Frequency and semantic prototypicality in L2 Spanish learners' dative constructions

Abstract

Usage-based approaches to psycholinguistics posit that the initial stages of grammar acquisition involve the most frequent and semantically prototypical exemplars of a construction. Our study extends such inquiry to Spanish by analyzing the verbs used with the indirect object clitic *le* in a written corpus. Compared to native writing, learner writing relied to a higher extent on the most frequent verbs, which in turn carried the most archetypical meanings for the different uses of *le* (e.g., *decir, gustar,* and *dar* for the epistemic, psychological, and benefactive uses, respectively). Beyond lending support to usage-based psycholinguistics models, these findings elucidate for language educators the disparities between learner and native writing and illustrate corpus analyses as a useful tool for language acquisition research.

Keywords: corpus, acquisition, frequency, prototypicality, construction

1. Background

Usage-based approaches to psycholinguistics posit that linguistic knowledge comprises a network of thousands of form-meaning pairings called constructions, which may involve fixed, semi-fixed, or completely abstract elements that carry certain conventionalized meanings (Bybee, 2010; Goldberg, 1995, 2006; Hoffmann & Trousdale, 2013). Under such models, acquisition of these constructions is highly sensitive to two factors: *frequency* and *semantic prototypicality.* Here, *frequency* refers to how often a learner is exposed to a particular form in the language input. Findings from previous psycholinguistic studies indicate that frequency plays

a strong role in the processing and production of language (e.g., Ellis, 2002, 2016a) across many different levels, including syntax (Saffran, 2002), multiword strings (Shantz, 2017), individual vocabulary words (Kirsner, 1994), morphology (Lehtonen & Laine, 2003), diachronic change in phonology (Schuchart, 1885) and synchronic variation in pronunciation (Bybee, 2001). Furthermore, linguistic forms within each of these levels of language are commonly found to follow a Zipfian distribution (Piantadosi, 2014), in that the highest-frequency items account for a disproportionately large percentage of the tokens overall (Zipf, 1929). For instance, in the Oxford English Corpus, the 100 most common words account for 50% of tokens, and the 1,000 most common account for 75%. However, this distribution falls off steeply, such that 7,000 words are needed to account for 90% of word tokens in the corpus, and 50,000 words for 95% (Oxford Dictionaries, 2019). Within a given grammatical construction, high-frequency forms would be easier to recall and thus facilitate grammatical acquisition by providing a readily accessible exemplar for learners to latch on to while they gradually abstract morphosyntactic patterns from the input (Ellis, 2016b).

The second major factor said to influence the acquisition of constructions under the usage-based approach is *semantic prototypicality*, which refers to how archetypal or exemplary of its category a token is (Ellis, 2013). In the same way that robins and crows are more prototypical examples of the category "birds" than ostriches or penguins, verbs like *move* and *put* fit more prototypically into a verb object locative construction like [SUBJECT VERB OBJECT OBJECT$_{path/location}$] than verbs like *sneeze* or *punch*. For instance, "Felipe moved the box into the office" would constitute a more prototypical instance of the construction than "Felipe slid the box into the office." Starting off with the form that carries the most archetypal meaning of a new grammatical

construction may give learners a "hold" onto the meaning typically conveyed by that construction, thus facilitating acquisition in the long run.

Putting these two factors together, the usage-based approach argues that the first exemplars of a given construction that language learners acquire are the ones that are most frequent and representative of its general meaning, as this facilitates initial acquisition of the abstract schema before learners ultimately move on to less common or archetypal uses of that construction. The usage-based approach's emphasis on frequency and semantic prototypicality as drivers of acquisition may be contrasted to approaches that instead invoke mechanisms such as proposed principles for second language processing (e.g., VanPatten, 2015), hierarchies of learnability (e.g., Pienemann & Lenzing, 2015), or parameters of Universal Grammar (e.g., White, 2015).

One study that illustrates the roles of frequency and semantic prototypicality in second language acquisition comes from Ellis and Ferreira-Junior (2009), who analyzed a longitudinal corpus from seven L2 English learners as they performed a range of speaking tasks with native speakers over a total of 234 sessions covering a five-year span. The authors focused their analysis on three constructions: the verb object locative construction [SUBJECT VERB OBJECT $\text{OBLIQUE}_{\text{path/location}}$] (e.g., *the man puts the apple on the table*), the verb locative construction [SUBJECT VERB $\text{OBLIQUE}_{\text{path/location}}$] (e.g., *the boy goes to the store*), and the ditransitive construction [SUBJECT VERB OBJECT OBJECT_2] (e.g., *the woman gives the girl an apple*). The authors found that, for each of these constructions, the most commonly occurring exemplar in the learners' input was much more frequent than the other items, as per Zipf's power law (1929). Furthermore, these high-frequency exemplars had meanings that were highly semantically prototypical for their respective constructions, as judged by English native speaker raters. Analyzing the participants' second language speech over a five-year period allowed the

authors to examine how these frequent and semantically prototypical exemplars made up the vast majority of instances of the construction at the initial stages of acquisition before learners moved on to less frequent and less prototypical exemplars. To illustrate this, for the ditransitive construction [SUBJECT VERB OBJECT $OBJECT_2$], participants would produce *give* before other verbs both because *give* occurs in the linguistic input with higher frequency and because its meaning aligns more closely with the basic meaning of the ditransitive construction in general (i.e., of indicating object transfer). Only later would participants move away from high use of this entrenched exemplar and develop more abstract schemas for the construction, allowing for the use of a wider variety of verb types with less prototypical meanings (e.g., *write somebody a letter*, *bake somebody a cake*, etc.).

Ellis and Ferreira-Junior's (2009) findings on the critical role of frequency and semantic prototypicality are corroborated both by first language acquisition research (e.g., Ambridge et al., 2013) and by other second language acquisition studies using a variety of methodologies such as free-association tasks (e.g., Ellis, O'Donnell, & Römer, 2014), classroom-based approaches (e.g., Madlener, 2015; Year & Gordon, 2009), and computational simulations (e.g., Ellis & Larsen-Freeman, 2009). Furthermore, this line of inquiry has been extended to target languages beyond English (e.g., Bernolet, Hartsuiker, & Pickering, 2007, for L2 Dutch and German; Izquierdo, 2007, for L2 French; Williams & Kuribara, 2008, for L2 Japanese). However, little work has extended this line of inquiry to the acquisition of Spanish in particular, with Navarro and Nicoladis (2005) and Zyzik (2006) standing out as rare exceptions.

Our study aims to fill this gap in extant research on frequency and semantic prototypicality by analyzing L2 learners' production of Spanish constructions involving the dative clitic *le*. One reason for choosing this construction is that Spanish dative clitics can convey several related but

arguably distinct meanings, allowing for an examination of semantic prototypicality for different senses of the same linguistic form. For this initial study we examine four broad categories of [*le* VERB] constructions that are defined based on the co-occurring verb:[1]

- *Epistemic use:* this category comprises verbs like *say, ask, explain, promise,* etc. that involve communication with an interlocutor. Specific examples of this category include *decir* ('say'), *explicar* ('explain'), *advertir* ('advise'), *preguntar* ('ask'), and *contar* ('tell').

- *Psychological use:* these are verbs that denote mental states of an emotional sort, involving an experiencer (i.e., the individual experiencing the mental state) and a theme (i.e., the content or object of the mental state) (Belletti & Rizzi, 1988). Examples include *gustar, encantar, apetecer* (all meaning roughly 'appeal/be pleasing to'), *interesar* ('interest'), *importar* ('be important to'), and *fascinar* ('fascinate').

- *Benefactive uses, dative required*: these are verbs that indicate that an action was performed for the benefit of someone else or involved someone as a recipient. For this category, the dative argument is required or at least implied as a consequence of the verb's inherent meaning. Examples include *dar* ('give'), *ayudar* ('help'), *ofrecer* ('offer'), *mostrar* ('show'), and *regalar* ('[to] gift').

[1] Note, however, that the delineation between different uses of the Spanish dative can vary in specificity. On the more specific end of the spectrum, one approach would say that *Pablo nos preparó sandwichitos de miga a todos* ('Pablo fixed us all tea sandwiches') and *Pablo le mandó un diccionario a Gabi* ('Pablo sent Gabi a dictionary') constitute different datives with correspondingly different underlying structures (Cuervo, 2003).

- *Benefactive uses, dative not required*: these are verbs that usually do not have a dative argument but can nevertheless take on an indirect object clitic to indicate that somebody is indirectly affected in some way. To illustrate this, in English a similar intuition holds in examples like *The computer broke on me*, where a typically intransitive verb like *break* suddenly takes on a dative argument. Instances of this category in Spanish include examples like *le cerró la puerta* (‘closed the door for him/her’), *le robó la llave* (‘stole the key from him/her), and *le apareció un fantasma* (‘a ghost appeared on him/her’).

These four broad categories chart out the semantic domain of the dative with sufficient specificity for investigating our current research questions, stated as follows:

1. Compared to native speakers’ use of [*le* VERB], do L2 Spanish learners rely to a larger extent on items with the *highest frequency* for that construction?

2. Compared to native speakers’ use of [*le* VERB], do L2 Spanish learners rely to a larger extent on items with the *most semantically prototypical meaning* for that construction?

Based on findings from Ellis and Ferreira-Junior's (2009) study mentioned previously, we predict that L2 Spanish learners, when compared to native writers, will use disproportionately more verbs that are high in frequency and semantically prototypical for their respective category of use of the *le* construction.

2. Methods

Our source corpus is CEDEL2 (*Corpus Escrito del Español como Segunda Lengua*, 'Written Corpus of L2 Spanish'; Lozano, 2009; Lozano & Mendikoetxea, 2013), which contains 552,401 words from essays written by 1,700 adult L1 English learners of Spanish. These were submitted online from about 1,500 different institutions around the world, 85% of which were in Anglophone countries. CEDEL2 participants were roughly equally distributed across different proficiency levels (ranging from beginner to advanced), as per a standardized Spanish grammatical placement test using multiple-choice questions in a fill-in-the-blank format (University of Wisconsin, 1998). As a native writer control, CEDEL2 also contains a subcorpus of 200,326 words from comparable essays from 660 Spanish native speakers, elicited online in the same manner and using the same essay prompts.

For this study, Spanish learners were divided into lower and higher proficiency groups, based on a median split on scores for the grammar placement test scores mentioned above. Specifically, learners with a score of 30 or below (from a maximum of 43) were assigned as "lower proficiency," and learners with a score above 30 were assigned as "higher proficiency." Information about each of these three writer groups and their essay submissions can be found in Table 1. We used placement test score as the basis for our grouping as an arguably more objective measure of Spanish proficiency than self-ratings, though we note that these aligned with the placement score results in terms of group means.

We note also that the writers in this study differed in their knowledge of additional languages beyond English and Spanish. Namely, the native writer group had a higher proportion of writers with additional languages beyond English and Spanish than the high-proficiency L2 writers, who in turn had a higher proportion of additional languages than the low-proficiency L2 writers. Furthermore, the native writers reported a higher mean proficiency in their additional languages

than high-proficiency writers, who, in turn, had a higher mean proficiency in their additional languages than the low-proficiency writers. In this way, our study is admittedly confounding Spanish proficiency with general multilingual proficiency to some extent. This may be somewhat problematic given that multilingualism has been associated with higher levels of metalinguistic awareness (Bialystok, 1987) which may in turn facilitate acquisition of additional languages (Thomas, 1988), Additionally, a look at the five most commonly reported additional languages showed that Romance languages (particularly French, Italian, Portuguese, and Catalan) were more common among writers in the high-proficiency than low-proficiency L2 groups. To our knowledge, these languages also have indirect object clitic constructions similar to Spanish *le* (i.e., French *lui*, Italian *gli*, Catalan *li*, Portuguese *lhe*). As such, Spanish proficiency as defined for the purposes of our study may, in a sense, inadvertently capture proficiency in Romance languages more generally. In this way, multilingualism wasn't exactly controlled across the writer groups. Nevertheless, in order to maximize our sample size and avoid potentially unbalanced data loss across groups, we chose to analyze the corpus as is, opting for a larger if perhaps noisier sample over a cleaner but smaller sample (i.e., if trilingual participants were excluded).

Insert Table 1 about here.

We limit our analysis to the third person clitic form because the first and second person counterparts *me*, *te*, *nos*, and *os* overlap in form with the accusative clitic forms, complicating our analysis. Additionally, to cut down on the amount of manual coding required, for this initial study we focus only the singular form *le*. Furthermore, so as to be able to extract relevant examples more easily, we focus on pre-verbal uses of *le* because post-verbal uses (e.g., *quiero decirle* 'I want to tell him') would falsely detect unrelated words that end with the same text

string (e.g., *calle* 'street'). We first extracted all lemmas following each instance of *le* using automatic concordance software (*AntConc,* Anthony, 2018). All instances of [*le* VERB] were then categorized into the four uses described above (*epistemic; psychological; benefactive, dative required; benefactive, dative not required*) as well as a fifth, "other" category that we do not analyze here, which includes:

- Cases where *le* is used as an accusative pronoun clitic, a dialectal variant known as *leísmo* that occurs largely in Spain. Examples of this include phrases like *le abrazó (*'hugged him/her') or *le vio* ('saw him/her').
- Instances of the causative construction [*le* HACER VERB] (e.g., *le hizo trabajar* 'made him/her work') under the intuition that this represents a different function than the dative constructions under analysis.
- Obvious typos (e.g., *le [sic] carro*, where the apparent intention was *el carro* 'the car')

All coding was performed by the author blind to group, i.e., so that there was no indication of which of the three participant groups the verb to be coded was from. Minor typos (e.g., *enganar* instead of *engañar*, *continuo* instead of *continúo,* etc.) were corrected so as to prevent data loss. Then, for each of the three groups, the proportions of occurrence of different verbs were calculated (overall and separately for each of the four categories) along with type-token ratios (i.e., the number of different verb types relative to the total number of occurrences of verbs).

For each of the verbs, two additional indices were also calculated. The first was the frequency of the verb in the Spanish language overall, based on the *Corpus del Español* (Davies, 2016), which comprises 5.5 billion words collected from the Internet from 2012 to the present. Note that these frequencies come from the language in general and are not calculated solely from instances of the [*le* VERB] construction. Rather, such construction-specific data about the verb comes from

our second index—Mutual Information between *le* and the verb—which measures the degree to which encountering one word helps to anticipate the occurrence of another word (for more information, see Pothos & Juola, 2007). Our intuition is that verbs with high Mutual Information with *le* constitute more archetypal instances of this construction, because they co-occur with *le* more frequently and appear without *le* less frequently.

One issue with a direct comparison across the three writer groups is that they differed in the essay prompts chosen for the corpus submission, as illustrated in Figure 1. This poses a problem because different prompts might lead to more/less use of *le* (or perhaps to a more/less diverse use of *le* verbs). For instance, an essay about the writer's personal experiences might involve more dialogue or interactions between individuals than an essay about one's home region, thus inducing more (and/or more diverse uses of) *le* constructions.

Insert Figure 1 about here.

This potential confound related to the essay prompts was addressed via stratified bootstrapping. *Bootstrapping* is a statistical method in which random subsamples are repeatedly taken from a whole sample group and analyzed separately. Aggregated measures from these subsamples provide a more robust estimation of the parameter of interest than the original sample taken as a whole. This is due to the fact that outliers would be picked only rarely in our subsamples because these outliers, by definition, occur less frequently in the whole sample. To illustrate this, if the verb *prestar* 'lend' occurred only rarely for a certain writer group, then it would occur even more rarely in a subsample of the original data sample, a difference that would be magnified when thousands of subsamples are taken and aggregated.

Stratified refers to the fact that the subsamples are taken in such a way that they adhere to a desired breakdown *vis à vis* different categories of observation from the original sample (in this case, the corpus essay prompts). In this way, we can achieve an identical number of observations for each essay prompt across group. For instance, if our three writer groups had 30, 40, and 25 essays for a given essay prompt, then we would take 20 essays from each of the groups. Using the same number of essays from this prompt would lead to comparability across groups. Meanwhile, deliberately using fewer than 25 essays would leave out some observations from the smallest group, so that different subsamples could be taken across iterations of bootstrapping.

For each subsample, we calculated how Zipfian the verb distributions were by measuring Shannon entropy (Grignetti, 1964). Recall that a distribution is more Zipfian if the most frequent tokens are very frequent and the lower-frequency tokens are very infrequent (i.e., there is a steep drop-off between the most frequent tokens and the less frequent tokens). Shannon entropy provides an index of how steep a distribution is, such that a lower entropy score indicates a more Zipfian distribution (Piantadosi, 2014).

The resampling and analyzing process was repeated 10,000 times such that, for each of the three writer groups, a distribution of entropy scores was created, with one score for each of the 10,000 bootstrapped samples. Finally, writer groups' entropy scores were compared using two-way independent samples t-tests (with $\alpha = .05$). Stratified bootstrapping was performed using the *boot* package (Ripley & Canty, 2017) for the R programming language (R Core Team, 2013).

3. Results

Table 2 presents, for each writer group, the proportion of essays that used *le* as well as the range of total *le* uses within individual essays.[2] In all, these indicate that *le* did not occur in many essays overall, and that when it did occur, it was only used a few times. This is unsurprising given their relatively short length. More importantly for our purposes, it shows that lower-proficiency learners used *le* to a lower degree overall than the other groups.

Insert Table 2 about here.

Table 3 presents statistics about the verbs used with *le*, aggregated across the four analyzed verb categories (*epistemic; psychological; benefactive, dative required; benefactive, dative not required*). For the L2 writer groups (and especially for the lower-proficiency group), lower type/token ratios indicate that learners' use of *le* was split between fewer verbs; lower mean verb frequencies indicate that learners use verbs that are less frequent in the language overall; and higher Mutual Information scores indicate that learners' verbs had stronger co-occurrences with *le*.

Insert Table 3 about here.

Figure 2 shows the breakdown of verbs across writer groups, both aggregated across and separately for the four verb categories under analysis. In the legend (top-right panel), the verbs used by lower- and higher-proficiency learners are indicated in black and gray, respectively. As can be seen, for the L2 groups (and especially the lower-proficiency group), fewer verbs

[2] Note that this and all subsequent analyses exclude a single essay from the lower-proficiency learner group which had 99 instances of *gustar* ('appeal') from an essay of only 507 words. Further examination of this essay showed it to be comprised mostly of a list of formulaic phrases (*Jennifer Lopez es Latina… Le gusta Sprite. Le gusta Mountain Dew. Le gusta Vanilla Coke.* [etc.]). This was the only outlier excluded.

constitute a larger proportion of total uses. This suggests that, the lower one's proficiency, the higher the degree to which a few given verbs dominate the instances of *le* constructions.

Insert Figure 2 about here.

Turning to the results of the stratified bootstrapped analysis, Figure 3 shows a density plot of Shannon entropies (indexing the degree to which verb frequency distributions are Zipfian) for the different writer groups. As can be seen, learners (and especially lower-proficiency ones) showed lower entropy scores—and thus, more Zipfian distributions—than native writers. This was confirmed through two-way t-tests: native writers had significantly higher entropy scores (i.e., a less Zipfian distribution) than high-proficiency learners: $t(65) = 3.65$, $p < .004$. In turn, high-proficiency learners had significantly higher entropy scores than lower-proficiency learners: $t(65) = 6.66$, $p < .001$.

Insert Figure 3 about here.

Turning to our second research question regarding semantic prototypicality, higher average Mutual Information scores for low- (6.74) and high-proficiency (5.54) Spanish learners compared to native speakers (4.43) suggests that lower-proficiency writers typically stick to verbs that are more commonly associated with *le.* To break this down by the specific semantic meanings associated with the different uses of [*le* VERB] constructions, Figure 4 shows the three writer groups' total uses of the ten most frequent verbs overall as well as the ten most frequent verbs for each category of *le* use.

Insert Figure 4 about here.

The graphs in Figure 4 indicate that low-proficiency writers typically use just a few of the verb types from each construction, and that these are the verbs that carry the most generic meaning for

that construction: *decir* ‘say’ for the epistemic category (comprising 75% of tokens for lower-proficiency writers), *gustar* ‘appeal’ for the psychological category (95% of tokens), and *dar* ‘give’ (45%) and *ayudar* ‘help’ (36%) for the dative-required benefactive category. The only exception to this came for benefactives with an optional dative, for which case the lower-proficiency L2 group showed very few instances overall and the higher-proficiency L2 learner group showed distributions that were, if anything, less Zipfian than the native writers'.

4. Discussion

For our first research question (regarding the role of frequency in L2 learners' *le* constructions), results from visual inspection of the verb breakdowns by writer group suggest that, more so than native writers, Spanish learners tend to rely on fewer verbs for most of their productions of [*le* VERB]. The results of the stratified bootstrapped analysis indicated that these findings of a more Zipfian distribution in lower-proficiency writers were not an artifact of different sample sizes for different essay prompt choices across groups, but rather were sustained when equal, stratified subsamples were taken. This aligns with our predictions for the first research question.

Contrary to our intuitions, the mean frequencies (i.e., in the language overall) of the verbs used with *le* were higher for native speakers than for L2 learners, and in turn higher for higher-proficiency speakers than for lower-proficiency speakers. Although this contradicts our initial predictions—as well as previous findings from corpus studies (e.g., Ringbom, 1998; Cobb, 2003; Kyle & Crossley, 2015)—that less proficient speakers generally use higher frequency-verbs (because these are more common and thus presumably easier/simpler to acquire), a closer look at the corpus data explained this disparity. In Spanish, many instances of *le* with a high-frequency

verb are rather idiomatic and carry meanings that are not immediately transparent to learners. For instance, in certain phrases with *hacer* (‘do’), the bulk of the meaning is carried by a collocate, as in *hacerle caso* (‘obey him/her') or *hacerle falta* (‘be missed by him/her’). Similar cases can be found with the verb *ser* ‘be’ (as in *le es infiel* ‘is unfaithful to him/her’) and collocates with *poner* ‘put’ (as in *Esta película le pone rostro a los secuestros* ‘This movie “gives face” to the kidnappings’). In this way, words with high frequency in the language overall are not necessarily simple or easy for language learners to learn. This suggests that future researchers should broaden their granularity of analysis beyond the single-word level so as to also account for such relatively uncommon and highly idiomatic multi-word phrases.

For our second research (regarding the role of semantic prototypicality in L2 learners' *le* constructions), higher Mutual Information scores for the L2 writer groups as well as the finding that the ten most common verbs for each category of the *le* construction tended to involve the most generic meaning for that respective category suggest that semantic prototypicality does play a role at initial stages of acquisition of this construction. That said, the “benefactive, dative not required” category showed a slightly different pattern than the other uses of *le*: there were many fewer tokens in this category overall across all groups (following intuitively from the fact that the dative is optional for these verbs to begin with), and the most common verb (*hacer* 'do') came mostly from native writers, breaking the pattern wherein lower-proficiency writers were the ones who supply the majority of tokens for the most common verb for each category. This might be due to the previously-mentioned non-Englishlike nature of Spanish constructions with the verbs *hacer* 'do,' as shown in phrases like, e.g., *hacerle gracia* ('be funny to him/her’, lit. ‘make him/her funniness’).

Our results so far indicate that, in the instances of [*le* VERB] analyzed here, Spanish learners rely on verbs with high frequencies (for that construction, though not the language overall) and with semantically prototypical meanings to a higher degree than native speakers do. However, we do not wish to imply that this deviation from native speaker linguistic behavior is necessarily a bad thing: after all, learners' idiosyncratic language behavior may suit their needs during the early stages of morphosyntax acquisition, by facilitating what would otherwise be the cognitively difficult task of using newly-learned grammatical constructions in conjunction with uncommon verbs. It is also worth remembering that Spanish learners are capable of both breaking and exploiting the patterns we observe here, in deliberate and perhaps self-conscious ways. As an example of *breaking* the pattern, one Spanish learner essay from the CEDEL2 corpus reads: *le encajuelaron (encajuelar = encerrar a alguien en la cajuela de su auto) y le robaron todas sus pertenencias* 'they entrunked [sic] (entrunk = to lock somebody in their car trunk) and stole his/her belongings.' This rather uncommon verb *encajuelar* contains no entry in the Royal Spanish Academy's online dictionary (Real Academia Española, 2018), though it is defined in a dictionary titled *Essential Mexican vocabulary* (Macazaga y Ordoño, 1999). The non-native writer's use of such a highly dialectal form with an explicit definition is somewhat poetic or stylish in light of the author's positioning as a learner rather than a native speaker. As an example of *exploiting* the pattern, even formulaic essays like our single excluded outlier's repetitive use of *le gusta* (e.g., *[A Jennifer López l]e gusta cantar y bailar. Le gusta la television. No le gusta la trompeta. [etc.]* 'Jennifer Lopez likes singing and dancing. She likes television. She doesn't like the trumpet.') carry a certain charm when we recognize how much can ultimately be communicated by a writer who, in opting for such repetitive forms, is clearly making no secret of their status as a second language learner. This charm might be called a

privilege of the non-native speaker, who can at times enjoy certain stylistic opportunities that a native speaker can't (Kramsch, 1997).

5. Conclusion

In sum, Spanish learners' uses of the [*le* VERB] construction in the CEDEL2 corpus were oriented more heavily than native writers' towards the verbs that were most frequent and had the most prototypical meanings for that construction. As suggested by summary statistics and confirmed with a stratified bootstrapped analysis, their most common verbs for a given construction were disproportionately more frequent, such that a select few of learners' verbs make up a lion's share of instances of the construction. This effect was more pronounced for lower proficiency learners than for higher proficiency learners. Additionally, when compared to native writers, learners used a higher proportion of verbs whose meaning was semantically prototypical for the different uses of *le*, such as *decir* ('say'), *gustar* ('appeal'), *ayudar* ('help'), and *dar* ('give'). These findings contribute to theoretical debates in linguistics by lending support to usage-based approaches that emphasize the roles of usage frequency and semantic prototypicality in the early stages of grammar acquisition (e.g., Ellis, 2016a), contrasting with other approaches that place less emphasis on the statistics of the second language input (e.g., Pienemann & Lenzing, 2015; VanPatten, 2015; White, 2015).

Beyond debates in psycholinguistics, the findings from this study are highly relevant for foreign language educators in illustrating the ways that second language learner writing differs from that of native speakers. If the goal of any teacher, curriculum designer, or tutor is ultimately to help a student write like a native speaker, then identifying the ways that language learners differ from

native speakers would be a critical first step in this endeavor (Wolfe-Quintero, Inagaki, & Kim, 1998). Previous controlled longitudinal classroom studies (e.g., Madlener, 2015; Year & Gordon, 2009) show that an understanding of the statistical properties of language can be leveraged to help shape second language input as well as opportunities for output in the classroom environment so as to best facilitate acquisition. Going beyond the generic recommendation to use a richer vocabulary, one specific suggestion that might be made from this study is to deliberately use this richer vocabulary in conjunction with grammatical constructions, particularly when the vocabulary doesn't necessarily align with the archetypal meaning of the construction.

On a final note, this project serves as an example of how simple and user-friendly tools can be used to analyze second language learner output. The software programs we used for coding and visualizing our data have been available for more than two decades, and the past few years have seen an explosion in the programs available for automatically analyzing corpora. Among the free (or relatively inexpensive) software programs currently available are VocabProfile (Cobb, 2013; Heathley, Nation, & Coxhead, 2002); Linguistic Inquiry and Word Count (LIWC; Pennebaker, Booth, & Francis, 2007); Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004; McNamara & Graesser, 2012; McNamara, Graesser, McCarthy, & Cai, 2014); and the Tool for the Automatic Analysis of Lexical Sophistication (TAALES; Kyle & Crossley, 2015), among others. These would be of interest both to researchers and to educators for facilitating tasks like grading essays automatically (e.g., Crossley & McNamara, 2013) and identifying the linguistic features that students can change to make their writing more native-like (e.g., Cobb, 2017). The potential value of such research methodologies would only increase in light of the growing availability of Spanish corpus data (Moreno-Fernández, 2018).

6. Limitations and Future Directions

This preliminary study is not without its limitations. First, the verbs were coded manually into different categories by the author. Although this was done blind to participant group (such that the relevant across-participant comparisons would not be affected), this introduces an element of human error to the study. This harkens back to a classic tradeoff in corpus research between data quantity vs. quality—that is, reliability of its parsing and cleaning/coding (Roland, Dick, & Elman, 2007). One way to circumvent this issue would be to involve additional coders and assess measures of reliability between them, or to automate coding completely by exploiting co-occurrences of words in similar contexts such that words with the same general meaning (e.g., "say," "give," or "seem") would cluster together (e.g., Lee, Goldsmith, & Jacobs, 2015).

Another limitation is that the three writer groups varied between each other in several factors other than just Spanish proficiency, such as age and prior knowledge of languages beyond English and Spanish (which may affect essay-writing skills, e.g., due to increased metalinguistic awareness; Klein, 1995; Thomas, 1988). Furthermore, the mean lengths of the three groups' essays varied, which might qualitatively affect the content of the essays, for instance, in the tone or level of detail used by the author. One way to address these potential confounds would be to include them as factors when stratifying subsamples for the bootstrapped analysis (e.g., by taking the same proportion of essays from bilinguals vs. trilinguals for each of the Spanish native and low-/high-proficiency learner groups). However, each additional stratum in our resampling method would require a correspondingly larger original corpus dataset to maintain sufficient sample sizes while ensuring parity between groups. As data collection or the CEDEL2 corpus is currently ongoing, hopefully this will become more feasible in the future.

We recognize also that we only examine instances where *le* was used, and not cases where *le* should have been used but wasn't. As such, there is the possibility that the Spanish learners in our sample actually had a vocabulary that was as varied as the native writers' but did not employ it fully in the dative construction. However, an analysis of all verbs in the corpus would require a much more exhaustive approach that would lie beyond the scope of this initial study. More importantly, to the extent that we wish to see what the initial stages of grammatical acquisition look like rather than simply measuring learners' vocabulary in general, analyzing the verbs *not* used with *le* would seem to fall outside the bounds of our immediate research question.

Finally, it is worth noting that online corpora are not foolproof measures of a language user's proficiency. For instance, lower proficiency writers may consciously avoid using *le* in cases where they are unsure that they are using it in the right way, a phenomenon of avoidance that has been widely discussed in the language acquisition literature (e.g., Chiang, 1980; Gass, 1980; Li, 1996; Maniruzzuman, 2008; Schachter, 1974; Zhao, 1989). In this way, evidence of absence is not absence of evidence. Conversely, learners might go out of their way to use more impressive vocabulary because they are aware that their linguistic skills are being assessed in a way that the native speakers' aren't. This is related to the so-called "John Henry effect," wherein members of an experimental group may try to compensate for a perceived disadvantage in such a way that the results are ultimately distorted (Saretsky, 1972). On a more epistemic level, Mackey and Gass (2005) note that the presence of a certain form in a text does not mean that the writer has necessarily acquired that form; conversely, a form's absence does not mean that the writer has not acquired that form.

References

Ambridge, B., Pine, J. M., Rowland, C. F., Chang, F., & Bidgood, A. (2013). The retreat from overgeneralization in child language acquisition: Word learning, morphology, and verb argument structure. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(1), 47-62.

Anthony, L. (2018). *AntConc* (Version 3.5.6) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/software

Belletti, A., & Rizzi, L. (1988). Psych-verbs and θ-theory. *Natural Language & Linguistic Theory*, *6*, 291–352.

Bernolet, S., Hartsuiker, R., & Pickering, M. (2007). Shared syntactic representations in bilinguals: Evidence for the role of word-order repetition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 931–949.

Bialystok, E. (1987). Influences of bilingualism on metalinguistic development. *Interlanguage studies bulletin (Utrecht)*, *3*(2), 154-166.

Bybee, J. (2001). *Phonology and language use*. Cambridge, UK: Cambridge University Press.

Bybee, J. (2010). *Language, usage and cognition*. Cambridge University Press.

Cambridge Michigan Language Assessments (2013). *Michigan English Language Assessment Battery Speaking Rating Scale.* Retrieved September 24, 2018 from http://www.cambridgemichigan.org/wp-content/uploads/2014/11/MELAB-RatingScale-Speaking.pdf

Chiang, T. (1980). *Error analysis: A study of errors made in written English by Chinese learners* (MA dissertation). National Taiwan University, Taipei, Taiwan.

Cobb, T. (2003). Analyzing late interlanguage with learner corpora: Quebec replications of three European studies. *Canadian Modern Language Review*, *59*(3), 393-424.

Cobb, T. *(*2013*). Web VocabProfile. Retrieved from* http://www.lextutor.ca/vp

Cobb, T. (2017). *Broad- versus narrow-band lexical frequency profiling*. March 18th-21st, American Association for Applied Linguistics 2017. Portland, Oregon, USA.

Crossley, S. A. & McNamara, D. S. (2013). Applications of text analysis tools for spoken response grading. *Language Learning & Technology*, *17*, 171–192.

Cuervo, M. C. (2003). *Datives at large* (Doctoral dissertation). Massachusetts Institute of Technology, Boston, Massachusetts. Retrieved from http://hdl.handle.net/1721.1/7991

Davies, M. (2016) *Corpus del Español: Two billion words, 21 countries*. Available online at http://www.corpusdelespanol.org/web-dial/

Duchon, A., Perea, M., Sebastián-Gallés, N., Martí, A., & Carreiras, M. (2013). EsPal: One-stop shopping for Spanish word properties. *Behavior research methods*, *45*(4), 1246-1258.

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in second language acquisition*, *24*(2), 143-188.

Ellis, N. C. (2013). Construction Grammar and second language acquisition. In Thomas Hoffmann & Graeme Trousdale (eds.), *The Oxford Handbook of Construction Grammar* (pp. 365-378). Oxford: Oxford University Press. 365–378.

Ellis, N. C. (2016a). Frequency in language learning and language change. In H. Behrens & S. Pfänder (Eds.), *Experience counts: Frequency effects in language* (pp. 239-256). Berlin: de Gruyter.

Ellis, N. C. (2016b). Online processing of Verb–Argument Constructions: Lexical decision and meaningfulness. *Language and Cognition*, *8*(3), 391-420.

Ellis, N. C., & Ferreira–Junior, F. (2009). Construction learning as a function of frequency, frequency distribution, and function. *The Modern Language Journal*, *93*(3), 370-385.

Ellis, N. C., & Larsen-Freeman, D. (2009). Constructing a second language: Analyses and computational simulations of the emergence of linguistic constructions from usage. *Language Learning, 59*, 90-125.

Ellis, N. C., O'Donnell, M. B., & Römer, U. (2014). Second language verb-argument constructions are sensitive to form, function, frequency, contingency, and prototypicality. *Linguistic Approaches to Bilingualism*, *4*(4), 405-431.

Gass, S. (1980). An investigation of syntactic transfer in adult second language learners. In R. Scarcella & S. Krashen (Eds.), *Research in second language acquisition*. Rowley, MA: Newbury House.

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford University Press on Demand.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, 36*(2), 193-202.

Grignetti, M. C. (1964). A note on the entropy of words in printed English. *Information and Control*, *7*(3), 304-306.

Heathley, A., Nation, I.S.P., & Coxhead, A. (2002*). Range and Frequency Programs. Retrieved from* http://www.victoria.ac.nz/lals/staff/paul-nation.aspx

Hoffmann, T., & Trousdale, G. (Eds.). (2013). *The Oxford handbook of construction grammar*. Oxford University Press.

Izquierdo, J. (2007). *Multimedia environments in the foreign language classroom: Effects on the acquisition of the French perfective and imperfective distinction* (doctoral dissertation). McGill University, Montreal, Canada.

Kirsner, K. (1994). Implicit processes in second language learning. In N. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 283-312). San Diego, CA: Academic Press.

Klein, E. C. (1995). Second versus third language acquisition: Is there a difference? *Language learning, 45*(3), 419-466.

Kramsch, C. (1997). Guest column: The privilege of the nonnative speaker. *Publications of the Modern Language Association of America, 112*(3), 359-369.

Kyle, K. & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly* 49(4), pp. 757-786. doi: 10.1002/tesq.194

Lee, J., Goldsmith, J. & Jacobs, S. (2015). *Learning linguistic structure: Integrating big data analysis and visualization.* Poster presented at Research Computing Center conference, University of Chicago. Accessed online at https://mindbytes.uchicago.edu/2015/posters/rcc2015-poster.pdf

Lehtonen, M., & Laine, M. (2003). How word frequency affects morphological processing in monolinguals and bilinguals. *Bilingualism: Language and Cognition, 6*(3), 213-225.

Li, J. (1996). Underproduction does not necessarily mean avoidance: Investigation of underproduction using in Chinese ESL learners. In L.F. Bouton (Ed.), *Pragmatics and language learning: Monograph series 7* (pp. 171-187). University of Illinois at Urbana-Champaign.

Lozano, C. (2009). CEDEL2: Corpus Escrito del Español L2. In C. Callejas et al. (Eds.), *Applied Linguistics Now: Understanding Language and Mind / La Lingüística Aplicada Hoy: Comprendiendo el Lenguaje y la Mente* (pp. 197-212). Almería: Universidad de Almería.

Lozano, C., & Mendikoetxea, A. (2013). Learner corpora and SLA: the design and collection of CEDEL2. In A. Díaz-Negrillo, N. Ballier and P. Thompson (Eds.), *Automatic Treatment and Analysis of Learner Corpus Data* (pp. 65-100). Amsterdam: John Benjamins.

Macazaga y Ordoño, C. (1999). *Vocabulario esencial mexicano*. Accessed online at https://editorialcosmos.com/vocabulario-esencial-mexicano/

Mackey, A., & Gass, S. M. (2005). *Second Language Research: Methodology and Design.* Mahwah, NJ: Lawrence Erlbaum Associates.

Madlener, K. (2015). *Frequency effects in instructed second language acquisition*. Berlin: De Gruyter Mouton.

Maniruzzaman, M. (2008). *Avoidance behaviour in EFL Learning: A study of undergraduates.* Accessed http://www.articlesbase.com/languages-articles/avoidance-behaviour-in-efl-learning-a-study-of-undergraduates-297436.html

McNamara, D. S., & Graesser, A. C. (2012). Coh-Metrix: An automated tool for theoretical and applied natural language processing. In P. McCarthy & C. Boonthum-Denecke (Eds.), *Applied natural language processing: Identification, investigation and resolution* (pp. 188-205). Hershey, Pennsylvania: IGI Global.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix.* Cambridge University Press.

Moreno-Fernández, F. (Chair) (2018). *Corpus de lengua española hablada.* Panel held at Hispanic Linguistics Symposium, October 25-27. Austin, TX.

Navarro, S., & Nicoladis, E. (2005). Describing motion events in adult L2 Spanish narratives. In *Selected Proceedings of the 6th Conference on the Acquisition of Spanish and Portuguese as First and Second Languages* (pp. 102-107).

Oxford Dictionaries (2019). *What can the Oxford English Corpus tell us about the English language?* Accessed at https://en.oxforddictionaries.com/explore/what-can-corpus-tell-us-about-language/

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007*). Linguistic inquiry and word count: LIWC* [Computer software]. Austin, TX: liwc.net.

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, *21*(5), 1112-1130.

Pienemann, M. & Lenzing, A. (2015). Processability Theory. In B. VanPatten and J. Williams (Eds.), *Theories in Second Language Acquisition: An Introduction*, 2nd ed. Mahwah, NJ. Routledge.

Pothos, E. M., & Juola, P. (2007). Characterizing linguistic structure with mutual information. *British Journal of Psychology*, *98*(2), 291-304.

R Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Accessed at http://www.R-project.org/.

Real Academia Española (2018). *Diccionario de la lengua española*. Accessed at http://dle.rae.es/

Ringbom, H. (1998). Vocabulary frequencies in advanced learner English: A crosslinguistic approach. In S. Granger (Ed.), *Learner English on Computer* (pp. 41-52). New York: Addison Wesley Longam.

Ripley, B. & Canty, A. (2017). *'boot': Bootstrap Functions.* R package version 1.3-20

Roland, D., Dick, F., & Elman, J. (2007). Frequency of Basic English Grammatical Structures: A Corpus Analysis. *Journal of Memory and Language.* 57(3), 348-79.

Saffran, J. R. (2002). Constraints on statistical language learning. *Journal of Memory and Language*, *47*(1), 172-196.

Saretsky, G. (1972). The OEO P.C. Experiment and the John Henry Effect. *The Phi Delta Kappan*, *53*(9), 579–581.

Schachter, J. (1974). An error in error analysis. *Language Learning* 24, 205-214.

Schuchart, H. (1885). *Über die Lautgesetze: gegen die Junggrammatiker*. Berlin: R. Oppenheim.

Shantz, K. (2017). Phrase frequency, proficiency and grammaticality interact in non-native processing: Implications for theories of SLA. *Second Language Research, 33*(1), 91-118.

Thomas, J. (1988). The role played by metalinguistic awareness in second and third language learning. *Journal of Multilingual & Multicultural Development, 9*(3), 235-246.

University of Wisconsin. (1998). *The University of Wisconsin College-Level Placement Test: Spanish (Grammar) Form 96M*. Madison, WI: University of Wisconsin Press.

VanPatten, B. (2015). Input Processing in Adult SLA. In B. VanPatten and J. Williams (Eds.), *Theories in Second Language Acquisition: An Introduction*, 2nd ed. Mahwah, NJ. Routledge.

White, L. (2015). Linguistic Theory, Universal Grammar, and Second Language Acquisition. In B. VanPatten and J. Williams (Eds.), *Theories in Second Language Acquisition: An Introduction*, 2nd ed. Mahwah, NJ. Routledge.

Williams, J. N., & Kuribara, C. (2008). Comparing a nativist and emergentist approach to the intial stage of SLA: An investigation of Japanese scrambling. *Lingua, 118*, 533–553.

Wolfe-Quintero, Inagaki, K., S., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity (Report No. 17)*. Honolulu: University of Hawai'i, Second Language Teaching and Curriculum Center.

Year, J., & Gordon, P. (2009). Korean speakers' acquisition of the English ditransitive construction: The role of verb prototype, input distribution, and frequency. *The Modern Language Journal, 93*(3), 399-417.

Zhao, R. (1989). A discourse analysis of relative clauses in Chinese and English: An error in 'An error in error analysis.' *IDEAL, 6*, 105-17.

Zipf, G. K. (1929). Relative frequency as a determinant of phonetic change. *Harvard studies in classical philology, 40,* 1-95.

Zyzik, E. (2006). Transitivity alternations and sequence learning: Insights from L2 Spanish production data. *Studies in Second Language Acquisition*, *28*(3), 449-485.

Table 1 Summary statistics for the three writer groups compared in this study.

	L2 – Lower Proficiency	**L2 – Higher Proficiency**	**Native**
Sample size	825 (541 female)	780 (569 female)	796 (582 female)
Mean age (standard deviation)	28.56 (13.89)	22.55 (9.64)	31.14 (10.44)
Mean age started learning Spanish (standard deviation)	16.81 (9.41)	16.67 (10.05)	N/A
Mean placement test score from max. of 43 (standard deviation)	20.17 (5.50)	37.61 (3.51)	N/A
Mean self-rated Spanish proficiency on a six-point scale	2.68 (1.07)	4.33 (0.86)	N/A
Participants with languages other than English or Spanish	13.70%	38.08%	72.61%
Reported additional languages	French (37%), German (15%), Chinese (6%), Italian (5%), Japanese (4%), 21 other languages (33%)	French (52%), German (9%), Italian (8%), Portuguese (7%), Catalan (7%), 24 other languages (17%)	French (45%), German (17%), Catalan (10%), Italian (9%), Portuguese (5%), 21 other languages (14%)
Mean self-rated proficiency in additional languages (standard deviation)	2.93 (1.44)	3.39 (1.51)	4.07 (1.65)
Mean essay word count (standard deviation)	390.381 (150.06)	159.48 (162.31)	265.98 (168.14)

Table 2 Proportion of essays that used le *and range of total number of* le *uses within essay for each of the three writer groups.*

	L2 – Lower Proficiency	**L2 – Higher Proficiency**	**Native**
Total essays that used *le* (percentage)	75 (9.09%)	218 (27.56%)	176 (22.11%)
Range of total uses of *le* within one essay	0-7	0-17	0-9

Table 3 Summary statistics for each group's le *construction usage, aggregated across the four analyzed verb categories.*

	L2 – lower proficiency	**L2 – higher proficiency**	**Native**
Total instances of *le*	117	401	306
Number of distinct verbs used with *le*	12	80	123
Type/token ratio	0.10	0.20	0.40
Mean verb frequency (instances per million words)	33.52	83.23	135.63
Mean Mutual Information between *le* and verb	6.74	5.54	4.43

Figure 1. Breakdown of the three writer groups' essay submissions by prompt.

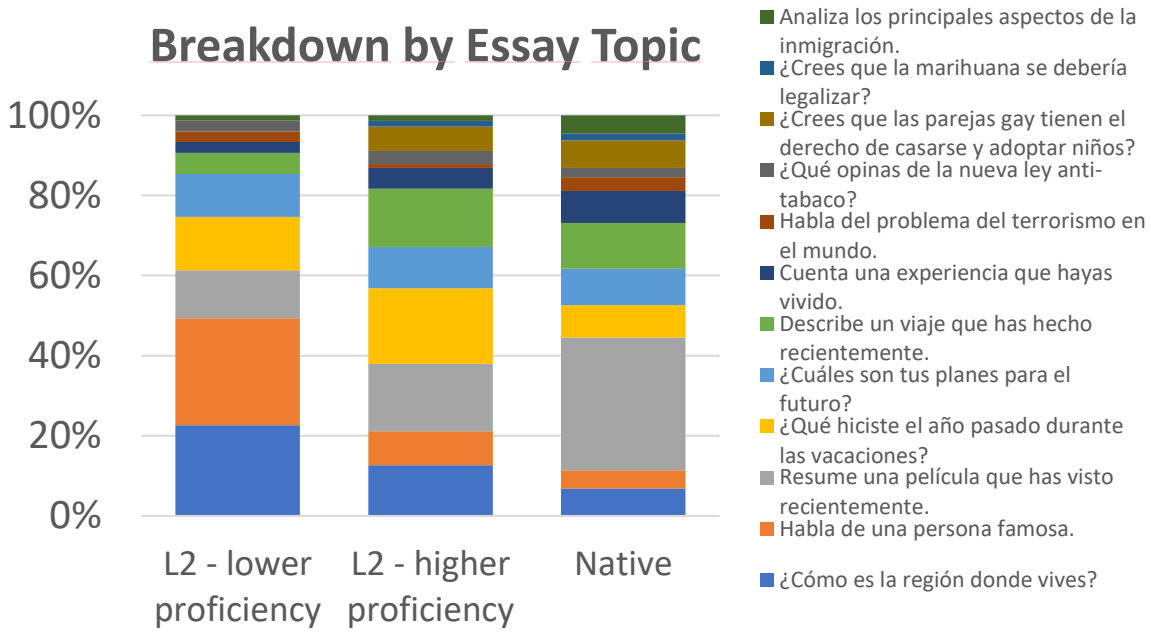

Figure 2. Breakdown of le *construction verb use by writer group, shown compiled across and separately for the different categories. The legend (top-right panel) indicates the verbs used by the lower proficiency group in black and the verbs used by the higher proficiency group in gray.*

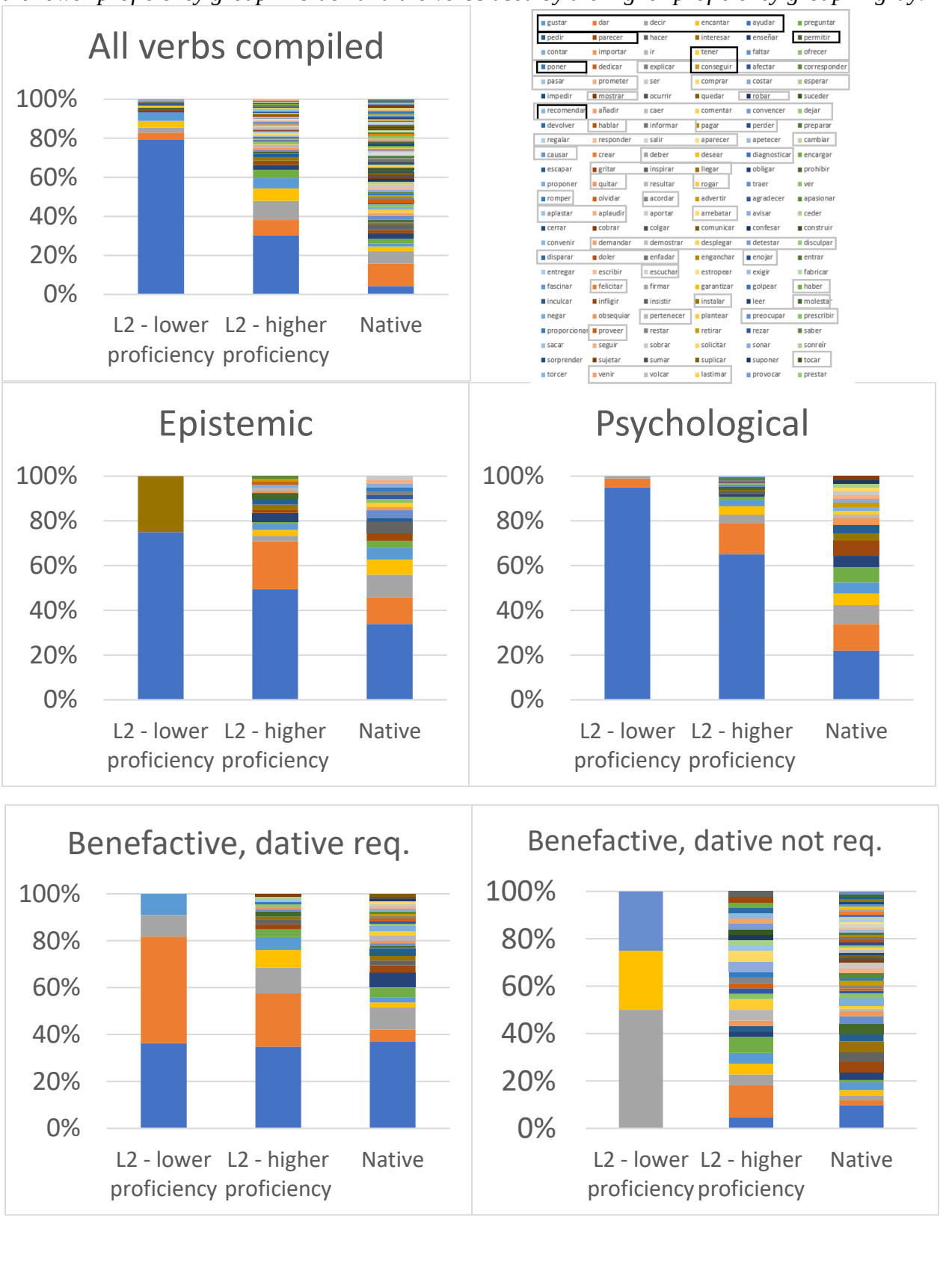

Figure 3. Density plot of Shannon entropies (an index of how Zipfian the le *construction's verb distribution is) for different writer groups under the stratified bootstrap analysis.*

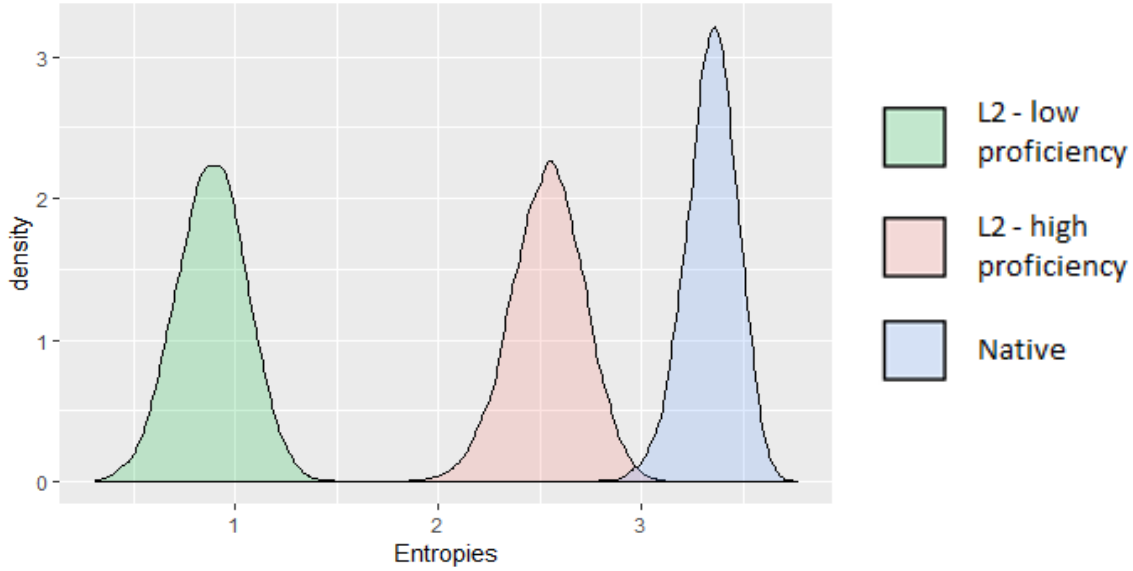

Figure 4. Each writer group's total uses of the ten most frequent verbs (overall across groups), shown aggregated across as well as separately for the different le *construction categories.*