Statistical analyses of learner corpus data

Stefan Th. Gries
University of California, Santa Barbara
& Justus Liebig University Giessen

Sandra C. Deshors
Michigan State University

Abstract

This chapter provides an overview of the different kinds of statistical analyses that are relevant to, and figure prominently in, analyses of learner corpus data. We begin by outlining a variety of core issues and challenges for such statistical analyses, which include individual variation, corpus structure, task effects, repeated measurements, and the multifactoriality of SLA phenomena. Next, we offer discussion and exemplification of statistical methods with a particular focus on regression approaches, because they provide a unified treatment for many if not most current LCR kinds of statistical questions and have evolved to a degree that can address all of the above challenges. We then point to important future directions for the field to take in order to deal with the ever increasing complexity of learner corpora.

1 Introduction

Learner corpus research (LCR), i.e. the field studying learner/non-native speaker performance using corpora, has established itself as a vibrant sub-field of general corpus linguistics. LCR inherits from corpus linguistics the notion that virtually everything a (learner) corpus researcher studies must be operationalized on the basis of frequencies of (co-)occurrence in corpora; thus, the analysis of such data requires statistical analysis (since statistics is the science teaching us how to 'make sense of' frequency data). However, LCR adds to corpus linguistics the connection to research on second/foreign language acquisition (SLA/FLA) with all the complexities that these entail for the kinds of statistical analyses that are ultimately required. While we will not exemplify in detail the wide range of issues that much statistical work in LCR exhibits, it is instructive to consider the state of the field as discussed in Paquot & Plonsky (2017). Their survey of 378 published LCR studies shows that the field's statistical sophistication is only slowly increasing and much remains to be done for LCR to become relevant to SLA/FLA research. For instance, currently, about 90% of all statistical analyses in LCR are one of the following: chi-squared tests (23%), the log-likelihood ratio G^2 (10%), *t*-tests (20%), simple correlation (17%), analysis of variance (14%) and regression (6%). Even this alone indicates that much of LCR is still monofactorial in nature, which is highly problematic given how that (i) guarantees that the real complexity of the phenomena studied will be underestimated and (ii) increases the probability of false positives in LCR research, i.e. findings that appear important and/or significant in a monofactorial analysis of aggregate data but would not in a more appropriate multifactorial analysis (see Gries 2018 for discussion). In the next section, we discuss some of the central issues that affect and complicate the statistical analysis of learner corpus data.

2 Core issues and topics

For a long time, much corpus-linguistic work has underanalyzed the complexity the statistical analysis of corpus data requires. Unfortunately, the analysis of learner corpus data is often even more complex; the following is a brief overview of the many interrelated challenges that scholars using learner corpus data need to confront; for didactic purposes, we present these in two groups but this implies no particular theoretical commitment or ranking.

The first kind of challenges result from learner corpus data being observational data rather than data from the carefully-controlled experiments typical of SLA research, in which many potential sources of noise/variability are controlled and where, often, the balanced number of responses per experimental condition is known by design in advance:

there is considerable *variability between corpora* and *variability within corpora*. Regarding the former (between-corpus variability), different corpora differ from each other in many ways: (i) corpora even from within the same compilation project can differ in terms of how the data are collected, from whom the data are collected, how exactly the collected data were generated, how much annotation/metadata is available for them, etc., and ideally all of this information would be controlled for at least statistically (if one cannot afford controlling it by, say, including only corpora whose metadata suggest a high degree of comparability). Regarding the latter (within-corpus variability), corpora have an internal structure one needs to consider. In simple cases, this can just arise from the fact that a researcher compiles their own corpus from existing corpora, e.g., by studying learner data from the combination of both spoken and written data such as the International Corpus of Learner English (ICLE, Granger et al. 2009) and the Louvain International Database of Spoken English Interlanguage (LINDSEI, Gilquin, De Cock, & Granger 2010); in such a case, each file/speaker would be nested into a mode and into an L1 (but the L1s in turn could be crossed with, i.e. observed in, both the written and the spoken corpus data).

corpus data often pose problems with regard to what is called *dispersion*, the degree to which an expression/structure is evenly distributed throughout a corpus (see Gries to appear). This is important because, if an expression is distributed clumpily/unevenly in a corpus, then most of the data points for that expression are only provided by a potentially small minority of the speakers while most speakers provide only few data points, which leads to problems with the generalizability of the results. A second and more extreme case is that extreme clumpiness can of course even result in the fact that most files/speakers do not contain the element in question even once: For instance, Hasselgård & Johansson (2011) study the use of the word *quite* in native and non-native writing and in their data, more than 80% of the files do actually not contain *quite* at all. This is important because it raises the question what to do with speakers/files not exhibiting a particular expression? Shall one assume that not using *quite* was a conscious choice (because the speakers preferred *rather*, *very*, … each time) or shall one consider the possibility that (some) speakers did not even know the word *quite*? Obviously, whichever one chooses will affect the interpretation of the results.

corpus data are notorious for their *Zipfian distributions* (Ellis et al. 2016) and, often, high degrees of *multicollinearity* (Tomaschek et a. 2018). The former (Zipfian distributions) means that, for most expressions/structures, many or even most frequencies of (co-

)occurrence will be quite low, whereas (very) few frequencies of (co-)occurrence will be quite high; that is, very few types may account for many, if not most, tokens, while many types will each only be instantiated with very few tokens. This kind of distribution can make statistical analysis difficult, because it means that for many word types data will be problematically sparse. The latter (multicollinearity) means that, often, predictors in corpus data – causes – will be correlated with each other in ways that complicate the analysis of unbalanced data sets.

The second kind of challenges is related to the *multivariate and multilevel structure of the data*:

for many phenomena, there will be a *repeated-measurements structure* just as in experiments, but much less balanced/regular: Some, but not all, speakers may contribute 2+ data points to the corpus, but different speakers contribute differently many (and many none); similarly, there may be multiple data points for each lexical item one studies. That also means there are often temporal effects in corpus data due to priming/persistence effects (speakers' preferences to re-use expressions/structures). Speakers will be nested into corpora and L1s, but might be crossed with tasks, might be observed multiple times over time (in longitudinal learner corpus data), etc. In some studies, some of the above might be predictors of interest (e.g., we may be interested in how different tasks affect an outcome), in some they may be statistical controls (e.g., we may be interested in temporal development, but must partial the effect of different tasks out of the temporal changes).

all linguistic phenomena are *multifactorial* in nature, i.e. every statistical analysis needs to be multifactorial to account for multiple potential causes of an effect at the same time *and* to control for already known causes of an effect so that a suggested new cause can be shown to add to our understanding. To this, we must add the fact that …

LCR studies often have an implicit *multilevel design*: many variables – predictors and controls – operate on different levels. For instance, many variables are observation-level variables (i.e., they describe an individual usage event involving specific words and/or constructions) and may require difficult-to-operationalize constructs involving concreteness ratings, meaningfulness ratings, age-of-acquisition data, phonological or orthographic neighborhood densities, etc.). The same is true of variables located at the speaker-level (capturing speaker-specific variability), which, too, are often hard to operationalize (e.g., when it comes to constructs such as aptitude, intelligence, motivation, proficiency, …). Other variables operate at the level of the corpus (e.g., annotation preferences adopted in one compilation project), and yet others are located at the level of tasks (e.g., task-specific and mode-specific effects). Variability can be observed at each of these levels and must be carefully separated from the predictors of interest to avoid anti-conservative results (i.e. results that lead researchers to reject a null hypothesis that should not be rejected).

In sum, observational LCR data and the typically more experimental data in SLA/FLA research can be located on a 'multidimensional continuum' as indicated in Figure 1.

observational/corpus	dimension	experimental
low	artificiality/control	high
collinear & Zipfian	distribution	controlled & equal/balanced
harder	statistical analysis	simpler

Figure 1: Simplistic comparison of observational and experimental data in LCR and SLA/FLA research respectively

Arguably, the observational data of LCR come with a higher degree of ecological validity compared to the more controlled data in SLA/FLA experiments, but because of the characteristics of these data, they also require quite sophisticated statistical analyses, certainly more complex than what is currently practiced in most LCR work (recall Paquot & Plonsky 2017 from above).

Next, we discuss practical implications and guidelines for statistical analyses of LCR data; in keeping with the fact that the vast majority of LCR statistical applications is hypothesis-testing in nature, we concentrate on such applications ourselves, too.

3 Main research methods (and tools)

3.1 Regression as a multi-purpose tool: the set-up of the data

Given the above data/research situation, the one central research method/tool that would benefit LCR most is proper mixed-effects / multilevel regression modeling. While that claim may seem simplistic, much statistics in LCR are already instances of regression modeling even if practitioners might not be very aware of that fact:

> *t*-tests, simple ANOVAs, linear regression, and Pearson's *r* are all cases of linear regression models: they all involve a numeric dependent/response variable and, while they differ in terms of whether the independent/predictor variable is binary (*t*-test), categorical (simple ANOVA) or numeric (linear regression / *r*), their results are identical; chi-squared tests are closely related to log-likelihood/G^2-values, which – either as significance tests or as association measures computed from 2×2 co-occurrence tables – correspond to results from generalized linear regression or multinomial models, as do in fact association measures such as *MI*, *t*-scores, *z*-scores, or ΔP-values.

However, the vast majority of statistical applications in LCR, no matter whether they are conceived of as traditional tests or as regression models, are too simplistic: Usually, they are monofactorial (i.e. excluding both other relevant predictors and required controls) and often also fraught with other problems (such as multiple testing without corrections etc.). In what follows, we describe the ways in which appropriate regression modeling would boost the discipline's analytical power and as an example to motivate our discussion, we will use a scenario in which a researcher is interested in how learners of English use the two variants, or constructions, in (1):

(1) a. the owl's beak *s*-genitive: possessor*'s* possessum
b. the beak of the owl *of*-genitive: posssessum *of* possessor

In our scenario, the researcher has corpus data from learners of two L1s, say German and Mandarin Chinese, which cover both spoken and written data (say, from LINDSEI and ICLE), which means each speaker is represented with only one file in only one mode/corpus. Our hypothetical researcher is interested in the following two potential causes of genitive choices (and, for the sake of simplicity, we ignore the fact that even a study focusing only on these two linguistic predictors would have to include more linguistic predictors to ensure that these predictors have an impact even when everything else is controlled for, see Gries 2018):

> the animacy of the possessor (because animate possessors prefer *s*-genitives);
> the length difference between possessors and possessums (because of a general short-before-long preference).

First, a proper statistical analysis in this scenario requires the data to be in the case-by-variable format, where every genitive – each *of* and each *s* – is represented by its own row in, most likely, a spreadsheet software and where every column represents one variable – predictor, control, or a 'data organization' variable (such as the files/speaker or corpus or L1 a data point belongs to / is from, see below). This means, one must not use data aggregated over many speakers but, for each genitive we need columns for

> the dependent variable: which genitive was observed (*of* or *s*);
> the animacy of the possessor (e.g., the simplest possible case of just *animate* vs. *inanimate* or any more complex hierarchy; in (1), *the owl* could be classed as *nun-human/animate*);
> the length difference between possessor and possessum (e.g., measured in characters or syllables; in (1), one character).

However, to address (i) variability between corpora, (ii) variability within corpora, (iii) the genitives' dispersion throughout the corpora, and (iv) speaker-specific effects (e.g., speakers who behave very differently from the overall average), we also need columns for

> the corpus the example is from (ICLE vs. LINDSEI);
> the L1 of the speaker of the current example (German or Mandarin Chinese; if native speakers were included, then of course also English);
> the file in which corpus the example is from (as a proxy for the speaker the genitive is from).

This allow us to do what LCR knows it needs but does not always do, namely "keep the group perspective […], while at the same time taking individual variability into account" (Granger et al. 2015:2). Then, to address lexically-specific effects (e.g., possessors that invariably take only one genitive or possessors that learners use fairly consistently non-nativelike), we need columns for

> the lexical item that is the possessor (perhaps in lemma form, *owl* in (1));
> the lexical item that is the possessum (perhaps in lemma form, *beak* in (1)).

In addition, we need to control for temporal effects such as speakers' tendency to re-use previously used structures, which, by the way, are the reason why one should not randomly sample on the level of the individual genitives (if one wants to study not all data points found in a corpus) – instead, one needs to randomly sample on the level of files or conversations or speakers so that all genitives from one conversation are included and allow the study of temporal effects. Thus, we minimally need columns for

> the genitive choice that precedes the current one (*of* or *s*) to account for priming effects operating while the student wrote the essay/participated in the conversation, and maybe the line number in which each genitive was found (so we can determine how far apart in 'text time' the two constructional choices happened).

However, the above does not exhaust the range of possibilities. While the above annotation scheme provides for every single genitive the speaker who produced it, with the right corpus annotation, it is still possible and often necessary to include additional speaker-level information into the analysis to better accommodate individual differences (see Dörnyei 2005). Individual differences relevant to much SLA research include, but are not limited to, personality, aptitude, cognitive predisposition and learning styles, motivation, etc., all of which are complex constructs. For instance, with regard to aptitude, we might distinguish grammatical sensitivity, phonological decoding ability, memory capacity, and inductive learning ability (Caroll & Sapon 1959); obviously, in LCR we will usually also be interested in including speakers' proficiency level (however operationalized, see Chapter 30). For instance, with regard to motivation, we might distinguish intrinsic and extrinsic motivation, and so forth, see Wulff & Gries (to appear) for more of an overview.

Unfortunately, most learner corpora do not provide much in terms of speaker-level information other than maybe age, sex, a global proficiency score (e.g. CEFR), or the length of exposure to English. Naturally, this absence of metadata is problematic for LCR practitioners wanting to benefit from recent work in SLA/FLA research. Some such information can be approximated, as when lacking proficiency scores are operationalized via lexical or formulaic diversity and/or syntactic complexity scores computed directly from the corpus files (see Wulff & Gries, to appear, for an example), but more fine-grained information would of course be better; for our scenario let us consider that we have a column with proficiency scores for each speaker.

With such data, most morphosyntactic LCR analyses – of over-/underuse, of alternations, … – would boil down to formulating a regression model that codifies, with regard to the sources of the data, the structure of the corpus data (e.g., how speakers/files are nested into corpora) and, with regard to the predictors in question, the hypothesized causal or correlational structure of the data. One example – not the most comprehensive one possible or necessary – is the model defined in Figure 2 in a format that corresponds to how this could be formulated within the R statistical language and environment (R Core Team 2018). (We are using R because (i) it is more powerful than SPSS, (ii) it is freely available), (iii) it is the leading platform for the development of new statistical tools, and (iv) it is a full-fledged programming language, which means it can be used for more than just statistical analysis (e.g. corpus processing, Gries 2016) and which means that exchange of code makes analyses perfectly replicable.)

Next, we briefly discuss how the results from such a regression model can be interpreted.

3.2 Regression results and their interpretation

The analysis of a regression model such as that in Figure 2 usually returns four parts of results. First, there will be some overall statistics describing the quality of the model (see 1. in Figure 3), where *quality* refers to the degree to which the model fits the data. For regression models, quality should be assessed by the following statistics:

some form of an R^2-value that quantifies the quality of the model (typically, $0 \leq R^2 \leq 1$, with higher values being better);

a classification accuracy in percent stating how many of the studied cases the model classifies correctly; ideally, this would be complemented by corresponding accuracy values from cross-validation, i.e. a process where the regression model is computed from test data (often 90% of the original data) and then applied to training data (the remaining 10% of the original data);

scores such as C ($0.5 \leq C \leq 1$), *precision* (e.g., how many of the genitives predicted to be *s*-genitives are really *s*-genitives?), and *recall* (how many of all the *s*-genitives were found?), and the baseline against which accuracies are compared.

GENITIVE ~	the dependent variable is modeled as a function of
1 +	an overall intercept
ANIMACY +	a main effect of the critical predictor ANIMACY
LENGTHDIFF +	a main effect of the critical predictor LENGTHDIFF
ANIMACY:LENGTHDIFF +	an interaction of the critical predictors
L1 +	a main effect of the speaker's L1
ANIMACY:L1 +	an interaction term allowing the effect of ANIMACY to be different in the two L1s
LENGTHDIFF:L1 +	an interaction term allowing the effect of LENGTHDIFF to be different in the two L1s
ANIMACY:LENGTHDIFF:L1 +	an interaction term allowing the effect of ANIMACY: LENGTHDIFF to be different in the two L1s
PROFICIENCY +	a main effect of the speaker's proficiency level
PROFICIENCY:L1 +	an interaction term allowing the effect of PROFICIENCY to be different in the two L1s
LASTGENITIVE +	a main effect of priming, i.e. which genitive was produced last time
(1|CORPUS/FILE) +	an adjustment for the fact that each file/speaker, which is nested into one corpus, may behave differently from others
(1|POSSESSOR) +	an adjustment for each possessor lemma
(1|POSSESSUM)	an adjustment for each possessum lemma

Figure 2: One possible regression model for the genitive scenario

Second, for the so-called *random effects*, i.e. the adjustments made for the corpora, files/speakers, and lexical items that are treated as if they were randomly sampled from a larger population, we would obtain estimates of their variability, with larger estimates indicating that the variables in question exhibit a large amount of variability (that needs to be accounted for). As exemplified in Figure 3, one might obtain a large estimate for FILE and a small one for CORPUS, which would mean that, once speaker-specific variability is taken into account (with FILE), differences between corpora are less relevant; thus, this integrates corpus comparisons into such analyses. In addition, any lexically-specific effects could be reflected in the variances for POSSESSOR and POSSESSUM.

Finally, for the *fixed effects*, i.e. the variables whose values in the sample (the corpus data) we consider to exhaust the possible ranges of values in the population (the world), we obtain two kinds of results. First, one *p*-value for the overall significance of each predictor that could be dropped from the model (see 3a). Second, for each coefficient estimated, we get (see 3b)

an estimate or coefficient indicating how much and in what direction a change in a variable affects the probability of the predicted genitive (in a certain statistical situation);
an estimate of that coefficient's uncertainty (a standard error);
a significance test for the coefficient: does it differ significantly from 0 or not?

1. Overall results			
Nagelkerke R^2	0.672	*C*	0.87
classification accuracy	0.87	baseline	0.55
precision	0.9	recall	0.8

2. Random effects variances	intercept
CORPUS	0.1
FILE	0.8
POSSESSOR	0.5
POSSESSUM	0.2

3a. Significance of fixed-effects predictors	LR-test	*df*	p_{deletion}
ANIMACY : LENGTHDIFF : L1	15.47	1	<0.0001
PROFICIENCY	5.2	3	0.158
[…]			

3b. Coefficients (intercept=0.81, predicted: *s*-genitive)	coefficient	std. error	*z*	*p*
ANIMACY$_{\text{animate} \rightarrow \text{inanimate}}$	-1.89	0.12	-15.75	<0.0001
LENGTHDIFF	-0.06	0.021	-2.86	0.0042
ANIMACY$_{\text{animate} \rightarrow \text{inanimate}}$: LENGTHDIFF	0.1	0.23	0.43	0.6672
L1$_{\text{Chinese}\rightarrow\text{German}}$	0.2	0.06	3.33	0.0009
ANIMACY$_{\text{animate} \rightarrow \text{inanimate}}$: L1$_{\text{Chinese}\rightarrow\text{German}}$	0.17	0.07	2.43	0.0151
LENGTHDIFF:L1$_{\text{Chinese}\rightarrow\text{German}}$	0.1	0.2	0.5	0.6171
ANIMACY$_{\text{animate} \rightarrow \text{inanimate}}$: LENGTHDIFF : L1$_{\text{Chinese}\rightarrow\text{German}}$	0.4	0.057	7.02	<0.0001
PROFICIENCY$_{\text{B1}\rightarrow\text{B2}}$	0.3	0.22	1.36	0.1738
PROFICIENCY$_{\text{B1}\rightarrow\text{C1}}$	0.17	0.15	1.13	0.2585
[…]				

Figure 3: Possible regression results for the genitive scenario

To some degree, one can interpret such a model on the basis of its fixed-effects coefficients: The value of -1.89 for ANIMACY means that, in a certain condition, inanimate possessors make *s*-genitives less likely (the negative sign) than animate possessors; similarly, the value of -0.06 for LENGTHDIFF means that, in a certain condition, an increase of length difference by 1 makes *s*-genitives less likely (the negative sign). However, this is much less straightforward once more complex models are interpreted; for instance, Figure 3 reveals a significant three-way interaction of ANIMACY : LENGTHDIFF : L1, i.e., the two learner varieties differ in how ANIMACY and LENGTHDIFF affect their genitive choices – such findings should be visualized by plotting

> for numeric predictors (such as LENGTHDIFF), a regression line whose location and slope reveal how LENGTHDIFF values are related to changes in the predicted probabilities of *s*-genitives;
> for categorical predictors (such as L1), points whose location reveal the predicted probabilities of *s*-genitives for the different levels of L1.

and of course in the case of an interaction, one might have to plot several regression lines (e.g. one for every level of a categorical predictor that a numeric predictor interacts with) or several sets of points that cover all predicted values arising from the interaction of 2+ categorical predictors.

Within the R environment, we recommend using effects plots (see Fox 2013, Fox & Hong 2009), which show effects of predictors with all other predictors in the model controlled (by holding them constant at typical values). If the predictors ANIMACY and LENGTHDIFF were significant and did not participate in interactions, such plots might look like Figure 4.

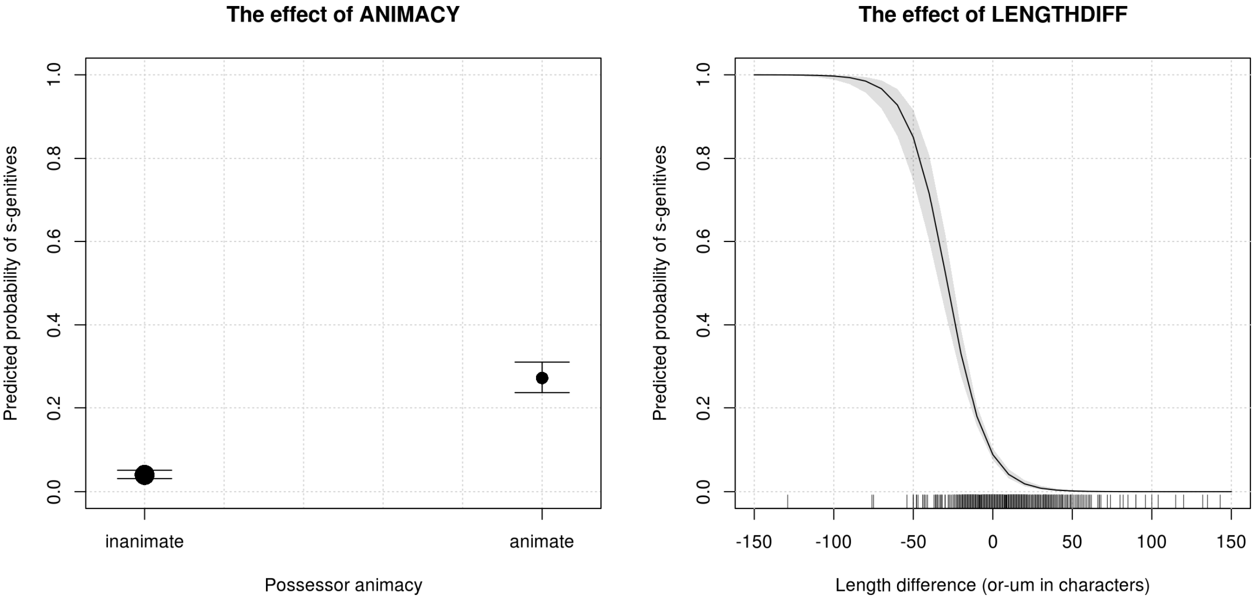

Figure 4: Effects plots for two hypothetical main effects

The left panel shows that *s*-genitives are more likely with animate possessors (as expected) with the point sizes and error bars representing the frequencies of the animacy levels and the 95% confidence intervals respectively. The right panel shows that *s*-genitives are less likely the longer the possessor is relative to the possessed (also as expected) with rugs on the *x*-axis representing the observed length differences and the grey band representing the 95% confidence band.

While this example might appear specialized – a constructional alternation with no native speaker data included and no other speaker-level controls – and while we did not discuss all possible predictors and their effects, the above is a methodological blue print of how, currently at least, the majority of LCR studies can be conducted (in better ways) to inform SLA/FLA research. The above regression-based logic extends to all kinds of alternation cases, be they morphosyntactic like here, lexical, pragmatic or something else; it extends to cases where more than two competing expressions are studied, to cases where the dependent variables are accuracies of uses of expressions (e.g., in percent) or ordinal proficiency levels, and even to cases of mere over-/underuse frequencies. In fact, most studies counted by Paquot & Plonsky (2017) as involving a regression-like procedure can not only be recast in the above framework, but actually improved because of how, now, many more 'quirks' of learner corpus data are accommodated: speaker-specific effects, lexically-specific effects, repeated measurements, corpus structure, multifactoriality, and more.

For instance, many classical over-/underuse studies such as Aijmer (2005), Hasselgård & Johansson (2011), or Laufer & Waldman (2011) could benefit from being reframed in this way, if only to address, e.g., the issue of within- and between-corpus variability, namely the repeated measurements of some speakers but no occurrences from others. As another instance, consider Durrant & Schmitt (2009), who compare native and non-native writers' use of collocations. They determine frequencies and compute association strengths – *MI* and *t* – of Adj-N and N-N pairs in comparable corpora of native and non-native English. They also (laudably!) compute text-based results as opposed to pooling all data, but even their analysis could benefit from a regression reframing: For example, their analysis of the association measures proceeds on the basis of grouping collocations into seven bins of association strengths, which loses much the information contained in the originally numeric scores and would not have been required in a regression-based framework.

It is important to point out that there is a variety of reasons why regression modeling will not always be a viable option. Most of these reasons involve the size and the structure of the data one is trying to analyze. Obviously, regression modeling requires a certain sample size depending on the number of variables one wishes to study; simplistically speaking, the more predictors, the larger the sample size should be. In addition, even large samples can prove problematic for regression modeling when the data are unevenly distributed in a way that leaves few data points for many combinations of predictors, when the predictors are highly correlated with each other, when some predictor combinations are perfectly predictive of the response, and/or when the response variable is extremely skewed (see Gelman & Hill 2008). However, there are few hard-and-fast rules and it takes exploration and experience to determine when other methods are likely to be more useful; alternative methods used in linguistics are especially conditional inference trees and random forests (see Baayen et al. 2013).

4 Representative corpora and research

Lester, N.A. (2019). *That*'s hard Relativizer use in spontaneous L2 speech. *International Journal*

of Learner Corpus Research, *5*, 1-32.
Lester (2019) is a study on the realization of the relativizer *that* in non-subject-extracted relative clauses (NSRCs) as in (2).

(2) a. This is the communicator that Lieutenant Reed forgot on the planet
b. This is the communicator ___ Lieutenant Reed forgot on the planet

This paper embodies the approach of regression-based studies in LCR, in particular the MuPDAR (Multifactorial Predication and Deviation Analysis with Regressions, see Gries & Adelman 2014, Gries & Deshors 2014) approach discussed in another chapter in particular. Lester extracted all NSRCs manually from the Louvain Corpus of Native English Conversation (LOCNEC) corpus and the German and Spanish components of the LINDSEI corpus. First, he documents over-/underuse patterns using generalized linear mixed-effects modeling of the type discussed above: The Spanish learners produced significantly fewer NSRCs and fewer tokens without *that* than the German learners, but were closer to the native speaker performance; the German learners produced more NSRCs and more tokens without *that* than the native speakers.

He then did a MuPDAR analysis of these data: Both the native and non-native speaker data were annotated with regard to a dozen variables, many of which are familiar from the regression discussion above. They included whether *that* was absent or present, the L1 of the speaker, a code for each speaker, the task type in which the example was produced (monologic vs. dialogic), the length of the relative clause subject, self-priming from previous relative clauses, etc. He then fitted a generalized additive mixed model – a regression model good at handling curvature or non-linearity in how numeric predictors affect the response – to the native speaker data, trying to predict whether or not a *that*-relativizer would be used.

Next, he determined that the resulting regression model yielded good prediction accuracy using cross-validation and then applied the model to the non-native speaker data to generate for each non-native speaker NSRC (i) a prediction of whether a native speaker would have produced *that* or not and (ii) a score of how much the learner choices differed from what a native speaker would have produced.

Finally, he fit a second regression model to predict the deviation scores based on the same predictors as the first regression. To mention just a few of the results: he found that

Spanish learners used *that*-relativizers more nativelike than German learners;

learners' choices became less nativelike as the relative clause subjects (*Lieutenant Reed* in (2) above) became longer;

learners' choices became less nativelike in more disfluent speech (i.e. under higher processing load);

learners from both L1s exhibited self-priming effects (Spanish learners more so than German learners).

Lester concluded that his study is a "first to reveal a tendency for learners to omit optional grammatical markers in complex environments," a finding that differs from other LCR findings, e.g., Wulff et al.'s (2018) on *that* as a complementizer and that suggests that "the costs of producing different types of constructions in L2 can lead to different, construction-specific communicative strategies."

Wulff, Stefanie, Nick C. Ellis, Ute Römer, Kathleen Bardovi-Harlig, & Chelsea LeBlanc et al. 2009. The acquisition of Tense-Aspect: converging evidence from corpora and telicity ratings. *The Modern Language Journal*, *93*, 354-369
Wulff et al. (2009) is concerned with the aspect hypothesis, the assumption that language learners of both L1 and SL/FL learners acquire tense and aspect morphology in ways that are influenced by the inherent semantic aspect of verbs and the events they describe: Using Vendler's aspectual categories, perfective past morphemes are used first with telic predicates (achievements and accomplishments) whereas progressive markings are used first with atelic predicates (activities).

First, Wulff et al. explored the frequencies and associations between verbs and tense-aspect marking in native speaker data (from the spoken component of the British National Corpus and in learner data (from the Michigan Corpus of Academic Spoken English, Swales et al. 2002). Regarding the verbs' frequencies, they found Zipfian distributions such that, for each tense-aspect pattern studied, (i) a few highly frequent types account for very many tokens whereas (ii) very many types are extremely infrequent. Regarding the verbs' co-occurrence preferences, they used multiple distinctive collexeme analyses (Gries & Stefanowitsch 2004), an extension of a measure of association for 2×2 co-occurrence tables (such as G^2) to more than two elements, to determine which verbs are significantly attracted to which tense-aspect morphemes. Here, too, they found Zipfian distributions such that each morpheme has only a few verbs strongly attracted to it and the verbs attracted to each tense-aspect morpheme come in straightforward semantic groups. This case study of Wulff et al. therefore shows a strong confirmation of the aspect hypothesis for both native and non-native speakers.

As a validation, Wulff et al. then collected telicity ratings for 86 verbs from 20 subjects in a rating task. They computed average telicity ratings for all verbs (*be* and *love* scored lowest, *end* and *finish* scored highest in telicity) and then reported

> the results of independent *t*-tests to determine whether, in both native and non-native language, verbs associated with past tense in the corpus data (from case study 1) received significantly higher telicity ratings, which they did;
> the results of Pearson's *r* correlations for whether the five most frequent past and progressive verbs were significantly correlated with their mean telicity ratings; they found correlations in the expected direction, but only significantly so for progressive verbs.

This study clearly confirms the aspect hypothesis and points to why it might hold: The tense-aspect patterning is strong in the input in terms of both frequency and contingency and so learners pick up on it and use it themselves. At the same time, this supports our earlier point of the relevance of proper regression modeling. In particular, case study 2 could have benefited from a more comprehensive regression-based approach (rather than just simple *t*-tests): In a single unified regression model, one could have determined whether the telicity ratings of the subjects were predictable from all verbs' overall frequencies, their associations to all tense-aspect morphemes, and a predictor for L1 vs. L2. Nonetheless, Wulff et al. is a nice study at the interface of learner corpus research and (more experimental) SLA work.

5 Future directions

The above overview leads to straightforward future directions. Regarding corpus compilation, we need more detailed metadata for our corpora. Put differently, while learner corpus compilers usually get told that researchers want bigger corpora and (more) longitudinal corpora – and of course we want all that, too – we also urge compilers towards including more and more detailed annotation regarding speaker and task characteristics so that LCR can begin to implement the above regression-based strategies and bridge the gap to much work happening in SLA/FLA.

Regarding statistical analysis of learner corpus data, the main desideratum is to have more studies that conform to the above blue print. While the number of more advanced analyses has grown , as witnessed by publications and paper presentations at, say, the LCR conference, progress is too slow, given how much faster both general linguistics and corpus linguistics are evolving quantitatively; a more widespread adoption of the kinds of methods that are now routinely used in other linguistic sub-disciplines is simply a must, especially given how much of a quantum leap this would mean for the field. This would be especially true if this adoption was coupled with a concomitant increase in awareness of some additional tools, such as the relevance of exploring non-linear relationships (as in Lester 2019), user-defined contrasts (see Fox & Weisberg 2019: Sectoin 4.7) to test specific hypotheses directly, and model diagnostics (to determine whether our models need to be tweaked or whether other statistical approaches are required, see Fox & Weisberg 2019: Chapter 8). Such other approaches include *robust statistics* (useful for data that violate the assumptions of many parametric statistical methods, see Larson-Hall & Herrington 2009, Wilcox 2012), but also general *machine-learning kinds of methods* such as random forests or, on the exploratory side of things, *association rules* (see Hastie, Tibshirani, & Friedman 2009 or James et al. 2013). We admit that this raises the desired level of complexity of the field of LCR to a whole new level and few linguists, including ourselves, had that kind of training, but given the kinds of data learner corpora offer and their complexity, the field as a whole will have to move things up a notch or two in order to address the challenges the data and our theories about the data pose to us.

6 Further readings

Desagulier, Guillaume. 2017. *Corpus linguistics and statistics with R*. Berlin & New York: Springer.
This textbook covers both the processing of data with R (corpora and tabular spreadsheet kinds of data) as well as various statistical analysis techniques. As for the former, Desagulier covers regular expressions and programming basics to generate concordances, frequency lists etc. from corpora (see Gries 2016 for a book-length treatment); as for the latter, while he discusses some hypothesis-testing approaches (though not much in terms of regression modeling, for that see Baayen 2008 or Gries 2013), his book is particularly strong when it comes to exploratory methods.

Larson-Hall, Jennifer & Richard Herrington. 2009. Improving data analysis in second language acquisition by utilizing modern developments in applied statistics. *Applied Linguistics* 31(3). 368-390.
This article makes two suggestions about how to improve statistical analyses in SLA research.

The first suggestion is seemingly modest but nonetheless vital and promotes more and more insightful visualization of SLA data, in particular preferring box plots over bar plots and regression lines involving curvature over straight ones (for the kinds of non-linear effects often found). The second suggestions involves, as alluded to above, a greater reliance on robust statistics, i.e., a family of methods that is geared towards handling data that violate the assumptions of many kinds of parametric tests (e.g., by exhibiting non-normality, outliers, etc.).

Bestgen, Yves. 2017. Beyond single-word measures: L2 writing assessment, lexical richness and formulaic competence. *System* 69(1). 65-78.
In this paper, Bestgen explores the degree to which association measures (specifically, *MI* and *t*) computed on learner bigrams using frequencies from native speaker use in the British National Corpus correlate with the quality of learner texts. Using simple and multiple regression models as well as cross-validation, he shows that formulaic richness measures, in particular *MI*, outperform lexical diversity scores in predicting measures of learner text quality.

References

Aijmer, K. (2005). Modality in advanced Swedish learners' written interlanguage. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition, and foreign language teaching* (pp. 55-76). Amsterdam & Philadelphia: John Benjamins.

Baayen, R. H. (2008). *Analyzing linguistic data: a practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Baayen, R. H., L.A. Janda, T. Nesset, A. Endresen, & A. Makarova. (2013). Making choices in Russian: Pros and cons of statistical methods for rival forms. *Russian Linguistics, 37*, 253-291.

Carroll, J. B. & S. Sapon. (1959). Modern Language Aptitude Test (M.L.A.T.). New York: The Psychological Corporation.

Dörnyei, Z. (2005). *The psychology of the language learner: Individual differences in second language acquisition*. New York and London: Routledge.

Durrant, P. & N. Schmitt. (2009). To what extent do native and non-native writers make use of collocations. *International Review of Applied Linguistics, 47*, 157-177.

Ellis, N. C., U. Römer, & M. Brook O'Donnell. (2016). *Usage-based approaches to language acquisition and processing: cognitive and corpus investigations of Construction Grammar*. *Language Learning* 66 (Suppl. 1, Language Learning Monograph Series). New York: John Wiley.

Fox, J. (2003). Effect displays in R for generalised linear models. *Journal of Statistical Software, 8*, 1-27.

Fox, J. & J. Hong. (2009). Effect displays in R for multinomial and proportional-odds logit models: Extensions to the effects package. *Journal of Statistical Software*, *32*, 1-24.

Fox, J. & S.Weisberg. (2019). *An R companion to applied regression*. 3rd ed. Los Angeles & London: Sage.

Gelman, A. & J. Hill. (2008). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

Gilquin, G., S. De Cock, & S. Granger. (2010). *Louvain International Database of Spoken*

English Interlanguage (CD-ROM + Handbook). Presses universitaires de Louvain, Louvain-la-Neuve.

Granger, S., E. Dagneaux, F. Meunier, & M. Paquot. (2009). *International Corpus of Learner English v2* (Handbook + CD-Rom). Presses universitaires de Louvain, Louvain-la-Neuve.

Granger, S., G. Gilquin, & F. Meunier. (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.

Gries, St. Th. (2013). *Statistics for linguistics with R*. 2nd rev. and ext. ed. Berlin & Boston: De Gruyter Mouton, pp. 359.

Gries, St. Th. (2016). *Quantitative corpus linguistics with R*. 2nd rev. & ext. ed. London & New York: Routledge, Taylor & Francis Group, pp. 274.

Gries, St. Th. (2018). On over- and underuse in learner corpus research and multifactoriality in corpus linguistics more generally. *Journal of Second Language Studies*, *1*, 276-308.

Gries, St. Th. (to appear). Analyzing dispersion. In M. Paquot & St. Th. Gries (Eds.). *Practical handbook of corpus linguistics*. Berlin & New York: Springer.

Gries, St. Th. & A. S. Adelman. (2014). Subject realization in Japanese conversation by native and non-native speakers: exemplifying a new paradigm for learner corpus research. In J. Romero-Trillo (Ed.), *Yearbook of Corpus Linguistics and Pragmatics 2014: New empirical and theoretical paradigms* (pp. 35-54). Cham: Springer.

Gries, St. Th. & S. C. Deshors. (2014). Using regressions to explore deviations between corpus data and a standard/target: two suggestions. *Corpora*, *9*, 109-136.

Gries, St. Th. & A. Stefanowitsch. (2004). Extending collostructional analysis: a corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics*, *9*, 97-129.

Hasselgård, H. & S. Johansson. (2011). Learner corpora and contrastive interlanguage analysis. In F. Meunier, S. De Cock, G. Gilquin, & M. Paquot (Eds.), *A taste for corpora: In honour of Sylviane Granger* (pp. 33-61). Amsterdam & Philadelphia: John Benjamins.

Hastie, T., R. Tibshirani, & J. Friedman. (2009). *The elements of statistical learning: data mining, inference, and prediction*. 2nd ed. Berlin & New York: Springer.

James, G., D. Witten, T. Hastie, & R. Tibshirani. (2013). *An introduction to statistical learning with applications in R*. Berlin & New York: Springer.

Larson-Hall, J. & R. Herrington. (2009). Improving data analysis in second language acquisition by utilizing modern developments in applied statistics. *Applied Linguistics*, *31*, 368-390.

Laufer, B. & T. Waldman. (2011). Verb-noun collocations in second language writing: a corpus analysis of learners' English. *Language Learning*, *61*, 647-672.

Lester, N. A. (2019). *That*'s hard: Relativizer use in spontaneous L2 speech. *International Journal of Learner Corpus Research* 5(1). 1-32.

Paquot, M. & L. Plonsky. (2017). Quantitative research methods and study quality in learner corpus research. *International Journal of Learner Corpus Research*, *3*, 61-94.

R Core Team. (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Simpson, R. C., S. L. Briggs, J. Ovens, & J. M. Swales. (2002). The Michigan Corpus of Academic Spoken English. Ann Arbor, MI: The Regents of the University of Michigan.

The British National Corpus, version 3 (BNC XML Edition). (2007). Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium.

Tomaschek, F., P. Hendrix, & R. H. Baayen. (2018). Strategies for addressing collinearity in multivariate linguistic data. *Journal of Phonetics*, *71*, 249-267.

Wilcox, R. (2012). *Modern Statistics for the Social and Behavioral Sciences: A Practical Introduction*. Boca Raton, FL: Chapman and Hall/CRC Press.

Wulff, S, N. C. Ellis, U. Römer, K. Bardovi-Harlig, & C. Leblanc. (2009). The acquisition of Tense-Aspect: converging evidence from corpora and telicity ratings. *The Modern Language Journal*, *93*, 354-369.

Wulff, S. & St. Th. Gries. (to appear). Exploring individual variation in Learner Corpus Research: some methodological suggestions. In B. Le Bruyn & M. Paquot (eds.), *Learner corpora and second language acquisition research*. Cambridge: Cambridge University Press.

Wulff, S., St. Th. Gries, & N. A. Lester. (2018). Optional *that* in complementation by German and Spanish learners. In A. Tyler, L. Huan, & H. Jan (Eds.), *What is Applied Cognitive Linguistics? Answers from current SLA research* (pp. 99-120). Berlin & Boston: De Gruyter Mouton.