Using variationism and learner corpus research to investigate grammatical gender marking in additional-language Spanish

Aarnes Gudmestad, Amanda Edmonds, Thomas Metzger

PRE-PRINT VERSION
Gudmestad, A., Edmonds, A., & Metzger, T. (2019). Using variationism and learner corpus research to investigate grammatical gender marking in additional-language Spanish. *Language Learning, 69*(4), 911-942.

Using Variationism and Learner Corpus Research to Investigate Grammatical Gender Marking in Additional-Language Spanish

Abstract

The current study responds to the call for increased dialogue among different areas of additional-language research. Specifically, we bring together learner corpus research (LCR) and variationist approaches to second language acquisition in order to advance LCR by modeling interlanguage development and variability and by conducting an analysis that moves away from descriptive error analysis. To accomplish this goal, we analyze grammatical gender marking in additional-language Spanish. The data come from LANGSNAP, a longitudinal learner corpus. We demonstrate through the use of a generalized linear mixed model – a multivariate, quantitative analysis – new information about the development and variable use of gender marking in additional-language Spanish. Moreover, we hope to contribute to greater interactions among researchers working in different strands of additional-language research.

Keywords: variability, learner corpus research, variationist approaches, regression, longitudinal, grammatical gender, Spanish

Learner corpus research (LCR), "although still quite young […] has already undergone remarkable developments," such as the move from mostly English language corpora to the availability of learner corpora in various languages (Granger, Gilquin, & Meunier, 2015, p. 2). Despite this growth, scholars have identified areas that are in need of further research and

reform. Key examples include the need for investigations of interlanguage development and variability, as well as a call for analyses that move beyond univariate descriptions of learner errors (e.g., Ädel, 2015; Gries, 2015b; Myles, 2005). One possible way to address these needs for research and reform is to look outside LCR to other approaches and paradigms. Myles (2015), for example, advocates for greater connections between LCR and the field of second language acquisition (SLA).[1] In the current study, we respond to this call by bringing together LCR and one area of SLA research – namely, variationism – that we believe is well suited to help advance LCR (see *Background*). More specifically, our analysis of a case of Type I variation (Rehner, 2002, see definition in the following section) further advances LCR by modeling development and variability in interlanguage, thus moving away from descriptive error analysis, which has tended to be univariate (i.e., the analysis of a dependent variable and a single independent variable). Through the use of a generalized linear mixed model – a multivariate, quantitative analysis – we offer new information about the development and variable use of gender marking in additional-language Spanish. This analysis models data from the learner corpus LANGSNAP, which contains longitudinal data of additional-language learners' production collected over a period of 21 months. In this vein, we attempt to open a dialogue between LCR and variationist SLA and we hope to contribute to greater interaction among researchers working in different areas of additional-language research.

Background

In this section we first introduce LCR and variationism. Next, we offer a brief overview of research on grammatical gender marking in additional-language Spanish.

LCR

LCR is an area of additional-language research that sprouted from corpus linguistics and began in the 1980s. Given the importance of linguistic data for research on additional-language learning and pedagogy, its main goals have been to contribute to these areas of scholarship by creating corpora of learner-language production that are available in electronic form and by conducting analyses of these data (Granger et al., 2015). As this research strand has developed over the years, there has been a shift from a dominance of corpora that consist of written essays, cross-sectional data, and learners of English to those that reflect different task types, longitudinal data, and other languages (cf. Alonso-Ramos, 2016). While LCR has contributed valuable knowledge about learner varieties, we highlight here three issues that are in need of further investigation.

The first concerns interlanguage variation.[2] Variability, meaning that "learners do not consistently make use of a single form or pattern, but rather show a preference for the use of one form among others that they are using during the same period" (Ellis, 2008, p. 117), is a prevalent characteristic of additional-language production that has been shown to be systematic (Tarone, 1988). While LCR has explored first-language influence as a source of variability, other issues remain understudied (Ädel, 2015). In fact, despite its prevalence in interlanguage, Ädel (2015, p. 420) writes, "[l]earner corpus research into variability is such an underexplored area that the fundamental issue involves taking variability into account in the first place."

The second issue in need of attention is additional-language development. Most corpora "allow the study of L2 [second-language] learners' language use at a single point in time (Callies, 2015, p. 37), which means that LCR has been limited in the contributions it has made to the understanding of language change among learners (Myles, 2005). This limitation, of course, is not unique to LCR. Ortega and Iberri-Shea (2005) write, "Why is longitudinal research

essential to advancement of knowledge in the field of SLA? The simple but uncontestable answer is that many questions concerning second language learning are fundamentally questions of time and timing" (p. 27). Thus, research that does not investigate language development longitudinally is missing a key element of development – time. The availability of longitudinal corpora, however, has been increasing (Granger et al., 2015), so there is potential for new research on additional-language development within LCR (see Meunier & Littré, 2013, and Huensch & Tracy-Ventura, 2017, for examples of longitudinal LCR research). With the current study, we aim to contribute to the growing body of longitudinal investigations in LCR.

The third pertains to analysis. LCR scholars have tended to conduct two types of analysis: contrastive interlanguage analysis and computer-aided error analysis (Callies, 2015, pp. 39-41). Contrastive interlanguage analysis focuses on comparisons between learners and native speakers in order to identify differences between the two and among learners of various first-language backgrounds. A principal goal of this type of analysis is to uncover whether the ways in which learners differ from NSs are unique to certain populations or common among learners, regardless of language background. Computer-aided error analysis utilizes corpus tools to identify incorrect features of learner production, allowing researchers to categorize and count errors. Although these two types of analyses are common in LCR, scholars have identified limitations. Myles (2005) argues that they are descriptive analyses lacking in explanatory power (p. 380). Another weakness is that LCR analyses tend to be univariate, but, as pointed out by Gries (2015b, p. 175), "nothing in linguistics is truly monocausal, so this perspective is impoverished." In recent years, however, there has been a push to broaden the analytical tools employed by LCR researchers, and multifactorial statistical analyses are being conducted more often (e.g., Gries, 2015b; Meunier & Littré, 2013; Wulff, 2017). In the next section we highlight

the characteristics of variationist SLA that we believe can help to move these three issues in LCR forward.

Variationist SLA

According to Ortega (2015, p. 251), systematicity and variability are "paramount in disciplinary understandings of interlanguage," although different frameworks may give more attention to one of these two characteristics. Variationism, a branch of sociolinguistics (Labov, 1966), is concerned with understanding systematic variability and its evolution. It examines the relationship between language and society and attempts to identify the linguistic and social factors that impact the variation and change of linguistic phenomena within and among communities of speakers. Variationists identify *linguistic variables*, or phenomena that vary in language, that are realized by at least two forms, called *variants*. For instance, in Spanish, future-time reference is a *linguistic variable* that is realized by at least three verb forms (*variants*): inflectional future, periphrastic future, and present indicative. The *envelope of variation* refers to the contexts where there is variation among two or more forms, rather than categorical use of a single form. Broadly defined, the envelope of variation for Spanish future-time reference is an action or state that is posterior to the utterance moment. Researchers identify a multitude of internal (linguistic) and external (social, extra-linguistic) factors that condition the use of the linguistic variable, in order to explain variable patterns and language change. For instance, future-time reference has been shown to vary by the distance between a future event and the moment of speaking, a linguistic factor, and geographical region, an extra-linguistic factor (e.g., Kanwit & Solon, 2013, who compared native speakers of Spanish from Valencia, Spain and Mérida, Mexico, in addition to additional-language learners).

Although variationism began with investigations on native speakers, it was later extended to additional-language learning (e.g., Preston, 1989). Variationist SLA examines two types of variability: Type I and Type II (Rehner, 2002). Beginning with Type II variation, which is more frequently studied, this type of variation concerns the linguistic variables that vary sociolinguistically among native speakers (e.g., future-time expression in Spanish). Given variable use among native speakers, learners' input with respect to any such linguistic variable will show variation, and the target of acquisition will necessarily be variable as well (Geeslin with Long, 2014). The variationist approach has also been applied to linguistic structures that are categorical among native speakers, signifying that the input and the target of acquisition are presumably not variable (e.g., plural marking in English; Young, 1991). This is called Type I variation and acknowledges the fact that learners may use structures in a variable manner, even if variation is not instantiated in the (native-speaker) input. On the whole, variationists have carried out little research on Type I variation, with the vast majority of variationist SLA having focused on Type II variation, leading to a wealth of studies on how learners approximate norms that are variable (Geeslin with Long, 2014). This imbalance between studies of Type I and Type II variation can be explained by a strong interest in socially conditioned variation within sociolinguistic research on native speakers. However, the small body of research applying the variationist paradigm to Type I variation (e.g., Bayley & Langman, 2004; Berdan, 1996; Young, 1991, 1996) demonstrates concretely and convincingly that variationism's conceptual and methodological tools have the potential to enable researchers to provide intricate explanations of developmental trajectories with respect to interlanguage variation. For example, Berdan (1996) investigated the use of verbal negation in additional-language English (a case of Type I variation). He examined longitudinal data from the well-studied learner Alberto (Schumann,

1975). Previous univariate analyses documenting the frequency of use of *don't* + verb and *no* + verb did not show evidence of development over time. However, when Berdan conducted a variationist analysis of the data and fit a regression model, he found that, among other factors, subject noun phrase and the interaction between time and style constrained use. Thus, Berdan's investigation not only showed evidence of systematicity in the use of *don't* + verb and *no* + verb but also that Alberto's use of these forms changed over time.[3] The current study attempts to further expand variationist research on Type I variation with an analysis of grammatical gender marking in additional-language Spanish.

Just as with scholarship on native speakers, variationist SLA seeks to understand the internal and external constraints on variation and how these factors change over time. Thus, additional-language development

> can be understood through changes in the frequency of use of a given form, changes in the linguistic and extra-linguistic factors that constrain that use, and, more subtly, changes in the distribution of the rates of use across the categories of those constraining variables. (Geeslin with Long, 2014, p. 191)

A common analytical tool in this area of research is a regression model. This multivariate statistical analysis allows scholars to model the complex array of linguistic and extra-linguistic factors that explain the systematicity in variable use, resulting in detailed, data-driven accounts of learner varieties and their evolution.

Thus, variationism has appealing characteristics that have the potential, we believe, to be fruitfully applied to LCR.[4] First, variation is at the heart of variationist SLA, so bringing a variationist perspective to LCR should help to give greater attention to interlanguage variability. Second, the methodological tools that are common in variationism provide a flexible and detailed

approach to modeling development. These models, which are generally carried out using regression analyses, allow the researcher to analyze multiple independent factors at the same time. Thus, these tools respond to the need in LCR for more sophisticated models (Gries, 2015b) that can offer explanations, rather than only descriptions, of interlanguage and its evolution (Myles, 2005).[5]

Grammatical Gender Marking in Spanish

In Spanish every noun has either feminine or masculine gender. Gender is determined by biological sex for some nouns (*madre*$_{\text{fem}}$ 'mother', *padre*$_{\text{masc}}$ 'father'), but this is not the case for most (*revista*$_{\text{fem}}$ 'magazine', *libro*$_{\text{masc}}$ 'book'). All determiners and adjectives agree in gender with the noun they modify (*un*$_{\text{masc}}$ *viaje*$_{\text{masc}}$ *fabuloso*$_{\text{masc}}$ 'a trip fabulous'), though not all modifiers are overtly marked for gender (e.g., the same form of the adjective *interesante*$_{\text{masc/fem}}$ 'interesting' is used to modify both masculine and feminine nouns). The canonical inflectional endings for gender with nouns and modifiers are the masculine '-o' and the feminine '-a.' Despite this feature, there are exceptions, such that masculine nouns ending in '-a' (e.g., *problema* 'problem') and feminine nouns ending in '-o' (e.g., *mano* 'hand') exist. In addition to the canonical endings, there are not only other noun endings that tend to be linked to a particular gender (e.g., '-e' for masculine nouns and '-ción' for feminine nouns; Teschner & Russell, 1984), but also endings that are not strongly connected to one gender (e.g., '-s'). Native speakers have been shown to be (near-)categorical in their gender-marking behavior in language production (e.g., Alarcón, 2011; Montrul, Foote, & Perpiñan, 2008).[6] To our knowledge, other than a small set of nouns whose gender varies by region (e.g., *azúcar* 'sugar'), there is no evidence of sociolinguistic variation with this linguistic phenomenon, meaning that there is no indication that either linguistic or extra-linguistic factors condition how native speakers mark

gender on modifiers. Consequently, within variationist SLA, variability in gender marking by learners of Spanish is a case of Type I variation.

Grammatical gender in Spanish has received much attention in SLA research (cf. Alarcón, 2014). While psycholinguistic studies on the processing of grammatical gender account for a large portion of this work (e.g., Halberstadt, Valdés Kroff, & Dussias, 2018; Sagarra & Herschensohn, 2012), we limit our review to investigations on production among adult learners, since this is the focus of the current study. Production studies include examinations of oral picture-description tasks (Alarcón, 2011; Montrul et al., 2008; White, Valenzuela, Kozlowska-MacGregor, & Leung, 2004), oral conversations (Finnemann, 1992; Franceschina, 2001a; 2001b), an oral elicited production task (Bruhn de Garavito & White, 2002; White et al., 2004), oral interviews (Fernández-García, 1999), a written elicited production task (Alarcón, 2010), and written compositions (Schlig, 2003).

Although some researchers have conducted mixed ANOVAs allowing for the simultaneous examination of multiple independent variables (Alarcón, 2010, 2011; Montrul et al., 2008), the tendency in this body of work has been to perform univariate error analyses. Collectively, previous studies have identified patterns of use according to a set of commonly studied factors: noun gender, noun ending, modifier type, and noun class. The most consistent finding is for noun gender, with learners showing higher accuracy rates with masculine compared to feminine nouns (e.g., Finnemann, 1992; Montrul et al., 2008; Schlig, 2003; White et al., 2004). Some studies have also found that learners exhibit fewer errors with nouns that have canonical -o/-a endings than other types of endings (Alarcón, 2011; Fernández-García, 1999). Regarding modifier type (also called domain), some investigations (e.g., Alarcón, 2010; Bruhn de Garavito, 2002; Fernández-García, 1999; White et al., 2004) have observed higher accuracy

rates with determiners (at times considered to reflect the gender assigned to the noun by the speaker) than with adjectives (at times considered to be indicative of a speaker's ability to make morphosyntactic agreement), while others show rates to be similar between determiners and adjectives (Alarcón, 2011; Montrul et al., 2008). Results for noun class have been less consistent, with production studies finding higher accuracy with semantic (or biological) gender than grammatical (or arbitrary) gender (Alarcón, 2010; Fernández-García, 1999) and others observing the opposite trend (Bruhn de Garavito & White, 2002). In addition, Finnemann (1992) showed that learners' gender marking is more accurate with singular nouns than plural ones. He also suggested that when there is distance between the noun and the modifier, the "demands on memory and processing capacity" increase, which could explain a decrease in accuracy (p. 129).

Despite a wealth of research on the expression of gender in additional-language Spanish, several questions remain. Although there is a general consensus regarding noun gender, the other factors reviewed in this section have shown differences across studies. Furthermore, certain potentially relevant factors have not yet been addressed. For example, despite the fact that research on gender-marking behavior in Spanish has examined various tasks, there has yet to be a systematic comparison of task type in production (but see Montrul et al., 2008, for a comparison of production and comprehension). We also note that prior investigations on development are limited. Bruhn de Garavito and White (2002) and Alarcón (2010) investigated learners at different instructional levels and Alarcón (2011) and Montrul et al. (2008) included proficiency as a variable in their analysis. Although these studies demonstrated that accuracy increased as learners became more proficient, SLA research on Spanish has yet to offer a detailed account of the developmental trajectory of gender-marking behavior in longitudinal data. Additional research, therefore, is warranted. Specifically, in the present investigation, we

offer an analysis of variability in gender marking in a longitudinal learner corpus. In so doing, we bring together the fields of variationist SLA and LCR: Specifically, we draw on strengths of variationist SLA in order to explain patterns of (Type I) variation in gender-marking behavior as they change across time in the learner corpus LANGSNAP, whereby also helping to advance LCR.

The Current Study

In order to demonstrate how establishing a dialogue between two areas of additional-language research (namely, LCR and variationist SLA) can contribute knowledge to the understanding of interlanguage variability and development, we conduct an investigation of variable gender marking in additional-language Spanish. Since linguistic phenomena that are frequent in language lend themselves well to a corpus-based study (Tracy-Ventura & Myles, 2015), gender marking is well suited for this pursuit given how often it occurs in language use. We seek to answer the following research question: How does the variability present in the marking of grammatical gender in a Spanish learner corpus change over time?

Method

Corpus. The data come from the LANGSNAP corpus (http://langsnap.soton.ac.uk/, e.g., Mitchell, Tracy-Ventura, & McManus, 2017). During the initial phase of the project (see Tracy-Ventura & Huensch, 2018, for details on subsequent phases of the project), data were collected six times over a period of 21 months. At each data-collection point, the participants completed two oral tasks, a semi-structured interview and a picture-based narration, and a written argumentative essay. For the current study, we have examined all tasks from three of the six data-collection points. Pre-stay was the first data-collection period in the corpus (May 2011). The data for the third in-stay visit (henceforth, in-stay) were gathered at the end of the academic

year abroad, which was a year after pre-stay (May 2012). The second post-stay (henceforth, post-stay) was the final data collection and took place about 21 months after pre-stay (January 2013).

We report on 21 of the 27 British learners of Spanish in the corpus who were enrolled in an undergraduate degree program in Spanish. Their average age was 20.8 years (*SD* = 1.6, range: 20-25). Fifteen were women and six were men. They had been studying Spanish between two and 14 years (*M* = 5.4, *SD* = 3.4). In terms of their first language, 19 participants listed English, one listed Polish, and one indicated that English and Polish were her first languages. Eighteen participants had studied other languages (French, German, and/or Italian); two had not studied other languages, and one did not provide information on this issue. During the period of data collection the participants spent an academic year in a target-language environment (16 in Spain and five in Mexico). While abroad, nine were exchange students, 10 worked as teaching assistants and two were workplace interns. At the onset of the project, the participants completed an elicited-imitation (EI) task for a measure of initial proficiency (see Bowden, 2016, for details). Their scores ranged from 50 to 108 out of a possible 120 points (*M* = 86.1, *SD* = 12.7).

Data Coding and Analysis

We began our coding by isolating as a separate token every instance of a determiner and/or adjective that modified a referent (*K* = 16,357).[7] The current dataset consists of modifiers that show overt gender marking and that modify a noun located in the immediate clause or one of the preceding 10 clauses (*k* = 11,846). Examples are provided in (1), with the noun in bold and modifiers underlined.

(1) *En vez de vivir en <u>un</u>*$_{masc}$ ***entorno***$_{masc}$ *que no sea <u>seguro</u>*$_{masc}$ (pre-stay, essay, participant 150) 'Instead of living in an$_{masc}$ environment$_{masc}$ that is not safe$_{masc}$.

Creo que no tuve <u>*muchas*</u>$_{\text{fem}}$ ***problemas***$_{\text{masc}}$ (post-stay, interview, participant 165).

'I believe that I don't have many$_{\text{fem}}$ problems$_{\text{masc}}$.'

The dependent variable was whether the gender marking of a modifier was targetlike (the same gender as the noun) or non-targetlike (a different gender from the noun). We chose targetlikeness as the dependent variable in order to build directly on previous literature on grammatical gender in additional-language Spanish that has focused on accuracy. We note that this decision departs from traditional variationist research in which the categories of the dependent variable are the forms or variants of a linguistic structure (see the previous discussion in the *Variationist SLA* subsection). Additionally, we coded for 11 independent, fixed-effects variables, provided in Table 1. All coding was done by hand using the publicly available transcripts.[8] It will be recalled that noun gender, noun ending, modifier type, noun class, and noun number have been shown to be important for gender marking in language use in previous, largely univariate, studies on additional-language Spanish.[9] Moreover, since there is evidence that general proficiency helps to account for targetlike gender marking (Alarcón, 2011; Montrul et al., 2008), we included the initial proficiency score obtained on the EI task for each participant as a fixed effect. It is worth noting that analyzing the EI score as a continuous variable for individuals, rather than grouping learners into dichotomous groups, responds to calls in SLA for reform in the ways that researchers examine proficiency (Bowden, 2016; Leal, 2018). We also explore five new variables: syllable distance, noun log-frequency (language), noun frequency (individual), task, and time. First, in light of Finnemann's (1992) suggestion that increased distance between a modifier and noun would lead to lower rates of targetlike behavior, we measure the distance between nouns and modifiers by counting the number of syllables between the two. For the second and third variables, we explore the role of noun frequency in the use of

gender marking. Previous research by Sabourin, Stowe, and de Haan (2006) has reported that additional-language learners of Dutch show higher rates of targetlike gender marking with higher frequency nouns on a grammaticality judgment task, suggesting that noun frequency may impact gender-marking behavior. However, there are different ways of operationalizing and measuring frequency (e.g., Linford, Long, Solon, & Geeslin, 2016); in the current project, we include two measures. First, the variable noun log-frequency (language) corresponds to the log of the frequency (per million words) with which each noun in our dataset occurred in Modern Spanish, as determined by the *Corpus del español* (Davies, 2016-). We analyzed the natural logarithm of noun frequency in order to account for the skew in the distribution of frequency scores. The second frequency variable – noun frequency (individual) – acknowledges that individual learners use specific nouns with varying frequency. Given evidence that "every usage event may slightly redefine a person's internal language system" (Tummers, Heylen, & Geeraerts, 2005, p. 228), we sought to include a measure that would capture how frequently each individual overtly marked gender with a given noun. To do so, we identified the full inventory of nouns each individual used with a modifier overtly marked for gender. Given that individual production may change both across time and across task, we then organized each inventory by task and time, calculating how frequently a given noun was used with a gender-marked modifier in each task at each time by each individual. For example, for the noun *cosa* 'thing,' we calculated how often each individual used this noun with a gender-marked modifier in each of the three tasks at each of the three data-collection points, giving us the potential of a total of nine different scores for this noun for each participant. The fourth new variable was task. Since there is evidence of variability in language behavior across tasks due to characteristics such as planning time (Ellis & Yuan, 2004) and discourse topic (Medina-Rivera, 1999), we examined possible differences among the written

essay, oral narrative, and oral interview. The final fixed effect was time; we analyzed pre-stay, in-stay, and post-stay data together in one statistical model. Lastly, we included one random intercept for participant.[10] This factor enabled us to account for variability at the level of the learner while simultaneously examining variability in aggregate data (see Meunier, 2015, for a discussion for the ability to analyze group and individual variability in a single statistical model).

Table 1.

Independent fixed-effects variables

Variable	Categories or description
Noun gender	feminine, masculine
Noun ending	canonical -o/-a, predictive endings (e.g., *-e/-ción*), deceptive -a/-o, other
Syllable distance	# of syllables between the noun and modifier [continuous]
Modifier type	adjective, determiner
Noun class	grammatical, semantic
Noun frequency (individual)	# of times each participant used a noun in a given task and a given data-collection point with a gender-marked modifier [continuous]
Noun log-frequency (language)	log of the # of occurrences per 1 million words of a noun [continuous], determined by the *Corpus del español* (Davies, 2016-)
Noun number	singular, plural
Task	interview, narration, essay
Initial proficiency	initial proficiency of each participant measure by an EI test

	[continuous]
Time	pre-stay, in-stay, post-stay

We began the analysis by identifying the average targetlike rates of gender marking for the group and for individuals at each data-collection point. These data are presented using boxplots with an overlay of individual data points (see Larson-Hall, 2017, for a discussion of the utility of both data-rich and data-accountable graphics in SLA). We then analyzed the data using the statistical softwares R and SAS.[11] First, with R, we used chi-square methods to see whether any of the independent variables were highly correlated. Then we explored the potential importance fixed effects described in Table 1 using bootstrapping, which allowed us to narrow down the set of fixed effects we examined further in this study (see Mannan, 2017, for an explanation of this technique; the code we used in R for chi-square methods and bootstrapping is provided in the Online Supporting Documentation). Subsequently, we ran one generalized linear mixed model (see the Online Supporting Documentation for the code we used in SAS). This statistical test analyzed multiple factors simultaneously in the data and determined what factors influenced gender-marking behavior. For the dependent variable and the nominal fixed effects listed in Table 1, the categories of a variable were compared to a reference point. For the dependent variable, targetlike behavior was the reference point. The reference points for the nominal fixed effects were as follows: masculine (noun gender), canonical -o/-a endings (noun ending), determiner (modifier type), grammatical (noun class), singular (noun number), essay (task), and pre-stay (time). Syllable distance, noun frequency (individual), noun log-frequency (language), and initial proficiency are continuous factors so there was no reference point. After identifying the fixed effects that significantly predicted targetlikeness, we explored interactions

between time and the other significant fixed effects. We used a top-down (also called backward selection) strategy to incrementally remove non-significant and spurious interactions. This approach (i.e., coupling bootstrapping and a top-down strategy) ensures that we do not overfit the model with spurious effects. Because the bootstrapping step considers many distinct datasets, we avoid including variables that may have shown misleading significance from the full dataset. Similarly, by removing interactions incrementally, we avoid the inclusion of interactions that might be highly correlated with one another and are thus inappropriate for consideration in the model. After fitting the model, we once again examined correlations between fixed-effect estimates – this time with those that reached significance only – in order to ensure there were no strongly correlated covariates. We also examined multiple comparisons among levels of the interaction and the nominal factors with more than two categories (time, noun ending, and task) using the Tukey-Kramer method, which adjusts confidence interval widths appropriately to reduce Type-I error rate when making multiple comparisons. Furthermore, we identified the percentage of correctly predicted observations, the Somers' D, the Index of Concordance C, McFadden's R^2, and the Bayesian Information Criterion (BIC) for the model – five metrics that show whether the model does a good job of fitting the data.[12] In general terms, this analysis allows us to model the variability present in learners' gender-marking behavior and to determine whether this variable use changes over time.

We see several advantages to this analysis. A mixed-effects model is a regression analysis that allows for the simultaneous examination of several independent variables (cf. Gries, 2015a, 2015b, Gries & Deshors, 2014).[13] Although a mixed ANOVA also has the ability to consider multiple independent variables in a single test (see Alarcón, 2010, 2011; Montrul et al., 2008 for analyses of gender marking in Spanish having used mixed ANOVAs), a regression has

at least two advantages over a mixed ANOVA. Foremost, a dependent variable in an ANOVA represents a mean value (e.g., average rate of targetlikeness), which leads to data reduction because it involves transforming multiple observations into a single response. By reducing observations (i.e., instances of gender marking), we lose information about the variability in use of the dependent variable and among participants. We believe this level of detail in the analysis is crucial for a complete understanding of interlanguage variability and additional-language development. Second, regressions allow for both continuous and categorical independent variables, whereas ANOVAs can include only categorical ones. Other strengths of this type of analysis are that it permits us not only to uncover the systematicity in the variable use of gender marking but also to explain these variable patterns probabilistically. Furthermore, although we report the findings of the fixed effects in aggregate terms, the statistical model contains a random effect for participant, which enables us to account for variability among individual learners. Finally, to our knowledge, the current study is the first to conduct a variationist analysis of Type I variation within LCR.

Results

We begin by presenting the average rate of targetlike use for the individual participants and for the group at pre-stay, in-stay, and post-stay. In Figure 1, the y-axis corresponds to the percentage of targetlike use. Each dot represents the average targetlike rate of use for one participant at one of the data-collection points, whereas the group means are shown with diamonds, and the standard deviations with arrows. The median is indicated by the bold, horizontal line. We see that the learners exhibited high rates of targetlike gender marking at each data-collection point, with the group averages staying over 90 percent. Targetlikeness increased from pre-stay to in-stay and this gain was maintained at post-stay. Moreover, the standard

deviation for individuals' rate of targetlikeness decreased between pre-stay and in-stay, suggesting that there was less variability among the learners after an academic year abroad.

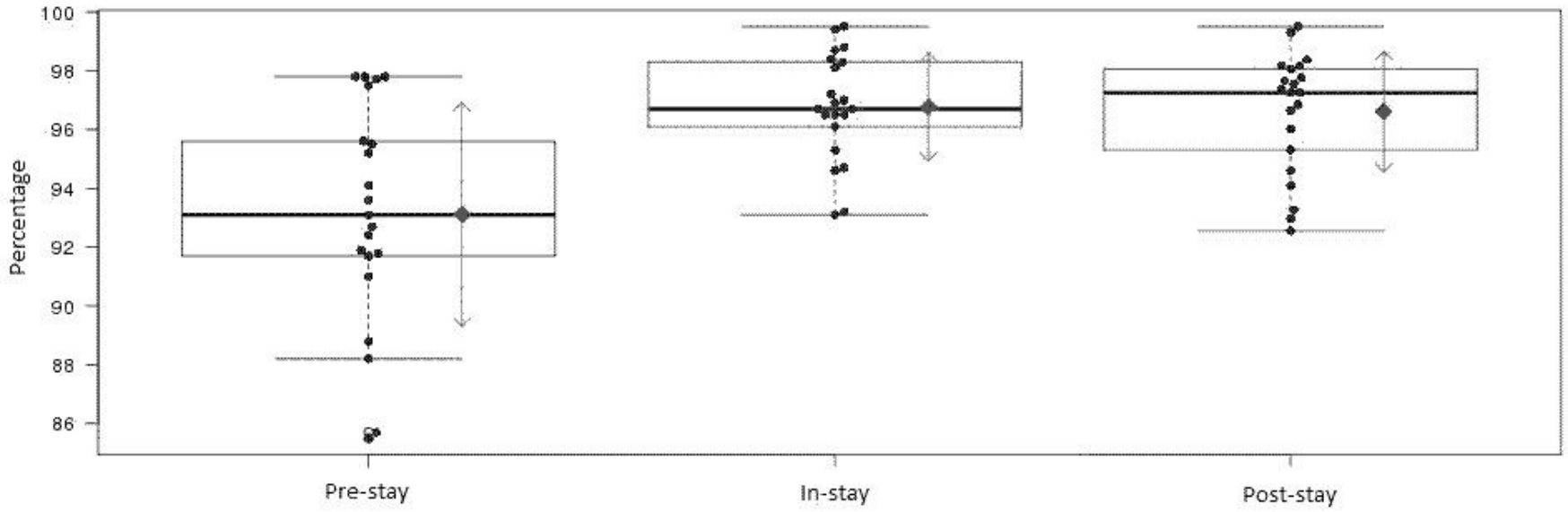

Figure 1. Targetlike rates of gender marking over time for individuals and the group

Next, we present the findings from the generalized linear mixed model. We note that chi-square methods did not show any strong correlations between the fixed effects, but two of the factors provided in Table 1 were not selected as important during the bootstrapping phase: noun log-frequency (language) and noun number. Moreover, although noun class appeared to be important during bootstrapping, it was not significant when we conducted the mixed-effects model. Following this observation, we removed noun class from the model and re-ran it.

The fixed effects included in the final model are available in Table 2 and the details of the random effect are in Table 3. Regarding the latter, the random effect is not tested for significance; rather, it represents each participant's personal shift in targetlike log odds if she/he is viewed as coming from a larger population of learners. The parameter estimate provided in each table demonstrates whether there was an increase (signified by a positive parameter estimate) or decrease (designated by a negative parameter estimate) in the log odds of targetlike use compared to the reference point. The *p*-value shows whether a parameter estimate was significant; we let $\alpha = 0.05$ and determined significance when $p < \alpha$.

Beginning with the fixed effects, eight main factors and one interaction significantly predicted targetlike gender marking (see the Online Supporting Documentation for the frequency distribution of the nominal fixed effects). For noun ending, the log odds of targetlike use was significantly lower with deceptive noun endings, compared to those ending in canonical -o/-a; the comparisons between canonical -o/-a nouns and those with predictive endings or other endings did not reach significance. For those nominal fixed effects (such as noun ending) that have more than two categories, the model results provide information as to the potential differences among the non-reference categories. In other words, it is possible to determine whether participants behaved differently on deceptive, predictive and other endings (without

reference to canonical -o/-a endings). To do so, we examined the confidence intervals. Overlap between the confidence intervals on non-reference point categories indicates that two categories have statistically similar log odds of targetlikeness. In other words, behavior on those two categories would not be considered different. This is precisely what we see when we compare other noun endings to predictive noun endings. However, the confidence interval of deceptive noun endings was lower than the intervals for other and predictive endings. Since there was no overlap in the values, we note that the log odds of targetlike use were lower with deceptive noun endings, when compared to other and predictive endings, meaning that these learners were less likely to be targetlike with deceptive noun endings versus all other noun ending categories.

For the factor task, which had three categories, these learners were significantly less targetlike on both oral tasks than on the written essay. By examining the confidence intervals of the narrative and interview tasks, we see overlap in the values. This overlap indicates that the log odds of targetlike use were similar for the two oral tasks. The variable of noun gender was significant, revealing that this participant group exhibited higher log odds of targetlikeness with masculine, compared to feminine nouns. The next finding concerns the variable noun frequency (individual). It will be recalled that this factor examines the frequency with which each participant used a given noun with a modifier that was overtly marked for gender in each task and at each data-collection point. The significance of noun frequency (individual) showed that learners' log odds of targetlike gender marking significantly increased as their frequency of use of individual nouns increased. For syllable distance, learners were found to be less likely to produce targetlike gender marking as the distance between the noun and modifier increased. The factor of modifier type showed that the participants were significantly more targetlike with determiners than adjectives, whereas the variable of initial proficiency found that targetlike

gender marking increased significantly as learners' initial proficiency score increased. The final significant fixed effect was time: This participant group was significantly more targetlike at in-stay and post-stay than at pre-stay. Because there is overlap in the confidence intervals of in-stay and post-stay, we are able to see that the probability of targetlike use was similar at these two data-collection points.

Finally, the interaction between noun ending and time was the sole significant interaction in the model. This interaction revealed that deceptively marked nouns at post-stay were significantly more probable to be targetlike than canonical -o/-a endings at pre-stay. One other comparison approached significance: the greater log odds of targetlikeness with deceptively marked nouns at in-stay compared to nouns with canonical -o/-a endings at pre-stay. The other effects for this interaction were not significant. Furthermore, there is overlap among all of the confidence intervals, which indicates that the remaining categories are not significantly different from each other.

Table 2.

Results for the fixed effects in the regression model

Effect	Estimate	SE	Df	t-value	*p*-value	Confidence intervals	
						Lower	Upper
Intercept	2.3947	0.6718	19	3.56	0.0021	0.9887	3.8008
Noun ending [canonical -o/-a]							
deceptive	-2.5867	0.2408	60	-10.74	<.0001	-3.0684	-2.1050

Table 2.

Results for the fixed effects in the regression model

Effect	Estimate	SE	Df	t-value	*p*-value	Confidence intervals	
						Lower	Upper
other	-0.2542	0.1755	60	-1.45	0.1528	-0.6052	0.0969
predictive	-0.1201	0.1872	60	-0.64	0.5236	-0.4946	0.2544
Task [essay]							
narrative	-0.6699	0.1696	40	-3.95	0.0003	-1.0126	-0.3273
interview	-0.5459	0.1372	40	-3.98	0.0003	-0.8232	-0.2687
Noun gender [masculine]							
feminine	-1.3597	0.1101	20	-12.35	<.0001	-1.5894	-1.1300
Noun freq. (individual)	0.1069	0.0240	11808	4.45	<.0001	0.0599	0.1539
Syllable number	-0.01963	0.0053	11808	-3.72	0.0002	-0.0300	-0.0093
Modifier type [determiner]							
adjective	-1.1372	0.0958	20	-11.87	<.0001	-1.3370	-0.9375
Initial prof.	0.02184	0.0075	19	2.90	0.0092	0.0061	0.0376
Time [pre-stay]							

Table 2.

Results for the fixed effects in the regression model

Effect	Estimate	SE	Df	t-value	*p*-value	Confidence intervals	
						Lower	Upper
in-stay	0.9032	0.1494	40	6.05	<.0001	0.6013	1.2052
post-stay	0.7887	0.1721	40	4.58	<.0001	0.4409	1.1366
Noun ending*time [canonical -o/-a*pre-stay]							
decep.*post-stay	0.9953	0.4789	119	2.08	0.0398	0.04705	1.9436
decep.*in-stay	0.6821	0.3533	119	1.93	0.0559	-0.0176	1.3817
other*post-stay	-0.4937	0.3013	119	-1.64	0.1039	-1.0902	0.1028
other*in-stay	-0.2329	0.2615	119	-0.89	0.3749	-0.7506	0.2848
predic.*post-stay	0.03989	0.3536	119	0.11	0.9104	-0.6603	0.7400
predic.*in-stay	-0.3118	0.3099	119	-1.01	0.3164	-0.9254	0.3018

Note. The reference point for the dependent variable is targetlike use. The reference points for the nominal fixed effects are in brackets.

Table 3. *Results for the random effect in the*

regression model

Effect	Participant	Estimate	SE
Intercept	150	0.4131	0.2065
Intercept	151	0.0667	0.2039
Intercept	152	-0.0008	0.2080
Intercept	155	0.0702	0.2505
Intercept	156	-0.2561	0.1889
Intercept	157	-0.2750	0.2134
Intercept	158	0.4474	0.2500
Intercept	160	0.0709	0.1976
Intercept	161	0.2281	0.2212
Intercept	162	0.5575	0.2556
Intercept	163	-0.3173	0.1967
Intercept	164	0.04519	0.2093
Intercept	165	-0.2031	0.1871
Intercept	166	-0.0451	0.1830
Intercept	167	-0.8089	0.1717
Intercept	168	0.2568	0.2431
Intercept	169	-0.0066	0.1977
Intercept	170	0.1143	0.2104
Intercept	171	-0.1036	0.2350
Intercept	172	-0.3523	0.1770
Intercept	173	-0.1577	0.2228

In terms of model assessment, this model predicted targetlike behavior with 95 percent accuracy. Somers' D, which measured the agreement between probability and targetlike behavior, was 0.5977. For this metric, 0 would indicate no agreement between targetlike behavior and fitted probability and 1 would indicate perfect agreement. The Index of Concordance C, which quantified how well the model discriminates between targetlike and non-targetlike behavior, was 0.7989. For this metric, 0 would indicate that the model does not differentiate between targetlike and non-targetlike behavior at all and 1 would indicate that it does so perfectly. The McFadden's R^2 measurement is used rather than the more commonly used R^2 because the usual R^2 is not mathematically appropriate to use to describe a model with a binary response. However, it is measured on a similar scale (from 0 to 1) as R^2; we obtained $R^2_{McFadden} = 0.1384$, indicating a moderate to good fit (see Smith & McKenna, 2013, for information on this metric). The BIC compared our model with a null model based on their log-likelihood (a measure of the probability of observing the data in the model). In order to avoid overfitting, it penalizes models with potentially extraneous parameters. With the BIC, a difference between the two models that is greater than 10 is considered to be very strong evidence against the model with the higher BIC (Kass & Raftery, 1995). Our model has a better (lower) BIC than the null model (i.e., a model with only the dependent variable).:The BIC for our model is 3838.07 and the BIC for the null model is 4393.53. Even accounting for the penalty of including eight fixed effects and a random effect for each participant,[14] the model we report in the current study has a higher probability of being true compared to a null model. Therefore, these five metrics collectively provide evidence to support the observation that the mixed-effects model does a good job of fitting targetlike gender marking and, moreover, the BIC indicates that the resulting model avoids overfitting with potentially extraneous factors.

Discussion

We start this section by answering our research question: How does the variability present in the marking of grammatical gender in a Spanish learner corpus change over time? We then reflect on contributions that the current study has made to additional-language research more generally. In order to respond to the research question, we conducted a variationist investigation of gender marking in Spanish, an example of Type I variation. This approach permitted us to uncover new knowledge about gender-marking behavior, specifically with regard to variability in use and how this variability evolves. We highlight here three main observations that have come out of this investigation and that provide a concrete illustration of the insights that can emerge when variationist SLA and LCR are brought together. First, while this participant group exhibited high rates of targetlike use, the findings from the mixed-effects model demonstrated that the variability in their gender-marking behavior is complex. Regarding the fixed effects, six linguistic factors (noun gender, noun ending, syllable distance, modifier type, noun frequency (individual), and initial proficiency) and two extra-linguistic factors (task, time) worked together to impact targetlike use. The interaction between time and noun ending was also significant. In other words, this regression enabled us to see that multiple factors simultaneously conditioned learners' variable use of gender marking, despite a targetlike rate of use of more than 90 percent for this group of learners. Moreover, we note that some prior studies have indicated that only targetlike rates of use that were lower than 90 percent may be considered to show "evidence of some sort of deficit or failure with gender assignment and agreement" (Montrul et al., 2008, p. 536). Whereas this cutoff seems to suggest that the rates of targetlike use observed in the present study may be less worthy of investigation, we argue that the complexity of the various factors influencing learners' gender marking offers evidence to the

contrary. In other words, targetlike rates of use – including those that are close to 100 percent – are limited in what they reveal about interlanguage development and variability, because they do not enable researchers to see intricate interlanguage patterns (cf. Berdan, 1996). This suggests that scholars would do well to investigate high levels of targetlikeness, as multivariate analyses of such data can reveal important findings about late points along the developmental trajectory for a particular linguistic phenomenon (for another example, see Donaldson's (2017) analysis of verbal negation in French, an example of Type II variation, among near-native speakers).

Second, in addition to investigating independent variables that have been shown to be important for gender marking in previous research, the present investigation included various factors that had not been examined previously in research on language use and gender marking in Spanish. Beginning with factors that have been studied beforehand, our results for noun gender (Montrul et al., 2008; White et al., 2004) and modifier type (Alarcón, 2010; Bruhn de Garavito & White, 2002) were consistent with other scholars' findings. However, while the general observation in previous research is that the learners are more targetlike with nouns that have canonical -o/-a endings (e.g., Fernández-García, 1999), our results for the main effect for noun ending indicated that canonical -o/-a endings did not lead to more targetlike gender marking, when compared to all types of noun endings. Instead, learners were only found to be more targetlike on nouns with canonical -o/-a endings, as compared to nouns with deceptive endings. At the same time, learners were found to be significantly less targetlike on nouns with deceptive endings, as compared to nouns with canonical -o/-a, predictive, and other endings. Our findings for noun class differed from those in previous research as well (e.g., Alarcón, 2010; Bruhn de Garavito & White, 2002) in that we did not find this factor to be influential. Also counter to Finnemann (1992), noun number did not condition use for the participant group under

investigation. Furthermore, the decision to analyze the data using a regression allowed for the inclusion of continuous independent variables. Our results for initial proficiency support previous work that examined learners in discrete proficiency groups (Alarcón, 2011; Montrul et al., 2008), such that targetlike gender marking increases as general proficiency improves. The findings for syllable distance offered support for Finnemann's (1992) proposal that rates of targetlike use would diminish as distance between the noun and modifier increased. Concerning frequency, this is, to the best of our knowledge, the first investigation of the production grammatical gender marking in additional-language Spanish that investigates noun frequency, and we did so in two ways. Whereas noun log-frequency (language) did not appear to impact use and was removed after the bootstrapping step, noun frequency (individual) impacted use, indicating that the more often a participant used a noun with an overtly marked modifier, the higher the log odds of targetlikeness. This finding is consistent with the idea that each use of a given structure can be expected to influence a participant's interlanguage grammar (cf. Tummers et al., 2005). The observations for both noun frequency variables highlight the complexity of operationalizing the concept of frequency. Indeed, in the future, it may be worth considering other ways of measuring noun frequency. With respect to frequency in the input, Ellis, Römer, and O'Donnell (2016), for example, suggest that contingency between cue and interpretation is more important than absolute frequency. For the factor of time, we found that learners were more likely to use targetlike gender marking at in-stay and post-stay, compared to pre-stay. The last new variable that we introduced was task. Results for this factor showed that there was variability in targetlike use between the written essay on the one hand and the interview and narrative tasks on the other hand, such that the log odds of targetlikeness were higher on the essay than on the oral tasks.

Third, to our knowledge, the current study is the first to examine gender-marking production in additional-language Spanish using longitudinal data, leading to new insights about the developmental trajectory. The findings indicated that over the 21 months analyzed there is evidence of change as well as stability in interlanguage variability. Beginning with change, the rate of targetlike use increased from pre-stay to in-stay and the variability in targetlike frequency among individuals decreases, as evidenced by the standard deviation for mean rate of targetlikeness. As pointed out by Meunier (2015, p. 382), the possibility of charting individuals and variation within groups across time is one of the advantages of longitudinal corpora. We also observed change with one fixed effect, as evidenced by a significant interaction. The results for this interaction demonstrated that learners made significant gains in their targetlike gender marking with deceptively marked nouns (compared to canonical -o/-a endings) from pre-stay to post-stay. In other words, significant improvement was seen across time, specifically with nouns that ended with a deceptive -a (e.g., *problema*$_{\text{masc}}$ 'problem') or -o (e.g., *mano*$_{\text{fem}}$ 'hand'). We also observed evidence for stability in gender-marking behavior over time. Six of the fixed effects (task, noun gender, noun frequency (individual), syllable number, modifier type, and initial proficiency) did not interact with time, suggesting that the impact that these factors have on targetlike gender marking was stable over the 21-month period we investigated. Across time, this participant group was more likely to be targetlike in the essay, with masculine nouns, and with determiners. Their log odds of targetlikeness also increased the closer the modifier was to the noun, the higher their initial proficiency was, and the more often they produced individual nouns with overtly marked modifiers. Moreover, the gains made during the year abroad were maintained eight months later, after the participants' return to the United Kingdom.

In sum, the multifactorial analysis conducted in the present investigation offers detailed evidence of both the systematicity and complexity of interlanguage variability as it pertains to Type I variation and the morphosyntactic phenomenon of gender marking. What is more, although variationist SLA distinguishes between Type I and Type II variation, based on the whether the variation also appears to be found in NS use, the present investigation indicates a commonality between the two. Specifically, just as with sociolinguistically variable structures (Type II variation), learners vary their use of categorical phenomena like gender marking (Type I variation) according to a host of linguistic and extra-linguistic factors. We believe this offers a concrete example of how the variationist paradigm can reveal intricate details about variability and its evolution, even for presumably categorical structures.

In addition to new insights about variable gender marking in additional-language Spanish, the current study has offered a concrete example of how two areas of additional-language research – LCR and variationist SLA – can be fruitfully brought together in the analysis of interlanguage. More specifically, the bringing together of LCR and variationist SLA has allowed us to address three issues in need of further investigation and reform in LCR: a lack of attention to variation, a dearth of research on interlanguage development, and an overreliance on univariate analyses. In addressing each of these issues, we also contribute to current trends in SLA more generally. First, the current analysis on variability in gender-marking behavior has contributed to illuminating an understudied issue in LCR, namely interlanguage variation (Ädel, 2015). As mentioned earlier, Ortega (2015, p. 251) has identified systematicity and variability as two pillars of interlanguage analysis. Insofar as variationism places variation at the heart of its analyses, we believe that it provides a well-equipped conceptual and methodological toolbox for the study of variation in learner corpora. Furthermore, we have helped to move LCR forward by

conducting an analysis of interlanguage development, rather than an examination of learners at a single point in time (cf. Meunier & Littré, 2013; Myles, 2005). The need for analyses of longitudinal datasets is not restricted to LCR, having also been a need identified in SLA more generally. As stated by Ortega and Iberri-Shea (2005, p. 26), "it can be argued that many, if not all, fundamental problems about L2 learning that SLA researchers investigate are in part problems about 'time', and that any claims about 'learning' (or development, progress, improvement, change, gains, and so on) can be most meaningfully interpreted only within a full longitudinal perspective." Longitudinal corpora, such as LANGSNAP, allow researchers within LCR and SLA to adopt this perspective. Finally, using variationism to analyze learner corpora has allowed us to heed the call to incorporate statistical analyses that offer greater explanatory power than those traditionally employed in LCR (cf. Gries, 2015b; Myles, 2005). As previously discussed, the regressions that are typical of variationism allow for investigations of a complex array of linguistic and extra-linguistic factors in a single model. In the case of the current analysis, we opted for a mixed-effects model, which enabled us to characterize the intricate nature of the variability found in the use of gender marking. It also showed how this variation can be explained in aggregate terms by linguistic and extra-linguistic factors, while taking into consideration individual variability by including participant as a random effect. This push for methodological reform has parallels in SLA, as Plonsky and Oswald (2017) have also advocated for multivariate statistical analyses, in part because they more adequately capture the complexity of interlanguage than univariate ones.

The current study, thus, constitutes an example of an attempt to build bridges between different areas of research. We are, of course, not alone in such an endeavor. As mentioned in the introduction, Myles (2015, p. 309) has advocated for "bidirectional moves (more LCR in SLA

and more SLA theory in LCR).” In a similar vein, scholars have encouraged dialogue among different lines of inquiry in SLA. Mitchell, Myles, and Marsden (2013) argue

> [w]hile we believe these different research strands within SLA will retain their autonomy and individual impetus, it is however clear that continuing attempts to cross-refer between them […] will continue to prove a productive way of developing our understanding of the specific modular domains and how they interlink. (p. 288)

We see evidence that scholars are moving in this direction and believe there is value in continuing to do so. For instance, in an empirical investigation, Kanwit (2017) draws connections between concept-oriented approaches and variationism in order to offer new insights on the acquisition of future-time reference in Spanish (see Edmonds, Gudmestad, & Donaldson, 2017, for a related discussion). In another example, researchers have brought together psycholinguistics and variationism to further knowledge on language processing (see Boland, Kaan, Valdés Kroff & Wulff, 2016, for an introduction to a special issue on this topic). And, in a more global treatment of this issue, Ortega (2015) reviews several theories and approaches and identifies specific connections among them. Given how diverse additional-language research is, establishing conceptual and methodological links among different areas of scholarship has the potential to strengthen the theoretical understanding of additional-language learning. With this in mind, it is worth noting that we also see the bringing together of variationist SLA and LCR as advantageous for the former, since variationist SLA has yet to be extended to publicly available corpora. A prominent strength of LCR is that data are made accessible to additional-language researchers working within distinct theories and approaches (Tracy-Ventura & Huensch, 2018). By allowing scholars from different research strands to work on the same data, we believe that publicly available corpora may be a tool that can foster greater interaction among variationists

and other researchers and that may lead to scholars being able to establish important links among various areas of research.

Conclusion

In the current study we established a dialogue between variationist SLA and LCR and sought to demonstrate how traits of variationist SLA could contribute to LCR. We exemplified this type of research by investigating variability in the development of gender marking in additional-language Spanish. Using a generalized linear mixed model to analyze longitudinal data, we found that several factors conditioned learners' variable use over time. The analysis also offered evidence of stability and change in gender-marking variability. While the present investigation has contributed new insights regarding the developmental trajectory of gender marking in additional-language Spanish, future research on learners over a longer period of time is necessary, in order to capture points along the acquisitional path that are outside of the period examined here. Further scholarship is also needed on the role of noun frequency in gender-marking variability and development and on other factors that can help us to better understand individual variability. Moreover, we join with other researchers who advocate for continued interaction among various areas of research (e.g., Mitchell et al., 2013). By showing how a variationist study in LCR can advance knowledge of interlanguage development and variability with respect to a specific linguistic phenomenon, we hope to have offered a clear example of the gains that can be made in the field by initiating an exchange between different strands of research.

Notes

1. A distinction exists in the literature whereby LCR and SLA are two separate but related fields (e.g., Myles, 2015). We maintain this distinction in the current paper and use the term *additional-language research* as an umbrella term for both.
2. In the present investigation we use *variability* and *variation* interchangeably, though we recognize that some research differentiates between the two terms.
3. Berdan (1996) performed four regression models for the same dataset and concluded that the one we showcase here is the best fit for the data.
4. We recognize that other approaches may also help LCR to address these issues. In the current study, we offer variationist SLA as one possible example of an area of research that could contribute to LCR.
5. We note LCR and variationist SLA are alike in that their origins lie with work on native speakers in the areas of corpus linguistics and variationist sociolinguistics, respectively. Szmrecsanyi (2017) has made connections between corpus linguistics and variationist sociolinguistics, primarily focusing on native-speaker research.
6. These studies report very few instances of mismatches between the gender of the noun and the modifier.
7. In addition to the learner data, LANGSNAP provides data from 10 native speakers of Spanish. We examined their gender-marking behavior on the same tasks we analyzed for the learners and found one instance of non-targetlike use between a determiner and a noun (*una*$_{\text{fem}}$ *pan*$_{\text{masc}}$ ‘a$_{\text{fem}}$ bread$_{\text{masc}}$’, interview, participant 187).
8. Due to how labor intensive this coding was, we coded a subset of the participants and data-collection periods in the corpus. The possibility of developing (semi-)automatic coding of the factors under investigation is an avenue for future research.

9. The endings for the predictive and other categories of the noun ending factor were based on Teschner and Russell (1984). The determiner category consists of definite articles, indefinite articles, demonstrative determiners, indefinite determiners, and possessive determiners.
10. We also considered including a random effect for lexical item in the analysis. However, because our dataset consists of 1223 different lexical items, 494 of which only occur once in the dataset, including a random effect for lexical item was not feasible. Bolker, Brooks, Clark, Geange, Poulsen, Stevens, and White (2009, p. 128) explain that computing an integral for all individual random effects becomes "computationally infeasible" when there is a large number of random effects.
11. While we had intended to use R for the entire analysis, we switched to SAS for the regression model for two reasons. By using SAS for the generalized linear model, we were able to explicitly specify a covariance structure regarding the fixed effects that accounts for the fact that observations are correlated both between and within participants. Additionally, we can explicitly specify that individual participants are correlated with one another.
12. The percentage of correctly prediction observations was calculated in Excel by comparing the fitted output to the observed behavior. Somers' D, the Index of Concordance C, and the BIC are all automatically part of the model output in SAS. We used R for the McFadden's R^2 (see the Online Supporting Documentation for the code).
13. The most common type of regression that has been used in SLA variationist research is a logistic regression with fixed effects. More recently, mixed-effects models with fixed effects and a random effect for participant have become more commonplace.

14. The 30 parameters come from 21 random effects (one for each participant), eight categories for the main fixed effects, and one interaction effect.

References

Ädel, A. (2015). Variability in learner corpora. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 401-421). Cambridge: Cambridge University Press. https://doi.org/10.1017/cbo9781139649414.018

Alarcón, I. (2010). Gender assignment and agreement in L2 Spanish: The effects of morphological marking, animacy, and gender. *Studies in Hispanic and Lusophone Linguistics, 3*(2), 267-299. https://doi.org/10.1515/shll-2010-1076

Alarcón, I. (2011). Spanish gender agreement under complete and incomplete acquisition: Early and late bilinguals' linguistic behavior within the noun phrase. *Bilingualism, Language and Cognition, 14*(3), 332-350. https://doi.org/10.1017/s1366728910000222

Alarcón, I. (2014). Grammatical gender in second language Spanish. In K. L. Geeslin (Ed.), *The handbook of Spanish second language acquisition* (pp. 202-218). Malden, MA: Wiley Blackwell. https://doi.org/10.1002/9781118584347.ch12

Alonso-Ramos, M. (Ed.). (2016). *Spanish learner corpus research*. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.78.01alo

Bayley, R., & Langman, J. (2004). Variation in the group and the individual: Evidence from second language acquisition. *International Review of Applied Linguistics, 42*(4)*,* 303-318. https://doi.org/10.1515/iral.2004.42.4.303

Berdan, R. (1996). Disentangling language acquisition from language variation. In R. Bayley & D. R. Preston (Eds.), *Second language acquisition and linguistic variation* (pp. 203-244). Amsterdam: John Benjamins. https://doi.org/10.1075/sibil.10.09ber

Boland, J., Kaan, E., Valdés Kroff, J., & Wulff, S. (2016). Psycholinguistics and variation in language processing. *Linguistics Vanguard, 2*, 1-10. https://doi.org/10.1515/lingvan-2016-0064

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology and Evolution, 24*(3), 127–135. https://doi:10.1016/j.tree.2008.10.008

Bowden, H. W. (2016). Assessing second language oral proficiency for research. *Studies in Second Language Acquisition, 38*(4)*,* 647-675. https://doi.org/10.1017/S0272263115000443

Bruhn de Garavito, J., & White, L. (2002). The second language acquisition of Spanish DPs: The status of grammatical features. In A. T. Pérez-Leroux & J. M. Liceras (Eds.), *The acquisition of Spanish morphosyntax: The L1/L2 connection* (pp. 143-160). New York: Bilingual Press. https://doi.org/10.1007/978-94-010-0291-2_6

Callies, M. (2015). *Learner corpus methodology*. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 35-55). Cambridge: Cambridge University Press. https://doi.org/10.1017/cbo9781139649414.003

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

Donaldson, B. (2017). Negation in near-native French: Variation and sociolinguistic competence. *Language Learning, 67*(1), 141-170. https://doi.org/10.1111/lang.12201

Edmonds, A., Gudmestad, A., & Donaldson, B. (2017). A concept-oriented analysis of future-time reference in native and near-native Hexagonal French. *Journal of French Language Studies, 27*(3), 381-404. https://doi.org/10.1017/s0959269516000259

Ellis, N. C., Römer, U., & O'Donnell, M. B. (2016). *Usage-based approaches to language acquisition and processing: Cognitive and corpus investigations of construction grammar.* Malden, MA: Wiley.

Ellis, R. (2008). *The study of second language acquisition* (2nd ed). Oxford: Oxford University Press.

Ellis, R., & Yuan, F. (2004). The effects of planning on fluency, complexity, and accuracy in second language narrative writing. *Studies in Second Language Acquisition, 26*(1), 59-84. https://doi.org/10.1017/s0272263104026130

Fernández-García, M. (1999). Patterns of gender agreement in the speech of second language learners. In J. Gutiérrez-Rexach & F. Martínez-Gil (Eds.), *Advances in Hispanic linguistics: Papers from the 2nd Hispanic Linguistics Symposium* (pp. 3-15). Somerville, MA: Cascadilla Press.

Finnemann, M. D. (1992). Learning agreement in the noun phrase: The strategies of three first-year Spanish students. *International review of Applied Linguistics in Teaching, 30*(2), 121-136. https://doi.org/10.1515/iral.1992.30.2.121

Franceschina, F. (2001a). Against an L2 morphological deficit as an explanation for the differences between native and non-native grammars. *EUROSLA yearbook, 1*, 143-158. https://doi.org/10.1075/eurosla.1.12fra

Franceschina, F. (2001b). Morphological or syntactic deficits in near-native speakers? An assessment of some current proposals. *Second Language Research, 17*(3), 213-247. https://doi.org/10.1191/026765801680191497

Geeslin, K. L. with Long, A. Y. (2014). *Sociolinguistics and second language acquisition: Learning to use language in context*. New York: Routledge. https://doi.org/10.4324/9780203117835

Granger, S., Gilquin, G., & Meunier, F. (2015). Introduction: Learner corpus research – past, present and future. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 1-5). Cambridge: Cambridge University Press. https://doi.org/10.1017/cbo9781139649414.001

Gries, St. Th. (2015a). The most under-used statistical method in corpus linguistics: multi-level (and mixed-effects) models. *Corpora, 10*(1), 95-125. https://doi.org/10.3366/cor.2015.0068

Gries, St. Th. (2015b). Statistics for learner corpus research. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 159-181). Cambridge: Cambridge University Press. https://doi.org/10.1017/cbo9781139649414.008

Gries, St. Th. & Deshors, S. D. (2014). Using regressions to explore deviations between corpus data and a standard/target: Two suggestions. *Corpora*, *9*(1), 109–136. doi:10.3366/cor.2014.0053.

Halberstadt, L., Valdés Kroff, J. R., & Dussias, P. E. (2018). Grammatical gender processing in L2 speakers of Spanish: The role of cognate status and gender transparency. *Journal of Second Language Studies, 1*(1), 5-30. https://doi.org/10.1075/jsls.17023.hal

Huensch, A., & Tracy-Ventura, N. (2017). L2 utterance fluency development before, during, and after residence abroad: A multidimensional investigation. *The Modern Language Journal, 101*(2), 275-293. https://doi.org/10.1111/modl.12395

Kanwit, M. (2017). What we gain by combining variationist and concept-oriented approaches: The case of acquiring Spanish future-time expression. *Language Learning, 67*(2), 461-498. https://doi.org/10.1111/lang.12234

Kanwit, M., & Solon, M. (2013). Acquiring variation in future-time expression abroad in Valencia, Spain and Mérida, Mexico. In J. Cabrelli Amaro, G. Lord, A. de Prada Pérez, & J. E. Aaron (Eds.), *Selected proceedings of the 16th Hispanic Linguistic Symposium* (pp. 206–221). Somerville, MA: Cascadilla Proceedings Project.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 773-795. doi: 10.2307/2291091

Labov, W. (1966). *The social stratification of English in New York City*. Washington, D.C.: Center for Applied Linguistics. https://doi.org/10.1017/CBO9780511618208

Larson-Hall, J. (2017). Moving beyond the bar plot and the line graph to create informative and attractive graphics. *The Modern Language Journal, 101*(1), 244-270. https://doi.org/10.1111/modl.12386

Leal, T. (2018). Data analysis and sampling: Methodological issues concerning proficiency in SLA research. In A. Gudmestad & A. Edmonds (Eds.), *Critical reflections on data in second language acquisition* (pp. 63-88). Amsterdam: John Benjamins. https://doi.org/10.1075/lllt.51.04lea

Linford, B., Long, A., Solon, M., & Geeslin, K. L. (2016). Measuring lexical frequency: Comparison groups and subject expression in L2 Spanish. In L. Ortega, A. E. Tyler, H. I.

Park, & M. Uno (Eds.), *The usage-based study of language learning and multilingualism* (pp. 137-154). Washington, DC: Georgetown University Press.

Mannan, H. R. (2017). A practical application of a simple bootstrapping method for assessing predictors selected for epidemiologic risk of models using automated variable selection. *International Journal of Statistics and Applications, 7*(5). 239-249. doi: 10.5923/j.statistics.20170705.01

Medina-Rivera, A. (1999). Variación fonológica y estilística en el español de Puerto Rico. *Hispania, 82*(3)*,* 529-541. https://doi.org/10.2307/346322

Meunier, F. (2015). Developmental patterns in learner corpora. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 379–400). Cambridge: Cambridge University Press. https://doi.org/10.1017/cbo9781139649414.017

Meunier, F., & Littré, D. (2013). Tracking learners' progress: Adopting a dual 'corpus cum experimental data' approach. *The Modern Language Journal, 97*(Supplement), 61-76. https://doi.org/10.1111/j.1540-4781.2012.01424.x

Mitchell, R., Myles, F., & Marsden, E. (2013). *Second language learning theories* (3rd ed.). New York: Routledge. https://doi.org/10.4324/9781315617046-1

Mitchell, R., Tracy-Ventura, N., & McManus, K. (2017). *Anglophone students abroad: Identity, social relationships, and language learning*. New York: Routledge. https://doi.org/10.4324/9781315194851-2

Montrul, S., Foote, R., & Perpiñán, S. (2008). Gender agreement in adult second language learners and Spanish heritage speakers: The effects of age and context of acquisition. *Language Learning, 58*(3), 503-553. https://doi.org/10.1111/j.1467-9922.2008.00449.x

Myles, F. (2005). Interlanguage corpora and second language acquisition research. *Second Language Research, 21*(4), 373-391. https://doi.org/10.1191/0267658305sr252oa

Myles, F. (2015). Second language acquisition theory and learner corpus research. In S. Granger, G. Gilquin, & F. Meunier (Eds.). *The Cambridge handbook of learner corpus research* (pp. 309-331). Cambridge: Cambridge University Press. https://doi.org/10.1017/cbo9781139649414.014

Ortega, L. (2015). Second language learning explained? SLA across 10 contemporary theories. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (2nd ed.) (pp. 245-272). New York: Routledge.

Ortega, L., & Iberri-Shea, G. (2005). Longitudinal research in second language acquisition: Recent trends and future directions. *Annual Review of Applied Linguistics, 25,* 26-45. https://doi.org/10.1017/s0267190505000024

Plonsky, L., & Oswald, F. L. (2017). Multiple regression as a flexible alternative to ANOVA in L2 research. *Studies in Second Language Acquisition, 39*(3), 579-592. https://doi.org/10.1017/s0272263116000231

Preston, D. R. (1989). *Sociolinguistics and second language acquisition*. Oxford: Basil Blackwell.

Rehner, K. (2002). *The development of aspects of linguistic and discourse competence by advanced second language learners of French*. (Doctoral dissertation). Retrieved from Dissertation Abstracts International*, 63*, 12.

Sabourin, L., Stowe, L. A., de Haan, G. J. (2006). Transfer effects in learning a second language grammatical gender system. *Second Language Research, 22*(1), 1-29. https://doi.org/10.1191/0267658306sr259oa

Sagarra, N., & Herschensohn, J. (2012). Processing of gender and number agreement in late Spanish bilinguals. *International Journal of Bilingualism, 17*(5), 607-627. https://doi.org/10.1177/1367006912453810

Schlig, C. (2003). Analysis of agreement errors made by third-year students. *Hispania, 86*(2), 312-319. https://doi.org/10.2307/20062864

Schumann, J. H. (1975). *Second language acquisition: The pidginization process* (Unpublished doctoral thesis). Harvard University.

Smith, T. J., & McKenna, C. M. (2013). A comparison of logistic regression pseudo R2 indices. *Multiple Linear Regression Viewpoints*, *39*(2), 17-26.

Szmrecsany, B. (2017). Variationist sociolinguistics and corpus-based variationist linguistics; overlap and cross-pollination potential. *Canadian Journal of Linguistics, 62*(4), 685-701. https://doi.org/10.1017/cnj.2017.34

Tarone, E. (1988). *Variation in interlanguage*. London: Edward Arnold.

Teschner, R. V., & Russell, W. M. (1984). The gender patterns of Spanish nouns: An inverse dictionary-based analysis. *Hispanic Linguistics, 1,* 115-132.

Tracy-Ventura, N., & Huensch, A. (2018). The potential of publicly shared longitudinal learner corpora in SLA research. In A. Gudmestad & A. Edmonds (Eds.), *Critical reflections on data in second language acquisition* (pp. 149-169). Amsterdam: John Benjamins. https://doi.org/10.1075/lllt.51.07tra

Tracy-Ventura, N., & Myles, F. (2015). The importance of task variability in the design of learner corpora for SLA research. *International Journal of Learner Corpus Research, 1*(1), 58-95. https://doi.org/10.1075/ijlcr.1.1.03tra

Tummers, J., Heylen, K., & Geeraerts, D. (2005). Usage-based approaches in cognitive linguistics: A technical state of the art. *Corpus Linguistics and Linguistic Theory, 1-2,* 225-261. https://doi.org/10.1515/cllt.2005.1.2.225

White, L., Valenzuela, E., Kozlowska-Macgregor, M., & Leung, Y-K. I. (2004). Gender and number agreement in nonnative Spanish. *Applied Psycholinguistics, 25*(1), 105-133. https://doi.org/10.1017/s0142716404001067

Wulff, S. (2017). What learner corpus research can contribute to multilingualism research. *International Journal of Bilingualism, 21*(6), 734-753. https://doi.org/10.1177/1367006915608970

Young, R. (1991). *Variation in interlanguage morphology*. New York: Peter Lang.

Young, R. (1996). Form-function relations in articles in English interlanguage. In R. Bayley & D. R. Preston (Eds.), *Second language acquisition and linguistic variation* (pp. 135–175). Amsterdam: John Benjamins. https://doi.org/10.1075/sibil.10.07you

Appendix S1: R Code Used for Chi-Square Methods and Bootstrapping

R portion: investigate correlations and conduct bootstrapping

```
#Investigate strong correlations between independent variables
set.seed(314159)

spanish=read.csv("C:\\Users\\Tom Metzger\\Desktop\\spanish_with_score.csv")

#Investigate strong correlations between independent variables
library(vcd)
library(lme4)
mosaic(spanish$Task~spanish$Gender)
summary(lm(spanish$Frequency~spanish$Task))
mosaic(spanish$Task~spanish$NounMOR)
mosaic(spanish$Task~spanish$Endthree)
mosaic(spanish$Task~spanish$Nclass)
mosaic(spanish$Task~spanish$NounPronum)
mosaic(spanish$Task~spanish$ModQuant)
mosaic(spanish$Task~spanish$Modifier)
plot(spanish$Modifier,spanish$Endthree)

N=nrow(spanish)
#should be the number of observations in your data set.
#when bootstrapping resamples from the data, it will
#almost certainly have duplicates in the sample.
B=1000 #The number of resamples to do. This should be large.
#But since you're just using this to explore which
#effects are important, it's not the end of the world
#if you reduce this number to shorten the time.

fixed.effects=as.list(rep(NA,B)) #create empty object to store
#bootstrap effect estimates
for(i in 1:B){
  rows=sample(1:N,N,replace=TRUE) #pick which observations to resample
  spanish.subset=spanish[rows,]     #create the appropriate data set

  #fit the model with the effects you wish to check on
  # example: endthree, task, gender

  boot.model=glm(TLGender~Endthree+Task+Gender,
                 data=spanish.subset,family="binomial")

  #store the fixed effects from each distinct model.
  fixed.effects[[i]]=boot.model$coefficients
  print(i)
}
```

```
#bootstrap loop

#create plots of the fixed effect estimates. Here the order
#1-11 matches the order of the effects in the output.
#Look at a summary of one of the models to see what order
#they're in. Then each line of code extracts the 1st, 2nd,
#etc. effect.
par(mfrow=c(3,4))
hist(unlist(sapply(fixed.effects,"[",1)),main="Intercept")
hist(unlist(sapply(fixed.effects,"[",2)),main="Endthreenotapp")
hist(unlist(sapply(fixed.effects,"[",3)),main="Endthreeotherfemmasc")
hist(unlist(sapply(fixed.effects,"[",4)),main="Endthreeregoa")
hist(unlist(sapply(fixed.effects,"[",5)),main="Taskoral")
hist(unlist(sapply(fixed.effects,"[",6)),main="Taskwritten")
hist(unlist(sapply(fixed.effects,"[",7)),main="Gendermasc")
```

SAS Portion

```
PROC IMPORT OUT= WORK.spanish
            DATAFILE= "C:\Users\Tom Metzger\Dropbox\SAIG Summer
2018\spanish_with_score.csv"
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;

data spanish2;
set spanish;
if TLGender='target' then tl_gender2='1-target';
if TLGender='nontar' then tl_gender2='2-nontar';
run;

Title "Spanish Analysis with Time Interactions";
proc glimmix data=spanish2 method=quad;
class Endthree(ref="regoa") Gender(ref="masc") Task(ref="writ") Modifier(ref="art") PartID
Time(ref="pre");
model tl_gender2  = Endthree Task Gender Lexical_Score SyllmodN Modifier Time Eiscore
Endthree*Time
/link=logit dist=binary corrb ddfm=bw oddsratio s alpha=0.05;
random intercept / subject=PartID type=chol cl solution;
lsmeans Endthree Task Gender Modifier Time /adjust=tukey oddsratio cl;
lsmeans Endthree*Time /slicediff=Endthree adjust=tukey oddsratio cl;
/*output out=predictions pred(blup ilink )=pred;
/*lcl(blup ilink )=lcl
ucl(blup ilink )=ucl ;*/
run;
```

```
/*Not significant: SyllmodN*Time, Modifier*Time,
Gender*Time, Lexical_Score*Time, Task*Time */

Title "ROC Curve";
proc logistic data=predictions;
model tl_gender2 = ;
roc "GLIMMIX model" pred=pred;
roccontrast;
run;
```

R portion: compute McFadden's R squared

```
#Fit the null model and compute its log-likelihood
Lnull=logLik(glm(spanish$TLGender~1,family=binomial))[1]
#Obtain the log-likelihood for the fitted model that is automatically output in SAS
Lmod=3777.18/-2
#Calculate McFadden's R-squared
mcfadden=1-(Lmod/Lnull)
```

Appendix S2: Frequency Distribution of the Nominal Fixed Effects

Distribution of the categories of the significant, nominal effects

Frequency of occurrence of the categories of the significant, nominal effects

Fixed effect	#	%
Noun ending		
Canonical -o/-a	7333	61.9
Predictive	1663	14.0
Other	2343	19.8
Deceptive	507	4.3
Task		
Interview	7707	65.1
Narrative	1736	14.7
Essay	2403	20.3
Noun gender		
Masculine	6408	54.1
Feminine	5438	45.9
Modifier type		
Determiner	9117	77.0
Adjective	2729	23.0
Time		
Pre-stay	3800	32.1
In-stay	5136	43.4
Post-stay	2910	24.6

Note. Percentages may not add up to 100 due to rounding.