Intensifier usage by English learners on a language-exchange social media site

Khia A. Johnson
University of British Columbia
April 15, 2019

Do English language learners (ELLs) use variants that abound in speech and informal written communication, but are decidedly less common in formal writing? The study reported on in this paper examines the use of *so* as an amplifying intensifier (e.g. "that was *so* cool"), as compared to frequent alternatives such as *very*, *really*, *too*, and others (Recski, 2004; Tagliamonte, 2016). While common in everyday speech Tagliamonte (2016) found that *so* remained relatively rare in the formal writing of native English speakers, even among individuals who regularly employ it in more casual registers such as writing emails and text messaging. Given that these forms are less common in formal writing compared to speech and computer-mediated communication, it is possible that ELL intensifier-*so* usage reflects exposure to a community of native language users.

The present study addresses the question of intensifier-*so* usage with a corpus of ELL-produced text scraped from the Lang-8 language learning social media site (as in Brooke and Hirst, 2013). Lang-8 is a desirable place to look for the emergence of such constructions as the website is social in nature. As such, Lang-8 is a desirable place to look for less formal language. This may not be an issue, however, as ELLs are known to use intensifiers at higher rates in academic writing when compared to native English speakers (Hinkel, 2005). The increased rate of intensification in ELL essays likely reflects a lack of argumentative sophistication on behalf of the learner, and may not necessarily transfer into other styles of writing, such as the journal posts on Lang-8. Regardless, there is clear evidence that learners use intensifiers, which makes them a promising test case when studying emerging variants. Furthermore, Recski (2004) reports on relative frequencies for a wide variety of intensifiers from both a corpus of ELL essays, as well as a corpus of transcribed ELL speech. This provides two different comparison points for the relative frequency of intensifiers on Lang-8.

Specifically, the research questions in this study are as follows: do ELLs on Lang-8 use the intensifier *so* in online written language? To what extent do ELLs use *so* as an intensifier? As this is a newer development in the English intensifier system, what sociolinguistic factors contribute to an ELL's usage of *so* as an intensifier? Given the available user-level information, this study was designed around the following hypotheses. Lang-8 ELLs will be more likely to use *so* as an intensifier if:

1. They are younger (as in Tagliamonte, 2016).

2. They are female (as in Tagliamonte, 2016).

3. They currently reside in a location where English is both the majority and dominant language.[1] This is based on the assumption that living in an English majority/dominant location leads to more exposure to informal language.

4. They have more connections to native English users on Lang-8. This hypothesis rests on the assumption that a greater number of connections translates into more exposure to and informal interaction in English.

5. They have a larger number of journal entries. This indicates that they use the site more, and effectively practice in the language more.

6. The post is more recent, as this is a relatively recent development in the English intensifier system (Tagliamonte, 2016).

Methods

Data collection

The corpus used in this study was created by scraping the Lang-8 social media language learning website.[2] On the website, users can connect to one another, as on typical social media websites. The primary use of Lang-8 is to keep journals in the language being learned, and to correct and/or comment on entries by other users, to engage with or help them learn their language of choice. In this sense, Lang-8 is both a a social media platform and a place for language exchange and learning.

Scraping was done in February–March 2019.[3] Beginning with a group of hand-selected seed users, the scraper traversed the website by collecting data from individual users and checking their connections for additional potential users. Journal entries and metadata about the user were scraped if the user (i) reported their location, (ii) studies English, (iii) did not list English as a native language, and (iv) the journal entry was tagged for English. Further data cleaning and filtering was needed, though this provided a first pass and ensured that the scraping process was more efficient than it might otherwise be. The scraped data was saved to a text file with each line corresponding to a single user, potentially with many journal entries.

[1] In the sample described in the Methods section, the locations are the United States, Canada, the United Kingdom, New Zealand, Australia, Scotland, Ireland, and Jamaica.

[2] `https://lang-8.com`

[3] All code for this project can be found in the supplementary material.

Data preparation

The data was first reshaped such that each line of the data frame represented a single journal entry, the year it was published, and the accompanying user metadata. In addition to user information reported on the site. Journal entries were filtered out if the user reported their age to be under 18 or over 100 years. Users were filtered out if they did not report their sex on the website, or had fewer than four entries. At this point, journal entries were cleaned, tokenized, and tagged for part-of-speech. Cleaning involved removing non-English characters and nonstandard characters from the text. Journal entries were tokenized with the NLTK TweetTokenizer (Potts et al., 2018), such that case was not preserved and lengthened words (e.g. *soooooooo*) were shortened to a single extra character (e.g. *soo*).[4] Part-of-speech tagging was completed with the StanfordPOSTagger version 3.9.2 in NLTK (Madnani and Al-Rfou', 2018), using the Fast GATE Twitter model for English (Derczynski et al., 2013).

Using the part-of-speech tags, adjectives with preceding and following context were then identified based on the tag *JJ*, which does not include superlatives (*JJS*) or comparatives (*JJR*). While some adjectives were certainly missed in this process, the default tag in cases where the part-of-speech cannot be identified was not *JJ*. After visually inspecting the longest and shortest adjectives, items were filtered out if the adjective contained non-alpha characters other than "-", was shorter than 3[5] characters, or was longer than 25 characters. Again, using the part-of-speech tags, all adjectives immediately preceded by an adverb were identified, and the adverbs were checked against a list of common intensifiers, which is included in the supplementary material. Approximately 8.7% of adjectives were intensified (n = 42,686). This subset was saved to a new data file, and used in the analysis.

Analysis

This study was designed around the question of if and how ELLs use *so* as an intensifier, and what social factors predict usage. To this end, *so* (n = 13,080) will be considered among the other most commonly used intensifiers in this corpus: *very* (n = 18,037), *too* (n = 4,170), and *really* (n = 3,899). All other intensifiers had around or below 1,000 tokens (total n = 3,563).

These questions are addressed here using multinomial logistic regression with the *MNLogit* function in the *statsmodels* Python package (Seabold and Perktold, 2010).[6] The dependent variable in the model was intensifier (*very*, *so*, *really*, or *too*), with *other intensifiers* set as the pivot class. The following factors—each representing one of the hypotheses outlined in the introduction—are included in the model:

1. **Age.** Numeric variable ranging from 18 to 85, with a mean of 29.9 years.

2. **Female.** Binary variable with values *True* and *False*.

[4] These were further shortened to their standard spellings at a later point.

[5] The single exception to this was "ok".

[6] Note that *statsmodels* does not support mixed effects regression.

3. **English dominant location.** Binary variable with values *True* and *False*, where *True* indicates that the user's current location is and English majority/dominant location.

4. **Number of native English connections.** Numeric variable ranging from 0 to 1,219 with a mean of 23.9 connections.

5. **Number of journal entries.** Numeric variable ranging from 5 to 4,002 with a mean of 47.8 entries.

6. **Year.** Numeric variable ranging from 2008 to 2019.

Results

The full output of the multinomial logistic regression model is reported in Table 1. Each of the main effects will be summarized descriptively in this section. There was a small significant effect of age for two of the four intensifiers—*very* and *too*. This indicates that Lang-8 users are slightly more likely to use those intensifiers compared to the rest. Overall, the effects of age were minimal, and are depicted in Figure 1. Proportions were used in plotting age, as users varied substantially in the number of intensifiers used across all of their entries.

Female users are more likely to use the each of the four intensifiers over other intensifiers, when compared to male users. This effect is especially pronounced in the case of *so* and *really*, and to a lesser degree with *very*. While there is a higher proportion of *too* for male users, the difference between male and female is most pronounced in the *other* category. Figure 2 depicts the overall proportions for each of the top intensifiers and *other* by user-reported sex.

Users located in an English dominant location were somewhat less likely to use the intensifiers *very* and *so*. The effect of location was not significant for the other intensifiers, even though the proportions appear to have a larger difference in Figure 3, which plots overall proportions for the top intensifiers by the binary location variable.

Users with more native English speaking connections were slightly less likely to use the intensifiers *so* and *really*, and while the effect was significant for these two, in all cases the effect size was tiny. Figure 4 depicts the intensifier proportion by user as it varies with the number of connections to native English speakers. Users with more entries on Lang-8 were less likely to use the top four intensifiers. The effect size was small across the board, but significant for each of the intensifiers. This was true across the board.

Figure 1. Age by user intensifier proportion.

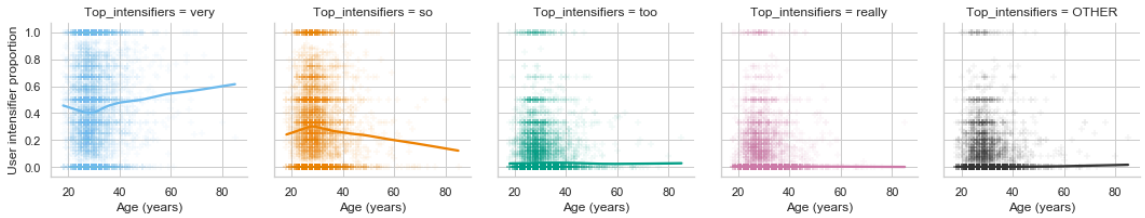


Table 1
*Full output of the multinomial logistic regression. Significance is indicated using * with a threshold of $p < 0.01$ in the rightmost column. The pivot class here is all other intensifiers.*

Intensifier	β	SE	z	$P > \|z\|$
Very ($n = 18,037$)				
Intercept	173.6389	16.359	10.614	0.000 *
Age	0.0092	0.003	3.020	0.003 *
Sex[Female]	0.2203	0.039	5.641	0.000 *
Location[Eng. dominant]	-0.3640	0.065	-5.622	0.000 *
Eng. L1 connections	0.0001	0.000	0.819	0.413
Entries	-0.0015	0.000	-9.644	0.000 *
Year	-0.0856	0.008	-10.546	0.000 *
So ($n = 13,080$)				
Intercept	117.1941	16.835	6.961	0.000 *
Age	-0.0005	0.003	-0.172	0.864
Sex[Female]	0.3579	0.040	8.911	0.000 *
Location[Eng. dominant]	-0.2070	0.066	-3.135	0.002 *
Eng. L1 connections	-0.0008	0.000	-3.519	0.000 *
Entries	-0.0010	0.000	-6.491	0.000 *
Year	-0.0576	0.008	-6.896	0.000 *
Too ($n = 4,170$)				
Intercept	76.5521	20.225	3.785	0.000 *
Age	0.0121	0.004	3.307	0.001 *
Sex[Female]	0.1944	0.048	4.041	0.000 *
Location[Eng. dominant]	-0.0928	0.079	-1.171	0.242
Eng. L1 connections	-0.0004	0.000	-1.403	0.161
Entries	-0.0007	0.000	-3.758	0.000 *
Year	-0.0381	0.010	-3.800	0.000 *
Really ($n = 3,899$)				
Intercept	67.8980	20.546	3.305	0.001 *
Age	-0.0048	0.004	-1.235	0.217
Sex[Female]	0.3193	0.049	6.535	0.000 *
Location[Eng. dominant]	0.0160	0.079	0.203	0.839
Eng. L1 connections	-0.0019	0.000	-5.117	0.000 *
Entries	-0.0002	0.000	-0.832	0.405
Year	-0.0337	0.010	-3.302	0.001 *

Figure 2. Overall intensifier proportions by sex.

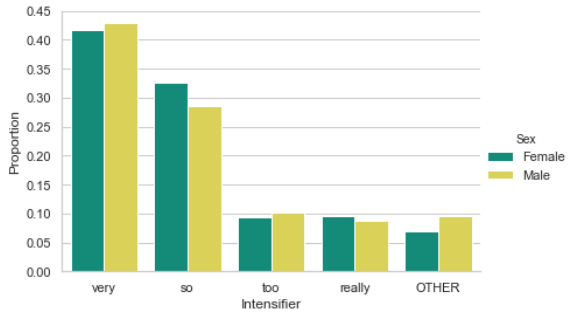


Figure 3. Overall intensifier proportions by location.

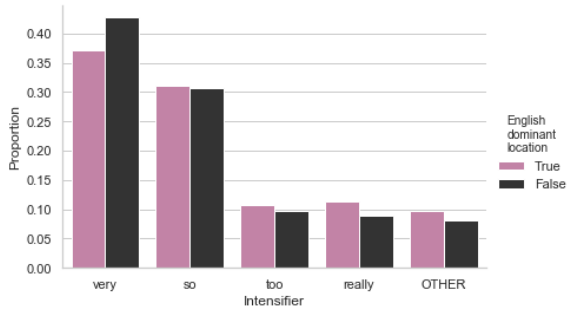


Figure 4. Number of L1 English connections by user intensifier proportion.

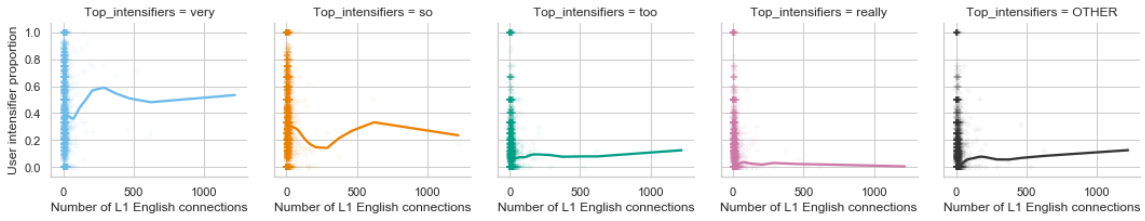
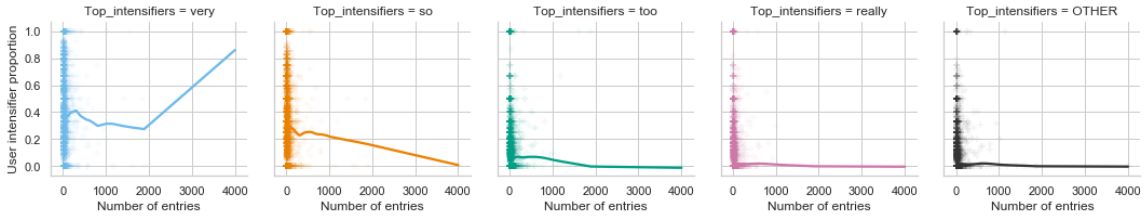


Figure 5. Number of entries by user intensifier proportion.



Each of the top four intensifiers was less likely in more recent years. This is clear in the case of *very*, as depicted in Figure 6, where the proportion decreases somewhat steadily over the course of the years represented. While the other top intensifiers—*so*, *too*, and *really*—appear to change less, they nonetheless have significant negative coefficients.

While not examined statistically here, ELLs combined the top intensifiers with a wide variety of adjectives. This is reflected in the top ten adjectives presented in Figure 7, with counts for collocations with each of the top intensifiers. While *so* most frequently combines with *much* and *many*, there are no major gaps in the top ten adjectives, as there are for others (e.g. very many, too happy, really much, really many, etc.).

Figure 6. Overall intensifier proportions by year.

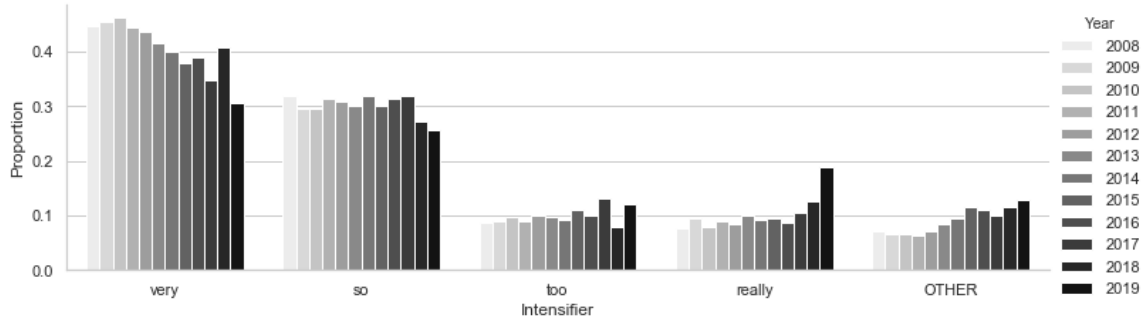
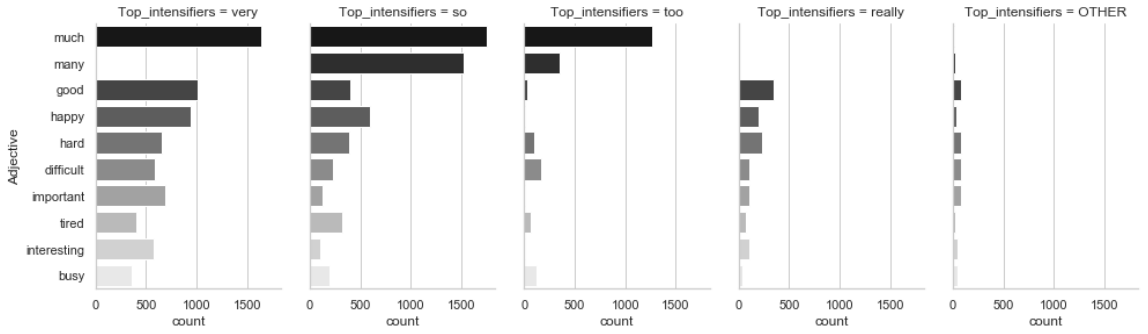


Figure 7. Overall top ten adjective counts by collocation with top intensifiers.



Discussion

This study addresses the question of if and how ELLs use intensifiers like *so* in online written communication. First and foremost, this study demonstrates that the use of intensifier-*so* is commonplace in online ELL writing, as it is second only to *very* in terms of frequency. While it is possible that ELL usage varies drastically from native English users, it seems more likely that intensifier-*so* has fully entered the lexicon, and that the study in Tagliamonte (2016) presented was describing a late 2000s phenomenon. This conclusion is problematic, however, as *so* was the second most frequent intensifier in every year reported here, including the years studied by Tagliamonte (2016).

While none of sociolinguistic predictors of intensifier-*so* usage have a particularly large effect size, all factors were significant for at least a subset of the intensifiers, except age. As such, there is no evidence to support the hypothesis that younger users would be more likely to use *so*. As hypothesized, female users were more likely to use *so* than male users. This suggests that the form may be at least somewhat newer or more innovative, as Tagliamonte (2016) indicates that females tend to lead changes. Finding an effect here suggests that future research on sex and intensifiers in ELL writing is a promising area.

In terms of exposure to a native English community, the results of this study contradict the hypotheses outlined in the introduction, that Lang-8 users in English dominant locations, and those with more connections to native English users, would be more likely to use so. These hypotheses was not supported, though there is a plausible explanation, as previous research demonstrates that ELLs with greater English proficiency generally use fewer intensifiers in writing than ELLs with lower proficiency (Hinkel, 2005). The result in this study suggests that exposure to native English speakers or users leads to less intensification, perhaps as a function of proficiency or imitation. While this explanation makes sense on the surface, it is worth highlighting that the overall rate of adjective intensification was lower than what Tagliamonte (2016) reports of native speakers for any register (e.g. formal writing, email)—8.7% here versus 10% for native English formal writing. If ELLs already use fewer intensifiers, then what role does this exposure play?

In a similar vein, users with more entries on Lang-8 were less likely to use the top intensifiers. This could also be a function of proficiency. Users with more posts have practiced writing English more, and have had more opportunity to receive corrections from native English users. As a result, it is possible that they intensify adjectives less than ELLs with lower proficiency. This conclusion is straightforward, but suffers from the same problem as before—the overall low intensification rate.

Lang-8 entries from more recent years are less likely to be intensified with the top intensifiers. This may also be a function of proficiency, such that over time the ELL user base on the site has become more proficient. This explanation is tentative, but supported by the fact that no users have been allowed to join the site since early 2017.[7]. Furthermore, the steepest drop has been for *very*, the most frequent intensifier overall, and ELLs are known to overuse this particular intensifier (Hinkel, 2003; Recski, 2004). As such, a drop for *very* suggests gains in proficiency. This is another area ripe for future work.

One area that this study leaves for future work is the role of native language. The vast majority of ELLs in this study reported Japanese and Mandarin as their native language, though there were sizable contingents reporting Russian, Korean, and Vietnamese as a native language(s). Future work should collect a corpus that is more balanced for different native languages, in order to investigate the role of native language.

While social factors modestly predict intensifier-*so* usage, it is clear that *so* is not new for ELLs. Promising directions for future work include sex and the various correlates of ELL proficiency. In this light, the primary contribution of this exploratory study is methodological, as it reiterates the feasibility of studying ELL language online.

[7] `http://blog.lang-8.com/`

References

Brooke, J. and Hirst, G. (2013). Native language detection with 'cheap' learner corpora. In *Twenty Years of Learner Corpus Research. Looking Back, Moving Ahead: Proceedings of the First Learner Corpus Research Conference (LCR 2011)*, volume 1, page 37. Presses universitaires de Louvain.

Derczynski, L., Ritter, A., Clark, S., and Bontcheva, K. (2013). Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 198–206.

Hinkel, E. (2003). Adverbial markers and tone in L1 and L2 students' writing. *Journal of Pragmatics*, 35(7):1049–1068.

Hinkel, E. (2005). Hedging, inflating, and persuading in L2 academic writing. *Applied Language Learning*, 15(2):29–54.

Madnani, N. and Al-Rfou', R. (2018). Natural Language Toolkit: Interface to the Stanford Part-of-speech and Named-Entity Taggers.

Potts, C., Klein, E., and Pantone, P. (2018). Natural Language Toolkit: Twitter Tokenizer.

Recski, L. J. (2004). "...It's really ultimately very cruel...": Contrasting English intensifier collocations across EFL writing and academic spoken discourse. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, 20(2):211–234.

Seabold, S. and Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

Tagliamonte, S. A. (2016). So sick or so cool? The language of youth on the Internet. *Language in Society*, 45(01):1–32.