Exploring Individual Variation in Learner Corpus Research: Some Methodological Suggestions

Stefanie Wulff
University of Florida

Stefan Th. Gries
University of California, Santa Barbara

Abstract

Second language acquisition is a complex process, and correspondingly, learner corpus research is increasingly turning to complex statistical methods. While traditional approaches mostly relied on simple frequency counts and monofactorial analyses, some recent studies employ more sophisticated statistics that permit the inclusion of more than one predictor variable at a time as well as more varied kinds of probabilistic/distributional information such as association strengths and dispersion values. One such recent approach is called MuPDAR (Multifactorial Prediction and Deviation Analysis Using Regression; Gries & Adelman 2014, Gries & Deshors 2014): two multifactorial regression analyses are carried out, one on native speaker data to be able to impute learner choices, one on learner data to explore their (non-)native choices, which yield (i) a comprehensive picture of native speaker behw2avior and (ii) a comprehensive model of learner behavior, namely fine-grained descriptions of which aspects of a target structure learners command in a native-like fashion and which aspects they struggle with, which gives rise to precise predictions of areas of difficulty and ease in L2 learning.

One intriguing aspect of MuPDAR is its extensibility: in its current form, the exploration of the learner data can be L1-specific, but in the present paper we will extend this approach towards also including speaker-specific effects. As such, this method should appeal to the growing number of researchers who are interested in investigating individual variation. After a brief overview of individual differences and individual variation research in SLA, and different levels of methodological complexity in learner corpus research, we present a case study of the genitive alternation in the Chinese and German sections of the *International Corpus of Learner English* alongside English native speaker data obtained from the *International Corpus of English* and illustrate different ways in which the MuPDAR approach could be extended to obtain models that license deeper interpretation at the individual speaker level. We close with a discussion of implications and desiderata for future corpus-based analysis and corpus design.

1. Introduction

In this paper, we would like to make a case for corpus-based analyses of individual differences and variation in second language acquisition (SLA), and specifically advocate for one corpus-based method, MuPDAR(F) (explained in detail below). After a compact overview of how individual differences and variation are defined in (non-corpus-linguistic) SLA in Section 1, we turn to corpus-based approaches to investigating learner language, presenting an argument in favor of MuPDAR(F)-type approaches to explore individual variation in Section 2. In Section 3, we present the results of a MuPDAR(F) analysis of the genitive variation in learner and native speaker data. In Section 4, we discuss the implications and desiderata for future corpus-based analyses of learner language as well as future corpus compilation projects.

1.1 Individual Differences and Variation in Second Language Acquisition Research

One of the first – and few to date – comprehensive discussions of individual differences in second language acquisition (SLA) research is offered by Dörnyei (2005) (see also Skehan 1986). He makes a case for including personality, aptitude, and motivation as key properties that differentiate individual learners, and he points to other research in SLA that also includes learning styles and learning strategies. Ultimately, "the concept of 'individual differences' is rather loose, containing certain core variables and many optional ones" (Dörnyei 2005:7).

A closer survey of SLA research confirms Dörnyei's position regarding the wide scope of how individual differences are defined. This aligns well with two by now established observations in SLA research that are shaping theory and methodology alike:

(i) SLA must be understood as an inherently complex process in which a multitude of factors, often in interaction with each other, drive and mold the acquisition process and determine ultimate attainment (VanPatten & Williams 2015).

(ii) Several of these factors are complex concepts themselves. For example, we can break down the concept of motivation minimally into intrinsic and extrinsic motivation, and many more complex definitions have been suggested (see Woodrow (2016) for a recent overview); similarly, the Modern Language Aptitude test defines aptitude as a combination of grammatical sensitivity, phonological decoding ability, memory capacity, and inductive learning ability (Carroll & Sapon 1959) and again, other definitions abound; see Singleton 2017).

Individual differences combine with other (both language-internal and language-external) factors such as the quality and quantity of the input received, the differences in form-function mappings between the first and the second language, the setting in which the learning takes place (for example either in an immersive or instructional context), or the specifics of the task at hand (e.g. timed vs. self-paced tasks, graded vs. non-graded assignments, etc.).

The interplay of individual differences with all these other factors inevitably gives rise to considerable variation in learner performance across comprehension, processing, and production. In fact, given this multi-layered complexity of SLA both in terms of "what goes in" as well as "what goes out", rather than being an alternative hypothesis to be justified, the emergence of individual variation must be considered the null hypothesis. In fact, since a majority of the cognitive and even language-external factors not only characterize each L2 learner in different ways, but native speakers just the same, there is a growing consensus that native speakers do not present a monolithic group. Instead, native speakers display significant amounts of individual variation as well, blurring, at times, the boundaries between native speakers and learners (and consequently rendering native speakers a questionable target norm, at least with regard to performance measures; see Hopp 2010; Street 2017).

A review of the empirical literature of the past decade or so indicates that both individual differences and individual variation are receiving increasing attention. Cognitive predispositions such as executive function and working memory (Wen, Mota & McNeill 2015), declarative vs. procedural memory (e.g. Hamrick 2015; Morgan-Short et al. 2014), aptitude (e.g. Granena 2013, 2016), and even resting qEEG (Prat et al. 2016) are among the most extensively studied individual differences. A good number of studies examine differences in personality types, learning styles, and learning strategies (e.g. Grey, Williams & Rebuschat 2015). A second large strand of research considers individual variation that emerges from the external circumstances and environment of the learners, such as the amount of input they receive (e.g. Unsworth 2016), the quality of the input (e.g. Rothman & Guijarro-Fuentes 2010), and the specific learning context (e.g. Collentine & Freed 2004, Grey et al. 2015). Additionally, and in accord with the line of reasoning outlined above, a growing number of studies also considers the relative impact of internal and external factors in combination (e.g. Chondrogianni & Marinis 2011; Courtney et al. 2017; Li 2013; Sun et al. 2016).

Corpus-based research, however, to date constitutes only a small portion of all studies in this area (see Möller 2017 for a recent exception). There might be several (inter-related) reasons for this. Firstly, corpus data are of limited use in the investigation of comprehension and processing-related individual variation, and much better suited to investigate individual variation in production. Secondly, the majority of learner corpora available to date do not include a wide array of meta-data for the learners who contributed to the corpus. Speaker information usually includes information about the speakers' L1, age, and gender, as well as some information about their globally measured L2 proficiency; cognitive measures such as measures of executive control, working memory capacity, etc. are harder to come by. A third reason for the rare application of corpus data might be a commonplace misconception among non-corpus linguists that corpora only contain massive pools of data collapsed over anonymized speakers, with no option to tie data points to the individual speaker who produced it.

While we acknowledge that most corpora would benefit from adding more comprehensive batteries of meta-data about the speakers captured in the corpus to allow meaningful discussion of individual differences, we would like to argue here that at least with regard to individual variation in production, corpus-linguistic methods offer a variety of advantages. For one, most corpora are designed such that data points can be traced back to individual speakers and whatever information is available about them. Furthermore, corpus linguistics is particularly well-equipped to deal with multifactorial phenomena. While corpus data are admittedly noisier/messier than experimental data, their ecological validity is higher and proper statistical control can help deal with the noise in instructive ways. In the following, we outline a corpus-linguistic approach that we believe to be particularly well-suited for the analysis of L2 production data and make some suggestions for how it can be employed to examine individual variation, specifically (see Gries 2018 for more detailed discussion than can be provided here).

1.2 The MuPDAR(F) Approach

One of the most recent methodological additions to both learner corpus research and indigenized-variety research is an approach called MuPDAR(F) (for multifactorial prediction and deviation analysis using regression/random forests) (Gries & Adelman 2014; Gries & Deshors 2014). This method is conceptually based on missing-data imputation: for every linguistic choice of a learner (or speaker of the target variety), one imputes what the native speaker (or speaker of the historical source variety) would have chosen given identical contextual conditions. While in the case study to be discussed below, we opted to compare the learners' choices with native speaker data, it is important to point out that MuPDAR does not require the reference data to be from native speakers – it is entirely up to the researcher to decide what a meaningful group for comparison should be.

MuPDAR(F) involves the following three steps:

fitting a regression/random forest $R(F)_1$ that predicts the choices that speakers of the source/reference level (typically, native speakers of the reference variety) make with regard to the phenomenon in question;

applying the results of $R(F)_1$ to the other/target speakers in the data (typically, learners or speakers of institutionalized second-language varieties) to predict for each of their data points what the native speaker of the source/reference variety would have done in their place;

fitting a regression/random forest $R(F)_2$ that explores how the other speakers' choices differ from those of the speakers of the source/reference variety: predictors that are significant in this regression are ones that help understand where the target variety speakers make choices that are not those of the source/reference variety.

This approach can be useful because it focuses primarily on probabilistic differences that result in different speaker choices. In other words, it identifies not whether a learner is predicted to choose a modal verb with a probability of 25% while the native speaker is predicted to choose that modal verb with a probability of 20% – rather, it focuses on cases where the individual discrete choices of each target speaker represented in the data differ from what we can predict a reference speaker will do.

In this paper, we apply the MuPDAR approach to the genitive alternation – *of* vs. *s*-genitives – in data from British English native speakers and Chinese and German second language learners of English. Section 2 explains our data, their annotation, and their statistical analysis. Section 3 discusses the results of our analysis; heeding to the point made above that individual differences can manifest at different levels of granularity, we separately present (i) results that apply to all speakers, (ii) results that distinguish between speakers from different L1 backgrounds, and (iii) results that distinguish individual speakers. Section 4 concludes with some implications of these results for both future analysis of this kind of data and learner corpus compilation.

2. Methods

As mentioned above, this paper is concerned with native speaker and learners' use of the genitive alternation as exemplified in (1).

(1) a. the squirrel's nuts *s*-genitive: possessor*'s* possessed
b. the nuts of the squirrel *of*-genitive: possessed *of* possessor

In this section, we discuss the data and their annotation for (in)dependent variables (Section 2.1) followed by a precise description of the statistical methods we used (Section 2.2).

2.1 Data

We retrieved examples of *of*- and *s*-genitives from three different corpora. For the learners, we retrieved all *of*- and *s*-genitives from the Chinese and the German sections of the ICLE (International Corpus of Learner English, version 2; see Granger et al. 2009 for details on the content of the corpus). Given the large overall number of matches, we decided to randomly sample 1,000 attestations of each variant from the full concordances; manual checking of these 2,000 hits led to 10 and 6 cases (involving plurals) being discarded from the Chinese and German data respectively. For the native speakers, we randomly sampled *of*- and *s*-genitives from all files of the ICE-GB (British component of the International Corpus of English), which resulted in examples being taken from 303 of the 500 files, which arguably constitutes a good representation of the target variety of the learners. The overall distribution of the data is represented in Table 1.

Table 1: The composition of our data set

	of-genitive	*s*-genitive	Total
L1: Chinese	872	118	990
L1: German	892	104	996
L1: English	817	183	1,000
Total	2,581	405	2,986

The matches were then manually annotated for a large number of variables; many of these have been argued to play a role in previous work (Gries & Wulff 2013), some others we added because they turned out to be significant predictors in our own pilot studies. The following is the list of predictors that we included in our analysis:

Genitive: *of* vs. *s*;
number of possessor (PossorNumber) and possessed (PossedNumber): *singular* vs. *regular plural* (with *s*) vs. *irregular plural* (e.g., *children* or *women*);
animacy of possessor (PossorAnim) and possessed (PossedAnim): inanimate vs. animate (but not human) vs. human;
syntactic modification of possessor (PossorBranch) and possessed (PossedBranch): *none* vs. *pre-modified* (e.g., *the natural environment*) vs. *post-modified* (e.g., *their right of choice of smoking*, where the underlined of is the genitive analyzed and *right* is post-modified by *of choice* vs. *pre- and post-modified* (e.g., *the richest source of protein of all veggies*);
complexity of possesor (PossorCompl) and possessed (Possedcompl): *simple* (e.g., non-modified nouns) vs. *intermediate* (nouns with adjectival or PP modification) vs. *complex* (nouns with clausal modification);
difference in length between possessor and possessed (LengthDiff): the difference between the number of characters of possessor minus possessed;
avoidance of adjacent identical surface forms (HorrorAequi): all genitives were annotated with regard to whether they contained additional genitives: *none* (e.g., *the parts of the Saudi desert*) vs. *of* (e.g., *the part of the map of Kent*) vs. *s* (*my neighbour's dung heap's odours*);
rhythmic alternation (RhythAltDiff): every phrase with a genitive and its other-genitive counterpart was coded for its sequence of stressed and unstressed syllables (*people's personalities* = suuusuu; and *personalities of people* = uusuuusu). From these, we computed a value whose absolute size increases with the number of stress clashes (sequences of stressed syllables) and stress lapses (sequences of unstressed syllables) and where positive and negative values indicate the criterion of rhythmic alternation would 'recommend' an *of*- and an *s*-genitive respectively; in addition, the higher the absolute value, the stronger that preference (see Wulff & Gries 2015 for details on how that value is calculated);
segment alternation (SegAltDiff): every phrase with a genitive and its other-genitive counterpart was coded for its two transitions from the end of one NP to the genitive marker and the genitive marker to the beginning of the next NP such that a CV/VC transition was scored as 0, a C_1C_2 transition (where $C_1 \neq C_2$) was scored as 1, and a C_1C_1 transition was

scored as 2 (*Isobel_'s_grief* = 1+1 and *grief_of_Isobel* = 0+0, i.e., the difference is 2). We then changed the sign of these differences such that positive and negative values indicate the criterion of ideal syllable structure would 'recommend' an *of*- and an *s*-genitive respectively, and the higher the absolute value of the difference, the stronger that preference;
first language (L1): the L1 of the speakers: *English* vs. *Chinese* vs. *German*.

These data were then used for the three-step MuPDAR approach.

2.2 Statistical Evaluation

In order to prepare the above data for a MuPDAR analysis, we did some initial data exploration. This included tabulating and plotting the data to determine whether variables needed to be transformed or variable levels needed to be conflated to avoid data sparsity, etc. As a result of this exploration,

the variables POSSORANIM and POSSEDANIM were recoded to only two levels: *animate* (conflating *animate* and *human*) and *inanimate*;
the variables POSSORBRANCH and POSSEDBRANCH were recoded to only three levels: *none*, *pre-modified* and *post-modified* (with or without additional pre-modification).

In addition, the numeric predictors LENGTHDIFF, RHYTHALTDIFF, and SEGALTDIFF were not just included as coded above, but as orthogonal polynomials to the second degree, in order to allow for curvature in their effect on genitive choices (in the first model) or nativelike choices (in the second model).

For the first step of the MuPDAR protocol, we decided against a mixed-effects regression model and chose a random forest (as in Deshors & Gries 2016). Random forests are well-known for being very good at detecting predictive structure in a data set while at the same time avoiding overfitting. This is crucial to MuPDAR(F) because step 2 involves imputing the choices that native speakers would have made. Thus, in this first step, the predictors were all the above; the response variable was GENITIVE; and we fitted 3500 trees where at each step, four variables were eligible to be used for the next split (this is the default setting for classification trees in randomForest::randomForest; see Liaw & Wiener 2015).

We then explored whether the random forest yielded a better-than-chance prediction accuracy for the native speaker training data – if that was not the case, the imputation that MuPDAR(F) requires would not be feasible (as will be shown below, we obtained an excellent prediction accuracy). We then applied the random forest to the learner data to get native-speaker predictions for each learner choice, and then compared whether the learner had made the native-like choice or not, which was captured in a variable NATIVELIKE (*no* vs. *yes*). In addition, we computed a variable called DEVIATION, which represents how much the learner choice differs from the imputed native-speaker choice: values of 0 indicate the learner made a native-like choice (i.e., NATIVELIKE is *yes*), values between 0 and -0.5 and between 0 and +0.5 indicate the learner used an *of*-genitive and an *s*-genitive respectively when the native speaker would have used the respective other variant, with higher absolute values representing higher degrees of 'nonnative-likeness'.

In a final step, we conducted a model selection process using generalized linear mixed-effects modeling to determine what predicts whether learners make native-like choices or not. The response variable was NATIVELIKE and the predictors were all of the above variables (now applied to the learner data) as well as, of course, L1. To identify the best model, we employed the following stepwise model-checking procedure: Our first model involved no predictors at all but just an overall intercept, which was allowed to vary for each speaker/corpus file. Then, we performed a bidirectional model selection process such that we checked at every step (i) what would improve the model best, the deletion or addition of which main effect or pairwise interaction (using *AIC* provided in the *R* output as the model selection criterion), (ii) whether said deletion or addition would raise overall multicollinearity *VIF*-values above 10 and (iii) whether said deletion or addition would introduce overdispersion problems. We discuss the results in the following section.

3. Results

3.1 Results of RF_1 and Its Application to the NNS

The result of the random forests analysis to the NS data yielded extraordinarily high prediction accuracies. The out-of-bag prediction accuracy reached 98.1% and a corresponding *C*-value of 0.99. Given these results, we applied the random forest model to the NNS data, which resulted in an expectedly lower, but still very good prediction accuracy (88%) and a good *C*-value of 0.82 (which exceeds the usual threshold value of 0.8). Also, examining the variable importance measures we obtained, we found those variables that previous studies have shown to be most important also turned out to be important in our analysis, including, among others, POSSORANIM, POSSORNUMBER, or LENGTHDIFF. From those predictions, we then added the two columns of NATIVELIKE and DEVIATION to the NNS data for subsequent analysis with R_2 and visualization.

3.2 Results of R_2

The model selection process as described above yielded a highly significant final model (LR-statistic=221.36, *df*=21, $p≈0$) with high R^2-values (in particular compared to some previous MuPDAR analyses): $R^2_m=0.721$, $R^2_c=0.75$. This final model achieved a classification accuracy of 90%, which, according to exact binomial tests, is significantly better than the baselines of the more frequent level of the NATIVELIKE variable and proportional random sampling of the levels of NATIVELIKE; the *C*-value for this model is 0.892. In the following sections, we discuss some of the results of R_2: for reasons of space, we only present a selection of instructive results in Sections 3.2.1 (results pertaining to all learners) and 3.2.2 (results distinguishing between learners of different L1s), before we turn to results pertaining to individual variation in Section 3.2.3.

3.2.1 Results at the Level of All Speakers

The first result is a main effect, namely that of LENGTHDIFF, which is represented in Figure 1 (with two panels). In (nearly all of) the following graphs, the *x*-axis represents one predictor (i.e., here LENGTHDIFF) while the *y*-axis represents the predicted probability of learners making the native-like choice; the left panel covers the complete range of LENGTHDIFF values whereas the right panel shows the same effect but zooms in to the central 90% of the values. The black points connected by a line are the regression line representing the model's predictions (with a grey-shaded 95% confidence band), and the jittered points around *y*=0 and *y*=1 represent the non-native-like and native-like choices, respectively (at *x*-coordinates representing the observed length differences); these are also represented by the rugs at the bottom and the top *x*-axis. The horizontal dashed line marks the predicted probability of 50% (i.e., the one where the prediction would flip), and the vertical dashed line marks the median LENGTHDIFF value.

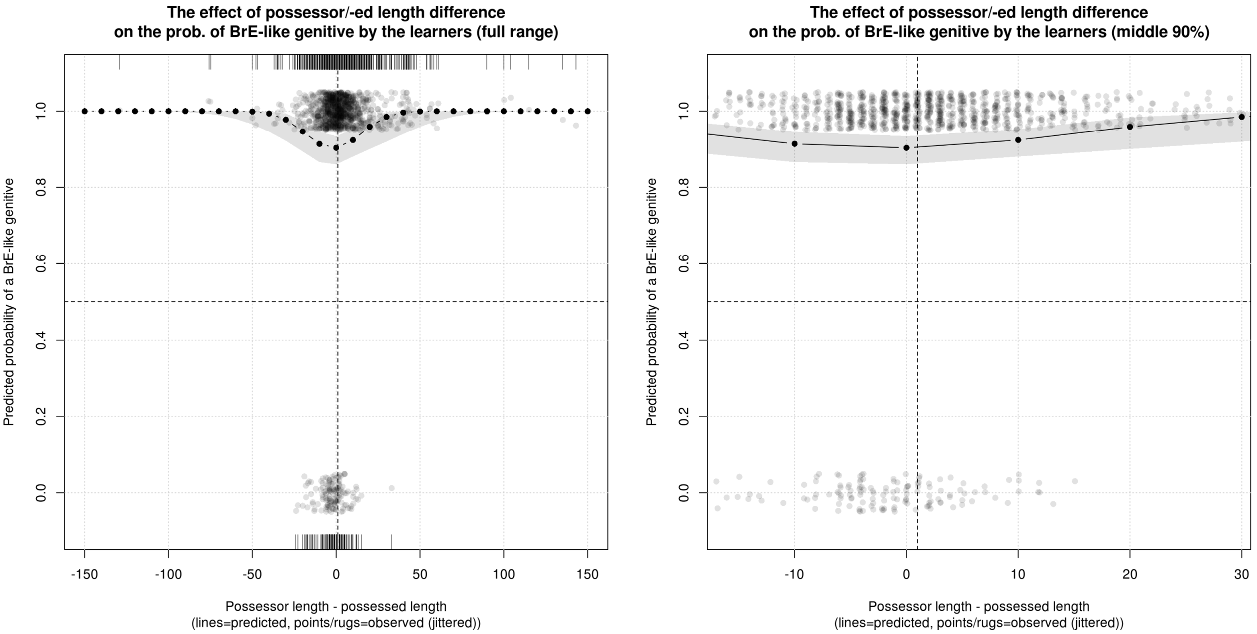

Figure 1: The effect of LENGTHDIFF (p<0.001)

The result is relatively straightforward: While the native-like choices very strongly outnumber the non-native-like ones (see the large number of jittered grey points around y=1 compared to the much smaller number around y=0), the learners are least native-like when the length differences between possessor and possessed are small, i.e. when LENGTHDIFF makes no strong 'recommendation' for a constructional choice that is compatible with short-before-long. This is evident from two observations: (i) the noticeable dip of the regression line (in particular in the left panel) around LENGTHDIFF=x=0 and (ii) the fact that the grey dots around y=0 representing non-native-like choices are not attested at all with larger absolute length differences, i.e. they do not exhibit the same wide spread as the grey dots around y=1 do.

Figure 2 represents the interaction of POSSEDNUMBER and POSSORANIM. The former variable is represented on the x-axis, the latter with different colors (as per the legend). The points' sizes are proportional to the frequencies with which the relevant combinations are attested in the data; in this case that means that inanimate singular possessors were the most frequent of the four combinations, whereas animate plural possessors were the least frequent. The error bars around the points are 95% confidence intervals, and the vertical dashed line represents the relative frequency of the variable levels on the x-axis, indicating here that there are many more singular than plural possesseds in our data.

The plot shows that when the possessor is inanimate (light blue points/line), then the learners make very native-like choices regardless of POSSEDNUMBER, which, upon checking the NS data, turn out to be predominantly *of*-genitives. However, when the possessor is animate, then the learner choices become much less native-like, especially with plural possesseds.

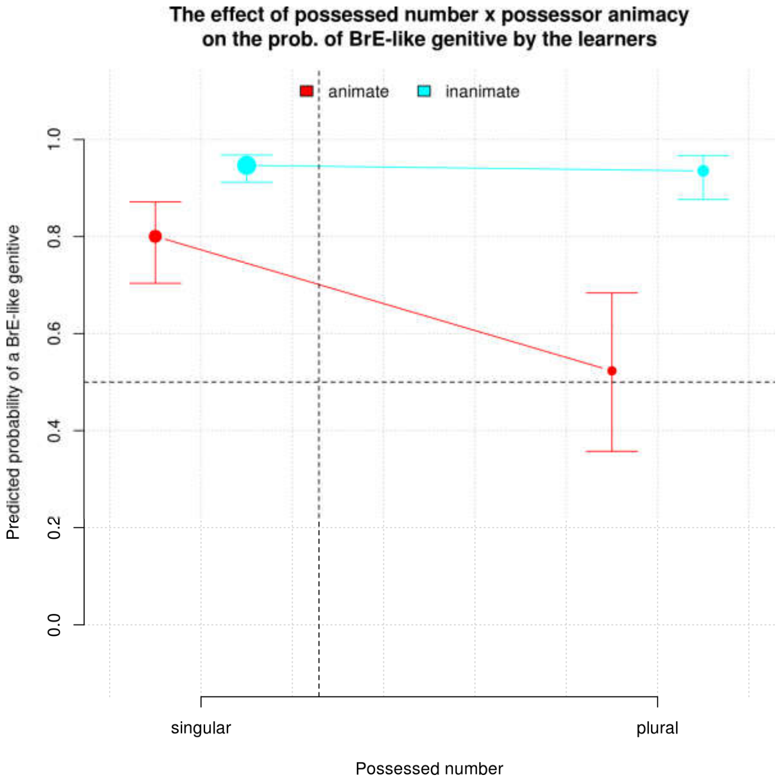

Figure 2: The effect of POSSEDNUMBER × POSSORANIM (p<0.037)

While space does not permit revisiting individual examples to see where this effect arises, the effect perfectly highlights the shortcomings of a simpler over-/underuse approach: if one computes the percentages of *of*- and *s*-genitives for animate possessors and plural possesseds for the native speakers and learners, we yield Table 2.

Table 2: Observed frequencies of genitives for animate possessors with plural possesseds

	POSSEDNUMBER: *plural*, POSSORANIM: *animate*		
	of-genitive	*s*-genitive	Total
NS	29 (65.9%)	15 (34.1%)	44 (100%)
NNS	57 (66.3%)	29 (33.7%)	86 (100%)
Total	86 (66.2%)	44 (33.8%)	130 (100%)

As Table 2 shows, the overall relative frequencies of genitives for animate possessors and plural possesseds are virtually identical. In a traditional over-/underuse account, we would conclude that, with animate possessors and plural possesseds, learners behave native-like – after all, the percentages of the genitives are nearly exactly the same. However, the more fine-grained resolution of multifactorial regression approaches in general and MuPDAR(F) in particular shows this to be false: With animate possessors and plural possessed, learners often make non-native-choices (as shown in Figure 2) by erroneously overusing *s*-genitives (which became obvious when we inspected the original data). (1)-(3) are a few examples from the data.

(1) … the smokers simply have no right to endanger other**s’** lives [Chinese learner]
(2) Employee**s’** savings are maximized by private mpf service providers … [Chinese learner]
(3) Teachers must set a good example and show their pupils that it is essential to see a person’s good qualities. [German learner]

The next effect – SegAltDiff × PossorNumber – is shown in Figure 3: SegAltDiff is on the *x*-axis, PossorNumber is symbolized by the two colors, which are used for predicted points, lines, confidence bands, and dots; again the left panel shows the whole range of SegAltDiff values while the right one shows the medial 90%.

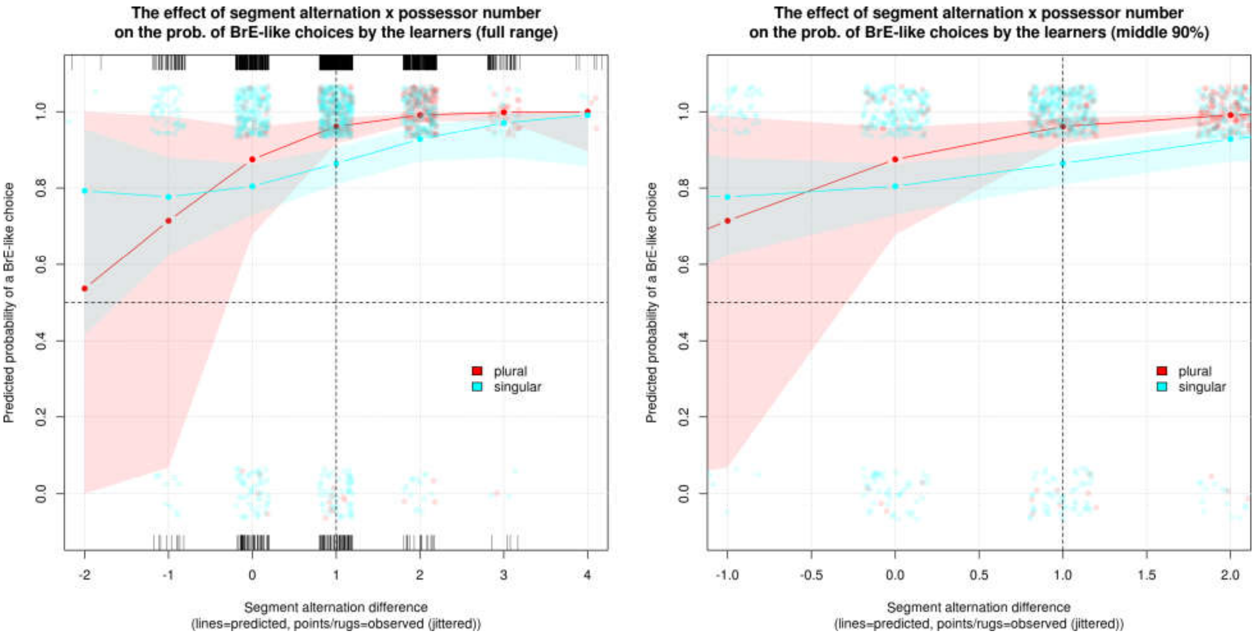

Figure 3: The effect of SegAltDiff × PossorNumber (*p*<0.123)

With singular possessors, learners make more native-like choices, and they do so regardless of SegAltDiff. However, with plural possessors, where plural *s* often leads to an avoidance of *s*-genitives by NS, the NNS perform more native-like by also avoiding the *s*-genitive. To fully understand this effect, it is important to focus on the right panel specifically, the one that covers the central 90% of the data, to see how (i) the red line is nearly consistently above the blue one and (ii) there are extremely few red points around *y*=0, indicating that while the polynomial effect of SegAltDiff in the regression curves down a bit in both panels on the left side, there are actually hardly any data points for such low SegAltDiff-values. A comparison with the NS data accordingly confirms what the bottom parts of Figure 3 already imply: the learners make most non-native-like choices when the possessor is singular and SegAltDiff makes no strong recommendation; (4)-(6) are a few examples from the data.

(4) The opening hour **of** the internet cyber café is 24 hours. [Chinese learner]
(5) … the consideration of the actual need **of** our society should be in a prior position. [Chinese learner]
(6) … use one car so that you use the full capacity **of** the car. [German learner]

In the next section, we turn to effects that distinguish between the learner groups.

3.2.2 Results at the Level of Speakers of a Certain L1

The first result pointing to differences between the Chinese and the German learners is L1 × PossorComplexity as represented in Figure 4.

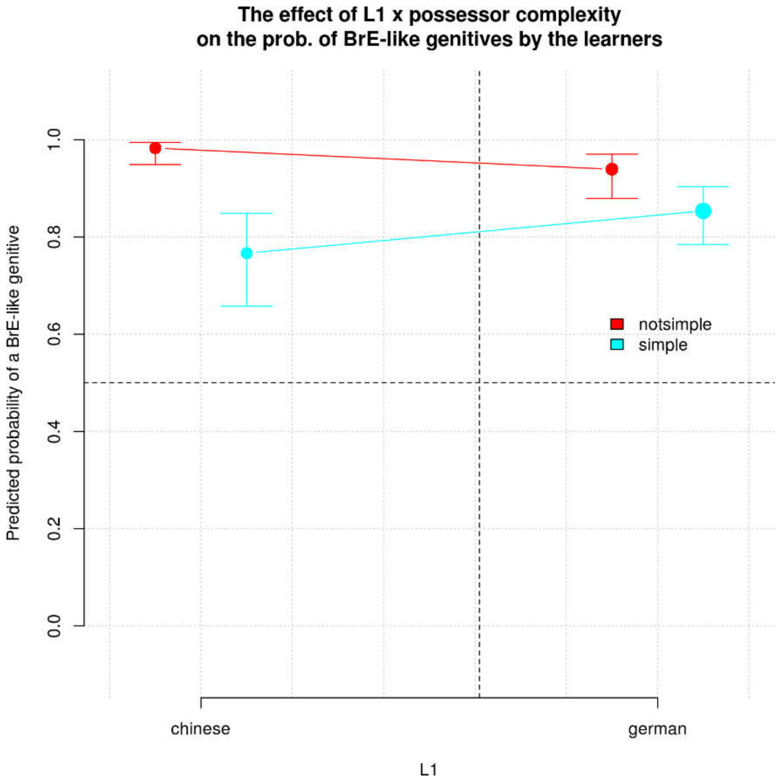

Figure 4: The effect of L1 × PossorComplexity (p<0.002)

When the possessor involves at least some degree of complexity, both learner groups make more native-like choices. However, when the possessor is simple, the Chinese learners become much less native-like than (i) they do when it is complex or (ii) the Germans. In other words, PossorComplexity affects the Chinese learners more, and an exploration of the data reveals that with simple possessors, they underuse the *s*-genitive. (7) and (8) are examples from the Chinese learner data.

(7) … the recycling industry is mainly driven by the awareness and the choice **of** consumers. [Chinese learner]

(8) They claim that public construction increases the expensing **of** the government [Chinese learner]

Finally, let us consider Figure 5, which reveals that the German learners' performance is not affected by whether the possessed is singular or plural, but the Chinese learners' performance is: with singular possesseds, they are less native-like than with plural ones because, as drilling down into the data shows, they underuse *s*-genitives.

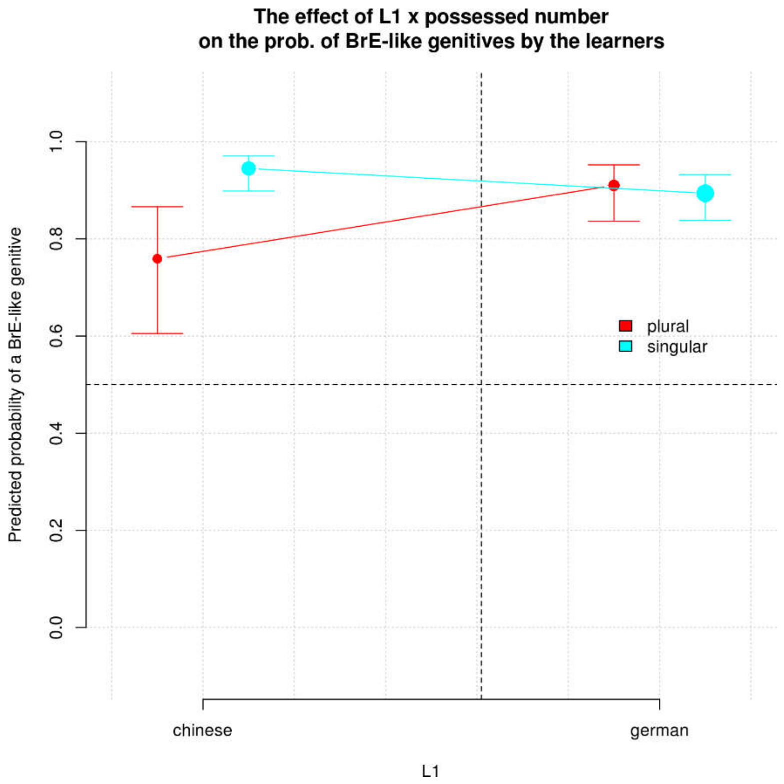

Figure 5: The effect of L1 × POSSEDNUMBER (p<0.001)

Let us now zoom in even more and begin to address the issue of individual variation.

3.2.3 Results at the Level of Individual Speakers

How does the MuPDAR(F) approach inform individual variation? There are again multiple levels of resolution that are available. The present approach is not unlike error annotation: consider applying the results from the random forest to the learner data an annotator who checks for each learner choice whether it is native-like or not, and so we can determine the error rates for each learner. Those can be plotted for a bird's eye view of the data and then be used to go back to each relevant learner's data for post-hoc exploration. Figure 6 is such a plot with genitive frequency on the *x*-axis and percentage of non-native-like uses on the *y*-axis; the red and blue dots are Chinese and German learners respectively, and the red and blue dashed lines represent means for both axes. That means the learners at the top of the plot struggle with genitives most and might merit further exploration.

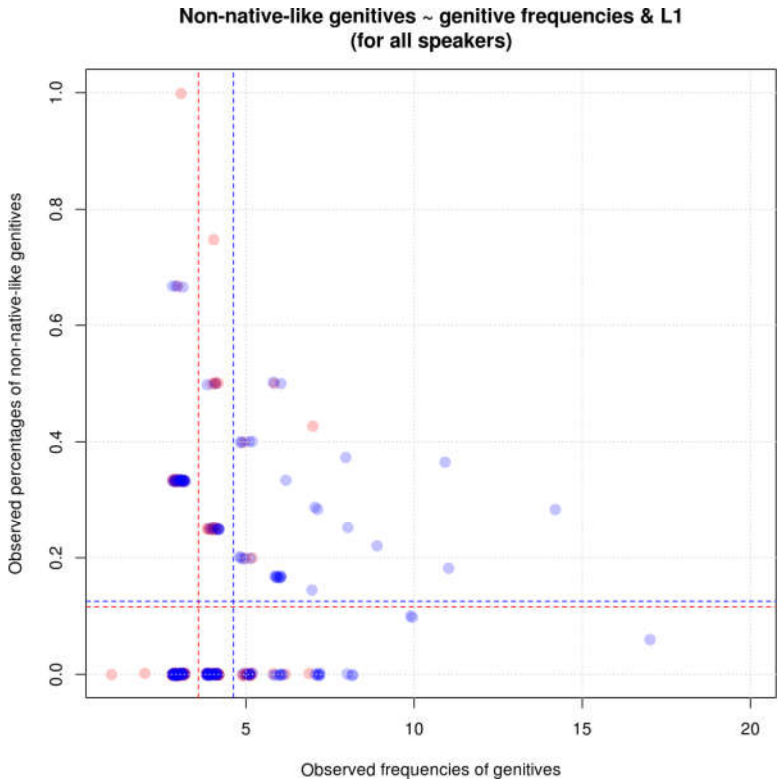

Figure 6: Percentages of non-native-like uses against genitive frequencies

A more fine-grained resolution would be to also take the severity of the non-nativeness and its directionality into consideration, which we captured in the variable DEVIATION. Such a plot makes it possible to see which learner errs in which direction and how much, as in Figure 7. All speakers below *y*=0 have more problems with overusing *of*, all speakers above *y*=0 have more problems with overusing *s*.

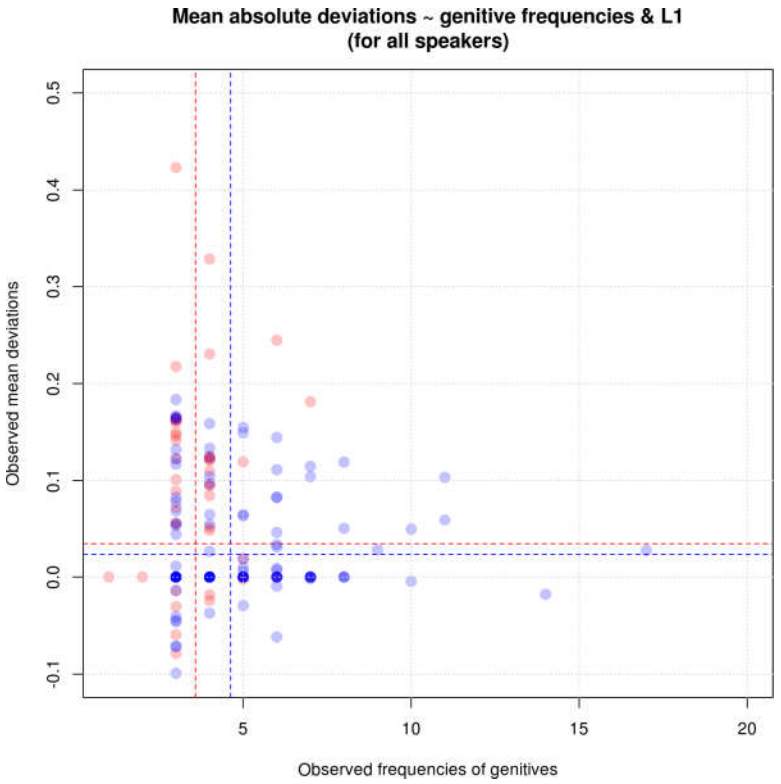

Figure 7: DEVIATION-values against genitive frequencies

Any of these types of results obtained for speakers could also be correlated with other information we have on the speakers, such as the kinds of individual differences discussed a lot in SLA research and surveyed above in Section 1.1. This can be done in various ways, but the two most obvious would be the following:

speaker-specific information such as working memory, executive function, age of arrival, etc. can become *a priori* predictors in either a traditional multifactorial regression or a MuPDAR kind of analysis to see, for instance, how they correlate with learners making more or less native-like choices;
such information can be explored *a posteriori* by correlating it after the fact with regression or MuPDAR results – if interesting correlations emerge, those can be interpreted tentatively and entered into subsequent statistical analyses of new data sets.

Finally, it is possible to drill down even deeper and check individual decisions of individual speakers. Figure 8 shows the individual deviation values for 10 speakers of their genitive choices (with dots) and their means (with ×). We can immediately see that some learners' values show they have problems with non-native *of*-genitives whereas for others, it is non-native *s*-genitives that are problematic.

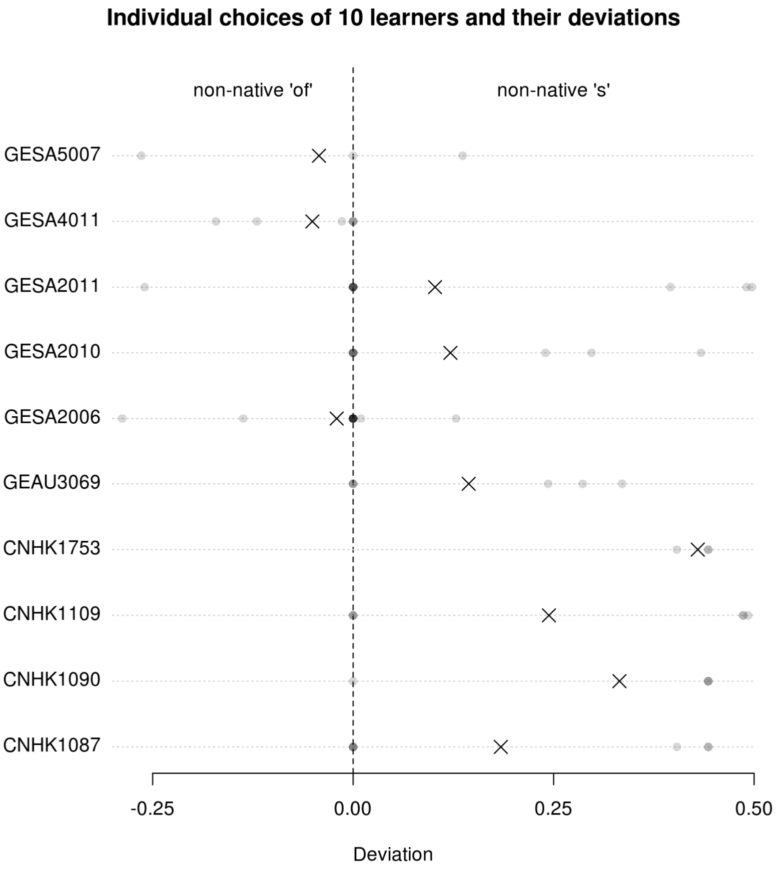

Figure 8: DEVIATION-values of 10 speakers

We could now examine those individual instances more deeply. In an attempt to at least very briefly demonstrate the degree to which the present approach can inform research on individual differences, we did a cursory inspection of the three German and the three Chinese learners with the highest error rates (as determined from the results visualized in Figure 8), which turned out to be revealing because most of the non-native-like uses of the Chinese learners involved one and the same possessor – the proper name *Hong Kong* – whereas there was no obvious pattern discernible in the German learner data; we will return to this finding and what we believe it implies below.

4. Discussion and Concluding Remarks

The above analysis has some important implications for the analysis of learner corpus data and design. In particular, our data show two things. First, there is quite some degree of individual variation in learner corpus data. That is less trivial than it sounds because even a cursory survey of learner corpus studies will reveal that most studies of the types discussed in Section 1.2 do not include individual variation systematically. In our data, the fact that we used a mixed-effects regression model in R_2 of MuPDAR(F) makes quite a difference when it comes to accounting for what the learners are choosing, which becomes very obvious if one compares that regression's results to one that does not involve random effects: Not only is the classification accuracy of the final fixed-effects regression model not significantly better than the baseline of always guessing the more frequent outcome, we also find that the interrater agreement between the mixed-effects and the fixed-effects regression model is somewhat strong, but not as high as one would wish for (κ=0.78). Obviously, a lot of information is lost when the individual learners' differences are not included.

That being said, other ways of including both individual and group variation should also become (more) standardly used in learner corpus research. For instance, for corpora for which proficiency information (or any other text/speaker-specific information) is easily available, then, if proficiency is expected to make a difference for the phenomenon in question, it could be added as a predictor in R(F)$_2$ just like any other coded characteristic of the data. If proficiency information is not easily available or only considered post-hoc, then it can be explored in heuristic ways based on text characteristics such as lexical diversity, average dispersion values of words, etc. For instance, as we discussed in the previous section, the Chinese speakers with the highest rates of non-native-like choices made many of those with one and the same lexical item, *Hong Kong*. A more general analysis of the learners' essays shows that the Chinese essays are characterized by a much higher degree of lexical repetitiveness (as measured by Yule's *K*):

> we computed a Yule's *K*-value for each essay and computed their averages for each L1 group: Chinese average = 134.4 vs. German average = 111.4, which is a highly significant difference (*t*=-9.005, *df*=1229, *p*<0.001);
>
> we conflated all Chinese and all German essays into one big file each and computed Yule's *K*-values for both, which supported the above impression: Yule's *K* for Chinese = 123.9 and Yule's *K* for German = 80.9.

Actually, in our data there is no correlation between the lexical repetitiveness values of all essays and their corresponding percentages of native-like genitives, and we are not implying there should be one – we are pointing out, however, that such kinds of differences between speakers/essays exist and that they *can* correlate with, or even have a causal effect on, learner performance (as seems to be the case in Wulff & Gries's 2015 exploration of prenominal adjective order). Thus, it is important to keep such factors controlled – if one does not do that, variability that may stem from different degrees of proficiency of speakers (lexical repetitiveness) may erroneously be considered variability due to different L1s (see Gries 2018 for much discussion).

Several observations that fell out from our analysis have implications for future corpus compilation projects, or initiatives to enhance existing corpora for that matter. For one, as we alluded to in the previous section, a potential addition to the MuPDAR(F) approach would be to either include a priori, or correlate a posteriori, any speaker information that one might deem relevant. The ICLE corpora include several variables that could be of interest (due to the way we sampled our data, we did not investigate those here), including the speakers' age and the length of exposure to English at school, university, or in an English-speaking country. Paying heed to the major trends in individual variation research in SLA outlined above, it appears that a majority of researchers would be interested in even more comprehensive profiles of each speaker including results of test batteries tapping into speakers' aptitude, motivation, and memory capacity, to name but a few (see Möller in this volume).

A maybe less obvious, but nonetheless crucial implication for corpus compilation regards the sometimes dramatic effects of differential task demands that will be manifest in the data. It is a long established fact that different tasks and task conditions impose varying degrees of cognitive burden on L2 speakers and thus dramatically impact learners' performance across linguistic domains, from lexical diversity to syntactic complexity etc. (see e.g. Robinson 2011 for details on effects of task complexity; Lambert, Kormos & Minn 2017 as an example study that examines effects of task repetition; or Ong 2014 for a study investigating effects of planning time and task conditions on metacognitive processes in L2 writing). In ICLE, one can distinguish texts based on at least some task conditions, including whether the text was produced in a timed or non-timed condition, whether or not the learner was allowed to use reference tools, and whether or not the text was part of an examination – while this is not a comprehensive list of task characteristics, it is much more information that most other learner corpora provide.

Next to cognitive impacts of different tasks and task implementations, task topics are known to trigger (more or less conscious) response strategies that can influence the results of a corpus-based analysis considerably, and in various ways – some of which need not be obvious right away (Paquot 2013, 2014). To give one example from this study, we return to the finding in our study that upon examination of the noun phrases involved in non-nativelike genitives, *Hong Kong* stood out as a highly favored noun phrase in the Chinese learner writing (while we could not discern any such trend in the German data). A qualitative analysis of the Chinese learner writing shows beyond a doubt that given a prompt that mentions *Hong Kong*, they are prone to referencing that noun phrases repeatedly. This not only (negatively) impacts lexical diversity scores (see Gries & Wulff 2013) and other performance measures such as overall word frequency and dispersion; what is more, since *Hong Kong* is such a culturally engrained concept, the Chinese learners consider it as more 'given', which in turn, and in fact suggesting some understanding of the *s*-genitive being correlated with discourse-givenness in native English, triggers their overuse of that noun phrase with the *s*-genitive specifically.

What is more, the task topic also triggers deployment of prefabricated language that learners have come to rote-learn as culturally desirable responses to being asked about their capital (we can infer that it must be rote-learned since various learners produce nearly verbatim passages). Shi (2004) found similar results when she investigated how task type affects the degree of lexical borrowing from source readings; among other things, she pointed out that Chinese students are significantly more likely to use material borrowed from source readings than the native English students. While such observations may be interesting from a cultural/anthropological point of view, it is likely undesirable for a linguistically-oriented analysis: since the automated production of prefabricated chunks artificially boosts proficiency measures of these learners, the data do not adequately reflect the learners' genuine ability to assemble utterances on their own. In conclusion, the ideal learner corpus would comprise both culturally specific as well as more generic prompts so that researchers may choose to examine the data for cultural key words etc., but not be confined to that type of data when the focus of analysis is not on cultural influences.

In conclusion, we hope that this paper has illustrated the usefulness of corpus-based analyses of individual variation and provides some guidance on how to approach individual variation phenomena using the MuPDAR approach. While MuPDAR is not suitable for the investigation of *all* relevant questions in learner corpus research, it is perfectly capable of being used for anything that can statistically be construed as an alternation phenomenon. This means, it can obviously handle any kind of actual alternation phenomenon (e.g., choice between two competing allomorphs, syntactic constructions, registers, modes, instructional styles, etc.). However, this also means that the frequency of any kind of lexical or grammatical choice can also be analyzed in a regression-modeling kind of approach that is conceptually extremely similar to a MuPDAR(F) analysis; see Gries (2018) for discussion of over- and underuse frequencies of the word *quite* using generalized linear mixed-effects modeling of the type used in the present paper and further extensions. We hope that the paper will stimulate vivid discussion among corpus developers and researchers on how to improve future corpus compilation projects with the goal in mind to make corpus linguistics a true alternative and complement to experimental studies.

References

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

Chondrogianni, Vicky & Theodoros Marinis. 2011. Differential effects of internal and external factors on the development of vocabulary, tense morphology and morpho-syntax in successive bilingual children. *Linguistic Approaches to Bilingualism* 1. 318–342.

Collentine Joseph & Barbara F Freed. 2004. Learning context and its effects on second language acquisition. *Studies in Second Language Acquisition* 26. 153-171.

Courtney, Louise, Suzanne Graham, Alan Tonkyn & Theodoros Marinis. 2017. Individual Differences in Early Language Learning: A Study of English Learners of French. *Applied Linguistics* 6(1). 824-847.

Deshors, Sandra C. & Stefan Th. Gries. 2016. Profiling verb complementation constructions across New Englishes: A two-step random forests analysis to *ing* vs. *to* complements. *International Journal of Corpus Linguistics* 21(2). 192-218.

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

Granena, Gisela. 2013. Individual differences in sequence learning ability and second language acquisition in early childhood and adulthood. *Language Learning* 63(4). 665-703.

Granena, Gisela. 2016. Cognitive aptitudes for implicit and explicit learning and information-processing styles: An individual differences study. *Applied Psycholinguistics* 37(3). 577-600.

Granger, Sylviane, Estelle Dagneaux, Fanny Meunier & Magali Paquot. 2009. *International Corpus of Learner English v2*. Louvain-la-Neuve: Presses universitaires de Louvain.

Grey, Sarah, John N. Williams & Patrick Rebuschat. 2015. Individual differences in incidental language learning: phonological working memory, learning styles, and personality. *Learning and Individual Differences* 38. 44-53.

Gries, Stefan Th. 2018. On over- and underuse in learner corpus research and multifactoriality in corpus linguistics more generally. *Journal of Second Language Studies* 1(2). 276-308.

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

Gries, Stefan Th. & Sandra C. Deshors. 2014. Using regressions to explore deviations between corpus data and a standard/target: two suggestions. *Corpora* 9(1). 109-136.

Gries, Stefan Th. & Stefanie Wulff. 2013. The genitive alternation in Chinese and German ESL learners: towards a multifactorial notion of context in learner corpus research. *International Journal of Corpus Linguistics* 18(3). 327-356.

Hamrick, Philip. 2015. Declarative and procedural memory abilities as individual differences in incidental language learning. *Learning and Individual Differences* 44. 9-15.

Hopp, Holger. 2010. Ultimate attainment in L2 inflection: Performance similarities between non-native and native speakers. *Lingua* 120(4). 901-931.

Lambert, Craig, Judit Kormos & Danny Minn. 2017. Task repetition and second language speech processing. *Studies in Second Language Acquisition* 39(1). 167-196.

Li, Shaofeng. 2013. The interactions between the effects of implicit and explicit feedback and individual differences in language analytic ability and working memory. *The Modern Language Journal* 97(3). 634-654.

Liaw, Andy & Matthew Wiener. 2015. Breiman and Cutler's Random Forests for classification and regression. R Package 4.6-12, URL cran.r-project.org/web/packages/randomForest/.

Möller, Verena. 2017. A statistical analysis of learner corpus data, experimental data and individual differences: Monofactorial vs. multifactorial approaches. In P. de Haan, S. van Vuuren & R. de Vries (eds.), *Language, learners and levels: progression and variation*, 409-439. Louvain-la-Neuve: Presses universitaires de Louvain.

Morgan-Short, Kara, Mandy Faretta-Stutenberg, Katherine A. Brill-Schuetz, Helen Carpenter & Patrick C.M. Wong. 2014. Declarative and procedural memory as individual differences in second language acquisition. *Bilingualism: Language and Cognition* 17(1). 56-72.

Ong, Justina. 2014. How do planning time and task condition affect metacognitive processes of L2 writers? *Journal of Second Language Writing* 23. 17-30.

Paquot, Magali. 2013. Lexical bundles and L1 transfer effects. *International Journal of Corpus Linguistics* 18(3). 391-417.

Paquot, Magali. 2014. Cross-linguistic influence and formulaic language: recurrent word sequences in French learner writing. In Leah Roberts, Ineke Vedder, & Jan Hulstijn (eds.), *EUROSLA Yearbook*, 216-237. Amsterdam and Philadelphia: John Benjamins.

Prat, Chantel S., Brianna L. Yamasaki, Reina A. Kluenda & Andrea Stocco. 2016. Resting-state qEEG predicts rate of second language learning in adults. *Brain and Language* 157-158. 44-50.

Robinson, Peter (ed.). 2011. *Second language task complexity: researching the cognition hypothesis of language learning and performance*. Amsterdam/Philadelphia: John Benjamins.

Rothman, Jason & Pedro Guijarro-Fuentes. 2010. Input quality matters: some comments on input type and age-effects in adult SLA. *Applied Linguistics* 31(2). 301-306.

Shi, Ling. 2004. Textual borrowing in second-language writing. *Written Communication* 21(2). 171-200.

Singleton, David. 2017. Language aptitude: desirable trait or acquirable attribute? Studies in *Second Language Learning and Teaching* 7(1). 89-103.

Skehan, Peter. 1986. *Individual differences in second-language learning*. New York and London: Routledge.

Sun, He, Rasmus Streinkrauss, Jorge Tendeiro & Kees de Boot. 2016. Individual differences in very young children's English acquisition in China: Internal and external factors. *Bilingualism: Language and Cognition* 19(3). 550-566.

Street, James A. 2017. This is the native speaker that the non-native speaker outperformed: Individual, education-related differences in the processing and interpretation of object relative clauses by native and non-native speakers of English. *Language Sciences* 59. 192-203.

Unsworth, Sharon. 2016. Early child L2 acquisition: age or input effects? Neither, or both? *Journal of Child Language* 43(3). 608-634.

VanPatten, Bill & Jessica Williams (eds.). *Theories in second language acquisition: an introduction*. New York: Routledge.

Wen, Zhisheng, Mailce Borges Mota & Arthur McNeill (eds.). 2015. *Working memory in second language acquisition and processing*. Bristol/Buffalo/Toronto: Multilingual Matters.

Woodrow, Lindy. 2016. Motivation in language learning. In Ruth Breeze & Carmen Sancho Guinda (eds.), *Essential competencies for English-medium university teaching*, 235-248. New York: Springer.

Wulff, Stefanie & Stefan Th. Gries. 2015. Prenominal adjective order preferences in Chinese and German L2 English: a multifactorial corpus study. *Linguistic Approaches to Bilingualism* 5(1). 122-150.