowledge-based hypothesis) or with frequent words (the opportunity hypothesis). The results support the knowledge hypothesis. When error proportions are expressed relative to the number of words within frequency bins, the dominant influence of word-frequency becomes evident. For each population (GRE NS/NNS and TOEFL NNS) we observed two clear trends. For words of every frequency band, as the writers' proficiency (essay score) increases, the proportion of NW misspellings they make with words of that frequency decreases. Within each proficiency level, the proportion of misspelled words is higher for rarer words, and tends to decrease as word frequency increases. The similarity of trends in the three populations suggests that for misspellings in college-level essays, writing proficiency might be a more important factor than native/non-native distinction.

This paper presented a broad overview of findings from the ETS spelling corpus. Our future work

will concentrate on more specific investigations, with more detailed categorizations of spelling errors, and also on investigation of spelling errors that span across multiple tokens.

References

Bebout, L. 1985. An error analysis of misspellings made by learners of English as a first and as a second language. *Journal of Psycholinguistic Research* 14 (6): 569–593.

Bestgen, Y., and S. Granger. 2011. Categorising spelling errors to assess L2 writing. *International Journal of Continued Engineering Education and Life-Long Learning* 21 (2/3): 235-252.

Botley, S., and D. Dillah. 2007. Investigating spelling errors in a Malaysian learner corpus. *Malaysian Journal of ELT Research* 3:74–93.

Carroll, J.B., P. Davies, and B. Richman. 1971. *The American Heritage word frequency book*. New York: American Heritage Publishing Co.

Chodorow, M., and J. Burstein. 2004. Beyond essay length: Evaluating e-rater's performance on TOEFL essays. *TOEFL Research Report* No. RR-73, ETS RR-04-04. Princeton, NJ: ETS.

Cook, V. J. 1997. L2 user and English spelling. *Journal of Multilingual and Multicultural Development* 18 (6): 474-488.

Damerau, F. 1964. A technique for computer detection and correction of spelling errors, *Communications of the ACM* 7 (3): 659-664.

ETS, 2011a. GRE®: Introduction to the Analytical Writing Measure. Available from (www.ets.org/gre/revised_general/prepare/analytical_writing)

ETS, 2011b. TOEFL® iBT® Test Content. www.ets.org/toefl/ibt/about/content

Flor, M. 2012. Four types of context for automatic spelling correction. *Traitement Automatique des Langues* (TAL), 53 (3): 61-99. (http://www.atala.org/IMG/pdf/Flor-TAL53-3.pdf)

Flor, M., and Y. Futagi. 2013. Producing an annotated corpus with automatic spelling correction. In *Twenty Years of Learner Corpus Research: Looking back, Moving ahead. Corpora and Language in Use*, eds. S. Granger, G. Gilquin and F. Meunier, 139-154. Louvain-la-Neuve: Presses universitaires de Louvain.

Flor, M., and Y. Futagi. 2012. On using context for automatic correction of non-word misspellings in student essays. In *Proceedings of The 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, 105-115, at the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT), June 3-8, 2012, Montréal, Canada. (http://aclweb.org/anthology-new/W/W12/W12-2012.pdf)

Graff, D., and C. Cieri. 2003. *English GigaWord 2003*. Philadelphia, PA: Linguistic Data Consortium. (http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T05)

Granger S., and M. Wynne. 1999. Optimising measures of lexical variation in EFL learner corpora. In *Corpora Galore*, ed. J. Kirk, 249–257. Amsterdam: Rodopi.

Hovermale, D. J. 2010. *An analysis of the spelling errors of L2 English learners*. Presented at the CALICO 2010 Conference, Amherst, MA, USA, June 10-12, 2010. (http://www.ling.ohio-state.edu/~djh/presentations/djh_CALICO2010.pptx)

Kukich, K. 1992. Techniques for automatically correcting words in text. A*CM Computing Surveys* 24 (4): 377-439.

Leacock, C., and M. Chodorow. 2003. C-rater: Automated Scoring of Short-answer Questions. *Computers and Humanities* 37 (4): 389-405.

Levenshtein, L. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady* 10:707-710.

Lunsford, A. A., and K.J. Lunsford. 2008. Mistakes Are a Fact of Life: A National Comparative Study. *College Composition and Communication* 59 (4): 781-806.

Mitton, R., and T. Okada. 2007. *The adaptation of an English spellchecker for Japanese writers*. Paper presented at the *Symposium on Second Language Writing*, 15-17 Sept. 2007, Nagoya, Japan. Available from (http://eprints.bbk.ac.uk/592)

Page, E. B. 1967. The imminence of grading essays by computer. *Phi Delta Kappan* 47 (5): 238-243.

Pollock, J., and A. Zamora. 1984. Automatic spelling correction in scientific and scholarly text. *Communications of the ACM* 27 (4): 358-368.

Ramineni C., C.S. Trapani, D.M. Williamson, T. Davey, and B. Bridgeman. 2012a. *Evaluation of the e-rater® Scoring Engine for the GRE® Issue and Argument Prompts*. Research Report RR-12-02, Princeton, NJ: Educational Testing Service. (http://www.ets.org/research/policy_research_reports/rr-12-02)

Ramineni C., C.S. Trapani, D.M. Williamson, T. Davey, and B. Bridgeman. 2012b. *Evaluation of the e-rater® Scoring Engine for the TOEFL® Independent and Integrated Prompts*. Research Report RR-12-06, Princeton, NJ: Educational Testing Service. (http://www.ets.org/research/policy_research_reports/rr-12-06)

Rimrott, A., and T. Heift. 2008. Evaluating automatic detection of misspellings in German. *Language Learning and Technology* 12 (3): 73–92.

Sukkarieh, J. Z., and J. Blackmore. 2009. C-rater: Automatic Content Scoring for Short Constructed Responses. In *Proceedings of the 22nd International Florida Artificial Intelligence Research Society Conference*, 290-295, Menlo Park, CA: AAAI Press.

Turba, T. N. 1981. Checking for spelling and typographical errors in computer-based text. *ACM SIGPLAN Notices* 16 (6): 51-60.